Population dynamics of *Vibrio cholerae* and its close relative *Vibrio metoecus* in an aquatic ecosystem

By

Paul C. Kirchberger

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

In

Microbiology and Biotechnology

Department of Biological Sciences

University of Alberta

Abstract

Vibrio cholerae, because of its role as the causative agent of cholera pandemics, is an extraordinarily well studied microorganism. Decades of research have uncovered a plethora of molecular mechanisms and a wealth of genomic information related to the organisms' lifestyle as a pathogen. *V. cholerae* is also easily kept in a laboratory setting and, as a common member of bacterial communities in brackish waters and coastal areas worldwide, is readily isolated in large numbers. These two properties make it an ideal organism to study the population dynamics and evolution of aquatic bacteria. However, decades of pathogenicity-focused research have left us with little understanding of even the baseline structure of a natural population of *V. cholerae*. Importantly, the ecology of numerous non-pathogenic lineages of this organism is largely unknown. In this thesis, I aim to test the hypothesis that various lineages of *V. cholerae* have evolved to occupy a number of different ecological niches, and develop a baseline understanding of the population structure and dynamics of this species.

By combining multi-locus sequence typing of over 400 *Vibrio* isolates and whole genome sequencing of selected strains from a pond ecosystem in the eastern United States, I discovered a highly clonal population structure dominated by only a few, phylogenetically distinct lineages. A larger number of non-dominant lineages exist in low abundance and undergo strong fluctuations in the span of a month. I also classify a particularly divergent lineage of *V. cholerae* as a new, phylogenetically and phenotypically distinct species, *V. metoecus*. This species represents the closest discovered relative to *V. cholerae* to date and is an example of recent ecological niche separation. Among a number of potentially ecologically relevant genes differentiating lineages of *V. cholerae,* I discovered an extraordinary diversity of Type VI secretion system (T6SS) associated effector and immunity genes. Based on evidence from extant genomes, I describe a mechanism of horizontal gene transfer and illegitimate recombination that leads to the evolution of complex arrays of effector and immunity genes

unique to each lineage of *V. cholerae* and *V. metoecus* at our sampling site. Finally, I develop a high-throughput sequencing method for a variable region of a gene exclusive to *V. cholerae* that allows the culture-independent study of its population structure at unprecedented scale. In addition to confirming the results of my isolation-based study, I uncover drastic shifts in the population structure of *V. cholerae* over the course of two years, including the invasion of a non-endemic strain after a seasonal depletion of the local *Vibrio* population. Furthermore, I demonstrate pervasive mosaic sympatry, with sampling sites 5m apart harbouring different strains of *V. cholerae*, perhaps mediated by their unique T6SS effector-immunity gene combination. Importantly, I also detect the presence of pandemic-related *V. cholerae* as a minor contributor to this population, sporadically rising to larger relative abundance. I provide evidence that this particular lineage is adapted to life on marine particles, while most *V. cholerae* lineages are generalists found free-swimming and particle associated. These results, taken together, imply that a mixture of mosaic sympatry and T6SS mediated interactions facilitate the coexistence of multiple lineages of *V. cholerae* in a single location. Some of these lineages show strong indication of niche-separation, yet a combination of intrinsic and extrinsic factors likely allows members of the species to diverge significantly while remaining roughly ecologically equivalent.

Preface

Chapters 2-4 as well as two appendices have been published in peer-reviewed journals and are reproduced here in unaltered form, save for minor changes in formatting such as continuous referencing throughout the thesis. Multiple authors have worked to these publications and the individual contributions to chapters included in this thesis are listed below.

Chapter 2:

Kirchberger PC, Orata FD, Barlow EJ, Kauffman KM, Case RJ, Polz MF, Boucher Y. 2016. A Small Number of Phylogenetically Distinct Clonal Complexes Dominate a Coastal *Vibrio cholerae* Population. Appl Environ Microbiol 82:5576-5586.

YB and MFP designed the study, YB and KMK performed sampling, FDO performed whole genome sequencing and core genome clustering, EJB provided multiple scripts used in the analysis. All typing of strains, biochemical assays and data analysis was performed by PCK. PCK, YB and RJC wrote the manuscript.

Chapter 3:

Kirchberger PC, Turnsek M, Hunt DE, Haley BJ, Colwell RR, Polz MF, Tarr CL, Boucher Y. 2014. *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. Int J Syst Evol Microbiol 64:3208-3214.

YB, CLT and MFP designed the study, YB, RRC, BJH, DEH and MT provided samples. Biochemical and sequence analysis was performed by YB and PCK. PCK and YB wrote the manuscript.

Chapter 4

Kirchberger PC, Unterweger D, Provenzano D, Pukatzki S, Boucher Y. 2017. Sequential displacement of Type VI Secretion System effector genes leads to evolution of diverse immunity gene arrays in *Vibrio cholerae*. Scientific Reports 7:45133.

PCK and YB designed the study based on preliminary data by DU and SP. DP performed genome sequencing. PCK performed all analysis, PCK and YB wrote the manuscript.

Chapter 5

High-throughput sequencing of a hypervariable protein-coding gene allows detailed tracking of *Vibrio cholerae* population dynamics and confirms the presence of O1-related strains in a cholera-free region

YB and MFP designed the study, YB and KMK performed sampling, FDO performed whole genome analysis, TN performed qPCR and assisted with some sample preparation. PCK developed all methods, performed sequencing and all analysis. PCK wrote the manuscript.

Appendix 1

Orata FD, **Kirchberger PC**, Méheust R, Barlow EJ, Tarr CL, Boucher Y. 2015. The Dynamics of Genetic Interactions between *Vibrio metoecus* and *Vibrio cholerae*, Two Close Relatives Co-Occurring in the Environment. Genome biology and evolution 7:2941-2954.

PCK assisted in phylogenetic analysis, identification of integron gene cassettes and data analysis.

Appendix 2

Labbate M, Orata FD, Petty NK, Jayatilleke ND, King W, **Kirchberger PC**, Allen C, Mann G, Mutreja A, Thomson NR, Boucher Y, Charles IG. 2016. A genomic island in *Vibrio cholerae* with VPI-1 site-specific recombination characteristics contains CRISPR-Cas and type VI secretion modules. Scientific Reports 6.

PCK performed phylogenetic analysis, identification of type VI secretion system associated genes and pathogenicity islands and wrote parts of the manuscript.

Acknowledgements

There are a few people that I would like to sincerely thank for making the last 5 years of my life so enjoyable. First and foremost, thank you to my PI Yan Boucher. Yan has been a great PI: Stating clear goals of what he expects from me, letting me come up with and pursue my own research ideas and generally trusting me to perform research independently. At the same time, this hands-off approach did not prevent him from always being available to help whenever I needed it and giving useful advice on whatever question I had. As the end of my PhD was approaching, I could not only rely on Yan for help with my research, but increasingly also for advice on my future career. My work with Yan has certainly been a synergy and I don't feel like it was only I that benefited. However, Yan made the last 5 years of my professional life such a fun and pleasant experience that I probably can't do him full justice in these acknowledgments. Needless to say, I'd gladly do the same thing all over again.

Thank you to Joleen Khey, the most wonderful person I've met here. Not only has she supported me emotionally through the good and bad times in my PhD, her scientific input has been invaluable. Barely a word in this thesis, in applications, proposals or professional correspondence has not received thoughtful and patient feedback from her and barely a scientific idea has not been bounced off her over the years. Naturally I blame her for every typo and grammar mistake included in this thesis.

Thank you to Sir Fabs Orata and Doctor Leen Labeeuw, for both their professional support and their friendship over the years. You two really got on my nerves a lot (and I hope I did so too), but I am glad I got to spend time with you.

Thank you to fellow graduate students Tania Nasreen and Tarequl Islam for being such pleasant labmates, Nora Hussain for being a great undergrad and hopefully great successor, Anna Bramucci for her calming presence in emergencies, Ido Hatam for teaching me to love and respect everyone, lunch group for all the lunches, PCR ghosts for not haunting me anymore, and thank you to my committee members Brian Lanoil and Bart Hazes and Rebecca Case and the rest of her lab members past and present. Thank you to Adam Parker and Victoria Collins for adopting me into their family and being perhaps the kindest people I've met in Canada, and also just plain fun to hang out with. Thank you to my parents for all the love and food packages and trying to ensure I'm not always dressed in rags. Lastly, obligatory thank you to Thomas Kickenweiz, who has unfortunately moved on to a better place (matrimony in Singapore).

Table of contents

List of Figures and Tables

Chapter 1

Introduction

1.1 *V. cholerae* as a model system to study microbial population dynamics and evolution

Population thinking lies at the heart of modern evolutionary biology. Populations, not individuals, dynamically change according to the forces of selection, drift and migration. Populations geno- and phenotypically deviate from one another in a process of diversification that can eventually lead to the origin of species. In sexually reproducing organisms such as animals, species are generally defined as groups of interbreeding populations, and the process of speciation is invariably linked to the evolution of reproductive barriers between populations (1). For microbiologists who study prokaryotic organisms where reproduction and sexuality are separate, even the existence of species is a contentious issue (2). Yet in the last decade, the notion that bacteria form monophyletic groups of ecologically similar strains has gained widespread acceptance (3). Two central differences in the biology of pro- and eukaryotes lead to the assumption that formation of such prokaryotic lineages might proceed fundamentally differently from eukaryotes: (I) Prokaryotes reproduce clonally and thus their populations lack the clear coherence that interbreeding of closely related organisms brings. (II) Rampant horizontal gene transfer (HGT), even between distantly related organisms, has the potential to erase any differences that might evolve between populations (4). Major hurdles in improving our understanding of prokaryotic evolution are a lack of knowledge on existing species diversity and population structure, often due to lack of culturability of many organisms, as well as limited knowledge of the particular organism's biology and genetics. *Vibrio cholerae*, a marine Gamma-Proteobacterium, lacks some of these characteristics that generally obstruct the study of prokaryotic evolution. It is readily isolated in large quantities from a number of environments (5). Also, due to the public health threat posed by the organism, its genetics and physiology are extraordinarily well studied. For the same reason, most research on *V. cholerae* is focused on a single clade in this diverse species, which contains all strains capable of causing pandemic outbreaks of the disease cholera. As a result, there is a dearth of information about the biology

of environmental strains, which represent the vast majority of the species. Thus, the immediate goal of my research is to form a baseline understanding of the diversity, population structure and niche adaptations in environmental, non-pathogenic *Vibrio cholerae*. In the longer term, I hope to establish and use *Vibrio cholerae* as a viable model organism to study the emergence of new bacterial lineages in an environmental setting.

1.2 **Patterns of bacterial diversity in nature**

The macroscopic world that we inhabit is at first glance neatly delineated. The reality of groups of similar organisms termed species is apparent (6). The success of numerical taxonomy in animals and other large creatures, where groups of similar organisms can be statistically delineated from one another based on their evolutionarily derived characteristics (7), can be attributed to the vast set of phenotypical traits immediately apparent to even the lay observer. Clusters of organisms based on such statistical analysis of pheno- and genotypic characteristics usually (although not always) overlap with that is both commonly and scientifically considered a species (8).

Conversely, despite the existence of large sets of selective and differential tests, prokaryotes as a whole lack the phenotypic characteristics that enable us to easily differentiate macrobes (9). Cell shape and physiological characteristics permit a degree of taxonomic categorization, yet the sheer diversity of bacterial taxa as well as trait variability within taxa often preclude the correct identification of bacteria on phenotypic basis alone (10). For example, based on its unique phenotype, the bioluminescent *Vibrio* VL426 was described as the species *Vibrio albensis* in 1896, but nearly 100 years later confirmed to fall firmly within the species *V. cholerae* (11). Similar fates have befallen numerous other Vibrios such as *V. gindha* (12) or *V. paracholerae* (13), all various strains of *V. cholerae.*

Studies by Woese and Fox revolutionized microbial systematics and opened the door for studying the ecology and evolution of natural bacterial populations: Phylogenies based on sequences of the 16S rRNA gene not only divided prokaryotes into two clearly separate “kingdoms” of bacteria and archaea (14), but nucleotide sequence based phylogenies of organisms within those two “kingdoms” formed clear clusters, similar to those observed in eukaryotes (15). Despite the more frequent occurrence of horizontal gene in protein coding genes, this clustered structure of bacterial diversity proved to be universal for others genes as well, especially when multiple independent molecular markers are concatenated (16). The

amplification of 16S rRNA sequences directly from DNA extracted from environmental samples, avoiding the cultivation of microorganisms altogether, ultimately demonstrated that natural bacterial communities are composed of clearly delineated groups, themselves composed of closely related sequences (17), corresponding to more or less closely related organisms rather than the continuum of forms dreaded by early taxonomists (18) .

While nowadays the existence of phylogenetic clusters of bacterial diversity is beyond debate, the categorization of these clusters is not. In typical patterns of nested diversity, clusters exist within clusters (19) , and without a concept to group them into sensible biological units, the logical conclusion would be to treat each individual strain as its own unit (4). This approach is obviously impractical, and therefore multiple attempts at making sense of bacterial diversity have been made.

Variation within a cluster of organisms considered to belong to a single species is generally accounted for by grouping similar individuals into populations (20). The term population can be defined through two main approaches: Statistical, where a population is simply defined as consisting of statistically indistinguishable individuals, and generative, which takes into account the forces responsible for the emergence of this variation (20). Traditionally, a biological population is recognized as comprising all individuals of a single species engaged in genetic exchange through mating (21). Through study of changes in allele frequencies over successive generations, a population is considered the main unit in which evolution can be observed (22). Thus, the study of evolution is invariably linked with the study of populations and their dynamics.

In eukaryotes, adaptive alleles tend to spread through populations relatively unlinked from their genetic background, as sexual reproduction is coupled with recombination between the parental genomes. Gene flow within a population leads to ecological and genomic coherence, while the prevention of gene flow between members of a population through reproductive or physical barriers of varying type leads to divergence. Thus, recombinatorial spread of alleles serves as a cohesive force that makes all individuals within a population more similar to each other. At the same time, reproductive barriers serve as a separating force differentiating members of a population from other individuals (6).

While these concepts may appear universally applicable, their translation into the study of microbial evolution is problematic. Even disregarding the problems in definition and delineation of microbial species (to be discussed later in this thesis), discerning populations within a bacterial “species” is difficult (19). In macroscopic organisms that are clearly incapable of traversing geographic barriers such as oceans or mountain ranges, populations are often visibly

separated and thus easy to discern. In bacteria, an oft repeated mantra is “everything is everywhere, and the environment selects” (23). Under the assumption that the ability to disperse is limitless in bacteria, the term population would be close to synonymous with species. On the other hand, biogeographic patterns in the distribution of bacterial diversity are observed at global (24) to microscopic scale (25).

1.3 **What is a bacterial population, what is a bacterial species?**

Perhaps due to the lack of agreed upon species definition and clear geographic boundaries, the term population in microbiology has been given wildly differing meanings by various researchers, ranging from a synonym for species to denoting specific phylogenetic clades or strains. Often, the term population is not defined at all despite its prominent use. Other times, populations are correctly defined as comprising only members of the same species, but spanning vast geographic areas. Accordingly, the “population structure” of *V. cholerae* and other *Vibrio* have been studied under various definitions of the word. Octavia et al. considered an international collection of strains the global *V. cholerae* population, synonymous with species. They divided this population into 4 subpopulations, corresponding to different clades of the species (26). In this case, the *V. cholerae* population is implicitly defined as a group that can be statistically differentiated from others (27). On the other end of the spectrum of population definitions, occasionally monoclonal groups of (*Alii*)*Vibrio fischeri* inhabiting a single crypt of the light organ of the bobtail squid *Euprymna scolopes* are considered a population (28). Between these two extremes ranges a variety of population definitions encompassing geographic regions of various size, such as the *V. parahaemolyticus* population of the Pacific Northwest of the United States (comprising strains isolated from clinical, water, sediment, mollusc and other samples(29)). The common factor in these population definitions is simply the shared species membership of the isolates, with neither of the terms properly defined. However, disregarding both geographic and temporal factors as well as the synonymous use of population and species can be problematic, as this thesis will later show. In the following paragraphs, I will therefore attempt to summarize current knowledge on the mechanisms of bacterial cluster formation in an attempt to come to an operational definition of both population and species that will be used throughout this thesis.

1.4 **Phenetic species concept**

Avoiding most theoretical considerations on how species form in the first place, microbial taxonomy currently uses a polyphasic approach to describe species (10). While this so-called phenetic species concept has been in use for several decades, the methods employed have been continuously updated according to recent advances in technology scientific knowledge. As such, the concept is likely to stay in use for the foreseeable future. The phenetic species concept is based on three connected tenets: (I) monophly (II) genomic coherence and (III) phenotypic coherence (30).

For demonstrations of monophyly (i.e. membership in a phylogenetic clade composed exclusively of conspecifics), initial single gene approaches (usually using the 16S rRNA gene) are being phased out in favour of multi-locus sequence analysis or whole genome phylogenies (31).

The long-standing primary method of inferring genomic coherence (indicating that on a nucleic acid basis, all members of a proposed species are more similar to each other than to other species), is DNA-DNA-Hybridization (DDH) (32). Generally, two genomes are considered to belong to the same species if reciprocal re-association values of their DNA after denaturation under controlled condition equal or exceed 70% (32). For *Vibrio*, DDH values of 80% corresponds more closely with observed phenotypic clusters and might thus be a more valid cutoff for that genus (33). Since DDH is exceedingly difficult and time-consuming to perform (34), multiple alternatives have been developed. Sequence identity of the 16S rRNA genes of two microbes below 97% (or even 98.7% (35)) has been shown to correlate with DDH values of 70% or below and are thus accepted as an alternative indication of genomic coherence (36). The opposite is not necessarily true: in *Vibrio* and many other genera, 16S rRNA identity above 97% or even identical 16S rRNA sequences have been observed to be shared between otherwise distinct species (33). Similarly, whole genome based average nucleotide identities (ANIs) of 95-96% correspond to DDH values of 70% and are also considered viable alternatives to DDH (36). DDH itself can be performed in-silico (37), and a variety of other methods such as tetranucleotide-identity, genomic-distance-matrices and conservation of core genes exist to provide taxonomists with a large number of alternatives to delimit species (38).

The requirement for phenotypic coherence states that the collection of strains belonging to a new species should differ from closely related species in at least one characteristic that is shared by all members of the species (30). As the first toolset available to bacterial taxonomists, a vast variety of different phenotypic tests has been conceived. Growth characteristics under

different conditions are easily performed, and biochemical tests that investigate the metabolic pathways used by bacteria are commercially available and commonly used (39). However, given the propensity for horizontal gene transfer of metabolic genes (40), finding exceptions to phenotypic peculiarities of a species should not come as a surprise. The extensive variation in biochemical properties found in the *Vibrio* genus does not allow for a clear differentiation of species, and interpretations based on carbon utilization are unreliable (33). The analysis of fatty-acid-methyl-ester content, thought to be species specific, is another popular technique much less susceptible to horizontal gene transfer between species. However, the lipid content of *Vibrio* cells does not allow for optimal differentiation of species, and additionally might be subject to change depending on culture condition of analysed strains (41). A high-throughput and efficient method demonstrating phenotypic coherence is Matrix-assisted laser-desorption time-of-flight mass spectrometry (MALDI-TOF), which generates species-specific peptide profiles mostly based on conserved ribosomal proteins. In the case of *Vibrio*, MALDI-TOF has been shown to largely correlate with DNA-based analyses in 97% of all tested species (42).

Despite its long and successful use in describing microbial biodiversity, polyphasic taxonomy and the phenetic species concept are increasingly seen as out-dated or at least impractical (34). The current arbitrary cutoffs for species delineation believed to be too broad. As a result, the diversity threshold of 96% ANI or 97% 16s rRNA identity groups bacteria into taxonomic units equivalent to orders in animals or plants (43). Therefore, the accurate comparison of diversity over large taxonomic distances is impossible (44). On the other hand, tightening the cutoffs would inflate the number of existing species and split long-standing “good” species into many different groups, causing potentially dangerous confusion in the medical field.

Another severe challenge to the phenetic species concept comes from advances in metagenomics. As more genomes of uncultured organisms are assembled from DNA extracted directly from water or soil, the number of clearly distinct “species” that are unable to be described due to the lack of cultured isolates is rising to problematic levels (45). A recent version of the vaunted universal tree of life showed that around half of all known bacterial diversity corresponds to organisms that have never and might never be grown in culture and thus can not be formally described as a species (46). Therefore, the phenetic species concept leaves prokaryotic taxonomy in the awkward situation where a large amount of life’s diversity can not be formally recognized. Thus, there exists the necessity to formalize a species concept that is based on a clear theoretical understanding of how species form based on the evolutionary record left in each organism’s genome (34).

1.5 **Ecotype theory**

According to ecotype theory, bacterial evolution proceeds by alternating periods of diversification from an initially clonal lineage through the acquisition of selectively neutral mutations, and periods of reduction to single clones by selective sweeps initiated by the acquisition of adaptive mutations (47). The resulting clusters of closely related organisms with shared ecology and ancestry are termed ecotypes. Multiple ecotypes can exist in groups that are generally accepted as bacterial species, which are considered more akin to genera of animals (4). The occurrence of niche-specific adaptive mutations (or horizontal gene transfer events) within these ecotypes regularly reduces within-ecotype diversity since clones carrying the mutant allele quickly outcompete all other members of the group. At the same time, successive genome-wide selective sweeps increase between-ecotype diversity, eventually diversifying and specializing these groups enough to be considered distinct species. Thus, the cohesive force forming bacteria into clusters would be provided by selective purges of diversity (47). Periodic selection resulting in clonal sweeps is easily observed in laboratory growth experiments (48, 49), but the occurrence of microdiverse sequence clusters in multi-locus sequence typing efforts of environmental bacteria is generally seen as the hallmark evidence for the theory (50). In culture-independent studies of *Vibrio* (among many other taxa), single-locus marker genes sequences similarly fall into discrete clusters differing by only few nucleotides, initially assumed to be caused by regular selective sweeps (17, 51).

Critics of the ecotype model have challenged the notion of such genome-wide selective sweeps since it largely disregards the occurrence of horizontal gene transfer as a disruptive factor in the cohesion of ecotypes, as well as classical factors in speciation: the roles of geography, migration and neutral evolution (52). In response, multiple variations on the basic theme of genome-wide selective sweeps have been proposed. The ‘Adapt Globally, act locally’ model of ecotype theory takes into account the horizontal gene transfer of universally adaptive alleles from one ecotype to others, causing selective sweeps by the recipient genomes in multiple ecotypes (53). The Geotype-plus-Boeing model posits that geographic isolation can lead to the formation of endemic sequence clusters with the same ecology in different locations (geotypes (54)), which are spread across the world through human transport. As a result, multiple separate sympatric lineages can belong to the same ecotype, at least until a selective sweep eliminates all but one lineage (50). The genetic drift model of ecotype theory takes into account the effect of drift in bacteria with small effective population sizes (for example human pathogens (55)), which, like the Geotype-plus-Boeing model, should result in multiple sequence

clusters in a single ecotype (50). The opposite outcome is provided by the Speedy Speciation model, which posits that a fast rate of ecotype formation (as opposed to the slow rate posited by the stable ecotype model) can lead to multiple ecotypes within a single sequence cluster. This model is thought to apply to adaptive radiation events, where empty niche space is combined with ecological innovation, allowing bacteria to rapidly speciate to fill a wide variety of niches (56). In unstable environments marked by rapid emergence and destruction of niches and the corresponding emergence and extinction of ecotypes, a species-less model is postulated. The result of this is similar to the speedy speciation model, with a single sequence cluster containing multiple recently emerged and soon to be extinct ecotypes (49, 50). The Nano-Niche model is postulated for bacteria in diverse habitats providing opportunity to adapt to minimally different niches (57). Here, multiple ecotypes differing in only few ecological characteristics occur in mosaic sympatry, with selective sweeps having the potential to span multiple ecotypes due to their overlap in niche-space.

1.6 A role for horizontal gene transfer in the emergence of species

The various ecotype models provide convenient theoretical explanations for the existence of almost any type of bacterial population structure. However, there exist very few actual examples of periodic sweeps of diversity affecting the whole genome in nature (58). Instead, studies of environmental populations of bacteria have often violated predictions of the ecotype theory. Single-cell genomics of bacteria such as the abundant marine cyanobacterium *Procholorococcus* have provided evidence for the coexistence of hundreds of genomically differentiated “subpopulations” existing in the same body of water (59). Similarly, metagenomic studies have uncovered the existence of so-called metagenomic islands, highly diverse genomic regions that exist in only a small proportion of a metagenomic “population”, which are considered to be unique to single cells (or cell lineages) (60). Such observations are at odds with the expectation of relative homogeneity in closely related cells due to frequent selective sweeps. Instead, a much larger emphasis is placed on the importance of horizontal gene transfer. While ecotype theory accepts horizontal gene transfer as a source of adaptive change leading to a genome-wide selective sweep (61), the rate of these events is considered too low to interfere with well separated ecotype clusters (62). However, multi-locus sequence typing based analysis of recombination and mutation rates in natural bacterial populations have uncovered a wide spectrum in the relative importance of recombination versus mutation (63). While some bacteria like *Salmonella* live a mostly clonal lifestyle, genetic diversity in other

species like *Neisseria gonorrhoeae* or *Helicobacter pylori* appears to be generated for the most part through horizontal gene transfer events (63). With horizontal gene transfer as the predominant driver of diversity in many species, it seems unlikely that the genome-wide selective sweeps of ecotype theory would be able to occur. However, horizontal gene transfer is not completely random: the efficiency of homologous recombination decreases logarithmically with increased divergence between incoming DNA and the genome of recipient bacteria. As a consequence, gene exchange between closely related organisms is frequent, while genes are only rarely transferred to more distantly related bacteria (64). This pattern is remarkably similar to what is observed in eukaryotes, where preferred genetic mingling with closely related organisms leads to the formation of clusters we term species. For that reason, a species concept akin to Mayr's classical biological species concept might be applicable to (some) species of bacteria after all (4).

1.7 **Gene sweep theory**

Several studies have found that some bacterial populations show evidence of gene- rather than genome-wide selective sweeps, similar to the spread of adaptive alleles in eukaryotic populations. The first exemplary study compared the genomes of two sympatric "populations" (defined as a group of individuals sharing genetic and ecological similarity coexisting in sympatry) of *Vibrio cyclitrophicus,* one adapted to life on large marine particles and the other more often found free-swimming (65, 66). Contrary to the expectation of the ecotype model, members of different populations did not show evidence of genome wide differentiation. Rather, the alleles that clearly denoted a genome as belonging to one or another population were concentrated in only a few genomic regions, while the majority of allelic diversity did not differentiate the populations. Furthermore, while between-population diversity was maximized within those few loci, they showed minimal divergence within a population (66).

This pattern can be explained by rampant horizontal gene transfer between most genomic regions of both populations, with a reduced between-population rate of recombination only observed in regions that confer adaptive advantages to one population or another. Niche-specific beneficial alleles therefore spread through a population by horizontal gene transfer without affecting the rest of the genome. These niche-specific alleles are detrimental to cells from different populations, thus differentiating populations at such loci while recombination at other loci also occurs between populations. Interestingly, signs of recent recombination events were increasingly observed between closely related genomes belonging to the same population,

while older recombination events appear to have occurred across population (66). This is interpreted as a sign of ongoing ecological separation where members of one population preferentially exchange genes with one another while mostly abstaining from between-population exchange – the hallmark of the biological species concept (2).

On a genomic level, diversification and specialization to niches are thought to result from the emergence of beneficial mutations causing either genome wide selective sweeps or by their spread through a population by horizontal gene transfer. The former is thought to increase divergence between ecologically specializing strains equally over the entire genome, the latter only at niche-adapted loci. With increasing specialization, the size and number of differentiating loci is thought to increase due to the accumulation of gene-specific sweeps, ultimately reaching the same endpoint as the genome-wide sweep model (67).

A recent study using a metagenomic time series tracked multiple bacterial “populations” (defined as consisting of all metagenomic reads falling within 95% nucleotide identity to a reference genome) over the course of several years and found what is interpreted as evidence for both competing theories (68). At least in that particular community, gene-specific sweeps appear to occur more often than genome-specific ones, with the type of sweep depending on the properties of particular populations. However, this apparent reconciliation is not universally accepted, since gene-specific selective sweeps can also be explained by the ‘adapt globally, act locally’ variation of ecotype theory (69).

Whether bacterial populations diverge through genome- or gene-wide specific sweeps is ultimately thought to be dependent on the relative strength of recombination versus selection. The emergence of a niche-specific beneficial allele perhaps initiates a genome-wide specific sweep, but in situations where recombination outweighs the selection, horizontal gene transfer spreads the allele throughout the entire population before the sweep is completed (58).

Theoretical considerations also allow for the formation of sympatric phylogenetic clusters even in the absence of natural selection by purely random forces, precluding ecotype theory. However, in the absence of a force that ensures the genomic coherence of clusters, they are expected to show only light divergence and to exist transiently, especially when taking into account the mixing effect of recombination between clusters (70). This prediction has been confirmed by observations in nature, where previously well separated allopatric “species” of bacteria merge into one after prolonged sympatry (71) and even form ecologically successful hybrids (72).

Thus, a biological species concept for bacteria is problematic, and finding a cohesive force (or multiple cohesive forces) capable of forming and retaining the observed patterns of bacterial diversity is one of the main goals in the study of microbial evolution. For sympatric populations to diverge meaningfully in face of horizontal gene transfer and selection, both forces must be reigned in to prevent homogenization due to selective sweeps or gene transfer (58).

1.8 **Phage as independent drivers of population diversity and dynamics**

The aforementioned models of periodic selection through either gene- or genome wide selective sweeps have been challenged by models of constant diversity in microbial populations/species. In models of constant diversity, the increasing evidence for genomic variation between single cells is not considered the result of neutral genetic drift between episodes of selection (60). Rather, the diversity observed in closely related strains is thought to be the adaptive result of the enormous selective pressure imposed on them by bacteriophages (73). Bacteriophages are the numerically strongest biological entities in the biosphere and, as predators of bacteria, are thought to lyse a fifth of the total marine bacterial biomass each day (74). The steady exposure of bacteria to viral elements should therefore impose a constant selective pressure on all members of a microbial population and thus play an important role as a selective force in its dynamics and evolution. Proponents of this model point to the fact that the most mutation-rich genomic regions of bacteria tend to be enriched in genes that are involved in lipopolysaccharide side chains, pili and flagella components – predominantly factors involved in phage attachment (60). The effect of this variation in genes influencing phage attachment is that phages are not necessarily generalists that are able to infect all members of a species, but rather specialists that attach to the lineage-specific binding sites of only a small subset of a species or population. Similar to Lotka-Volterra predator-prey interactions of classical ecological theory, the rise in number of one bacterial lineage within a population is closely followed by the rise in number of lineage-specific phages until phage lysis leads to the collapse of that lineage (75). The result of this “kill-the-winner” model is the prevention of periodic sweeps of successful clones, as competitively superior lineages would succumb to phage predation (73). An example of this has been observed in the bloom and bust cycles of the pathogenic *V. cholerae* serogroups O1 and O139 which are targeted by serogroup-specific phages (76). Through the killing of competitively superior strains, phage predation is thought to put selective pressure on suboptimal adaptation of bacterial lineages. The consequence of this would be a constant

diversity of many different bacterial strains using a non-overlapping set of substrates in an inefficient way, as generalism or improved substrate use would lead to extinction (77).

1.9 **Antimicrobial interactions can lead to coexistence**

Aside from frequency-dependent selective processes, microgeographic separation of strains is another important factor preventing selective sweeps in sympatry (58). Besides the existence of a finely structured habitat, a large factor in creating these conditions is the frequent occurrence of antimicrobial interaction between bacterial cells. Competition for limited space and resources is of course a fundamental fact of both macro- and microbiology (78-80). However, in asexually reproducing organisms, the lack of obligate recombination at each generation produces a much great number of closely related yet genetically independent and thus competing groups. This has led to the evolution of a staggering array of genes mediating antagonism between bacterial lineages. Such genes encode antimicrobial agents ranging from classic antibiotics that are released into the environment to inhibit the growth of competitors (81), bacteriocins that are produced by suicidal cells to the benefit of their clonal kin (82), 'toxins on a stick' contact-dependent-inhibition through autotransported membrane proteins (83) and, among the most recently discovered additions to the repertoire of bacterial weaponry, the Type VI secretion system (T6SS) of *Vibrio cholerae* and numerous other Gram-negative bacteria (84). All of these factors lead to varying degrees of kin-discrimination, allowing bacteria to differentiate closely related cells from foreign competitors. Through selective killing of non-kin, genetically divergent cells are separated from each other, limiting competitive interactions (85). This bacterial territoriality is extremely common in nature, and not only does this phenomenon exist on the level of species, but rather at a much finer scale as well (86). For example, in *Bacillus subtilis*, a large number of varied proteins create a seemingly unique antimicrobial profile for sympatrically occurring cells (87, 88). In the genus *Vibrio*, a vast network of antimicrobial interaction has been uncovered and mapped to correspond closely to genetic distances between populations (defined as ecologically and phylogenetically cohesive units) (89). In addition to observations in nature, modeling and laboratory experiments have shown that (in conditions that do not allow complete mixing of a population) antimicrobial interactions promote the coexistence of multiple strains of bacteria producing different toxins and anti-toxin molecules (90, 91). As such, competition between ecologically identical strains through the use of antimicrobials could allow for the coexistence of multiple equivalent clusters of organisms in sympatry.

1.10 **Thesis objectives**

For the purpose of this text, I will be using the operational definition of a population as the sum of all members of a single species in a single geographical location. This definition requires two points of clarification: What is the geographic range over which the population is spread, and what is a species. The former will be narrowly defined as the location of sampling, a single brackish-water pond at the coast of the Eastern United States directly connected to the ocean by a lagoon. As a species definition, I will be using the phenetic species concept to discriminate *V. cholerae* from other species. While my research is not based on proving an explicit hypothesis, I am working under the implicit hypothesis that a population of *V. cholerae* is divided in multiple distinct lineages that have evolved to occupy different niches, thereby allowing their coexistence. With these points clarified, **the main goal of my thesis is to develop a baseline understanding of the causes and extent of diversity existing in this operationally defined population of *Vibrio cholerae.***

This goal can be roughly divided into three main objectives, and the following four chapters will each contribute a small part in meeting them.

1.) What is the standing diversity of lineages observed at a single point in time in a population of *V. cholerae?*

2.) How does the composition of this population change over time?

3.) What are potential genomic factors that allow members of this population to coexist?

In achieving these objectives, I will have laid the groundwork for integrating the study of *V. cholerae* and its close relative *V. metoecus* into the greater framework of studying microbial evolution and hopefully provided a fruitful groundwork for future studies using these species not only as a specific topic of research, but rather as a broad model system in understanding the population dynamics and evolution of aquatic microbes.

As the preceding pages have demonstrated, the study of bacterial populations is a topic of considerable interest. Our understanding of all aspects of it has progressed dramatically over the last decade through the use of numerous model organisms. Why then reinvent the wheel by establishing basic parameters for what seems to be yet another “model” organism? *V. cholerae* can be argued to hold a special place in microbiology. The organism has been studied since the early days of the field, and due to its medical importance, a tremendous amount of effort has been put into characterizing the species (or rather its pandemic lineage) to minute detail. A vast body of knowledge exists on the physiology and molecular mechanisms that govern the lifestyle

of this organism and its interactions with the human host and its aquatic environment. Important aspects of bacterial lifestyle such as quorum sensing and type VI secretion system mediated competition have been integrated into detailed knowledge of the regulation and functions of genes in *V. cholerae* (92) (93). Furthermore, since the sequencing of the first *V. cholerae* genome in 2000 (94), hundreds of additional genomes have become available. While, as mentioned before, most of the research is focused on a small lineage of medical interest, this research presents an invaluable source of comparative information that can be much more directly applied to conspecific organisms than more distantly related *Vibrio* that have served as model organisms in other studies. Arguably, this broad body of knowledge is nearly unparalleled in other species and will facilitate insights that can not be gained from studying less "popular" organisms. Regardless of this pre-existing knowledge, other characteristics of *V. cholerae* greatly facilitate research with this species. Due to its ease of isolation and ubiquitous occurrence in brackish waters, lagoons and estuaries worldwide, international strain collections can enable biogeographic studies. Furthermore, through its varied lifestyle that involves both growth in and on various animal hosts as well as lengthy periods in environmental reservoirs, *V. cholerae* is exposed to a large number of different ecological challenges that offer many venues of evolutionary adaptation (95). The capability to easily take up DNA from a large variety of sources both facilitates such adaptations in general but also specifically the evolution of pathogenic strains from harmless environmental bacteria (96). An understanding of the population dynamics of *V. cholerae* that is not just narrowly focused on already pathogenic lineages is therefore interesting on more than on a basic research perspective. A thorough understanding of the emergence and spread of pandemic strains can only be achieved by studying the species as a whole, by understanding how its populations are structured and changing over time, by observing the distribution and migration of strains and their interaction with each other and the rest of their environment. A holistic understanding of the species *V. cholerae* can thus lead to an increased ability to predict and potentially prevent outbreaks of cholera caused by already known and newly emerging pathogenic lineages.

1.11 **Thesis outline**

Chapter 2 describes a culture-dependent study of *V. cholerae* and serves as the foundation for the following chapters, which expand on the initial insights that we obtained. First, I observed a clear delineation between classical *V. cholerae* strains and a closely related group of *Vibrio.* The description of that species in **Chapter 3** shows that among the diverse assembly of clades

within *V. cholerae*, at least one is ecologically and genomically divergent enough to be considered a new species, *V. metoecus.*

The observations from Chapter 2 and 3 allow us to clearly delineate the population of *V. cholerae* in Oyster Pond from the co-occurring population of *V. metoecus* using an eco-evolutionary definition: Groups of individuals sharing genetic and ecological similarities coexisting in sympatry. Since the observed lineages of *V. cholerae* occur at both sampling locations of my study (adjacent pond and lagoon) and the genomes of members of a single lineage isolated from both locations do not differ considerably, they can be considered to belong to the same population. Furthermore, comparisons with previously described strains of *V. cholerae* show that these lineages, while sharing alleles with *V. cholerae* from elsewhere on the globe, differ from them in them in their total set of alleles.

While all major lineages of *V. cholerae* exhibit a number of potentially ecologically relevant genes, among the most interesting genomic differences in the major CCs is the variation in the antibacterial effectors of their type VI secretion systems. This could provide a mechanism for competitive exclusion that would allow different lineages of *V. cholerae* to coexist in mosaic sympatry. **Chapter 4** will expand on the diversity of T6SS effector and immunity proteins and posits a hypothetical mechanism through which horizontal gene transfer of a novel effector-immunity gene pair can lead to the emergence of new *Vibrio* lineages and the maintenance of stable clonal complexes.

Another great point of interest regarding the population structure of *V. cholerae* concerns spatial and temporal dynamics of the organism. Only five lineages of *V. cholerae* could be found in both sampling months of Chapter 2, and the number of isolates varies considerably between those months. Furthermore, the lineages appeared to be clustered into groups that either fall into a Pond or a Lagoon category, with some additional clustering according to particle size. However, in both cases, the low sample size only provides a snapshot of the structure of both *V. cholerae* and *V. metoecus*. **Chapter 5** uses a novel protein coding marker gene sequencing approach to overcome the limitations of culture-based studies and provides a longer-term insight into the dynamics of *Vibrio* populations. I show that not only does the population undergo strong fluctuations in relative abundance of lineages over the course of a month, but extensive mosaic sympatry and evidence for the existence of different particle-size niches for various lineages. **Chapter 6** contains additional discussion of topics that have not been fully covered as well as future experiment

Chapter 2

A small number of phylogenetically distinct clonal complexes dominate a coastal *Vibrio cholerae* population

2.1 Abstract

Vibrio cholerae is a ubiquitous aquatic microbe in temperate and tropical coastal areas. It is a diverse species, with many isolates harmless to humans, but others highly pathogenic. Most notable among them are strains belonging to the pandemic O1/O139 serotype lineage, which contains the causative agents of cholera. The environmental selective regimes that led to this diversity are key to understanding how pathogens evolve in environmental reservoirs. A local population of *V. cholerae* and its close relative *Vibrio metoecus* from a coastal pond and lagoon system was deeply sampled during two consecutive months across four size fractions (480 isolates). In stark contrast to previous studies, the observed population was highly clonal, with 60% of *V. cholerae* isolates falling into one of five clonal complexes, which varied in abundance in the short temporal scale sampled. *V. cholerae* clonal complexes had significantly different distributions across size fractions and the two environments sampled, pond and lagoon. Sequencing the genomes of 20 isolates representing these five *V. cholerae* clonal complexes revealed different evolutionary trajectories, with considerable variations in gene content with potential ecological significance. Showing genotypic differentiation and differential spatial distribution, the dominant clonal complexes are likely ecologically divergent. Temporal variation in the relative abundance of these complexes suggests that transient blooms of specific clones could dominate local diversity.

2.2 Introduction

While long thought of as being specifically adapted to life as a pathogen in the human gut (97), numerous strains of *Vibrio cholerae* with varying degrees of virulence thrive in the brackish waters of lagoons and estuaries of the world (98) from the coast of Australia (99) to Iceland (100). The bacterium is believed to form close associations with aquatic invertebrates,

preferentially living a life attached to the chitinous surfaces of those animals (101, 102). It is also regularly isolated from marine vertebrates, algae, sediment or directly from the water column (95).

Perhaps owing to the wide variety of such macro- and microhabitats, *V. cholerae* displays a large degree of genetic diversity (26, 103-105). Horizontal gene transfer by transduction (106), transformation (107) or conjugation (96) is a major factor in the creation of this diversity. Horizontal gene transfer has not only conferred new phenotypes such as the ability to cause lethal bouts of diarrhea, commonly termed cholera (108), but also built patterns of geographic structure due to strains residing in different locations being exposed to genes specific to that environment (109). Variation in both the core- and accessory-genome of *V. cholerae* and the resulting phenotypic diversity has been linked to large-scale environmental factors such as pH, salinity and temperature as well as changes in turbidity and nutrient concentrations, often in seasonal patterns (110-112). The most studied lineage of *V. cholerae*, comprising of the O1/O139 serogroups responsible for pandemic cholera outbreaks, has evolved adaptations to life in the human gut (103, 113), though some of these adaptations might be exaptations from this lineage's association with zooplankton (114). Other (nonpathogenic) strains could prefer different niches - perhaps free swimming or attached to non-animal particles of various sizes, as seen in studies of closely related *Vibrio* species (65, 115, 116). Identification of differential spatial distribution of various strains within *V. cholerae* would indicate that such ecological differentiation might occur at the subspecies level. Ultimately, differentially adapted lineages within a species might be more meaningful units not only when considering the ecology of organisms, but also in tracking potentially harmful pathogens (117).

We conducted extensive sampling and multi-locus sequence analysis of *V. cholerae* and its closest relative, *Vibrio metoecus* (118), from a range of particle sizes in two connected water bodies (pond and lagoon) in a single coastal location in the Northeastern United States (Falmouth, MA). This revealed that co-occurring *V. cholera*e were organized into several abundant clonal complexes, while isolates that were not members of these complexes were rare. We demonstrate statistically significant differences in spatial distribution between the two species studied, but also between different clonal complexes within the *V. cholerae* population. Comparison of the genome sequences of isolates from the major *V. cholerae* clonal complexes unveiled genotypic divergence linked to phenotypes relevant to fitness in the coastal environment. These differences could provide a means of competitive exclusion by which diverse strains contesting for a mostly overlapping set of resources could coexist.

2.3 Material and Methods

2.3.1 Strain isolation, growth and DNA extraction

Environmental strains of *V. cholerae* and *V. metoecus* were isolated from Oyster Pond and Lagoon (Falmouth, MA, USA) on August 24th and September 14th 2009. Three samples were collected on each sampling date from the pond and the lagoon at a 0.5 m depth, with a distance of 5 m between samples. Strains were isolated from different size fractions obtained by sequential filtration of water. Each initial sample consisted of 100 L of water, which were filtered through a 63 µm nylon mesh net. Material collected in the net was transferred to a disposable 50 ml tissue-grinder tube using 20 ml of sterile-filtered local water and crushed. Two ml of the crushed material were diluted 1000-fold (equivalent of 10 ml of pond/lagoon water) and applied to a 0.22 µm filter, which was immediately plated on selective TCBS media (Becton Dickinson, Sparks, MD, USA). The water passing through the mesh net was collected and 10 ml pushed through a series of in-line 4.5 cm filters (Millipore Durapore 5 µm, 1 µm, 0.22 µm) in polypropylene casing using a syringe. All filters were extracted from their casings and immediately placed on TCBS media, before being incubated overnight at 37°C. The ability to utilize sucrose is found only in a few species of vibrios, including *V. cholerae* and *V. metoecus*, and produces yellow colonies on TCBS media. Yellow colonies were picked from TCBS plates and streaked on tryptic soy agar (Becton Dickinson) supplemented with 1% NaCl and incubated overnight. To ensure pure cultures, single colonies from these plates were re-streaked on TCBS and then tryptic soy agar once more, incubating overnight between inoculations.

2.3.2 Multi-locus sequence analysis

DNA extraction from each isolated strain and gene amplification for multi-locus sequence typing (MLST) analysis were performed as previously described (109). All genes were chosen due to their presence as a single copy in the genomes of both *V. cholerae* and V. *metoecus,* high sequence variation in the amplified product (maximizing phylogenetic resolution) and the presence of relatively conserved primer binding sites, allowing the amplification of genes from both species with a single set of primers. All sequences were amplified using *Taq*-Polymerase (Promega, following the manufacturer's instructions for PCR reaction mixture) with the following PCR conditions: 94°C at 2 min, followed by 35 cycles of 94°C at 30 sec, 50°C at 30 sec and 72°C at 1min with a final step of 72°C at 10 min. Sanger sequencing was performed for the forwards reads of each PCR product. Geneious 6.1.7 (119) was used for manual inspection of

reads based on multiple sequence alignments of all products. Low quality ends of sequences were trimmed and sequences edited to correct erroneous base calls where possible. In case of ambiguities, amplicons were re-sequenced.

The Geneious plugin seqpartitioner (http://flossbio.technology/seqpartitioner.html) was used to identify unique alleles and convert the dataset into an MLST format (tab-delimited table). Sequence types (STs) and clonal complexes (CCs) were then identified using eBurst (120). STs were defined as all isolates sharing seven identical alleles, and CCs as groups of closely related STs sharing six out of seven alleles. A nucleotide BLAST search against the NCBI nr (non-redudant) database was performed on each gene of every unique ST to determine if identical alleles or STs had been found in previous studies.

2.3.3 Diversity statistics

As clonal complexes can vary considerably in their sequence diversity (allele changes due to point mutation or recombination are treated equally), isolates were not clustered into operational taxonomic unit (OTUs) based on a simple DNA sequence identity cutoff, but rather assigned to OTUs based on eBurst group membership (see above paragraph). Two different OTU definitions were used, taking into consideration that singleton sequence types could expand into clonal complexes with deeper sampling: 1) Each of 85 unique STs was considered an OTU; 2) Each of 17 CC and all 13 singleton STs were considered OTUs. Rarefaction curves and Chao1 richness index calculations were then performed using each of these OTU definitions, based on the concatenated sequences of the seven genes from each isolate using mothur 1.31.1 (121).

2.3.4 Spatial distribution statistics

To investigate differential environmental distribution of the 17 CCs, a Bray-Curtis dissimilarity matrix of all CCs was calculated based on their relative abundance in different samples using Primer 6 (http://www.primer-e.com). This matrix was then used to create a UPGMA-clustered similarity profile in SIMPROF (122) to determine whether isolation of the various CCs from different fraction sizes of the pond and lagoon differed significantly from each other and from a purely random distribution. To overcome sampling bias, isolates representing each CC were randomly subsampled 100 times, limiting the sampling pool to 32 isolates from each of 8 sample types (four size fractions from both lagoon and pond). The sampling-pool size was determined by the sample type with the lowest number of isolates (Lagoon 1-5 µm size fraction with 32 isolates).

2.3.5 Recombination analysis

The ratio of number of recombination to mutation events (rho/theta) and the ratio of probabilities that a site would be altered by either recombination or mutation (r/m) was estimated from three independent runs of ClonalFrame (100,000 steps, with the first 50% discarded as burn-in) (27). Values were assessed independently for *V. cholerae* and *V. metoecus* based on datasets that included only unique sequences. Run convergence was confirmed using the Gelman-Rubin statistical test implemented in ClonalFrame, with values below 1.2 considered adequate. Rho/theta was also estimated empirically based on the following rationale (123, 124). Variation between alleles of the dominant (i.e., most numerous) and minor STs within a CC was counted as being caused by mutation when varying by one nucleotide and by recombination when varying by two.

Rho/theta was then calculated by dividing the absolute number of recombination events by mutation events. Additionally, the r/m value was calculated on the concatenated dataset of unique *V. cholerae* sequences using the gene conversion model of LDHat (125) using standard settings as implemented in RDP4.66 (126)

2.3.6 Multiple alignments and phylogenetics

Nucleotide sequences of the seven partially sequenced genes of all isolates were aligned with ClustalW using standard settings before being concatenated to create a larger alignment (127). Assembled genomes were aligned with mugsy using standard settings (128). Short locally collinear blocks (LCBs) (<500bp) were removed and the alignments converted into FASTA format using the Galaxy Web server (129). Using Geneious (119) alignments were then manually inspected for quality and all positions containing gaps removed. For the alignment of 20 sequenced genomes alone, LCBs were ordered according to the closed genome of *V. cholerae* N16961, but this step was omitted for the alignment of these genomes with reference strains. Maximum likelihood phylogenetic trees for both the genome and MLST datasets were constructed with RAxML 8.0 (130) using the GTR (general time reversible) model with gamma rate heterogeneity. Statistical support of branches from 100 rapid bootstraps was mapped on the best scoring tree. Additionally, a 50% majority-rule consensus tree based on the seven partially sequenced genes of 438 strains of *V. cholerae* was constructed from three independent runs of ClonalFrame (100,000 steps, with the first 50% discarded as burn-in) (27).

2.3.7 Gene content analysis of clonal complexes

Protein coding genes were clustered into families based on a 30% amino-acid sequence identity (131) using OrthoMCL v 2.0 (132). Unique gene content of different clonal complexes was analysed in Intella (https://www.vound-software.com), then identified through blastp (133) and classified into Clusters of Orthologous Groups (COGs) (134).

2.3.8 Carbon metabolism assays

Carbon metabolism profiles of the five persistent clonal complexes were created by growing the two most prevalent sequence types of each clonal complex on 96-well BIOLOG PM1 and PM2a plates according to the manufacturers instructions (Biolog, Hayward, CA, USA). This was replicated with two independent cultures of isolate. Change in colour of the medium from transparent to purple, indicating positive carbon use, was assayed using a Synergy H1 microplate reader (BioTek). Optical density was measured in comparison to a negative, with a change of 0.5 and above scored as strong use (++), 0.5-0.2 as weak use (+) and scores below as no use (−). Use had to be consistent between replicates to be scored.

2.3.9 NCBI Accession numbers

NCBI Accession numbers for partial *intI*, *plsX*, *mutS*, *recA*, *pgi*, *mdh* and *gppA* sequences of all STs are KX253430 - KX253546, KX253196 - KX253312, KX252845 - KX252961, KX252845 - KX252961, KX253079 - KX253195, KX253313 - KX253429 and KX253547 - KX253663 respectively.

2.4 Results and Discussion

2.4.1 *V. cholerae* populations can be locally dominated by a few clonal complexes

We isolated a total of 480 strains of *V. cholerae* and *V. metoecus* [438 and 42 respectively, with 397 strains isolated in August 2009 and 83 isolated in September 2009 (see Table 2.1 for a list of all isolates)]. For both August and September, we obtained isolates from four size fractions (>63, 5-63, 1-5 and 0.2-1 μm) of three water samples from each the Oyster Pond and its connected lagoon in Falmouth (MA, USA). To estimate population structure and diversity, we performed multi-locus sequence typing (MLST) by partially sequencing seven housekeeping

genes from all isolates. In MLST, isolates having identical sequences at all seven loci belong to a single sequence type (ST). When STs differ at only one locus out of seven, they are considered part of a clonal complex (CC), i.e., a group of closely related strains sharing a recent common ancestor (120).

The 438 *V. cholerae* strains isolated formed 17 CCs, composed of 72 unique STs, with an additional 13 STs found as singletons (not part of a CC) (Figure 2.1 and Figure 2.2).

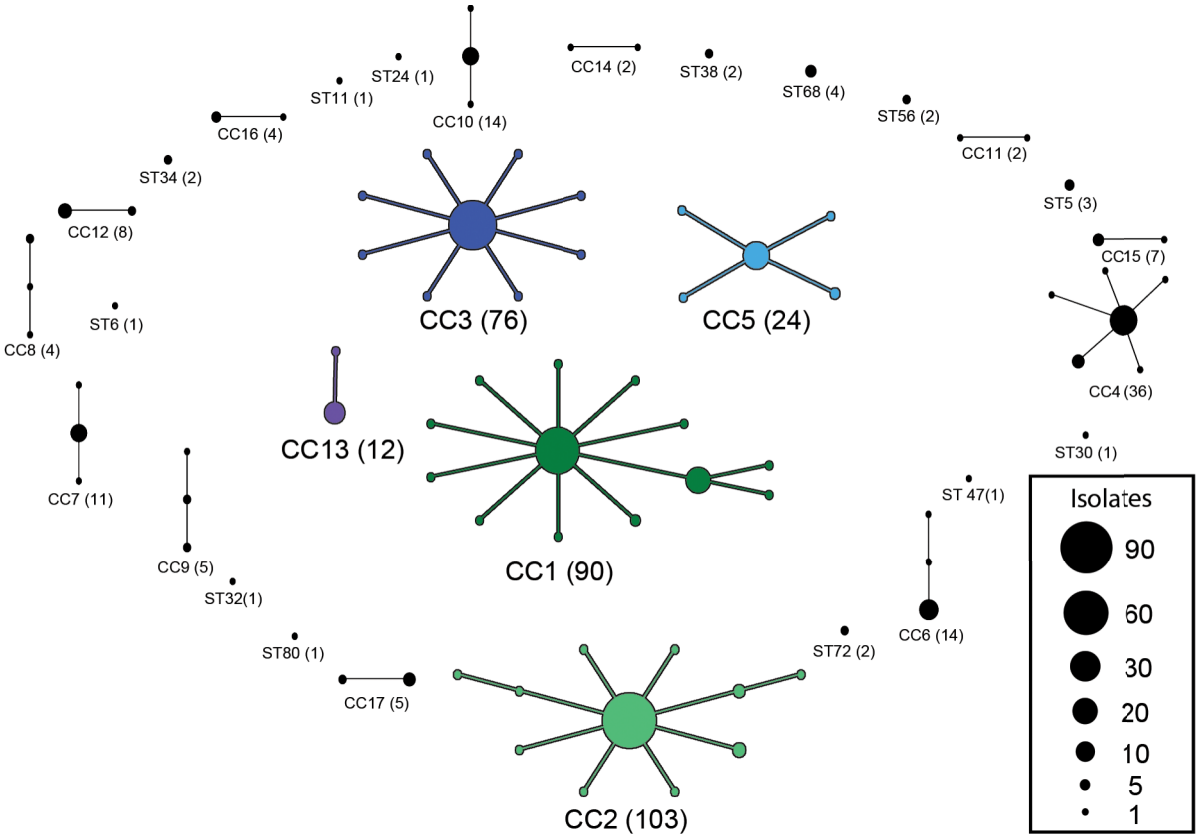


Figure 2.1: eBurst diagram of 438 *Vibrio cholerae* isolates from Oyster Pond (MA, USA) and connected lagoon. Dots represent sequence types (ST) corresponding to unique allele sets from seven partially sequenced housekeeping genes. STs differing in only a single allele are connected by lines and form a clonal complex (CC). Coloured CCs were isolated in both August and September while others were only found in one of the two months. Numbers of isolates are indicated in parentheses.

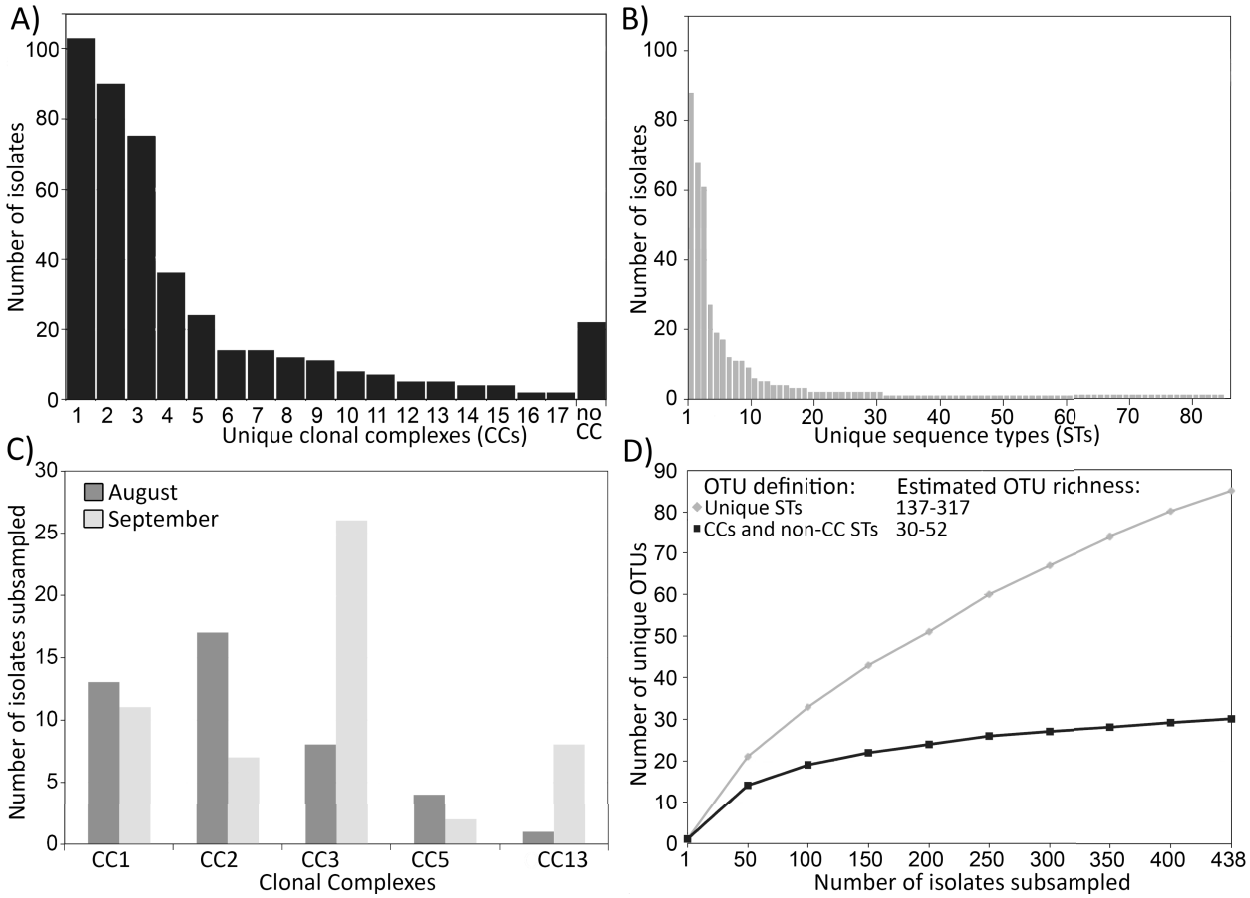


Figure 2.2: Diversity of *V. cholerae* in Oyster Pond and Lagoon. a) Number of isolates assigned to different clonal complexes; b) Number of isolates assigned to different sequence types; c) Number of isolates assigned to different clonal complexes found in August and September (2009). Isolates from August were subsampled to match sampling depth from September; d) Rarefaction curves and Chao1 richness estimation of sampling with different OTU definitions.

Chao1 richness estimation predicts the presence of a minimum of 137 and a maximum of 317 STs, suggesting the existence of a considerably higher number of STs than what was observed (Figure 2.2D). Using a slightly broader OTU definition corresponding to that of a clonal complex, which groups together isolates with at least six identical alleles out of seven and counts all 13 singleton STs as separate OTUs, Chao1 richness estimation predicts the existence of 30-52 OTUs, of which 30 have been observed (Figure 2.2D). Our dataset therefore appears to contain most if not all the clonal complexes present in our samples (60-100%) but only a portion of existing sequence types (30-60%).

Three *V. cholerae* CCs dominated our sampling site, containing a total of 263 isolates (CC1: 90, CC2: 103, CC3: 76), representing 60% of our sampling effort. Most CCs were comprised of a dominant ST, with other variant STs occurring only sporadically (Figure 2.1 and Figure 2.2). The proportion of *V. cholerae* STs found in CCs (87%) in our in-depth sampling of a single geographical location is much larger than what has been found in previous studies of environmental populations of this species (13-18%) (124, 135). This suggests a much more clonal structure and lower diversity than previously estimated for natural populations of *V. cholerae*. Our study comprises the largest dataset isolated to date, with 438 isolates grouping into 85 STs, with 72 STs found in 17 CCs. It also targeted a restricted geographical area (two water bodies, pond and lagoon, within 50 m of each other and connected by a channel) over a short timeframe (less than one month). The most recent comparable study sampled 109 isolates from a dozen sites (lagoon, channel, river and sea) in a 20 km long Mediterranean lagoon system over the course of half a year, discovering 78 STs of *V. cholerae,* with only 14 STs found in five CCs (135). Another study isolated *V. cholerae* from 15 sites (creek, river and harbor) on a 100 km stretch of the Californian coast near the San Francisco Bay area (124) and identified 113 STs and 8 CCs (comprised of three STs each at most) from 156 strains over the span of a year.

This notable difference in the degree of clonality between the populations observed here and the populations investigated in other studies is accompanied by a difference in the estimated ratio of recombination to mutation events. This value is usually given as r/m, the ratio of probabilities by which either recombination or mutation affect a site (which is a measure of the importance of recombination in the creation of nucleotide diversity), or rho/theta, the ratio of absolute number of recombination and mutation events. Keymer & Boehm (124), using various methods of assessing the impact of recombination and mutation, estimated rho/theta from 4:1 to 6.5:1, and r/m of at least 45:1. Similarly, Esteves et al. (135) estimated an r/m of 37:1, and Vos & Didelot (63) an r/m of 20:1 for various populations of coastal *Vibrio*. In stark contrast to that, our r/m estimates ranged from 1.6 (ClonalFrame) to 3.8 (LDHat) (124). Similarly, our rho/theta estimations ranged from 0.49 (ClonalFrame) to 1.9 (empirical estimate (123)), considerably below previous estimates for *V. cholerae* and closely related species.

Our observation of few dominant clonal complexes (low evenness, moderate rate of recombination) in the *V. cholerae* population we studied thus stands in contrast with the large number of highly recombinogenic singletons found in previous reports. It is possible that sampling depth had so far been insufficient to capture local population structure adequately.

Previous sampling efforts have been limited to ~10 isolates per water sample, adding over a dozen sites often separated by 1-10 km to compose a "population" of ~100-150 isolates. If the variation between strains present at each site is significant, such superficial sampling will yield artificially high evenness, with numerous STs of low abundance being observed. As we have sampled most of the diversity present at our two sites at the level of CCs, we obtained a more accurate measure of evenness in our sample. Furthermore, although sequences of some of the seven partially sequenced housekeeping genes in our study were often identical to sequences from previous studies found in the NCBI nr database (especially the commonly sequenced *recA* and *mdh* genes), none of the actual STs (all seven gene sequences from a single isolate combined) identified in this study have identical matches in public databases. This is consistent with previous findings that *V. cholerae* might be a globally panmictic species in which a number of individual alleles are found over a wide geographic range, with local recombination and selection creating locally-dominant variants with unique allele combinations (109). We do not believe that the differences we observed are due to technical errors: Our methods of minimizing PCR and sequencing errors do not appear to be particularly more or less stringent than in other studies, and the nature of the observed population structure is rather robust to the introduction of errors. Since the majority of observed STs belong to CCs and most exist at least in duplicates, an overestimation of STs due to the faulty identification of single nucleotide polymorphisms would have only moved our dataset away from an even more unexpectedly clonal appearance.

Comparison of the results presented here with those of previous studies thus suggests that while diversity is high on a large spatial scale (kilometers) (124, 135), it might be limited within a given environment (pond, lagoon, etc.), and as such sampling schemes might have a large influence on the inference of population structure. Although we can be confident that our extensive sampling allowed us to describe the local Oyster Pond and Lagoon population structure adequately, we cannot specifically identify the cause of this highly clonal structure. Infrequent and/or insufficient mixing with bacterial populations from the ocean could lead to a high degree of geographical isolation, resulting in the limited diversity observed. Population structure would then likely be temporally stable, and vary between sites with different degrees of isolation. On the other hand, if transient blooms of specific CCs forming locally create the clonal structure observed, we would expect temporal instability. There is some evidence supporting the latter hypothesis. Among the 17 *V. cholerae* CCs, only five consist of isolates from both August and September (CC1, CC2, CC3, CC5, CC13), with the remainder found only in a single month. For the five CCs found in both months, their relative abundance varies considerably

between months, indicating that temporal instability is likely, even on a short timescale of weeks or months (Figure 2.2C). Additionally, strong zooplankton association of specific CCs could even lead to diurnal fluctuation in observed diversity due to migration of host animals.

2.4.2 *V. cholerae* clonal complexes have different spatial distributions

V. cholerae as a species does not seem to display particular preferences for the environmental parameters explored in this study (i.e. filter size and location). However, the distribution of clonal complexes of strains from this species shows distinct spatial structuring (Figure 2.3).

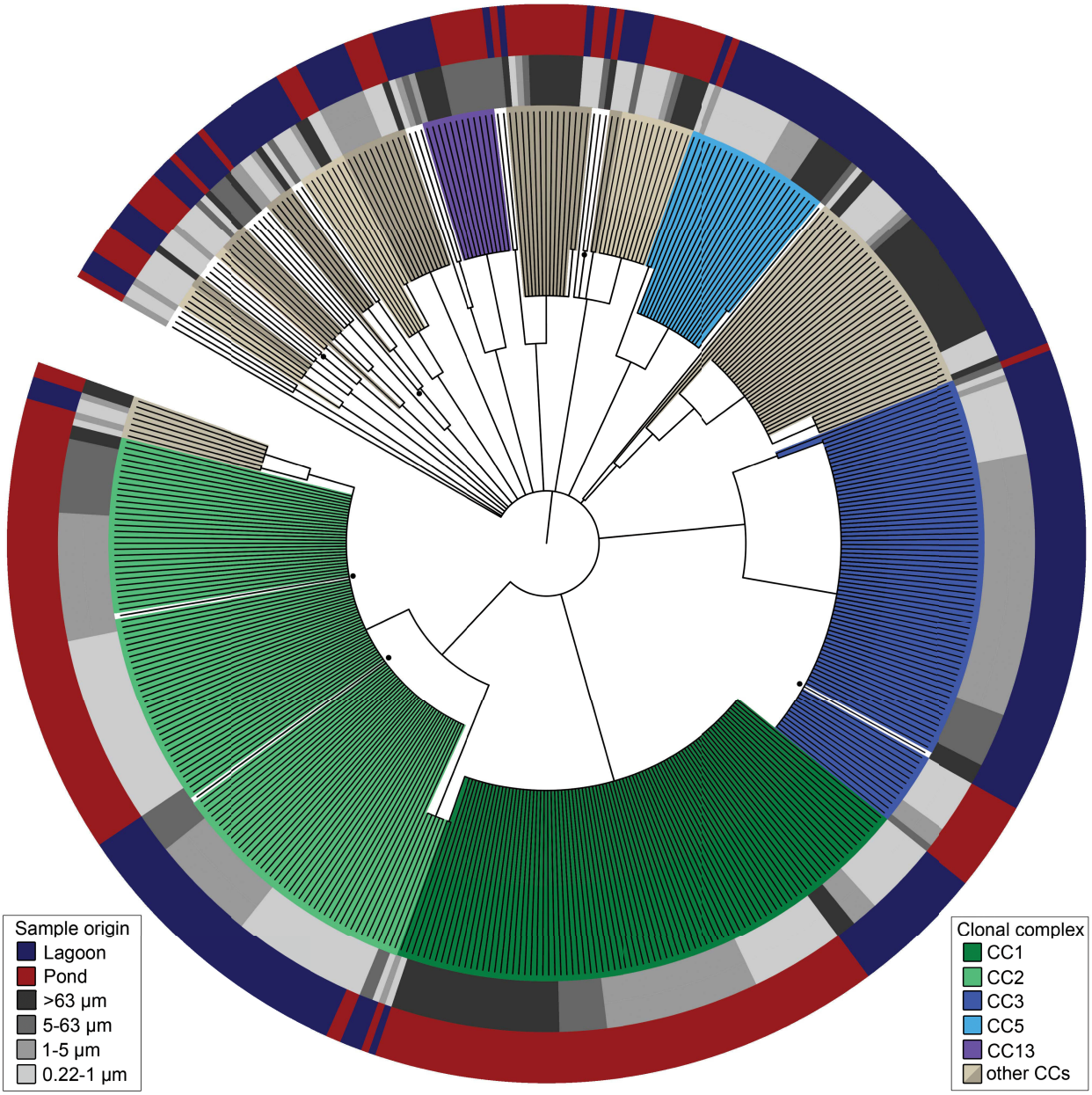


Figure 2.3: Phylogeny of *V. cholerae* isolates from Oyster Pond and Lagoon. Colored rings indicate sample origin with regard to isolation location (outer ring) and filter size (inner ring). Clonal complexes (CCs) are indicated with color shading of alternating intensity. CCs found in only one month are all highlighted in shades of brown, while those found in both August and September are each highlighted with a unique color. Phylogeny represents a 50% majority consensus tree from 3 independent runs of ClonalFrame (27), indicating the clonal backbone of a 3,062-bp alignment of seven partially sequenced housekeeping genes without the effect of recombination. Black dots on nodes represent sequences whose assignment to CCs by eBurst

differs with phylogenetic clustering. Branch lengths are adjusted to facilitate visualization and do not represent true phylogenetic distances.

Because of the poor resolution of phylogenetic trees (a consequence of the close phylogenetic relationship between isolates in the population studied), standard methods of inferring statistically significant environmental distribution patterns for clades of bacteria such as AdaptML (65) or Ecosim (136) could not be used. Instead, we opted for an approach where we treated each CC as a sample, comparing the similarity of CCs based on the spatial origin of isolates they contained (filter size and lagoon or pond). SIMPROF (122) was then used to test whether the calculated Bray-Curtis dissimilarities between UPGMA clustered CCs differed significantly from each other and from a random distribution. In order to avoid skewing of the data by the potential sampling of clonal expansions, we counted isolates of identical STs from the same origin/month as a single isolate.

This approach enabled us to find statistically significant differences in the environmental distribution of isolates from major *V. cholerae* clonal complexes (Figure 2.4).

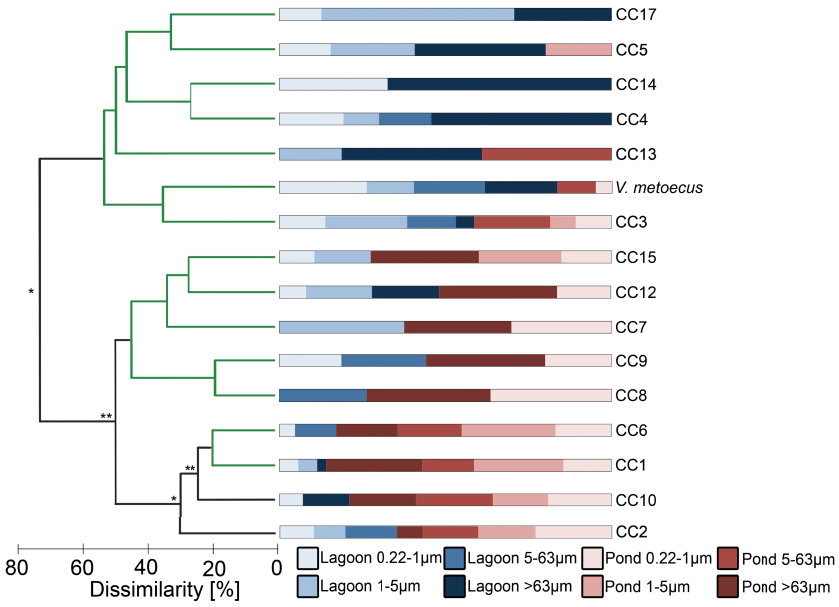


Figure 2.4: Comparison of the spatial distribution of *Vibrio* clonal complexes from Oyster Pond and Lagoon. CCs were clustered by UPGMA of a Bray-Curtis dissimilarity matrix based on the source of individual isolates. Green branches connect CCs that do not show statistically significant differences according to SIMPROF analysis ($p < 0.05$ *, $p < 0.01$ **).

Isolates of identical ST from the same origin/month were counted as a single isolate to avoid the inclusion of clonal expansions. Abundance numbers were derived from 100 random subsamples of isolates from each CC. Number of isolates subsampled (32) was based on the sample type with the lowest number of isolates.

CC1 and CC2 are mostly pond-dwelling (90% and 75% of isolates were found in the pond), with CC1 relatively evenly distributed across size fractions but CC2 mostly found in the smaller size fractions (98% of strains <63 μm). CC3 also has few isolates found in the largest size fraction (96% of strains <63 μm) and is relatively evenly distributed between the pond and lagoon. Isolates from CC5 are predominantly found in the lagoon (95%) with little preference for a specific size fraction. CC13 only had a modest number of isolates (12), but 10 of them were found in the >5 μm fractions. The abundance distribution among sample types is significantly different between CC1 and CC2, and both of them are significantly different from that of CC3, CC5 and CC13 (Figure 2.4). These data demonstrate that *V. cholerae* clonal complexes can display significant differences in their spatial distribution, both in terms of fraction sizes or water reservoirs. The Oyster Pond and Lagoon share similar chemical parameters, with a slightly higher dissolved organic carbon in the lagoon (Table 2.1). The lagoon waters also generally exhibit higher salinities than the pond (5-10 ppt vs. 0-5 ppt). CCs could differ in their growth rates at different salinities, although *V. cholerae* displays species-wide tolerances to salinities far exceeding those found in this lagoon (118, 137). A possible indirect influence of location in the different prevalence of these CCs is the composition of the prokaryotic microbiota of the pond and lagoon, as abundances of bacterial taxa have been shown to correlate stronger with each other than with abiotic factors or eukaryotes in marine environments (138). Another factor which could influence the spatial distribution of clonal complexes is predation by phages. Bacteriophages have been found to play a role in the seasonality of cholera, and could be a major factor influencing the abundance of specific clonal complexes in the Oyster Pond and Lagoon (139).

2.4.3 Different evolutionary trajectories for various clonal complexes

In order to find the possible genetic determinants of the different spatial distribution found for some of the *V. cholerae* CCs, we analyzed 20 genomes from the 5 dominant CCs recovered in sampling from both August and September (CC1, CC2, CC3, CC5 and CC13) that we sequenced in a previous study (140). A 3,410,640 bp whole genome alignment of these isolates

displays 98.6% average pairwise nucleotide identity, with 120,730 phylogenetically informative sites. Within single CCs, any two genomes show between 17 and 124 SNPs, with the exception of CC2, in which multiple regions on both chromosomes 1 and 2 display elevated SNP density. In comparison, every pair of isolates from different clonal complexes differs from each other by approximately 48,000-61,000 SNPs (Supplementary Table S2.2)

A phylogenetic tree based on a core 2,246,831 bp alignment with multiple reference strains of *V. cholerae* and *V. metoecus* (Figure 2.5) shows an early, well-supported node that separates clonal complexes more frequently found in the lagoon (CC3, CC5 and CC13) from CC1 and CC2, which both show stronger association with the pond, similar to the UPGMA clustering based on sample origin. CC13 is found most closely related to *V. cholerae* MZO-3, an O37 serogroup strain from Bangladesh. CC13 and MZO-3 together form the sister clade to pandemic *V. cholerae* O1/O139. CC1 and CC2 are sister clades, and are related most closely to environmental strains RC385 and VL426 ("biovar albensis").

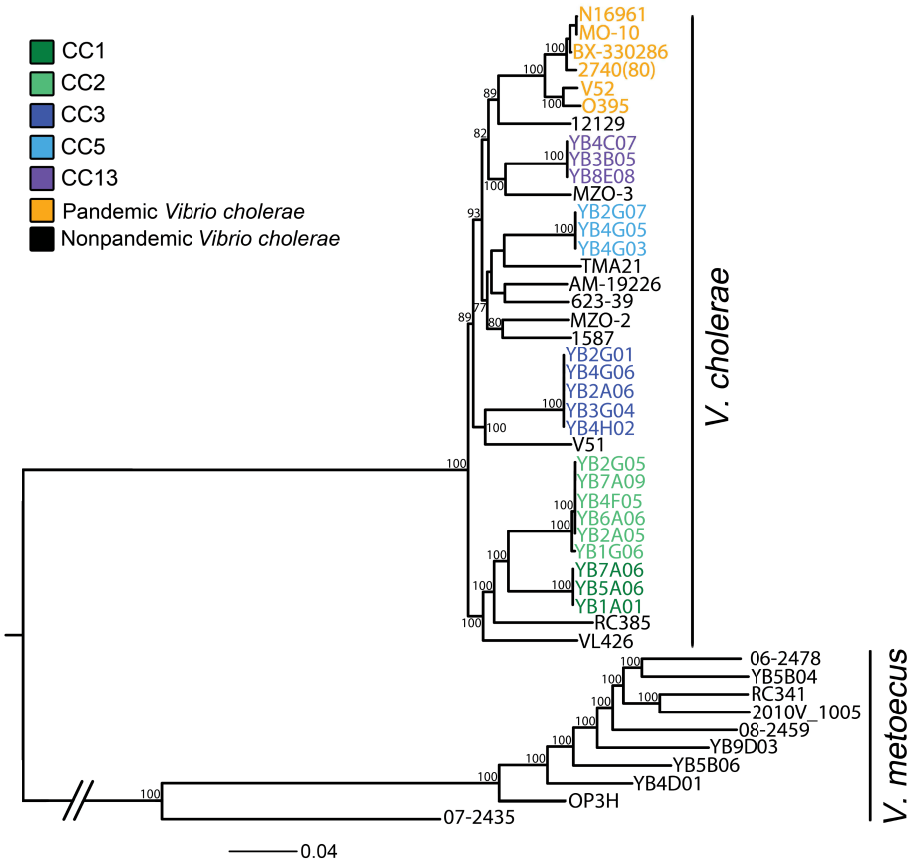


Figure 2.5: Phylogenomic analysis of *V. cholerae* and *V. metoecus* from Oyster Pond and Lagoon. Maximum likelihood phylogenetic tree based on a 2,246,831 bp core genome

alignment of sequenced Oyster Pond and Lagoon isolates with multiple reference strains. Numbers on branches indicate bootstrap support values above 75 derived from 100 bootstrap pseudoreplicates. The branch between V. cholerae and V. metoecus was truncated by 0.04 nucleotide changes (see legend on bottom) to ease viewing. For NCBI Accession numbers see Supplementary Table S2.3)

Single CCs contained a number of genes not shared with members of other CCs (CC1: 113, CC2: 94, CC3: 218, CC5:103, CC13: 229). A large number of these genes were hypothetical, yet around a third could be placed into clusters of orthologous genes (COG) categories and attributed putative functions (Figure 2.6 and Supplementary Tables S2.6 and S2.7, available online).

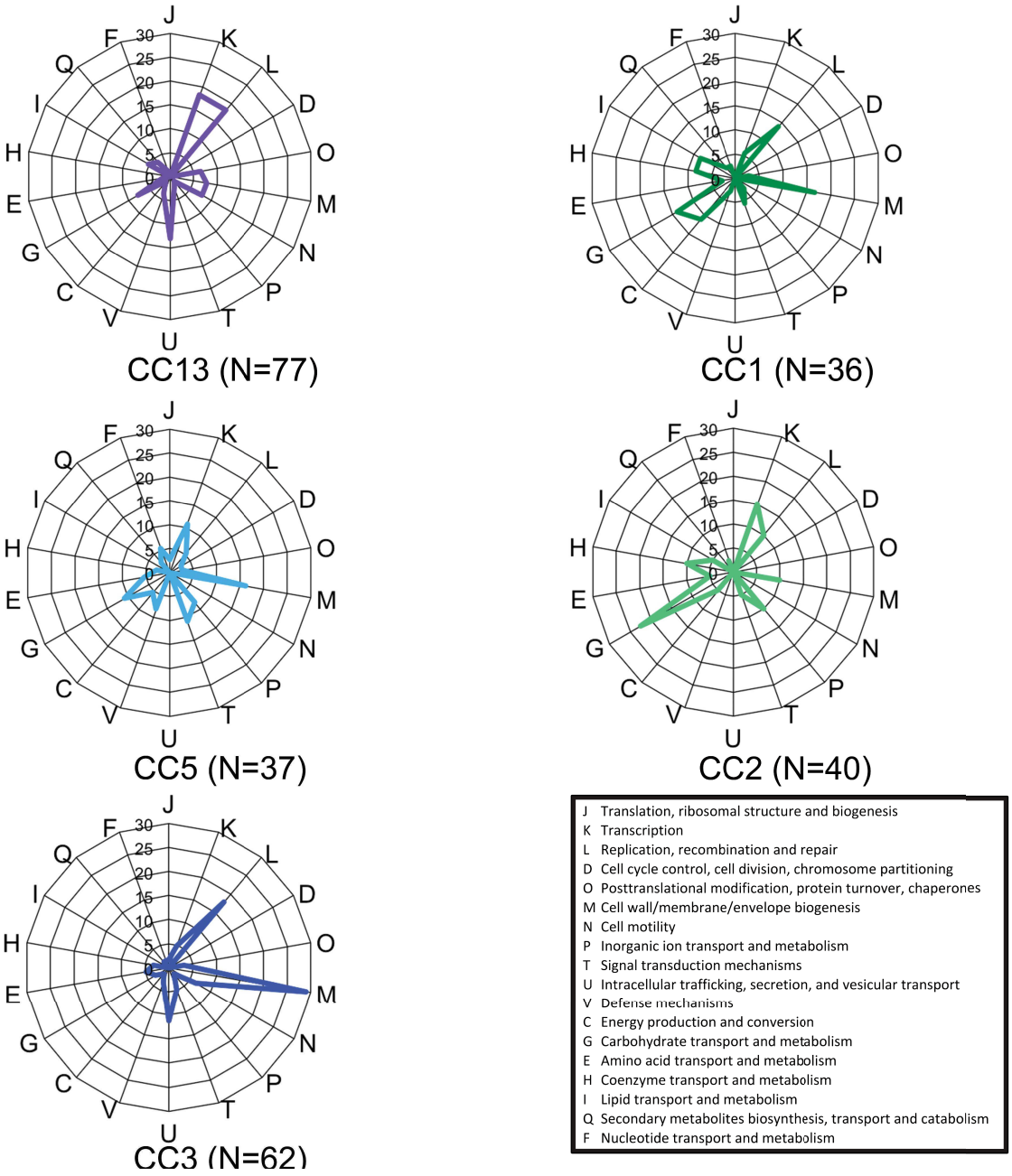


Figure 2.6: Cluster of Orthologous Genes (COG) classification of unique gene families in selected *V. cholerae* clonal complexes. Gene families were defined based on 30% amino acid identity. Radar graphs represent percentage of genes found in specific COG categories. Numbers in parentheses indicate number of genes in COG categories.

Unique sets of type 6 secretion system (T6SS) effector proteins were also identified for each CC, based on the nomenclature by Unterweger et al. (141) (Figure 2.7).

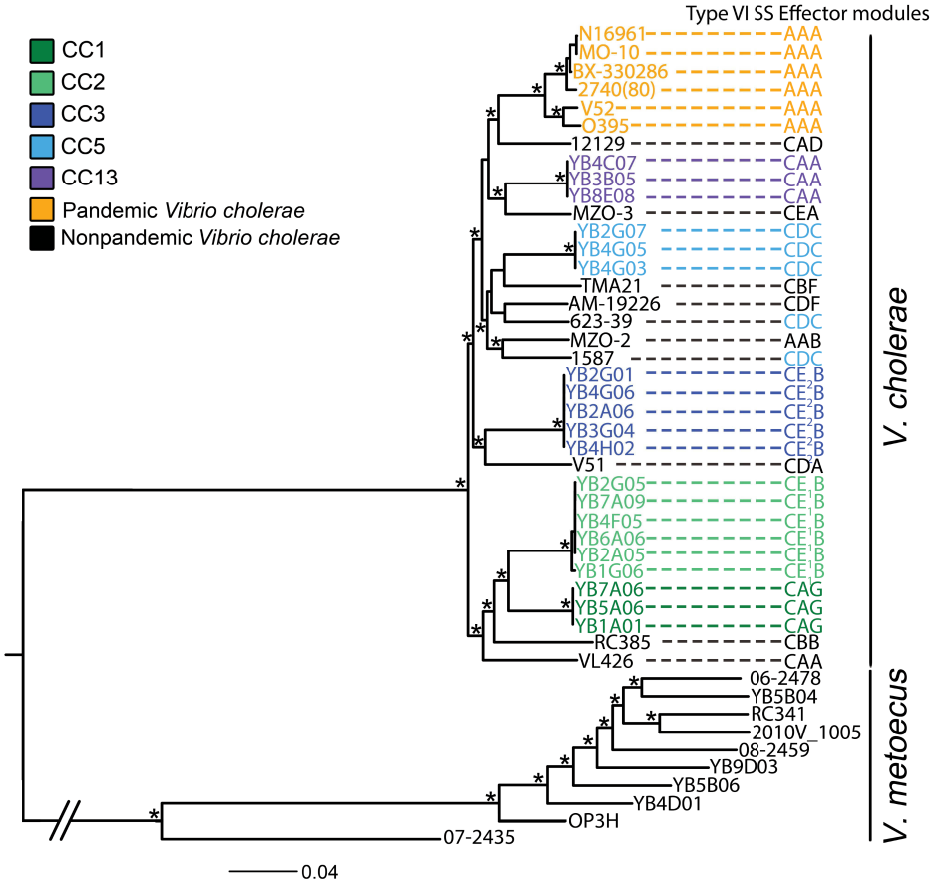


Figure 2.7: Phylogenomic analysis of *V. cholerae* and *V. metoecus* from Oyster Pond and Lagoon with T6SS effector types indicated. Maximum likelihood phylogenetic tree based on a 2,246,831 bp core genome alignment of sequenced oyster pond and lagoon isolates with multiple reference strains. Letters correspond to T6SS effector types as described Unterweger et al. (141), with numbers in subscript indicating subtypes differing by one or more amino acid exchange. Asterisks indicate bootstrap support values above 75 derived from 100 bootstrap pseudoreplicates. The branch between V. cholerae and V. metoecus was truncated by 0.04 nucleotide changes (see legend on bottom) to ease viewing. For NCBI Accession numbers see Supplementary Table S2.3.

For CC1, the most notable among these genetic differences (also displaying an obvious phenotypic effect) is the presence of the *luxCDABG* operon responsible for bioluminescence in vibrios. Isolates in this clonal complex predominantly stem from pond waters and could be considered specialized to this type of environment. Previous studies in Chesapeake Bay, a brackish water habitat similar to Oyster Pond, have found the presence of bioluminescence in around 50% of all isolated strains and, based on clustering by phenotypic traits, hypothesized the presence of the *lux* operon to be an ecologically relevant trait of environmental, non-toxigenic branches in the phylogeny of *V. cholerae* (112, 142). The bioluminescence trait has been linked to the colonization of zooplankton, which in an illuminated state makes easy prey for visually oriented predators, thus enabling bioluminescent bacteria to invade the nutrient rich gut regions of vertebrates predators (112, 143). CC1 strains also harbor the ability to uptake choline and convert it to betaine (through the action of the *betTIBA* operon, shared with sister taxon CC2, RC385 and VL426). This osmoprotectant would be beneficial in coastal waters with variable salinity (144).

CC2 shows the unique presence of the uronate isomerase *uxaC* and *uxuA*, *uxuB* and *uxuR* involved in the metabolism of D-galacturonate and D-glucuronate. Glucuronic acids are compounds produced in the liver of animals and often found in microbial lipopolysaccharides, making it a commonly used carbon source for a number of bacteria (145). The activity of this gene cluster is confirmed by isolates of this CC being uniquely able to use glucuronic acid and glucuronamide as a carbon source (Table 2.1). The ability to metabolize these substrates was previously identified as a trait differentiating *V. metoecus* from *V. cholerae* (118). CC2 uronate utilization genes display nearly 100% identity on the protein level with orthologs from *V. metoecus* (in which this gene cluster is part of the core genome), indicating a recent horizontal gene transfer event from that species.

Table 2.1: Differential carbon source use in *Vibrio cholerae* clonal complexes. ++ Strong use, + weak use, - no use, v differing results within clonal complex

Strains	Carbon source								
	L-Ornithine	L-Arginine	N-Acetyl-Neuraminic Acid	Gelatin	α-Glutaric acid-g-lactone	Glycine	Glycyl-L-Glutamic Acid	D-Glucuronic-Acid	Glucuronamide
N16961	++	++	++	+	-	++	++	-	-
Clonal Complex 13	v	++	-	++	-	++	++	-	-
Clonal Complex 5	-	+	++	++	++	-	++	-	-
Clonal Complex 3	-	+	++	++	-	++	++	-	-
Clonal Complex 2	-	+	-	++	++	++	v	++	++
Clonal Complex 1	++	+	-	++	++	++	++	-	-

CC13, together with its sister clade MZO-3, contains both a set of genes encoding for a pilus as well as an ADP-ribosyltransferase toxin. These are reminiscent of the two principal virulence factors of pandemic O1/O139 *V. cholerae*, the toxin co-regulated pilus and the cholera toxin (146). The pilus and toxin found in CC13 display strongest similarity to the type IV pilus and the heat-labile enterotoxin LT-A, both elements of diarrhea-causing enterotoxic *Escherichia coli* (ETEC) (147). Because most of their virulence factors are encoded on mobile genetic elements, any strain of *V. cholerae* is theoretically able to become a pathogen, yet almost all strains responsible for epidemic outbreaks of cholera are found in the O1/O139 lineage, with a much smaller number of outbreaks attributed to other serogroups such as O37 (148). CC13 represents the closest relative to strains of O37 or O1/O139 serotypes in our dataset (Figure 2.5).

Out of the twelve CC13 strains isolated, ten originated from particles >5 µm. An ability to attach to zooplanktons and other chitinous organisms could increase their potential for virulence, as a single copepod can contain up to 10^4 *V. cholerae* cells (102). CC13, O37 and O1/O139 strains might therefore form a clade of *V. cholerae* with ancestral adaptations that predispose them to a pathogenic lifestyle (148).

CC3 and CC5 both contain multiple capsular lipopolysaccharide (LPS) related genes (several dozens in the case of CC3) of unclear origin in single genomic regions. Horizontal transfer of LPS gene clusters is a frequent occurrence in *V. cholerae* (103), and perhaps a way to evade phage predation, which is dependent on attachment of virions to surface molecules (106). In a system underlying negative-frequency dependent selection by phages, so-called 'defense-strategists' which rise to high abundance not because of their ability to efficiently use resources

but due to their immunity to phage predation are thought to be able to stably coexist with well-adapted 'competition-strategists' (149).

2.4.4 Is the type 6 secretion system shaping the population structure of *V. cholerae*?

V. cholerae prevents eukaryotic predation (84) and kills other bacteria in a contact-dependent way by type 6 secretion system (T6SS) mediated injection of a combination of toxins into target cells. (141). In *V. cholerae*, three different loci encode T6SS toxin/immunity modules, coding for a toxin that is directly injected into target cells and a corresponding immunity protein that confers resistance against that particular toxin. Each distinct module has been assigned a one-letter code, leading to a three letters designation for a specific strain. For example, *V. cholerae* strains belonging to the O1/O139 pandemic lineage are all AAA. Difference at a single locus means that strains are incompatible and will likely kill each other (e.g., CAA vs. AAA), not possessing an immunity protein against the different toxins they each produce. Even if two strains possess the same toxin/immunity modules, differences in the sequences of the toxins and/or immunity proteins could make strains incompatible, a relationship expressed by a number subscript (e.g. CE_1B Vs. CE_2B). All strains in major clonal complexes found in Oyster Pond and Lagoon belong to the same incompatibility groups but none of the clonal complexes are compatible with each other (CC1: CAG, CC2 CE_1B, CC3: CE_2B, CC5: CDC and CC13: CAA). This presents an additional possibility to explain the clonality of the *V. cholerae* population and how multiple *V. cholerae* clades with at least partially overlapping ecological niches can be sustained in a single environment: A community of *V. cholerae* of the same clonal complex could conceivably monopolize the resources around them by killing incompatible invaders of the same species, even those that might theoretically be better adapted to life in that particular niche. Uptake of foreign DNA (natural competence is co-regulated with T6SS expression (93) from killed cells of other clonal complexes could then provide a means of quick adaptation by horizontal gene transfer (or an additional source of nutrients). The ephemeral small patches of resources prevalent in the aquatic environment of vibrios (150) would be particularly suited for such a system, as it could allow a limited number clonal complexes to grow to high densities, effectively exclude late-comers, use up all present nutrients and then continue to spread to other resources. This process could lead to a population dominated by a relatively small number of competing strains, with early colonizer of resources blooming for a period of time until their number is reduced by external factors such as phage predation.

2.4.5 Different habitats and evolutionary dynamics for *V. cholerae* and *V. metoecus*

While our sampling effort was primarily aimed at *V. cholerae*, we also gathered 42 isolates of the closely related *V. metoecus,* organized in 7 CCs and 32 STs (including 13 singletons). *V. metoecus* is rarely isolated (in fact, to date it has only been found in two environmental sites on the U.S. East Coast and few clinical samples) and could simply be more rare than *V. cholerae*. The effect of different isolation regimes on the recovery of *V. metoecus* is unknown, and thus a potential bias in isolation cannot be ruled out. The lack of sampling depth also makes it difficult to directly compare the population structure of *V. metoecus* with *V. cholerae*. Nonetheless, notable differences were observed between these two species.

V. cholerae, as a species, was isolated equally from lagoon and pond. *V. metoecus* however, which shares most of its phenotypic characteristics with *V. cholerae* (118), is found predominantly in the lagoon water (92% of isolates), suggesting an ecological differentiation at the species level (Figure 2.3). *V. metoecus* clades also display notably higher phylogenetic resolution for both whole genome based phylogenies and MLST (Figure 2.5 and Figure 2.8).

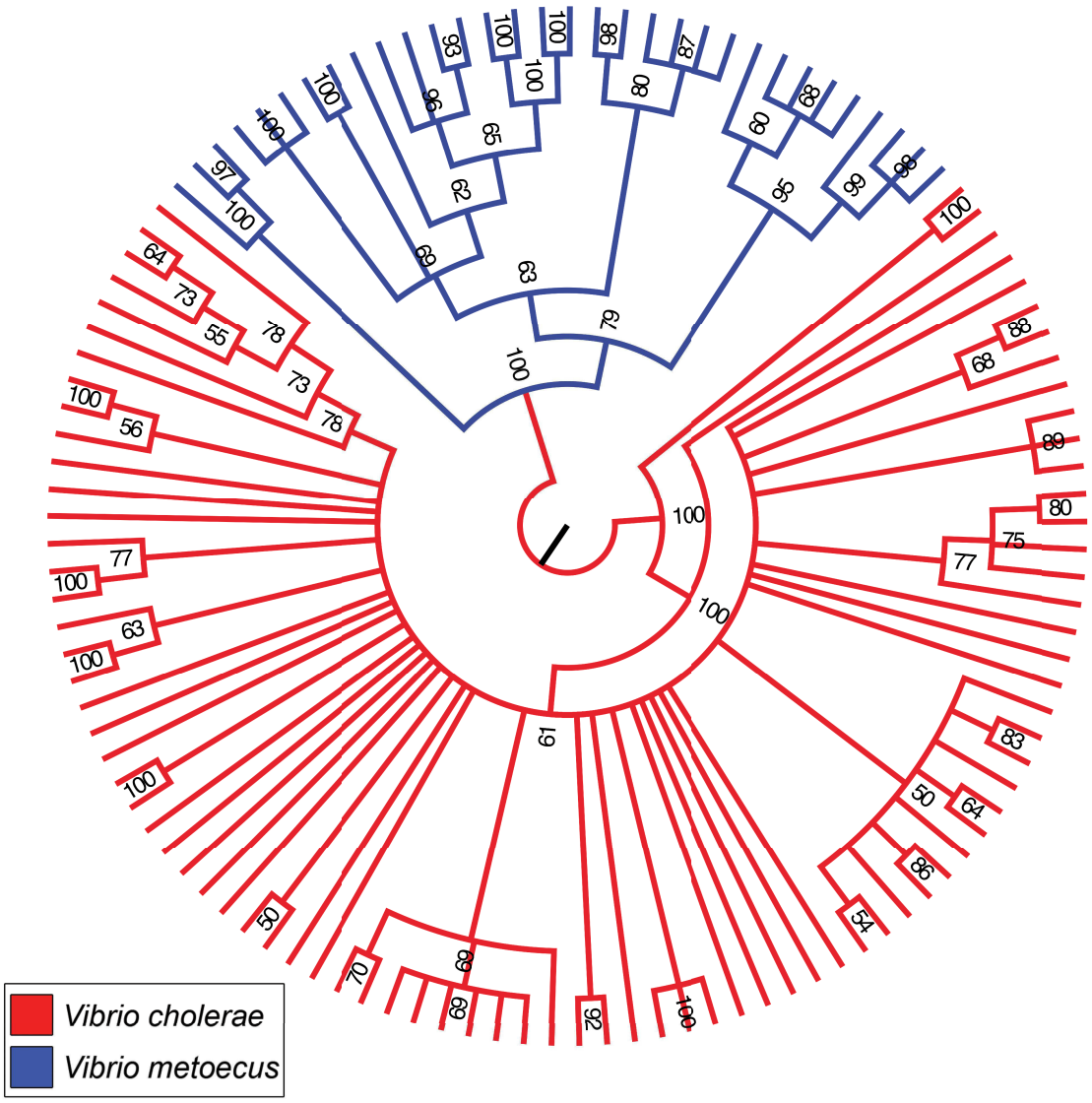


Figure 2.8: Cladogram of unique sequence types of *V. cholerae* and V. *metoecus*. Cladogram is based on a maximum likelihood phylogenetic tree based on a dataset of seven concatenated MLST sequences (3,062 bp), with branch lengths adjusted to ease viewing. Numbers on branches indicate bootstrap support values above 50 derived from 100 bootstrap pseudoreplicates.

The reason for this could lie in differential homologous recombination dynamics in those two species. An analysis of recombination/mutation rates using ClonalFrame (27) determined an r/m ratio of 1.6 for both *V. metoecus* and *V. cholerae*, while rho/theta was 0.49 for *V. cholerae* and only 0.22 for *V. metoecus*. This indicates that less than half the number of recombination events

in *V. metoecus* than in *V. cholerae* account for the same amount of introduced nucleotides. We have previously noted that *V. metoecus* receives considerably more DNA from *V. cholerae* than *vice versa*, presumably due to their sympatric occurrence where *V. cholerae* is the most abundant donor of nucleic acids (140). A situation where large *V. cholerae* populations co-occur with smaller *V. metoecus* populations could lead to a dynamic where the rarer *V. metoecus* more often takes up distantly related DNA while *V. cholerae* predominantly exchanges DNA with the more abundant members of it own species.

2.5 **Conclusions**

By performing the first deep sampling of *V. cholerae* in its natural environment, it was possible to infer the population structure for this species on a small geographical scale. The population found within a coastal pond and lagoon system exhibited moderate recombination rates and a mostly clonal structure with a few dominant clonal complexes. These clonal complexes exhibited significantly different spatial distributions across size fractions in the water column, as well as between neighbouring environments of pond and lagoon. Although they persisted for at least one month, their abundance changed considerably over that period. This suggests that *V. cholerae* is likely to form transient clonal complexes blooming locally, which are genotypically and phenotypically differentiated, displaying divergent spatial distribution patterns and potentially occupying various ecological niches. It has previously been argued that such spatial separation is a prerequisite for the differentiation of gene pools into what could eventually become recognizably different groups of bacteria (117).

The population structure of *V. cholerae* and other bacteria is affected by the geographic scale, timeframe and the depth of sampling (19, 63). Whether the bacteria in a single body of water like the Oyster Pond and Lagoon system is considered a population or merely represents a subpopulation in a larger coastal ecosystem can drastically alter the overall impression of its structure. The time span over which sampling occurs can also significantly affect attempts at determining population structure in bacteria. Our results suggest that a more in-depth understanding of the ecology of specific clonal complexes will require extensive sampling of several sites from specific geographical areas, as well as sampling over a temporal range.

Table S2.1: List of all isolates

V. cholerae isolate	CC	ST	*V. cholerae* isolate	CC	ST	*V. cholerae* isolate	CC	ST
PW0.2-OYP6-A01	1	52	PW1-OYP7-D12	1	59	PW63-OYP8-B04	1	71
LW0.2-OYP5-A06	1	57	PW1-OYP7-E07	1	59	PW1-OYP7-F09	1	73
LW0.2-OYP5-H04	1	57	PW1-OYP7-F07	1	59	PW63-OYP8-B02	1	74
LW0.2-OYP9-E11	1	57	PW1-OYP7-G02	1	59	PW1-OYP7-B07	1	75
LW1-OYP3-C11	1	57	PW1-OYP7-G03	1	59	LW0.2-OYP4-F05	2	8
PW0.2-OYP6-C09	1	57	PW1-OYP7-G05	1	59	LW0.2-OYP4-G04	2	8
PW0.2-OYP6-C10	1	57	PW1-OYP7-G06	1	59	LW0.2-OYP4-H09	2	8
PW0.2-OYP6-D11	1	57	PW1-OYP7-G08	1	59	LW0.2-OYP5-C02	2	8
PW0.2-OYP6-E02	1	57	PW1-OYP7-H04	1	59	LW0.2-OYP5-C09	2	8
PW0.2-OYP6-E03	1	57	PW5-OYP1-H03	1	59	LW0.2-OYP5-C11	2	8
PW0.2-OYP6-E08	1	57	PW5-OYP1-H07	1	59	LW0.2-OYP5-D12	2	8
PW0.2-OYP6-E11	1	57	PW5-OYP1-H08	1	59	LW0.2-OYP5-E07	2	8
PW0.2-OYP6-F12	1	57	PW5-OYP1-H09	1	59	LW0.2-OYP5-E10	2	8
PW1-OYP7-A05	1	57	PW5-OYP8-E01	1	59	LW0.2-OYP5-F07	2	8
PW1-OYP7-C10	1	57	PW5-OYP8-F10	1	59	LW0.2-OYP5-F08	2	8
PW1-OYP7-F08	1	57	PW63-OYP1-A01	1	59	LW0.2-OYP5-F09	2	8
PW5-OYP1-H11	1	57	PW63-OYP1-A12	1	59	LW0.2-OYP5-F11	2	8
PW63-OYP1-B05	1	57	PW63-OYP1-B08	1	59	LW0.2-OYP5-H06	2	8
LW0.2-OYP4-G03	1	59	PW63-OYP1-B12	1	59	LW0.2-OYP5-H11	2	8
LW0.2-OYP4-G09	1	59	PW63-OYP1-C01	1	59	LW0.2-OYP9-E06	2	8
LW0.2-OYP4-G10	1	59	PW63-OYP1-C02	1	59	LW1-OYP2-G05	2	8
LW0.2-OYP5-A05	1	59	PW63-OYP1-C03	1	59	LW1-OYP2-G06	2	8
LW0.2-OYP5-C04	1	59	PW63-OYP1-C04	1	59	LW1-OYP2-H01	2	8
LW0.2-OYP5-C07	1	59	PW63-OYP1-C09	1	59	LW1-OYP3-B06	2	8
LW0.2-OYP5-F04	1	59	PW63-OYP1-C11	1	59	LW1-OYP3-C01	2	8
LW0.2-OYP5-H05	1	59	PW63-OYP1-D01	1	59	LW1-OYP3-C04	2	8
LW1-OYP3-B02	1	59	PW63-OYP1-D03	1	59	LW1-OYP3-D04	2	8
LW1-OYP3-B08	1	59	PW63-OYP1-D08	1	59	LW1-OYP3-D06	2	8
LW1-OYP3-C05	1	59	PW63-OYP1-D11	1	59	LW1-OYP3-D10	2	8
LW63-OYP4-A03	1	59	PW63-OYP1-D12	1	59	LW1-OYP3-E04	2	8
LW63-OYP4-D07	1	59	PW63-OYP1-E01	1	59	LW1-OYP3-E05	2	8
LW63-OYP4-E01	1	59	PW63-OYP8-B05	1	59	LW1-OYP3-E07	2	8
PW0.2-OYP6-B08	1	59	PW63-OYP8-B07	1	59	LW1-OYP3-E08	2	8
PW0.2-OYP6-E04	1	59	PW63-OYP8-B08	1	59	LW1-OYP3-F02	2	8
PW0.2-OYP8-D09	1	59	PW63-OYP8-B09	1	59	LW1-OYP3-G09	2	8
PW1-OYP7-A06	1	59	PW63-OYP8-B11	1	59	LW5-OYP2-A05	2	8
PW1-OYP7-A08	1	59	PW0.2-OYP6-G01	1	62	LW5-OYP2-A08	2	8
PW1-OYP7-A10	1	59	LW0.2-OYP4-H07	1	63	LW5-OYP2-B05	2	8
PW1-OYP7-B10	1	59	PW1-OYP7-D10	1	63	LW5-OYP2-C03	2	8
PW1-OYP7-B11	1	59	PW0.2-OYP6-F10	1	65	LW5-OYP2-E06	2	8

PW1-OYP7-C12	1	59	PW1-OYP7-B03	1	66	PW0.2-OYP6-A06	2	8
PW1-OYP7-D04	1	59	PW63-OYP1-E08	1	67	PW0.2-OYP6-A07	2	8
PW1-OYP7-D09	1	59	PW1-OYP7-C09	1	69	PW0.2-OYP6-A08	2	8
V. cholerae isolate	CC	ST	*V. cholerae* isolate	CC	ST	*V. cholerae* isolate	CC	ST
PW0.2-OYP6-C01	2	8	PW5-OYP1-H05	2	8	LW1-OYP3-B10	3	9
PW0.2-OYP6-C04	2	8	PW5-OYP1-H06	2	8	LW1-OYP3-B11	3	9
PW0.2-OYP6-C05	2	8	PW5-OYP8-E02	2	8	LW1-OYP3-B12	3	9
PW0.2-OYP6-C11	2	8	PW5-OYP8-F11	2	8	LW1-OYP3-C02	3	9
PW0.2-OYP6-C12	2	8	PW63-OYP1-C08	2	8	LW1-OYP3-C08	3	9
PW0.2-OYP6-D04	2	8	PW63-OYP1-E04	2	8	LW1-OYP3-C09	3	9
PW0.2-OYP6-D05	2	8	LW0.2-OYP4-H11	2	12	LW1-OYP3-C10	3	9
PW0.2-OYP6-D12	2	8	PW0.2-OYP6-A04	2	13	LW1-OYP3-D03	3	9
PW0.2-OYP6-F01	2	8	PW0.2-OYP6-E12	2	13	LW1-OYP3-D12	3	9
PW0.2-OYP6-F02	2	8	PW0.2-OYP6-F03	2	13	LW1-OYP3-E01	3	9
PW0.2-OYP6-F05	2	8	PW0.2-OYP6-F08	2	13	LW1-OYP3-E09	3	9
PW0.2-OYP6-F06	2	8	LW0.2-OYP4-F10	2	14	LW1-OYP3-E10	3	9
PW0.2-OYP6-F11	2	8	PW0.2-OYP6-B02	2	14	LW1-OYP3-E12	3	9
PW0.2-OYP6-G02	2	8	PW0.2-OYP6-B12	2	14	LW1-OYP3-F03	3	9
PW0.2-OYP6-G03	2	8	PW0.2-OYP6-E09	2	18	LW1-OYP3-F07	3	9
PW0.2-OYP6-G05	2	8	LW5-OYP2-E01	2	20	LW1-OYP3-F08	3	9
PW0.2-OYP6-G06	2	8	LW0.2-OYP5-B11	2	23	LW1-OYP3-G05	3	9
PW0.2-OYP8-D05	2	8	LW1-OYP3-D09	2	28	LW1-OYP3-G11	3	9
PW0.2-OYP8-D11	2	8	LW5-OYP2-C01	2	33	LW1-OYP9-B01	3	9
PW1-OYP7-A09	2	8	PW1-OYP7-B05	2	35	LW1-OYP9-B07	3	9
PW1-OYP7-C02	2	8	PW0.2-OYP6-F04	2	60	LW1-OYP9-C01	3	9
PW1-OYP7-C07	2	8	LW0.2-OYP4-G06	3	9	LW1-OYP9-C06	3	9
PW1-OYP7-D05	2	8	LW0.2-OYP4-H03	3	9	LW1-OYP9-C07	3	9
PW1-OYP7-D06	2	8	LW0.2-OYP5-A07	3	9	LW1-OYP9-C08	3	9
PW1-OYP7-D08	2	8	LW0.2-OYP5-A08	3	9	LW1-OYP9-C09	3	9
PW1-OYP7-E10	2	8	LW0.2-OYP5-G08	3	9	LW5-OYP2-A06	3	9
PW1-OYP7-G01	2	8	LW0.2-OYP5-G10	3	9	LW5-OYP2-B02	3	9
PW1-OYP7-G07	2	8	LW0.2-OYP9-C11	3	9	LW5-OYP2-B03	3	9
PW1-OYP7-G09	2	8	LW0.2-OYP9-D04	3	9	LW5-OYP2-D04	3	9
PW1-OYP7-G10	2	8	LW0.2-OYP9-E04	3	9	LW5-OYP2-D10	3	9
PW1-OYP7-G11	2	8	LW1-OYP2-G01	3	9	LW5-OYP2-E04	3	9
PW1-OYP7-H01	2	8	LW1-OYP2-G03	3	9	LW5-OYP8-H08	3	9
PW1-OYP7-H02	2	8	LW1-OYP2-G04	3	9	LW5-OYP8-H11	3	9
PW1-OYP7-H03	2	8	LW1-OYP2-G10	3	9	LW63-OYP4-B11	3	9
PW1-OYP8-A02	2	8	LW1-OYP2-H02	3	9	LW63-OYP4-C11	3	9
PW1-OYP8-A03	2	8	LW1-OYP3-A01	3	9	PW0.2-OYP8-C08	3	9
PW5-OYP1-G06	2	8	LW1-OYP3-A02	3	9	PW0.2-OYP8-C10	3	9
PW5-OYP1-G08	2	8	LW1-OYP3-A07	3	9	PW0.2-OYP8-C12	3	9
PW5-OYP1-G10	2	8	LW1-OYP3-A09	3	9	PW0.2-OYP8-D01	3	9

PW5-OYP1-G11	2	8	LW1-OYP3-A11	3	9	PW0.2-OYP8-D07	3	9
PW5-OYP1-H01	2	8	LW1-OYP3-A12	3	9	PW0.2-OYP8-D08	3	9
PW5-OYP1-H02	2	8	LW1-OYP3-B03	3	9	PW0.2-OYP8-D10	3	9
PW5-OYP1-H04	2	8	LW1-OYP3-B07	3	9	PW1-OYP8-A07	3	9
V. cholerae isolate	CC	ST	*V. cholerae* isolate	CC	ST	*V. cholerae* isolate	CC	ST
PW1-OYP8-A08	3	9	LW63-OYP4-A12	4	82	LW1-OYP3-D05	7	49
PW1-OYP8-A10	3	9	LW0.2-OYP4-F01	4	83	LW1-OYP3-D08	7	49
PW5-OYP8-G08	3	9	LW63-OYP4-D09	4	84	LW1-OYP3-F09	7	49
LW1-OYP2-H11	3	10	LW0.2-OYP4-G05	5	1	LW1-OYP3-G06	7	49
PW0.2-OYP8-C06	3	21	LW0.2-OYP4-G12	5	1	PW0.2-OYP6-G08	7	49
LW0.2-OYP5-B09	3	25	LW0.2-OYP5-A02	5	1	PW0.2-OYP6-G09	7	49
LW0.2-OYP4-H02	3	26	LW0.2-OYP5-B01	5	1	PW63-OYP1-C05	7	49
LW1-OYP3-G04	3	27	LW0.2-OYP5-B03	5	1	PW0.2-OYP6-E10	7	53
LW0.2-OYP9-F05	3	29	LW0.2-OYP5-C06	5	1	LW1-OYP3-F05	7	85
PW5-OYP8-F12	3	31	LW0.2-OYP5-D06	5	1	LW5-OYP2-A04	8	48
LW0.2-OYP5-A04	4	76	LW0.2-OYP5-D07	5	1	PW63-OYP1-D09	8	48
LW0.2-OYP5-C01	4	76	LW0.2-OYP5-G05	5	1	PW0.2-OYP6-D09	8	50
LW0.2-OYP5-E04	4	76	LW0.2-OYP9-E05	5	1	PW0.2-OYP6-F09	8	51
LW0.2-OYP5-E05	4	76	LW0.2-OYP9-E07	5	1	PW0.2-OYP6-D03	9	40
LW0.2-OYP5-E06	4	76	LW1-OYP2-G07	5	1	LW0.2-OYP4-H10	9	43
LW0.2-OYP5-F06	4	76	LW1-OYP2-H05	5	1	PW63-OYP1-A09	9	43
LW1-OYP3-C07	4	76	LW1-OYP3-E06	5	1	LW0.2-OYP5-A10	9	45
LW5-OYP2-E03	4	76	LW1-OYP3-E11	5	1	LW5-OYP2-D07	9	45
LW63-OYP4-A04	4	76	LW63-OYP4-B03	5	1	LW0.2-OYP5-D04	10	37
LW63-OYP4-A05	4	76	LW63-OYP4-B04	5	1	PW0.2-OYP6-C03	10	37
LW63-OYP4-A07	4	76	LW63-OYP4-B05	5	1	LW63-OYP4-B02	10	39
LW63-OYP4-A08	4	76	LW63-OYP4-C09	5	1	PW0.2-OYP6-G07	10	39
LW63-OYP4-A09	4	76	LW63-OYP4-B01	5	2	PW1-OYP7-C05	10	39
LW63-OYP4-A10	4	76	LW63-OYP4-A01	5	3	PW63-OYP1-A04	10	39
LW63-OYP4-A11	4	76	LW0.2-OYP4-H06	5	4	PW63-OYP1-A06	10	39
LW63-OYP4-B07	4	76	PW1-OYP7-D07	5	4	PW63-OYP1-B01	10	39
LW63-OYP4-B10	4	76	LW1-OYP3-E02	5	7	PW63-OYP1-B02	10	39
LW63-OYP4-C01	4	76	LW0.2-OYP5-G03	6	46	PW63-OYP1-B03	10	39
LW63-OYP4-C03	4	76	LW5-OYP2-D05	6	46	PW63-OYP1-B11	10	39
LW63-OYP4-C06	4	76	PW0.2-OYP6-A05	6	46	PW63-OYP1-C10	10	39
LW63-OYP4-C08	4	76	PW0.2-OYP6-A09	6	46	PW63-OYP1-E07	10	39
LW63-OYP4-C10	4	76	PW0.2-OYP6-B04	6	46	PW5-OYP1-G01	10	44
LW63-OYP4-D02	4	76	PW0.2-OYP6-B05	6	46	LW0.2-OYP4-G08	11	58
LW63-OYP4-D03	4	76	PW1-OYP7-F10	6	46	LW0.2-OYP4-H04	11	64
LW63-OYP4-D05	4	76	PW5-OYP1-G03	6	46	PW0.2-OYP6-D06	12	41
LW63-OYP4-D06	4	76	PW63-OYP1-C12	6	46	PW0.2-OYP6-D08	12	41
LW63-OYP4-E06	4	76	PW63-OYP1-D02	6	46	LW0.2-OYP5-B07	12	42
LW0.2-OYP4-G07	4	79	PW63-OYP1-E05	6	46	LW0.2-OYP5-D02	12	42

LW0.2-OYP5-A03	4	79	PW63-OYP1-E12	6	46	LW0.2-OYP5-E02	12	42
LW0.2-OYP5-D05	4	79	PW0.2-OYP6-E06	6	54	LW1-OYP3-B09	12	42
LW0.2-OYP5-F12	4	79	PW1-OYP7-G04	6	55	LW63-OYP4-D04	12	42
LW63-OYP4-D11	4	79	LW1-OYP3-C06	7	49	PW63-OYP1-E10	12	42
LW63-OYP4-B12	4	81	LW1-OYP3-D02	7	49	LW1-OYP3-B05	13	17
V. cholerae isolate	CC	ST	*V. cholerae* isolate	CC	ST	*V. metoecus* isolate	CC	ST
LW63-OYP4-C07	13	17	LW5-OYP2-D09		56	LW0.2-OYP9-D09	1	32
LW63-OYP4-D08	13	17	LW5-OYP2-A07		68	LW0.2-OYP9-E01	1	4
LW63-OYP4-E10	13	17	LW5-OYP2-C04		68	LW63-OYP4-D10	1	26
PW5-OYP8-E08	13	17	LW5-OYP2-D02		68	LW63-OYP4-E03	1	24
PW5-OYP8-E09	13	17	LW63-OYP4-B08		68	LW63-OYP4-E04	1	4
PW5-OYP8-E10	13	17	LW1-OYP3-F12		72	LW0.2-OYP9-D03	2	30
PW5-OYP8-E11	13	17	LW5-OYP2-C05		72	LW0.2-OYP9-E03	2	14
PW5-OYP8-E12	13	17	LW5-OYP2-A12		80	LW0.2-OYP9-E10	2	12
PW5-OYP8-F01	13	17				LW1-OYP9-A10	2	3
PW5-OYP8-F08	13	17				LW5-OYP8-H07	2	3
LW63-OYP9-A01	13	36				LW63-OYP9-A06	2	3
LW63-OYP4-E08	14	16				LW0.2-OYP9-E12	3	8
LW0.2-OYP5-F10	14	22				LW0.2-OYP9-F07	3	8
PW0.2-OYP6-E05	15	15				LW63-OYP4-E07	3	29
PW1-OYP7-C03	15	15				LW63-OYP4-E09	3	29
LW0.2-OYP5-E12	15	19				PW5-OYP8-F09	3	8
LW0.2-OYP5-H10	15	19				PW5-OYP8-G05	3	21
LW1-OYP3-G10	15	19				LW0.2-OYP4-F04	4	20
PW63-OYP1-C06	15	19				LW0.2-OYP5-B10	4	19
PW63-OYP1-D07	15	19				LW1-OYP9-B09	5	7
PW0.2-OYP6-A10	16	77				PW0.2-OYP6-A03	5	18
PW0.2-OYP6-E01	16	77				PW0.2-OYP8-D04	5	7
PW0.2-OYP6-E07	16	77				LW1-OYP9-B03	6	13
PW0.2-OYP6-D07	16	78				LW5-OYP8-H05	6	22
LW0.2-OYP4-G02	17	61				LW0.2-OYP5-D09	7	6
LW0.2-OYP4-H08	17	61				LW0.2-OYP5-H08	7	11
LW1-OYP3-F04	17	61				LW0.2-OYP9-E08	7	11
LW63-OYP4-B09	17	61				LW63-OYP4-E05	7	11
LW1-OYP3-F10	17	70				LW0.2-OYP4-G11		23
LW0.2-OYP5-F02		5				LW0.2-OYP5-B04		1
LW0.2-OYP9-F06		5				LW0.2-OYP5-B05		15
LW1-OYP2-H04		5				LW0.2-OYP5-B06		9
PW1-OYP8-A01		6				LW0.2-OYP9-C12		2
PW1-OYP7-C01		11				LW0.2-OYP9-D06		17
LW5-OYP2-B04		24				LW0.2-OYP9-D11		31
LW0.2-OYP5-F05		30				LW1-OYP9-B08		5

LW63-OYP4-E02	32	LW1-OYP9-C03	16
LW0.2-OYP5-G01	34	LW5-OYP8-G09	10
LW0.2-OYP5-H09	34	LW5-OYP8-G11	28
LW5-OYP2-E05	38	LW5-OYP8-G12	27
PW5-OYP8-E03	38	LW5-OYP8-H03	23
PW0.2-OYP6-F07	47	LW63-OYP4-D01	25
LW0.2-OYP5-F01	56		

LW = Lagoon water, PW= Pond water. 0.2 = 0.2-1µm, 1 = 1-5µm, 5 = 5-63µm, 63 = >63µm filter size.

Table S2.2: Parwise comparison of genome differences from 5 dominant clonal complexes

Percentage of identical bases

Clonal Complex	CC2	CC2	CC2	CC2	CC2	CC2	CC3	CC3	CC3	CC3	CC3	CC5	CC5	CC5	CC13	CC13	CC13	CC1	CC1	CC1
Strain	YB1G06	YB2A05	YB2G05	YB4F05	YB6A06	YB7A09	YB2A06	YB2G01	YB3G04	YB4G06	YB4H02	YB2G07	YB4B03	YB4G05	YB3B05	YB4C07	YB8E08	YB1A01	YB5A06	YB7A06
CC2	YB1G06		99.9	99.9	99.9	99.9	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.6	98.6	98.6
CC2	YB2A05	99.9		100	100	100	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.6	98.6	98.6
CC2	YB2G05	99.9	100		100	100	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.6	98.6	98.6
CC2	YB4F05	99.9	100	100		100	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.6	98.6	98.6
CC2	YB6A06	99.9	100	100	100		98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.6	98.6	98.6
CC2	YB7A09	99.9	100	100	100	100		98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.6	98.6	98.6
CC3	YB2A06	98.3	98.3	98.3	98.3	98.3	98.3		100	100	100	100	98.4	98.4	98.4	98.4	98.4	98.4	98.3	98.3	98.3
CC3	YB2G01	98.3	98.3	98.3	98.3	98.3	98.3	100		100	100	100	98.4	98.4	98.4	98.4	98.4	98.4	98.3	98.3	98.3
CC3	YB3G04	98.3	98.3	98.3	98.3	98.3	98.3	100	100		100	100	98.4	98.4	98.4	98.4	98.4	98.4	98.3	98.3	98.3
CC3	YB4G06	98.3	98.3	98.3	98.3	98.3	98.3	100	100	100		100	98.4	98.4	98.4	98.4	98.4	98.4	98.3	98.3	98.3
CC3	YB4H02	98.3	98.3	98.3	98.3	98.3	98.3	100	100	100	100		98.4	98.4	98.4	98.4	98.4	98.4	98.3	98.3	98.3
CC5	YB2G07	98.3	98.3	98.3	98.3	98.3	98.3	98.4	98.4	98.4	98.4	98.4		100	100	98.4	98.4	98.4	98.3	98.3	98.3
CC5	YB4B03	98.3	98.3	98.3	98.3	98.3	98.3	98.4	98.4	98.4	98.4	98.4	100		100	98.4	98.4	98.4	98.3	98.3	98.3
CC5	YB4G05	98.3	98.3	98.3	98.3	98.3	98.3	98.4	98.4	98.4	98.4	98.4	100	100		98.4	98.4	98.4	98.3	98.3	98.3
CC13	YB3B05	98.3	98.3	98.3	98.3	98.3	98.3	98.4	98.4	98.4	98.4	98.4	98.4	98.4	98.4		100	100	98.2	98.2	98.2
CC13	YB4C07	98.3	98.3	98.3	98.3	98.3	98.3	98.4	98.4	98.4	98.4	98.4	98.4	98.4	98.4	100		100	98.2	98.2	98.2
CC13	YB8E08	98.3	98.3	98.3	98.3	98.3	98.3	98.4	98.4	98.4	98.4	98.4	98.4	98.4	98.4	100	100		98.2	98.2	98.2
CC1	YB1A01	98.6	98.6	98.6	98.6	98.6	98.6	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2		100	100
CC1	YB5A06	98.6	98.6	98.6	98.6	98.6	98.6	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	100		100
CC1	YB7A06	98.6	98.6	98.6	98.6	98.6	98.6	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	100	100	

Number of nucleotide differences

Clonal Complex	CC2	CC2	CC2	CC2	CC2	CC2	CC3	CC3	CC3	CC3	CC3	CC5	CC5	CC5	CC13	CC13	CC13	CC1	CC1	CC1
Strain	YB1G06	YB2A05	YB2G05	YB4F05	YB6A06	YB7A09	YB2A06	YB2G01	YB3G04	YB4G06	YB4H02	YB2G07	YB4B03	YB4G05	YB3B05	YB4C07	YB8E08	YB1A01	YB5A06	YB7A06
CC2	YB1G06		2249	2252	2258	2274	2258	58585	58582	58614	58590	58579	58230	58220	58223	59360	59370	59354	48679	48668	48677
CC2	YB2A05	2249		21	17	51	39	58543	58540	58574	58548	58537	58194	58184	58187	59394	59396	59380	48712	48717	48710
CC2	YB2G05	2252	21		30	50	36	58550	58547	58581	58555	58544	58190	58177	58181	59387	59397	59385	48707	48712	48705
CC2	YB4F05	2258	17	30		54	32	58536	58533	58567	58541	58530	58188	58178	58181	59390	59392	59378	48717	48722	48715
CC2	YB6A06	2274	51	50	54		52	58537	58543	58568	58542	58540	58192	58182	58185	59389	59399	59386	48718	48725	48718
CC2	YB7A09	2258	39	36	32	52		58551	58548	58576	58556	58545	58201	58185	58194	59401	59411	59397	48730	48735	48728
CC3	YB2A06	58585	58543	58550	58536	58537	58551		30	95	43	35	55492	55507	55501	56033	56049	56045	58895	58869	58866
CC3	YB2G01	58582	58540	58547	58533	58543	58548	30		105	43	25	55495	55512	55505	56031	56055	56047	58899	58873	58870
CC3	YB3G04	58614	58574	58581	58567	58568	58576	95	105		124	108	55486	55503	55503	56034	56046	56036	58905	58879	58876
CC3	YB4G06	58590	58548	58555	58541	58542	58556	43	43	124		46	55495	55510	55505	56037	56047	56043	58901	58875	58872
CC3	YB4H02	58579	58537	58544	58530	58540	58545	35	25	108	46		55497	55512	55505	56034	56050	56042	58899	58873	58870
CC5	YB2G07	58230	58194	58190	58188	58192	58201	55492	55495	55486	55495	55497		64	93	53087	53087	53080	59314	59299	59297
CC5	YB4B03	58220	58184	58177	58178	58182	58185	55507	55512	55503	55510	55512	64		65	53082	53084	53074	59317	59304	59300
CC5	YB4G05	58223	58187	58181	58181	58185	58194	55501	55505	55503	55505	55505	93	65		53080	53108	53093	59312	59300	59296
CC13	YB3B05	59360	59394	59387	59390	59389	59401	56033	56031	56034	56037	56034	53087	53082	53080		103	113	60913	60909	60905
CC13	YB4C07	59370	59396	59397	59392	59399	59411	56049	56055	56046	56047	56050	53087	53084	53108	103		61	60922	60916	60916
CC13	YB8E08	59354	59380	59385	59378	59386	59397	56045	56047	56036	56043	56042	53080	53074	53093	113	61		60921	60915	60913
CC1	YB1A01	48679	48712	48707	48717	48718	48730	58895	58899	58905	58901	58899	59314	59317	59312	60913	60922	60921		92	67
CC1	YB5A06	48668	48717	48712	48722	48725	48735	58869	58873	58879	58875	58873	59299	59304	59300	60909	60916	60915	92		47
CC1	YB7A06	48677	48710	48705	48715	48718	48728	58866	58870	58876	58872	58870	59297	59300	59296	60905	60916	60913	67	47	

Table S2.3: NCBI accession numbers of genomes used in this study.

Strain	Accession Number
Vibrio metoecus 06-2478	LCUD00000000
Vibrio metoecus 07-2435	LCUE00000000
Vibrio metoecus 08-2459	LCUF00000000
Vibrio metoecus 2010V-1005	LCUG00000000
Vibrio metoecus YB4D01	LBGO00000000
Vibrio metoecus YB5B04	LBGP00000000
Vibrio metoecus YB5B06	LBGQ00000000
Vibrio metoecus YB9D03	LBGR00000000
Vibrio metoecus OP3H	JJMN00000000
Vibrio metoecus RC341	ACZT00000000
Vibrio cholerae YB1A01	LBCL00000000
Vibrio cholerae YB1G06	LBFV00000000
Vibrio cholerae YB2A05	LBFW00000000
Vibrio cholerae YB2A06	LBFX00000000
Vibrio cholerae YB2G01	LBFY00000000
Vibrio cholerae YB2G05	LBFZ00000000
Vibrio cholerae YB2G07	LBGA00000000
Vibrio cholerae YB3B05	LBGB00000000
Vibrio cholerae YB3G04	LBGC00000000
Vibrio cholerae YB4B03	LBGD00000000
Vibrio cholerae YB4C07	LBGE00000000
Vibrio cholerae YB4F05	LBGF00000000
Vibrio cholerae YB4G05	LBGG00000000
Vibrio cholerae YB4G06	LBGH00000000
Vibrio cholerae YB4H02	LBGI00000000
Vibrio cholerae YB5A06	LBGJ00000000
Vibrio cholerae YB6A06	LBGK00000000
Vibrio cholerae YB7A06	LBGL00000000
Vibrio cholerae YB7A09	LBGM00000000
Vibrio cholerae YB8A08	LBGN00000000
Vibrio cholerae 12129	ACFQ00000000
Vibrio cholerae 1587	AAUR00000000
Vibrio cholerae 2740-80	AAUT00000000

Vibrio cholerae 623-39	AAWG00000000
Vibrio cholerae 877-163	LBNV00000000
Vibrio cholerae AM-19226	AATY00000000
Vibrio cholerae BX 330286	ACIA00000000
Vibrio cholerae MAK757	AAUS00000000
Vibrio cholerae MO-10	AAKF00000000
Vibrio cholerae MZO-2	AAWF00000000
Vibrio cholerae MZO-3	AAUU00000000
Vibrio cholerae N16961	AE003852; AE003853
Vibrio cholerae O395	CP001235; CP001236
Vibrio cholerae RC385	AAKH00000000
Vibrio cholerae TMA21	ACHY00000000
Vibrio cholerae V51	AAKI00000000
Vibrio cholerae V52	AAKJ00000000
Vibrio cholerae VL426	ACHV00000000

Chapter 3

Vibrio metoecus sp.nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens

3.1 Abstract

A Gram-negative, curved rod shaped bacterium with close resemblance to *Vibrio cholerae*, the etiological agent of cholera, was isolated over the course of several years from coastal brackish water (17 strains) and from clinical cases (two strains) in the United States. 16S rRNA gene identity with *V. cholerae* exceeds 98% yet an average nucleotide identity based on genome data of around 86% and multi locus sequence analysis of six housekeeping genes (*mdh, adk, gyrB, recA, pgi, rpoB*) clearly delineates these isolates as a distinct genotypic cluster within the *V. cholerae*-*V. mimicus* clade. Most standard identification techniques do not differentiate this cluster of isolates from *V. cholerae*. Only amplification of the *ompW* gene using *V. cholerae*-specific primers and a negative Voges-Proskauer test shows a difference between the two clusters. Additionally, all isolated strains differ phenotypically from *V. cholerae* in their ability to utilize *N*-Acetyl-D-galactosamine and D-glucuronic acid as sole carbon sources. Furthermore, they are generally unable to infect the slime mold *Dictyostelium discoideum*, a widespread ability in *V. cholerae*. Based on these clear phenotypic differences that are not necessarily apparent in standard tests and, average nucleotide identity and phylogeny of protein-coding genes, we propose the existence of a new species, *Vibrio metoecus* sp. nov. with the type strain OP3H^T (LMG 27764^T = CIP 110643^T). Due to its close resemblance to *V. cholerae* and the increasing number of strains isolated over the past several years, we suggest that *V. metoecus* sp. nov. is a relatively common *Vibrio* species that has been identified as atypical isolates of *V. cholerae* in the past. Its isolation from clinical samples also suggests strains of this species, like *V. cholerae*, are opportunistic pathogens.

3.2 Introduction

Bacteria of the genus *Vibrio* are a group of ubiquitous marine organisms, most notable among them *Vibrio cholerae*, the etiological agent of the diarrheal disease cholera. Due to their clear phylogenetic differentiation, it has been suggested that *V. cholerae* and its closest known relative, *V. mimicus*, should be placed in a different genus than the rest of the more than hundred described *Vibrio* species (33) We describe a novel species, *Vibrio metoecus* sp.nov., that is more closely related to *V. cholerae* than any other described *Vibrio*, yet can be differentiated using phenotypic and phylogenetic characteristics. We use multilocus sequence analysis (MLSA) of 19 strains of *V. metoecus* sp. nov. (Table 3.1), average nucleotide identity (ANI) values obtained from whole genome sequences and metabolic profiling of seven strains to clearly delineate *Vibrio metoecus* sp. nov. from other vibrios and demonstrate the species' close relatedness to *V. cholerae*.

A single *V. cholerae*-like bacterium was initially isolated from a water sample at the Chesapeake Bay estuary (MD, USA) in 1998 (151). While initially classified as *V. cholerae* based on its 16S rRNA sequence, a comparative genomic analysis revealed significant differences between this isolate and its closest relatives, *V. cholerae* and *V. mimicus*. The isolate was provisionally named "Vibrio metecus" (152). A global survey of the mobile gene pool content of *Vibrio cholerae* led to the isolation of additional strains from the Oyster Pond (Falmouth, MA, USA) (109) and since 2006, the bacterium has also been isolated from two human clinical cases as a seemingly opportunistic pathogen.

3.3 Material and Methods

3.3.1 Isolation of strains

Sixteen environmental *V. metoecus* sp. nov. and three environmental *V. cholerae* strains from Oyster Pond (OP4A, OP4C and OP8A) were cultured and isolated as described by Choopun and Boucher et al. (109, 151). Aliquots of water samples were filtered through hydrophilic 0.22μm pore size filters (Pall Scientific). The filters were placed on thiosulfate citrate bile salts sucrose (TCBS) plates (Difco) and incubated for two days at 30°C. Resulting colonies showing typical yellow *Vibrio cholerae* appearance were alternatingly re-streaked on tryptic soy broth agar (TSB, Difco), TCBS and a final time on TSB to obtain pure cultures. One additional

environmental and two clinical isolates of *V. metoecus* sp. nov. were obtained from the United States Centers for Disease Control and Prevention (Atlanta, GA).

3.3.2 Electron microscopy

For scanning electron microscopy, isolate 06-2478 was grown overnight at 37°C in TSB + 10g/l NaCl. Bacteria were filtered on 0.22μm pore size Isopore filters (Millipore) before fixing in 2.5% glutaraldehyde, 2% para-formaldehyde in 0.1M Phosphate Buffer for approximately 2h. Fixed samples were washed 3x10min in 0.1M phosphate buffer, 1x50% ethanol, 70% ethanol, 90% ethanol, 2x100% ethanol, 1x75:25 ethanol:HMDS (hexamethyldisilazane, Electron Microscopy Sciences), 50:50 ethanol:HMDS, 25:75 ethanol:HMDS, 100% HMDS, pelleted and left to dry overnight in a fume hood. Samples were mounted on SEM stub, sputter coated with Au/Pd and viewed in a Philips/FEI (XL30) Scanning Electron Microscope.

For transmission electron microscopy, a single colony of *V. metoecus* sp. nov. strain 06-2478 was taken from a TSB +10g/l NaCl agar plate that had been grown over night at 37°C and was subsequently fixed in 50ul 2.5% glutaraldehyde, 2% para-formaldehyde in 0.1M Phosphate Buffer for approximately 2h. A droplet of sample was placed on a Formvar coated grid (Electron Microscopy Sciences), blot-dried with filter paper after 30 seconds and covered with a droplet of 2% phosphotungstic acid (pH 7) for 15 seconds. Sample was viewed in a Philips/FEI (Morgagni) Transmission Electron Microscope with Gatan Digital Camera.

3.3.3 Phenotypic assays

Phenotypic characterization was performed on seven *V. metoecus sp. nov* (two clinical: 08-2459 and 06-2478; five environmental: OP6B, OP5A, OP4B, OP1E and OP3H^T), five *V. cholerae* strains (two clinical: N16961 and V52; three environmental: OP4A, OP4C and OP8A) and *V. mimicus* CAIM 602 This was done using API20NE (Biomerieux) and Biolog GN2 plates with cultures that had been grown on TSB agar plates over night at 37°C. *V. mimicus* CAIM GN2 data was taken from (153). Additional Voges Proskauer assays were performed in BBL MR-VP Broth (BD), test for lysine decarboxylase and ornithine decarboxylase activity in BBL Moeller Decarboxylase Broth (BD) with 10g/l L-lysine or L-ornithine respectively, according to the manufacturers' instructions. Test for tartrate utilization was performed using the ingredients and procedure from BBL Jordan's tartrate agar deeps (BD).

Permissive growth temperature was determined in LB Medium (Difco) with 1% NaCl added in a range of 4-45°C while permissive salinity levels were determined in LB Medium at 30°C in a range of 0-8 % NaCl. The ability to kill *Dictyostelium* amoebae, an assay designed for the detection of virulence-associated secretion genes in non-O1/non-O139 *V. cholerae*, was performed by co-plating of bacteria (*V. metoecus* OP1B, OP1D, OP1H, OP6H, OP3H^T, OP4B, OP4F, OP5A, OP5D, OP6A and OP6B as well as the five *V. cholerae* strains mentioned above) and *Dictyostelium* and observing plaque formation as described previously (84). All tests were performed in triplicate. The results of all phenotypic assays are given in the description section.

3.3.4 Phylogenetic analysis

Genotypic analysis was performed as described in Boucher et al. (109). All 19 *V. metoecus* sp. nov. and five *V. cholerae* isolates were grown in TSB overnight at 37°C. DNA extraction was performed using Tissue DNA Kit (QIAGEN) or Lyse-and-Go (Pierce). The partial 16S rRNA gene of OP3H^T was amplified using primers 27F and 1492R (154). Sequences were aligned with ClustalW (155) and Geneious 6.0.3 (Biomatters) was used to assess sequence variation. *V. cholerae*-specific internal transcribed spacer region (ITS) was amplified using primers pVC-F2 and PVCM as described previously (156). *V. cholerae*-specific PCR amplification of *ompW* was performed using primers ompW-F and ompW-R following the protocol by Nandi et al. (157). For MLSA, fragments of six housekeeping genes (*mdh, adk, gyrB, recA, pgi, rpoB*) were amplified and sequenced as described by Boucher et al. (109). This analysis was performed for seventeen strains of *V. metoecus* sp. nov., as well as three strains of *V. cholerae* (OPA4, OP4C and OP8A). Sequences for reference strains identified by BLAST search of available genomes using OP3H^T sequences as queries and retrieved from GenBank. Also included in the dataset was a single isolate of the yet to be formally described *Vibrio* sp. RC586 (152). Partial sequences of these six genes were concatenated (3593bp in total) and aligned using ClustalW (155) and a Maximum likelihood phylogenetic tree was constructed using RAxML 7.2.8-ALPHA with 100 bootstrap replicates and GTR Gamma substitution model with eight categories (130).

3.3.5 Whole genome sequencing and analysis

Whole genome sequencing of strain OP3H^T was performed by generating single-end pyrosequencing (GS FLX-Titanium; Roche Diagnostics, Indianapolis, IN, USA and single-end Illumina reads (GAIIe sequencer; Illumina, San Diego, CA, USA). Pyrosequencing reads were then assembled de novo by using Newbler version 2.5.3 (Roche Diagnostics). To correct

potential base-calling errors attributed to homopolymers, Illumina GAIIe reads were mapped to the Newbler contigs by using CLC Genomics Workbench version 4.5 (CLC bio, Quiagen).

ANI was calculated between the genome of strain OP3H^T and the previously sequenced genomes of strain RC341 (152) as well as *V. cholerae* N16916 and *V. mimicus* CAIM 602 (Table 3.1) using the ANIb algorithm implemented in JSpecies with standard parameters (36).

3.4 **Results and Discussion**

V. metoecus sp. nov. strains resemble *V. cholerae* in the majority of biochemical and growth characteristics. Giving their highly similar biochemical profiles, these two sister species are differentiated from *V. mimicus* by the same characteristics. Most phenotypic tests normally used to diagnose the identity of an unknown strain as *V. cholerae* yield identical results for *V. metoecus* sp. nov. However, there are some notable exceptions to this phenotypic similarity (Table 3.2). All strains of *V. metoecus* sp. nov. were able to grow using *N*-Acetyl-D-galactosamine and D-glucuronic acid as a sole carbon source. No *V. cholerae* strains have so far been reported to use either substrate. Five of the seven *V. metoecus* sp. nov. strains tested were able to utilize α-cyclodextrin and two tween-40; both abilities were absent in *V. cholerae*. Only one strain of *V. metoecus* sp. nov. displayed a slight ability to infect the slime mold *Dictyostelium discoideum* via virulence associated secretion genes, while strong virulence towards this slime mold is ubiquitous in *V. cholerae.* Furthermore, all tested *V. metoecus* sp. nov. strains were negative for the Voges Proskauer assay, which tests for the production of acetoin from the fermentation of dextrose. Incidentally, this is the only biochemical characteristic test that is shared between *V. metoecus* and *V.mimicus* while also differentiating them from *V.cholerae* (Table 3.2).

V. metoecus sp. nov. can not be differentiated from *V. cholerae* through the amplification and sequencing of the 16S rRNA gene (similarity of 16S rRNA gene sequences between *V. metoecus* sp. nov. type strain OP3H^T and *V. cholerae* strains ranges from 98.4 to 99.8%), or amplification of the ITS region using *V. cholerae* specific primers (both species yield strong products using this standard molecular diagnostic technique for *V. cholerae*) (158). Although these results indicated a very close phylogenetic relationship, the PCR amplification of *ompW*, a gene encoding an outer membrane protein, allowed the differentiations between strains of the two species, being positive for all *V. cholerae* and negative for all *V. metoecus* sp. nov. Also, MLSA allowed a clear phylogenetic delineation of *V. metoecus* sp. nov. from both *V. cholerae*

and *V. mimicus* as well as from the single isolate of the as of yet to be formally described isolate RC586 (152) (Fig. 3.1).

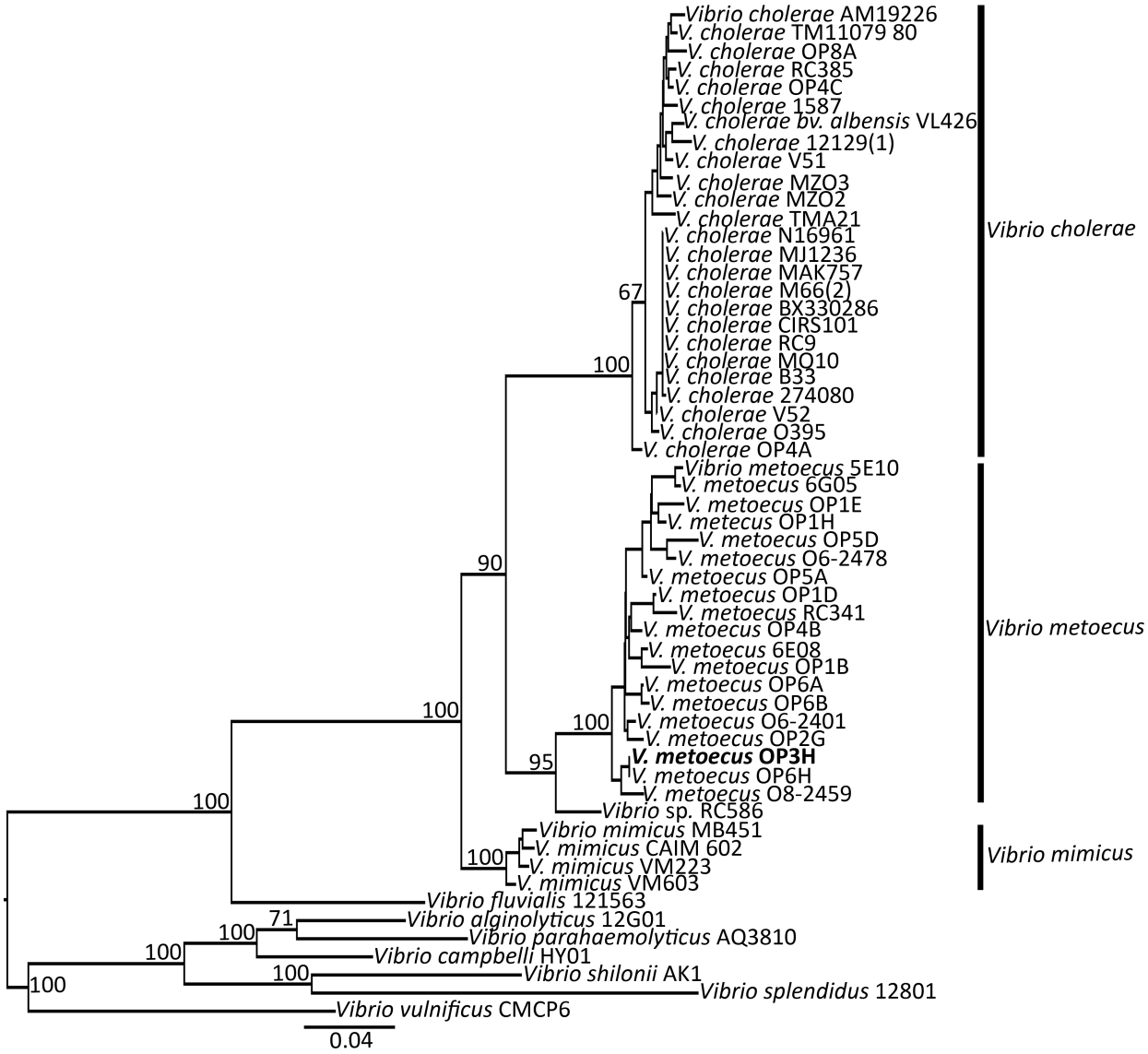


Fig. 3.1: Phylogenetic relationships of *V. metoecus* sp. nov. and its closest relatives based on a concatenated dataset of six partial gene sequences. The tree is derived from a 3593bp alignment composed of protein-coding housekeeping genes mdh, adk, gyrB, recA, pgi and rpoB. It was reconstructed by Maximum Likelihood phylogenetic analysis using RAxML 7.2.8 (GTR Gamma substitution model, eight rate categories). The *V. metoecus* type strain is bolded. Relevant bootstrap values >50% are shown. Bar represents substitutions per site. GenkBank Accession numbers of all sequenced genes are KJ638721-KJ638858.

This further emphasizes the utility of MLSA over the use of the 16S rRNA gene for the identification of *Vibrio* (159, 160). Futhermore, ANI between *V.metoecus* sp. nov strains OP3H^T and RC341 is 98.9%. The ANI between the type strain OP3H^T and *V. cholerae* N16961 is 86.6% while the ANI to *V. mimicus* CAIM 602 is 86.3%. This is far below the threshold for a taxonomic species, which is usually considered to be 95% ANI or above (161).

Thus, despite large overlap with *V. cholerae* in phenotypic characteristics and a 16S rRNA sequence similarity exceeding 98.4% (a value found both between different strains of a single species and between closely related species (18), we propose, based on ANI and phylogenetic analysis, that the isolates described in this study should be assigned to a new species, *Vibrio metoecus* sp. nov. We also note that due to their close resemblance in standard diagnostic characteristics, previous studies might have (mis)identified *V. metoecus* sp. nov. as atypical isolates of *V. cholerae.* The isolation of several strains over a period of more than 10 years from both environmental and clinical samples suggests that our novel species might represent a relatively common but overlooked species of *Vibrio* in North America and perhaps elsewhere.

3.4.1 Description of *Vibrio metoecus* sp. nov.

Vibrio metoecus (met.oe'cus. N.L. masc. n. metoecus (from Gr. n. metoikos), non-resident, stranger.

Gram staining negative, oxidase positive, curved-rod shaped bacterium, roughly 1.25-2μm in length and 0.4μm in width. It exhibits motility by means of a single polar flagellum (Fig. 3.2). DNA G+C content of type strain OP3H^T, based on whole genome sequencing, is 47.73%. The ability to utilize *N*-Acetyl-D-galactosamine and D-glucuronic acid as well as the lack of virulence associated secretion and *ompW* genes and negative Voges Proskauer reaction differentiates the *V. metoecus* sp. nov. type strain from its closest described relative *V. cholerae* (>98% identity of 16S rRNA gene). Growth occurs in LB-broth in a temperature range of 20-40°C, and at 30°C at NaCl concentrations below 8%. *V. metoecus* sp. nov. is capable of producing indole and β-glucosidase as well as reducing nitrate to nitrite; it is positive for glucose fermentation, lysine decarboxylase and ornithine decarboxylase. Produces acid from the fermentation of mannitol but not arabinose. Arginine dihydrolase, and urease negative. Finally, it forms *V.cholerae*-like yellow circular colonies on TCBS agar and flat, smooth, circular colonies of creamy-white colour on TSB agar.

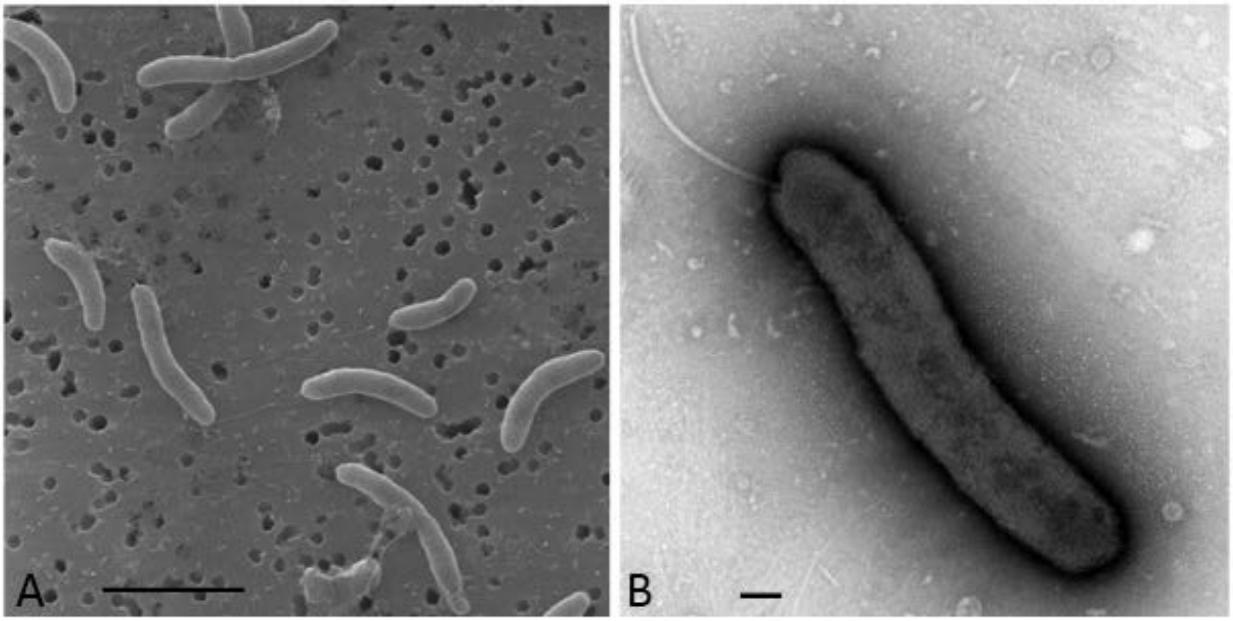


Fig. 3.2: Electron micrographs of *Vibrio metoecus* sp. nov. 06-2478 A) Scanning electron micrograph of bacteria grown in TSB broth, filtered on 0.22μm filters. Scale bar represents 2μm. B) Transmission electron micrograph of bacteria grown on TSB agar plate. Scale bar represents 0.2μm.

All tested strains are positive for carbon utilization from α-D-glucose, β-galactose, citrate, D-fructose, D-galactose, D-gluconic acid, D-glucose, D-glucuronic acid, D-maltose, D-mannitol, D-mannose, D-trehalose, dextrin, gelatin, gluconate, glycogen (except OP1E), malic acid, maltose, *N*-acetyl-D-galactosamine, *N*-acetyl-D-glucosamine sucrose and tartrate. Diversity between strains exists for utilization of α-cyclodextrin, α-ketoglutaric acid, β-methyl-D-glucose, D-psicose, inosine, L-glutamic acid, succinic acid, thymidine, tween 40 and uridine. Not utilized by any isolates are 2-aminoethanol, 2,3-butanediol, adipic acid, L-arginine, β-hydroxybutyric acid, capric acid, cellobiose, cis-aconitic acid, D-alanine, D-galactonic acid lactone, D-glucose-6-phosphate, D-mellibioise, D-serine, D,L-carnitine, D,L,α-glycerol phosphate, formic acid, g-aminobutyric acid, g-hydroxybutyric acid, gentiobiose, glucuronamide, glycyl-L-glutamic acid, hydroxy-L-proline, L-alaninamide, L-alanine, L-arabinose, L-aspartic acid, L-histidine, L-ornithine, L-phenylalanine, L-serine, L-threionine, lactulose, m-inositol, p-hydroxy-phenylacetic acid, phenylacetic acid, propionic acid, putrescine, pyruvic acid methyl esther, quinic acid, sebacic acid, succinamic acid, succinic acid mono-methyl ester, turanose, urea, urocanic acid and xylitol.

The OP3H^T (type strain, = LMG 27764^T = CIP 110643^T) was isolated from Oyster Pond, a brackish water coastal pond in Falmouth, Massachusetts, USA, in 2006.

The GenBank accession numbers for 16S rRNA of OP3H^T as well as *mdh, adk, gyrB, recA, pgi* and *rpoB* for all sequenced strains are KJ647312 (OP3H^T 16S) KJ638721 - KJ638743 (rpoB), KJ638744 - KJ638766 (recA), KJ638767 - KJ638789 (pgi), KJ638790 - KJ638812 (mdh), KJ638813 - KJ638835 (gyrB), KJ638836 - KJ638858 (adk). The GenBank accession number for the type strain OP3H^T genome is JJMN00000000

Table 3.1: Source and year of isolation of *Vibrio metoecus* sp. nov. and *Vibrio cholerae* strains

Sample	Isolation source/location	Year of Isolation	citation
06-2401	fish tank, OH, USA	2006	this study
06-2478	human stool, MS, USA	2006	this study
08-2459	human blood, NC, USA	2008	this study
5E10	water, Oyster Pond, MA, USA	2006	this study
6E08	water, Oyster Pond, MA, USA	2006	this study
6G05	water, Oyster Pond, MA, USA	2006	this study
OP1B	water, Oyster Pond, MA, USA	2006	this study
OP1D	water, Oyster Pond, MA, USA	2006	this study
OP1E	water, Oyster Pond, MA, USA	2007	this study
OP1H	water, Oyster Pond, MA, USA	2006	this study
OP2G	water, Oyster Pond, MA, USA	2006	this study
OP3H^T	water, Oyster Pond, MA, USA	2006	this study
OP4B	water, Oyster Pond, MA, USA	2006	this study
OP5A	water, Oyster Pond, MA, USA	2006	this study
OP5D	water, Oyster Pond, MA, USA	2006	this study
OP6A	water, Oyster Pond, MA, USA	2006	this study
OP6B	water, Oyster Pond, MA, USA	2006	this study
OP6H	water, Oyster Pond, MA, USA	2006	this study
RC341	water, Chesapeake Bay estuary, MD, USA	1998	Haley et al. (2010)(3)
RC586	water, Chesapeake Bay estuary, MD, USA	1998	Haley et al. (2010)(3)
CAIM 602*	Human ear, NC, USA	1981	Davis et al. (1981)(18)
V52**	human stool, Sudan	1968	Heidelberg et al.(2010)(19)
N16961**	human stool, Bangladesh	1971	Heidelberg et al.(2010)(19)
OP4A***	water, Oyster Pond, MA, USA	2006	this study
OP4C***	water, Oyster Pond, MA, USA	2006	this study
OP8A***	water, Oyster Pond, MA, USA	2006	this study

* This isolate represents a the type strain of *V. mimicus*

** These isolates represent clinical isolates of *V. cholerae*

*** These isolates represent environmental isolates of *V. cholerae*

Table 3.2: Differentiating characteristics of *Vibrio metoecus* sp. nov. from its closest relatives *Vibrio cholerae* and *Vibrio mimicus*

	V. metoecus sp. nov. (1-7)							*V. cholerae* (9-13)					*V.mimicus* (14)
Test	1	2	3	4	5	6	7	9	10	11	12	13	14
N-Acetyl-D-galactosamine*	+	+	+	+	+	+	+	-	-	-	-	-	-
D-glucuronic acid*	+	+	+	+	+	+	+	-	-	-	-	-	-
Acetoin production	-	-	n	n	-	n	-	+	+	+	+	+	-
Amplification of *ompW*	-	-	-	-	-	-	-	+	+	+	+	+	-
Dictyostelium virulence**	-	-	-	-	-	-	-	+	+	+	+	+	n
α-cyclodextrin*	-	+	+	+	+	-	+	-	-	-	-	-	-
Tween 40*	-	-	+	+	v	-	v	-	-	-	-	-	-
β-methyl-D-glucoside*	+	+	-	v	-	-	-	(+)	v	v	v	v	-
D-psicose*	+	+	-	(+)	-	-	(+)	d	(+)	v	(+)	-	-
α-ketoglutaric acid*	-	+	-	(+)	(+)	-	(+)	(+)	-	-	-	-	-
Succinic acid*	v	-	(+)	(+)	(+)	-	+	(+)	(+)	-	-	-	+
L-glutamic acid*	+	(+)	(+)	(+)	(+)	(+)	(+)	(+)	-	-	-	v	-
Inosine*	+	+	(+)	(+)	(+)	(+)	(+)	+	(+)	-	-	-	-
Uridine*	+	+	(+)	(+)	v	(+)	(+)	(+)	(+)	-	-	-	-
Thymidine*	+	-	-	-	v	-	-	-	v	-	-	-	-
Glycogen*	v	+	+	+	+	-	+	+	+	+	+	+	-
D-fructose*	+	+	+	+	+	+	+	+	(+)	(+)	+	(+)	-
D-galactose*	+	+	+	+	+	+	+	+	v	v	v	+	-
D-mannitol*	+	+	+	+	+	+	+	+	+	(+)	v	(+)	-
D-mannose*	+	+	+	+	+	+	+	+	+	(+)	+	+	-
Sucrose*	+	+	+	+	+	+	+	+	+	(+)	+	(+)	-
Succinic acid mono-methyl ester*	-	-	-	-	-	-	-	-	-	-	-	-	+
D-gluconic acid*	+	+	+	+	+	+	+	+	(+)	v	+	+	-
Succinic acid*	+	+	+	+	+	+	+	+	(+)	v	+	+	-
D,L,α-glyerol phosphate*	-	-	-	-	-	-	-	-	-	-	-	-	+
Tartrate utilization	+	+	+	+	+	+	+	+	+	+	+	+	-

Strains 1-8 are *Vibrio metoecus* sp. nov. 1: 08-2459, 2: 06-2478, 3: OP6B, 4: OP5A, 5: OP4B, 6: OP1E, 7: OP3H^T. 9-12 *Vibrio cholerae* 9: N16961, 10: V52, 11: OP8A, 12: OP4C and 13: OP4A. 14 *Vibrio mimicus* CAIM 602, results from Nishiguchi and Nair (2003). + Growth/positive test result, (+) weak growth/weakly positive, - no growth/negative test result, v variable results between triplicates, n not determined. * Refers to positive result on Biolog GN2

plate on specific substrate. ** of 11 tested *V. metoecus* sp. nov strains, only OP6H showed signs of low-level virulence towards *Dictyostelium discoideum*

Chapter 4

Sequential displacement of Type VI Secretion System effector genes leads to evolution of diverse immunity gene arrays in *Vibrio cholerae*

4.1 Abstract

Type VI secretion systems (T6SS) enable bacteria to engage neighboring cells in a contact-dependent competition. In *Vibrio cholerae*, three chromosomal clusters each encode a pair of effector and immunity genes downstream of those encoding the T6SS structural machinery for effector delivery. Different combinations of effector-immunity proteins lead to competition between strains of *V. cholerae*, which are thought to be protected only from the toxicity of their own effectors. Screening of all publically available *V. cholerae* genomes showed that numerous strains possess long arrays of orphan immunity genes encoded in the 3' region of their T6SS clusters. Phylogenetic analysis reveals that these genes are highly similar to those found in the effector-immunity pairs of other strains, indicating acquisition by horizontal gene transfer. Extensive genomic comparisons also suggest that successive addition of effector-immunity gene pairs replaces ancestral effectors, yet retains the cognate immunity genes. The retention of old immunity genes perhaps provides protection against nearby kin bacteria in which the old effector was not replaced. This mechanism, combined with frequent homologous recombination, is likely responsible for the high diversity of T6SS effector-immunity gene profiles observed for *V. cholera*e and closely related species.

4.2 Introduction

The family Vibrionaceae consists of over 100 related species of highly motile, heterotrophic bacteria that enzymatically convert inaccessible organic matter found in aquatic environments into carbon sources available to higher trophic levels of the ecosystem they inhabit (33). Numerous mostly harmless lineages of *Vibrio* coexist within niches, competing for largely similar resources(162). Among them are a few human pathogens of relevance, including *Vibrio cholerae,* the causative agent of the sometimes dramatic and lethal cholera diarrhea. More specifically, a single lineage of the *V. cholerae* species, comprised primarily of O1 and O139 serogroup strains (163), has adapted to effectively colonize the human gastrointestinal tract and

is responsible for all known cholera pandemics (103). Pandemic *V. cholerae* strains harbor the horizontally acquired genetic elements VPI-I and CTX-Φ encoding the toxin co-regulated pilus and cholera toxin respectively. These virulence factors enable pandemic strains to colonize the crypts of villi in the small intestine, causing watery purges of diarrhea and releasing billions of pathogenic bacteria into the environment (164) Thus, pathogenic *V. cholerae* lead a dual lifestyle: One that requires the ability to pursue, attach and colonize biotic surfaces in a relatively oligotrophic aquatic environment of low osmolarity, and another that requires the successful colonization of a eutrophic, biochemically challenging human intestine populated by a highly diverse commensal host flora (114).

In both of these competitive environments, V. *cholerae* is believed to actively employ their Type VI secretion system (T6SS), which is induced by chitin in the environment (93) and by bile salts in the gut (165). The T6SS is a membrane-spanning nanomachine capable of injecting toxin-tipped protein spears into adjacent eukaryotic and bacterial target cells (166, 167). The T6SS spear consists of Hcp multimers tipped by a VgrG (hetero)trimer and effector proteins with varying cytotoxic effects (168). For example, the VgrG-1 protein of some *V. cholerae* strains harbors a C-terminal domain that mediates crosslinking of cytoskeletal actin fibers in eukaryotic cells (such as predatory amoebae or macrophages), leading to cell rounding and death (169, 170). VgrG-3, on the other hand, displays antibacterial properties by degrading prokaryotic peptidoglycan, and is also an important factor in the colonization of the human intestine (171, 172). Additionally, so-called cargo effectors can be loaded onto the Hcp-VgrG spear, further expanding the toxic capabilities of the T6SS (173).

Unterweger et al. (141) found that a multitude of diverse T6SS effector-immunity (EI) gene modules are encoded in different *V. cholerae* genomes. Effector proteins are placed as cargo onto the T6SS-spear by an adaptor protein, while immunity proteins remain inside the cell and prevent intoxication by incoming cognate effectors (174). The resulting “poisoned” spear proves lethal to target cells that do not possess an EI module of the same type (141). Through this system, strains of *V. cholerae* are not only able to attack eukaryotes and bacteria belonging to different species, but also their perhaps strongest competitors, non-kin strains of the same species (175). Unterweger *et al.* (141) established a three-letter system for typing *V. cholerae* T6SS variants based on their EI modules. Different letters designate unique EI gene families (as defined by a 30% amino acid identity of immunity proteins) encoded in three genomic clusters: aux-1 (A and C), aux-2 (A-E) and the large cluster (A-G). In the case of the large cluster, the effector is a domain at the 3’ end of *vgrG-3,* not a separate gene (Figure 4.1).

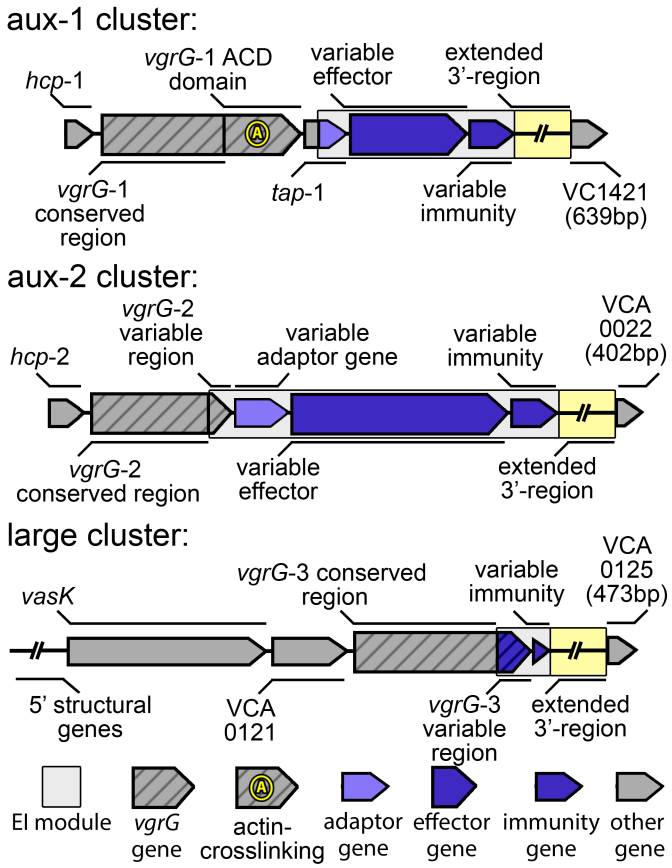


Figure 4.1: Schematic organization of *V. cholerae* T6SS clusters. Striped arrows denote genes encoding VgrG effectors; non-striped coloured arrows denote variable effector (large arrows) or immunity genes (small arrows); grey arrows indicate conserved genes based on cluster tags from the reference genome of V. cholerae N16961. Not pictured: VCA0122, the coding region of which is frequently interrupted by deletions. Boxed region denotes the EI module. Extended 3'-region is of variable length and gene content, and all coloured genes vary in length. Large cluster extends further upstream than shown in figure.

Strains with identical EI module composition belong to the same compatibility groups, whereas those that possess different EI modules are T6SS incompatible. For example, most *V. cholerae* from the lineage containing pandemic strains belong to the AAA type (note that the same letter for different clusters does not denote the same gene family) and can co-exist among each other (they are "compatible"). In contrast, strains of the AAA-type engage in T6SS-mediated competitive interactions with strains of different groups such as CAG or AAC, and are thus "incompatible" with them (141). The AAA genotype appears to be the most effective for intraspecies competition under laboratory conditions, as suggested by the ability of *V. cholerae* V52 (an O37 serogroup strain from the same lineage as O1 serogroup pandemic strains) to outcompete any strain with a different EI module combination (141). In addition to each *V.*

cholerae strain possessing three EI modules at conserved genomic clusters, recent studies have found further EI modules encoded on horizontally transferred genomic islands that potentially expand the T6SS mediated competitive abilities of *V. cholerae* even further (176, 177).

The large number of distinct EI modules may indicate an ongoing evolutionary arms race to succeed in intra- and interspecies competition and to overcome eukaryotic host defenses. Such intense selective pressure could facilitate not only rapid mutational divergence of effectors and immunity proteins of different lineages, but also horizontal gene transfer, either as a whole or in parts, giving rise to new variants.

To elucidate the evolutionary dynamics of the *V. cholerae* T6SS, we performed a systematic survey of the T6SS-harboring genomic regions in *V. cholerae* and its closest relatives among the Vibrionaceae. This led to the discovery of additional putative effector and immunity genes in the 3'-region of several *Vibrio* T6SS clusters. Additionally, we find evidence that insertion of distinct EI modules replaces old effectors, yet often retains the immunity genes, leading to an array of multiple different orphan immunity genes and the establishment of new types of mosaic T6SS regions. We also provide evidence that such modular insertion may have given rise to the unique combination of effector and immunity genes found in pandemic *V. cholerae* strains.

4.3 Material & Methods

4.3.1 Identification and annotation of T6SS clusters in *Vibrio* species

Initial screening for T6SS clusters in *Vibrio cholerae* and the closely related *Vibrio metoecus*, *Vibrio mimicus*, *Vibrio fluvialis*, and *Vibrio furnissii* was conducted by performing megaBLAST searches against all genomes of these species (*V. cholerae*: 548, *V. metoecus*: 10, *V. mimicus*: 10, *V. fluvialis*: 8, *V. furnissii*: 4) available on NCBI. Genes VC1421, VCA0022 and VCA0125 of *V. cholerae* strain N16961, each located downstream of one of the three T6SS clusters, were used as conserved, single copy query sequences to identify contigs containing aux-1, aux-2 and large T6SS region genes, respectively. T6SS regions were then located on the extracted contigs by mapping them against N16961 *vgrG*-1, *vgrG*-2 and *vgrG*-3 (VC1416, VCA0018 and VCA0123) in Geneious 8 (119). Homologous genes within those regions were then identified by extracting all ORFs >300 bp and conducting all-vs-all local BLASTP searches on translated sequences. Hits with a minimum of 30% protein sequence identity were considered homologous and annotated accordingly. To ensure completeness of our initial genomic survey, all identified

putative effector and immunity genes found in our first round of searches were then used as query for a second round of megaBLAST searches, and additional T6SS regions found by these searches were added to the dataset. This dataset was then trimmed to 95 genomes (Supplementary Table S4.1) by eliminating all genomes that did not show any nucleotide divergence in the identified T6SS clusters. Additional megaBLAST searches for the mobile antibacterial *tseH-tsiI* EI module (176) and T6SS cluster identified by Labatte et al. (177) revealed their presence in various *V. cholerae* strains.

4.3.2 Phylogenetic analysis

From the finalized selection of genomes, a pangenome k-mer SNP dataset was extracted using kSNP3.0 with an inferred optimal k-mer size of 19 (178). This dataset contained a set of 1,085,207 19-mers found in at least one genome (with absences in other genomes denoted as missing data).

A maximum-likelihood phylogenetic tree was then calculated using RAxML 8.0 using the GTRGAMMA substitution model, and statistical support was estimated based on 100 bootstrap replicates (130). This whole-genome rather than core-genome approach provides increased phylogenetic resolution when comparing a large number of closely related genomes (belonging to a single species) while also including more distantly related genomes in the dataset. In a core-genome approach, the inclusion of more distantly related genomes or incomplete draft genomes (only a small number of *Vibrio* genomes are complete) leads to a loss of resolution since phylogenetically informative characters needed for the differentiation of closely related genomes are removed due to their absence in some genomes. The whole genome approach becomes more problematic over longer evolutionary timeframes where the cumulative effect of horizontal gene transfer might trump the higher incidence of vertical descent in closely related genomes(179). To avoid potentially false inference of relationships in more ancestral branches of the tree, we collapsed bipartitions with bootstrap support lower than 70.

Alignments of single genes or gene regions were performed using the CLUSTALW (127) plugin in Geneious 8 using standard settings and then manually edited in Geneious 8. Phylogenetic trees were generated from these alignments using RAxML as described above.

4.3.3 Recombination analysis

In order to infer putative regions within T6SS clusters that have undergone horizontal gene transfer, recombination analysis was performed using RDP4 (126). Four different algorithms implemented in RDP4 were used: GENECONV, RDP, MAXCHI and CHIMAERA. Briefly, GENECONV (180) identifies regions of sequence pairs in an alignment with significantly lowered amount of nucleotide polymorphisms compared to the whole region. The RDP algorithm (126) performs a sliding window analysis along triplet sequences in an alignment and identifies regions of high sequence similarity incongruent with an UPGMA dendrogram created from the entire alignment. MAXCHI (181) and (in a modified form) CHIMAERA (182) identify putative recombination breakpoints by moving a bi-partitioned sliding-window along a sequence pair and detecting significant differences in sequence similarity between the two sides of the window.

Since different algorithms do not always identify the same recombination events, only events detected by at least three out of four algorithms were counted as valid. Detection of regions affected by homologous recombination and identification of recombination breakpoints was conducted on alignments of each cluster type separately (i.e. an alignment of A-type aux-1 clusters), as the presence of large non-homologous alignment regions impairs correct identification. For the same reason, regions of the same type that contained additional genes, such as multiple copies of immunity proteins that are not present in the majority of sequences, were also left out.

4.3.4 Whole genome sequencing and assembly

Isolation and DNA extraction of *V. cholerae* strains DL4211 and DL4215 from the Rio Grande estuary was described in a previous study (175). Whole genome sequencing was performed by Ambry Genetics (CA, USA) using 100bp paired-end Illumina HiSeq 2000 technology after following the TruSeq DNA sample preparation guidelines. De-novo assembly of reads into contiguous sequences was then conducted using CLC Genomics Workbench 5.0 (CLC Bio, Aarhus, Denmark). The two draft genomes were submitted to NCBI GenBank and given the accession numbers MOLL00000000.1 (DL4211) and MOLM00000000.1 (DL4215).

4.4 Results and Discussion

4.4.1 T6SS cluster structure is conserved in *V. cholerae* and its closest relatives

We identified and annotated T6SS clusters in the publically available genomes of *V. cholerae*, and its four closest relatives *V. metoecus* (118), *V. mimicus* (183), *V. furnissii* and *V. fluvialis (33)*. All investigated strains possess the same three clusters structure as previously described for *V. cholerae* (169) (Figures 4.1 and 4.2): one large cluster and two auxiliary clusters (here termed aux-1 and aux-2). The large cluster includes 17 (or 18, depending on the presence of the regulator protein VCA0122 (167)) structural genes that encode proteins forming the membrane-spanning machinery of the T6SS, which also contains the *vgrG*-3 gene and an immunity gene encoding a protein protecting against the antibacterial activity of the VgrG-3 C-terminus (172, 184). The aux-1 and aux-2 clusters share a common structure: an allele encoding the secreted Hcp protein, a VgrG protein (VgrG-1 and VgrG-2 in aux-1 and aux-2, respectively) followed by an adaptor protein, an additional effector and a cognate immunity protein (141) (Figure 4.1). VgrG-2 proteins, previously described as differing from VgrG-1 and VgrG-3 due to their lack of variable C-terminus (185), were found to also encode a variable region of around 60 amino acids in length. Although no known functional domain was identified in this region, the sequence of this variable C-terminus varies considerably between strains carrying different EI module types at the aux-2 cluster (with the exception of A and B-types, which carry similar C-termini) (Fig. 4.3a). Each specific combination of VgrG-2 C-terminus and effector-immunity protein pair is also accompanied by a specific putative adaptor protein (Fig. 4.4). In light of the functional link of the 3'-end of the aux-1 Tap-1 protein with cargo effectors (174) as well as the aux-2 A-type adaptor gene *vasW* with the A-type effector *vasX* (186), it is possible that both the VgrG-2 variable C-terminus and effector-specific adaptor proteins encoded in aux-2 are involved in the loading of effector proteins onto VgrG-2. For this reason, we include them in our definition of an EI module (Figure 4.1).

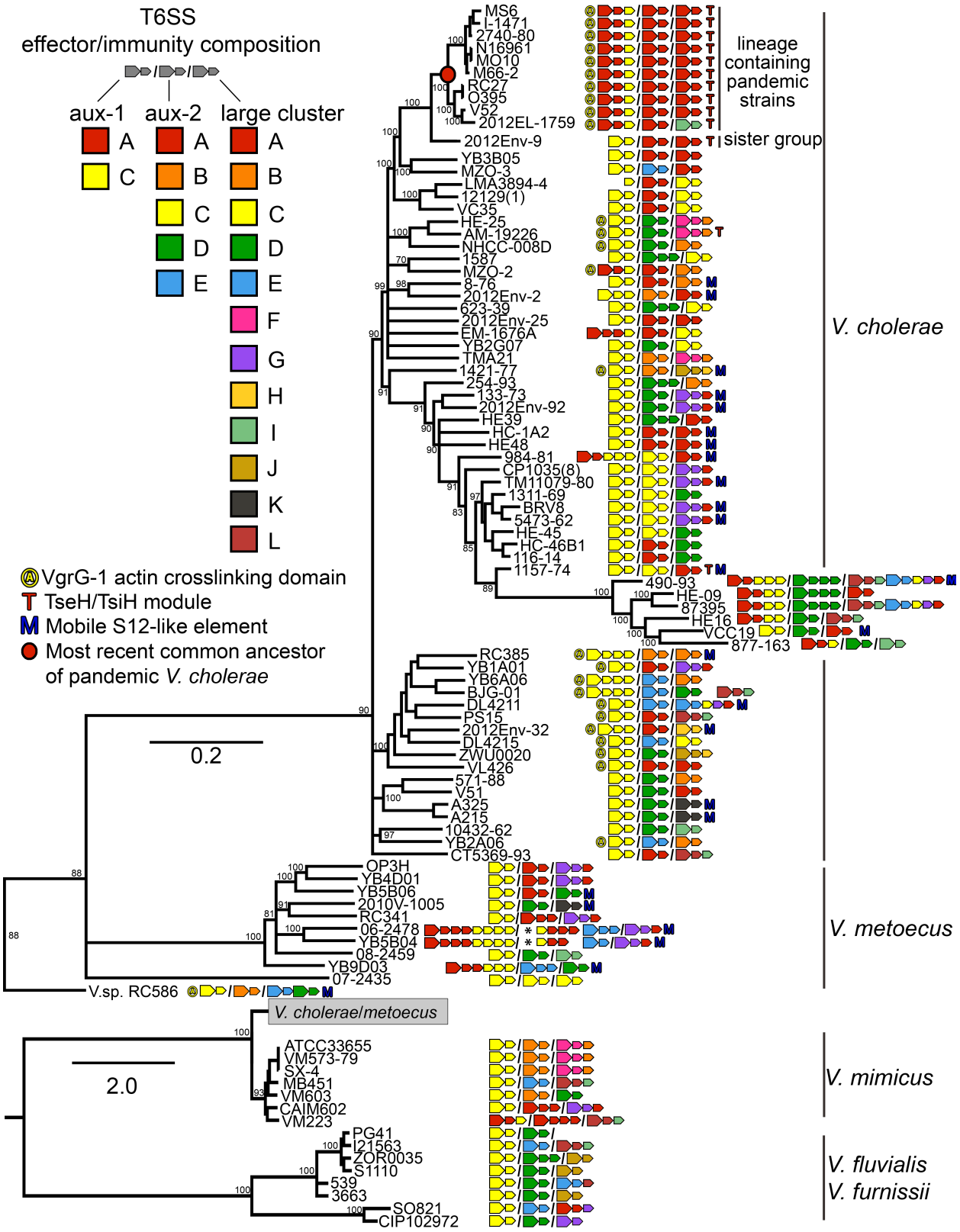


Figure 4.2: Whole-genome phylogeny and T6SS EI module composition of *Vibrio cholerae* and closely related species. Large arrows next to strain names indicate effector,

small arrows immunity genes. Auxiliary cluster 1, 2 and the large cluster are separated by slashes. Asterisks indicate transposons. The phylogenetic tree was calculated using the GTR+Gamma Maximum likelihood model implemented in RAxML based on a 1,085,207 pangenome 19-mer alignment (including not just characters shared by all, but by at least two genomes) created using kSNP3. V. cholerae/V. metoecus are visualized separately for better visibility of short internal branches. Statistical branch support was obtained from 100 bootstrap repeats. Bootstrap support for relevant bipartitions is indicated, and branches with support <70 were collapsed. Scale bar indicates substitutions/site. Accession numbers of genomes are listed in Supplementary Table S4.1.

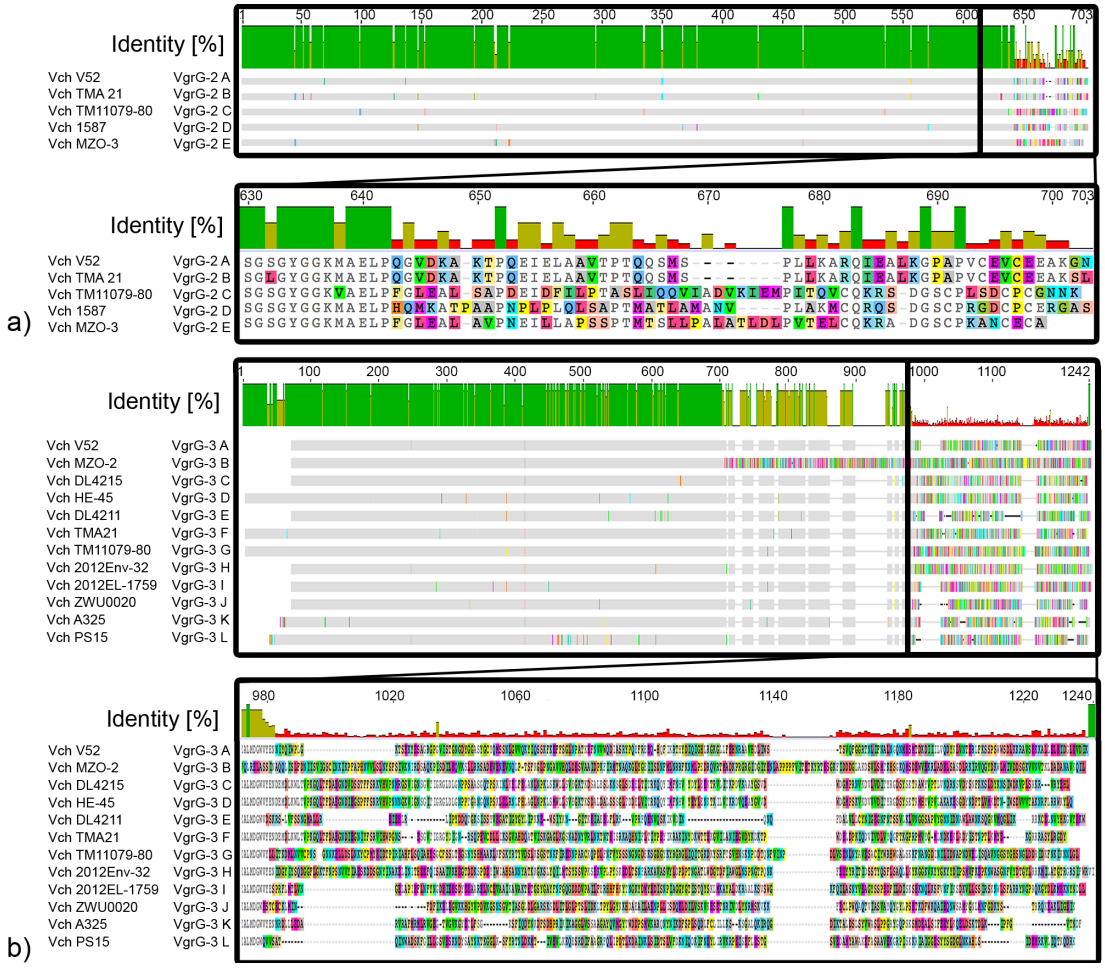


Figure 4.3: Amino acid alignment of VgrG-2 and VgrG-3. a) Alignment of VgrG-2 proteins found in aux-2 cluster encoding different types of cargo effectors b) Alignment of VgrG-3 proteins in large cluster encoding different variable C-terminal regions. Conserved sites are depicted as grey, while colour indicates amino acid change compared to the consensus. Variable C-terminal regions of VgrG-2 and VgrG-3 are enlarged. Vch = Vibrio cholerae.

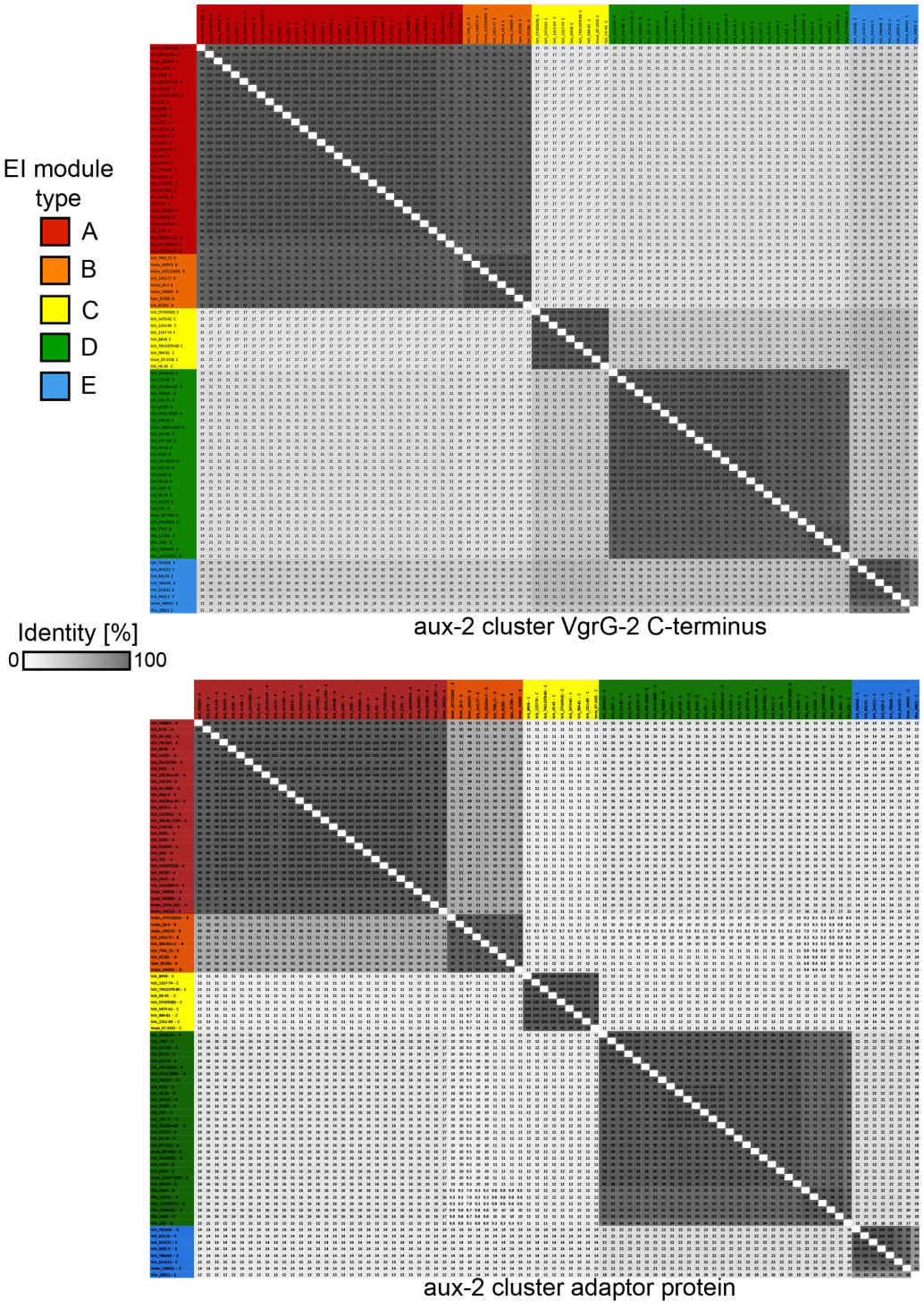


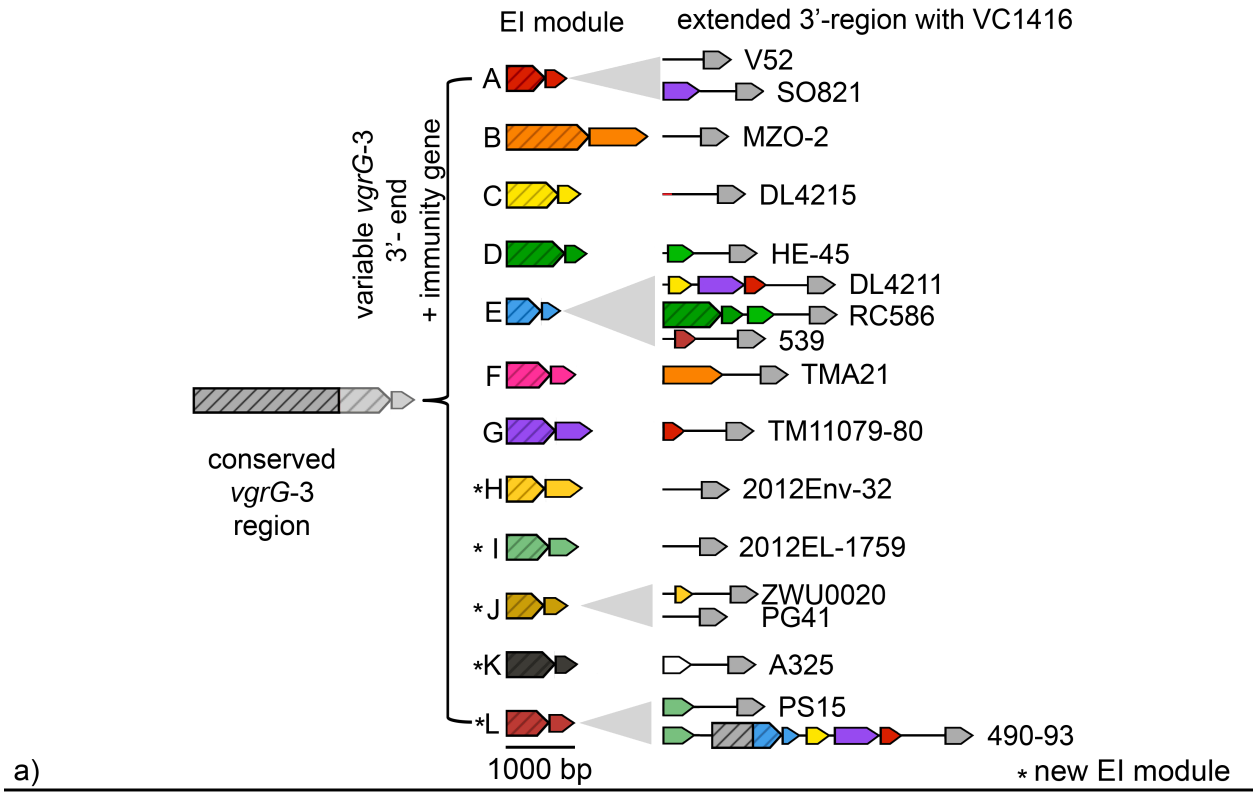
Figure 4.4: Amino acid identity of variable region of VgrG-2 and aux-2 adaptor proteins. Heat maps indicate percent identity of aligned protein sequences found in investigated strains. Colouring of strain names indicates type of EI module present in the aux-2 cluster of the respective strain. Pseudogenes with interrupted reading frames were not included in the comparison. Vch = Vibrio cholerae, Vmet = V. metoecus, Vmim = V. mimicus, Vfur = V. furnissii, Vflu = V. fluvialis, Vpar = V. sp. RC586.

We also discovered five novel variable 3'-ends of *vgrG*-3 encoding putative effector domains associated with the large T6SS cluster, along with their corresponding immunity protein coding genes (Fig 4.5a, Fig 4.3b and Fig 4.6). In accordance with Unterweger *et. al*'s nomenclature (141), these were named H-L, following the previously described A-G types. Only two of the novel effector domains corresponded to known proteins or could be assigned a putative function. The I-type effector contains a DUF3380/pfam11860 domain, which is annotated as a phage-derived peptidoglycan binding/muraminidase protein. The K-type effector contains a lambda phage derived lysozyme (cd00736/COG4678).

4.4.2 Multiple additional immunity genes can be present downstream of effector-immunity modules

The regions between the canonical EI modules and the conserved genes VC1421, VCA0022 and VCA0125 (downstream of aux-1, aux-2 and the large cluster, respectively) vary considerably in number and type of genes between closely related strains (Figure 4.2). Parts of these extended 3'-regions are homologous to previously described immunity genes, but not necessarily to those corresponding to the strain's cognate effector protein (Fig 4.5a). In other words, the regions downstream of the EI module in each T6SS cluster in many cases appear to consist of arrays of alternate immunity genes that could cumulatively confer not only resistance to a strain's own, but also to a number of different additional effectors.

The aux-1 cluster of *V. cholerae* may harbor one of two types of EI modules, either the A or C. In our analysis, the EI module previously identified as a B-type in strain LMA3894-4 (141) appears to be a divergent C-type resulting from a fusion of the *vgrG*-1 and C-immunity gene lacking an effector. While some strains contain up to three C-type immunity genes (Figure 4.2), their presence is independent of whether the strain contains an A or C-type effector gene. C-type immunity genes are in fact universal at the aux-1 cluster, either as part of the strains EI module or their extended 3' region. *V. metoecus*, thus far isolated exclusively from North American coastal environments and a small number of blood and stool samples from the United States(118), harbors multiple C-type and up to three A-type immunity genes. In some *V. metoecus* genomes, the total number of A- and C-type immunity genes in the aux-1 cluster can be as high as seven (Figure 4.2).


EI module
extended 3'-region with VC1416
variable vgrG-3
3'- end
+ immunity gene
A
V52
SO821
B
MZO-2
C
DL4215
D
HE-45
E
DL4211
RC586
539
F
TMA21
G
TM11079-80
conserved
vgrG-3
region
*H
2012Env-32
* I
2012EL-1759
*J
ZWU0020
PG41
*K
A325
*L
PS15
490-93
1000 bp
a)
* new EI module

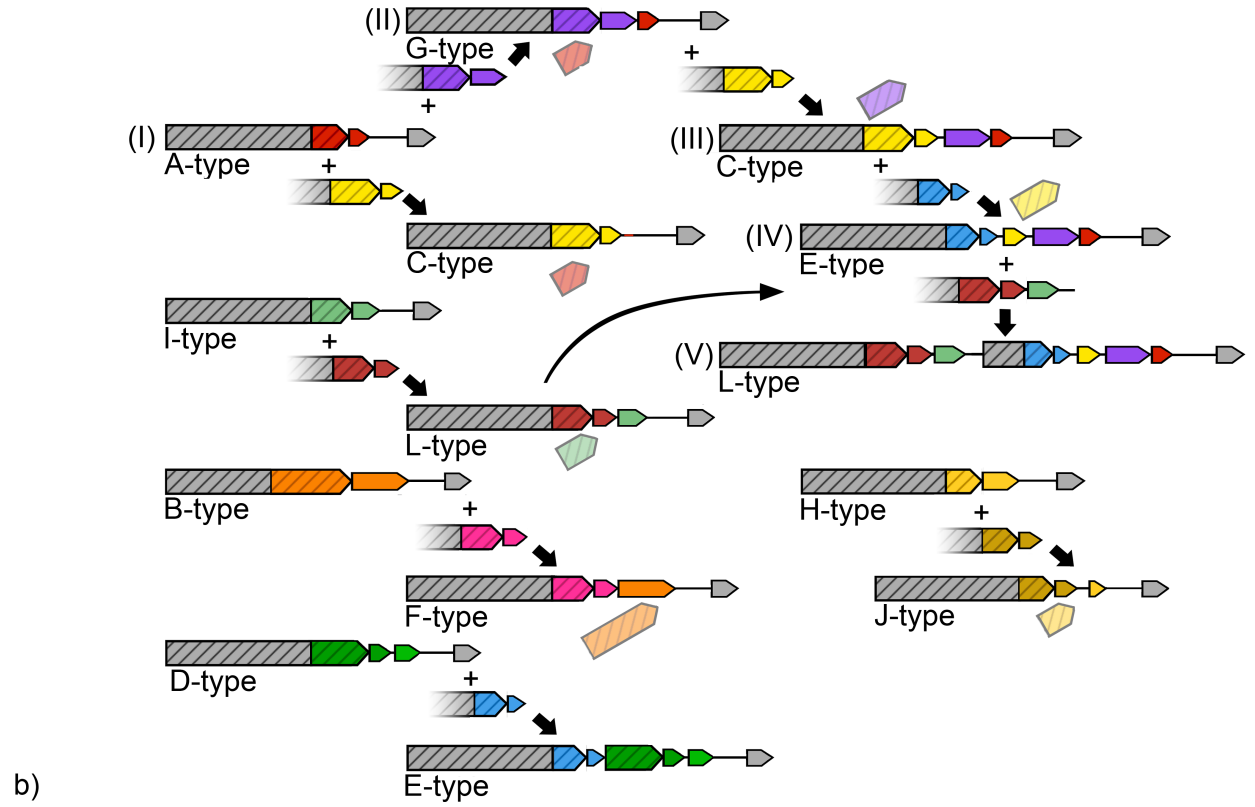

(II)
G-type
+
+
(I)
A-type
+
(III)
C-type
+
C-type
(IV)
E-type
+
I-type
+
(V)
L-type
L-type
B-type
+
H-type
+
F-type
J-type
D-type
+
E-type
b)

Figure 4.5: Organization and evolution of the *Vibrio cholerae* large T6SS cluster EI modules. a) Large cluster EI module organization and variability. Conserved 5'- region of vgrG-3 can be tipped with different effector encoding 3'-regions and cognate immunity genes. b) Recombinatorial reshuffling of vgrG-3. Cluster evolution proceeds by insertion of EI module and variable length of conserved vgrG-3 region into ancestral vgrG-3 gene. New EI module often (but not always) replaces original effector and shifts immunity gene(s) downstream. Integration of more complex EI clusters can result in larger clusters containing multiple EI components. Arrows indicate coding sequences, lines noncoding regions. Striped arrows denote genes encoding VgrG-3 effectors, non-striped arrows depict variable effector (large arrows) or immunity genes (small arrows). Colours indicate homology between either effector or immunity genes. Identically coloured effector and immunity genes are part of the same EI module. Grey arrows indicate conserved genes. Exact genomic locations can be found in Supplementary Table S4.2.

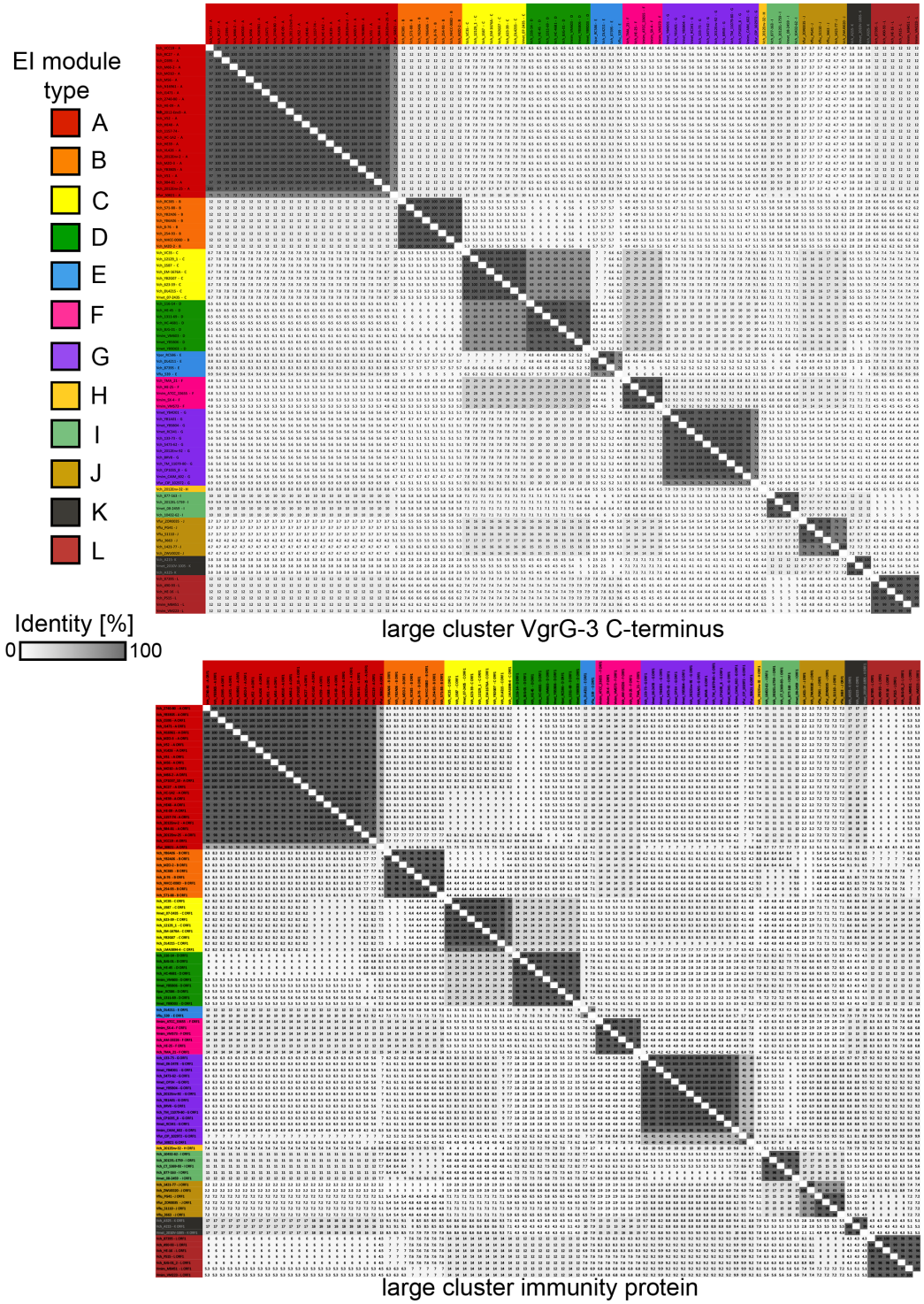


Figure 4.6: Amino acid identity of variable region of VgrG-3 and cognate immunity gene. Heat maps indicate percent identity of proteins found in investigated strains. Colouring of strain names indicates type of EI module present in the large cluster of the respective strain. Pseudogenes with interrupted reading frames were not included in the comparison. Vch = Vibrio cholerae, Vmet = V. metoecus, Vmim = V. mimicus, Vfur = V. furnissii, Vflu = V. fluvialis, Vpar = V. sp. RC586.

In the aux-2 cluster of *V. cholerae*, for which five EI pairs have been described (A-E), the number of immunity genes varies only for the D-type immunity protein. Similar to the aux-1 cluster, most genomes contain just a single aux-2 immunity gene (matching the effector found upstream), but several strains contain up to three. Additionally, *V. mimicus* strain CAIM602 possesses two A-type immunity proteins and *V. mimicus* VM223 harbours three. The most complex arrays are again observed in *V. metoecus*. The aux-2 region in two strains (YB5B04 and 06-2478, also containing seven immunity genes in the aux-1 cluster) appears to be disrupted by a transposon insertion: the extended 3'-region containing A- and C-type immunity genes and further genes typically found downstream of their E-type EI module are located in the vicinity of a transposase in a different region of their genomes (Figure 4.2).

The T6SS large cluster, like the auxiliary clusters, can contain additional immunity genes of a different type than the one found in the canonical EI module. In a few instances, effectors matching these additional immunity genes are also found in the extended 3'-region (Figure 4.5a). For example, the closely related *V. cholerae* isolates 87395 and 490-93 harbor an L-type EI module, followed by an extended 3'-region containing an I-type immunity gene, a short conserved region usually found downstream of normal effectors, a partial E-type VgrG-3 effector protein, the cognate E-type immunity gene, a C-type immunity gene, a G-type immunity gene and finally an A-type immunity gene (Figures 4.2 and 4.5).

4.4.3 Horizontal gene transfer of effector-immunity modules

Like in Proteobacteria in general (187), effector and immunity gene distribution mapped on the phylogeny of *Vibrio* shows both patterns of vertical inheritance as well as horizontal gene transfer (HGT) (Figure 4.2). The vast majority of aux-1 EI-modules in non-pandemic strains of *V. cholerae* are of the previously described C-type (141). Despite being rare in *V. cholerae* overall, aux-1 A-type EI modules are ubiquitous in the *V. cholerae* clade containing pandemic strains as well as a divergent clade containing strains 490-93 and 877-163, indicating independent acquisitions in the ancestors of these two clades and subsequent vertical inheritance (Figure 4.2). A few strains outside of these clades, as well as strains of *V. metoecus* and *V. mimicus,* also possess this type of module. The VgrG-1 protein containing an actin-crosslinking domain also displays a distribution pattern indicative of horizontal gene transfer: It is only found in few divergent lineages of *V. cholerae* (including the lineage containing pandemic strains) while virtually absent in *V. metoecus*, *V. mimicus*, *V. furnissii* and *V. fluvialis* (Figure 4.2). Instead, a

truncated version of VgrG-1 is present in most investigated strains. The truncated protein is fully functional in its antibiotic activity (174), yet lacks the C-terminal actin-crosslinking domain that is involved in cytotoxicity against human macrophages and predatory *Dictyostelium* slime molds (169). This disparate distribution of the A-type EI module as well as the VgrG-1 actin-crosslinking domain suggests that they were, like the major *V. cholerae* virulence factors *tcp* or *ctx* (108, 188), independently introduced into various lineages by horizontal gene transfer.

Aux-2 EI modules C and E as well as large cluster modules C, D, F and G show similar phylogenetically disparate distributions. For example, the aux-2 C-type and the large cluster D-type EI modules are predominantly found in a single, well-supported clade of *V. cholerae* (containing among others the atypical O1 serogroup strain TM11079-80) as well as occasionally in distantly related genomes (Figure 4.2). HGT is also apparent in single gene phylogenies of effector and immunity genes (Fig 4.7-4.10): alleles from closely related strains often fall into different gene clusters.

Disparate distributions are also found for additional recently discovered EI modules located outside of the three known *Vibrio* T6SS clusters. The *tseH-tseI* EI module previously described by Altindis et al. (176) rarely appears outside the *V. cholerae* pandemic group (Figure 4.2). This EI module is located 3.5kb upstream of the chromosomal integron region of *V. cholerae* and possesses antibacterial activity (176). An additional auxiliary cluster containing *hcp*, a *vgrG*-allele and what appears to be a novel EI module, was recently found encoded on a genomic island termed GI*Vch*S12 (177). EI modules of the GIVchS12-type are irregularly distributed as well (Figure 4.2).

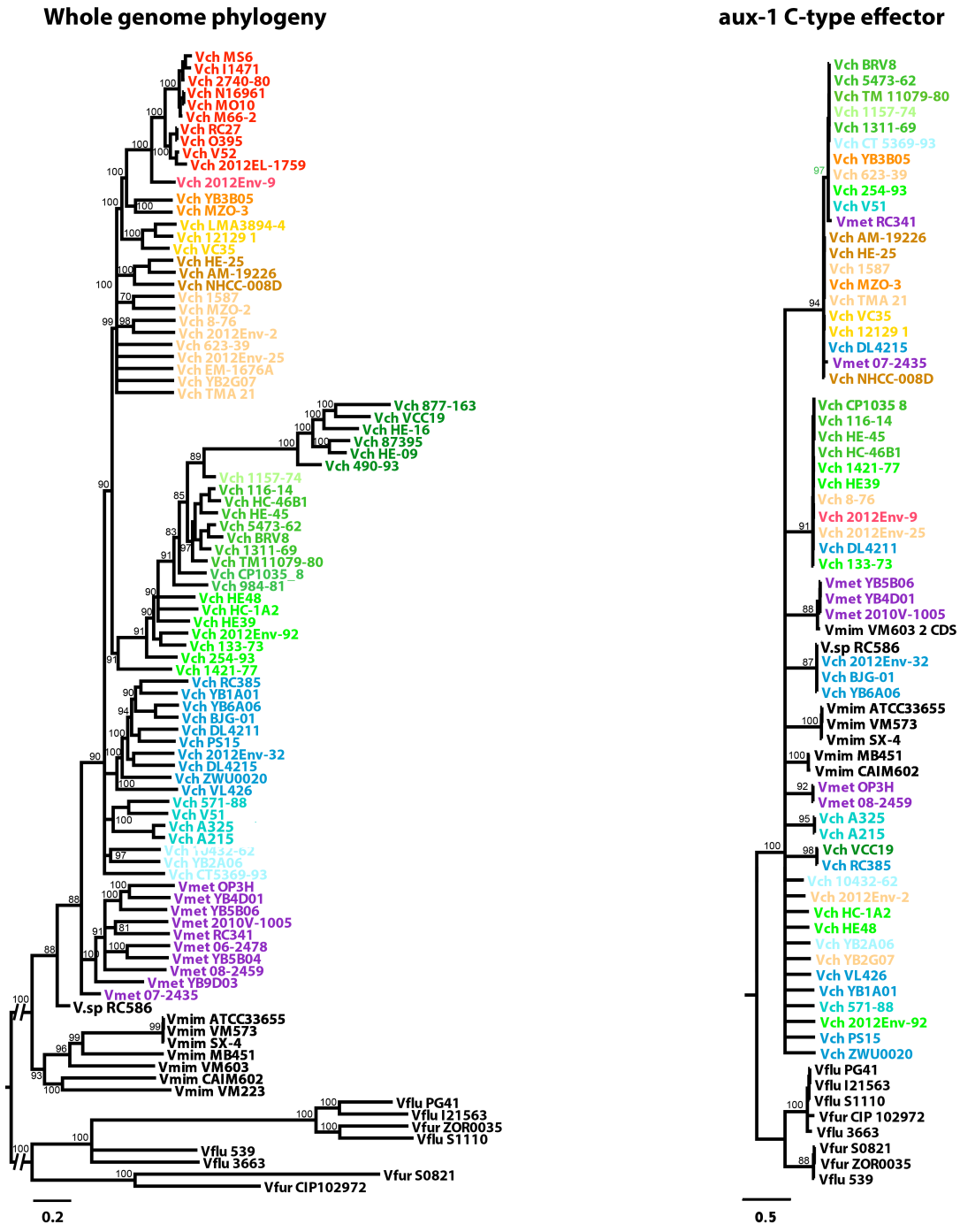


Figure 4.7: Incongruence between whole genome phylogeny and single gene phylogeny of aux-1 C-type effector. Whole genome tree corresponds to Figure 4.2. Related strains are colored similarly to ease comparison between tree topologies but do not necessarily correspond to monophyletic groups. Aux-1 C-type effector phylogeny was calculated using the GTR+Gamma Maximum Likelihood model implemented in RAxML based a 2,315bp alignment. Black dots indicate orphan immunity genes (i.e. genes not directly downstream of an effector). Statistical branch support was obtained from 100 bootstrap repeats. Branches with support <70

were collapsed, as such every branch has a bootstrap support of >70. Scale bar indicates substitutions/site. Double dashes through branches indicates cut by length of 4.0 substitutions/site. Vch = *Vibrio cholerae*, Vmet = *V. metoecus*, Vmim = *V. mimicus*, Vfur = *V. furnissii*, Vflu = *V. fluvialis*.

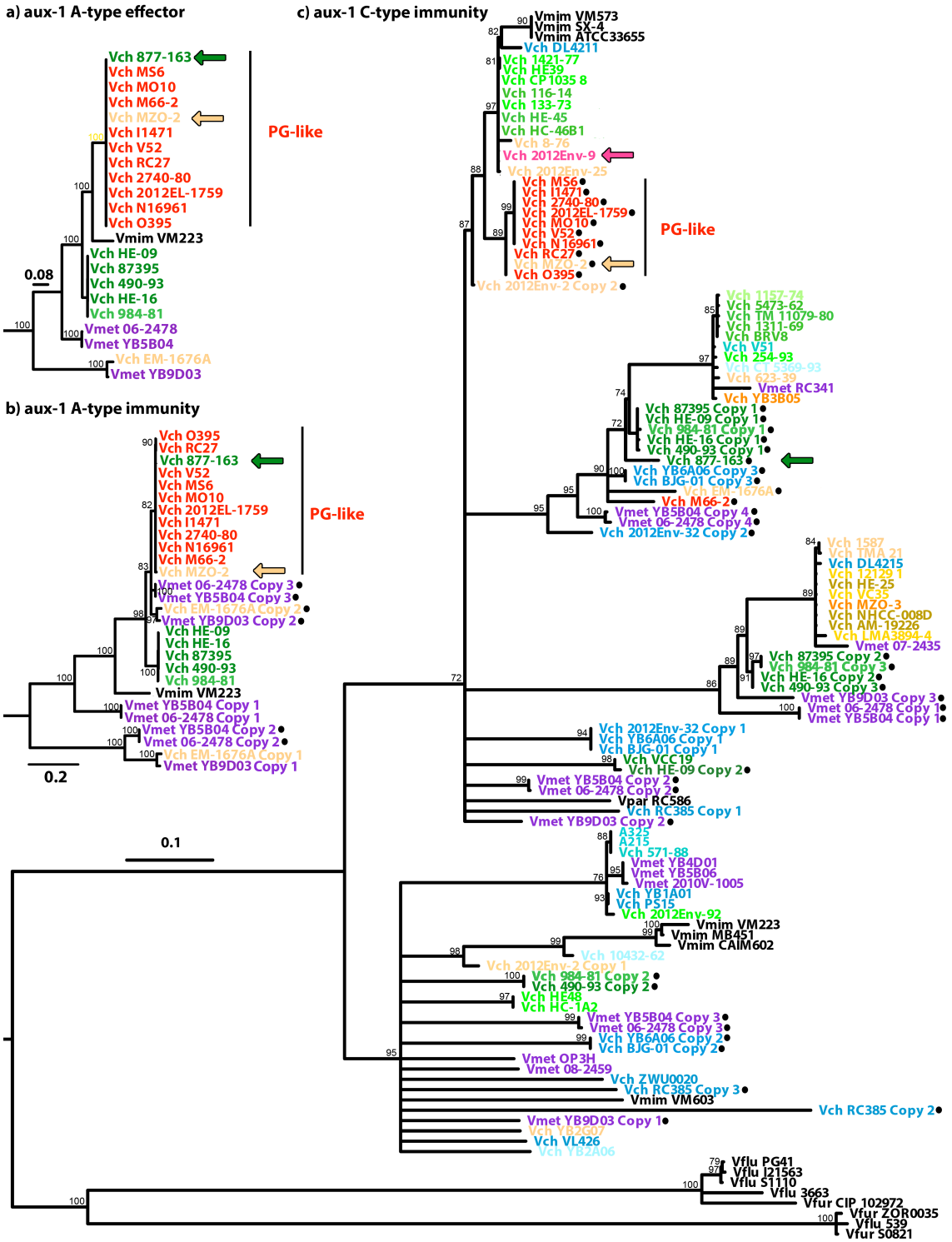


Figure 4.8: Single gene phylogenies of aux-1 a) A-type effector, b) A-type immunity and c) C-type immunity genes. Colouring of strains corresponds to whole genome tree of figure 4.7. Arrows indicate strains of interest discussed in text. Phylogenetic trees were calculated using the GTR+Gamma Maximum Likelihood model implemented in RAxML based on single gene alignments (aux-1 A-type effector: 2,040bp; aux-1 A-type immunity: 754bp, aux-1 C-type immunity: 740bp. Black dots indicate orphan immunity genes (i.e. genes not directly downstream of an effector). Statistical branch support was obtained from 100 bootstrap repeats.

Branches with support <70 were collapsed, as such every branch has a bootstrap support of >70. Scale bar indicates substitutions/site. Vch = *Vibrio cholerae*, Vmet = *V. metoecus*, Vmim = *V. mimicus*, Vfur = *V. furnissii*, Vflu = *V. fluvialis*.

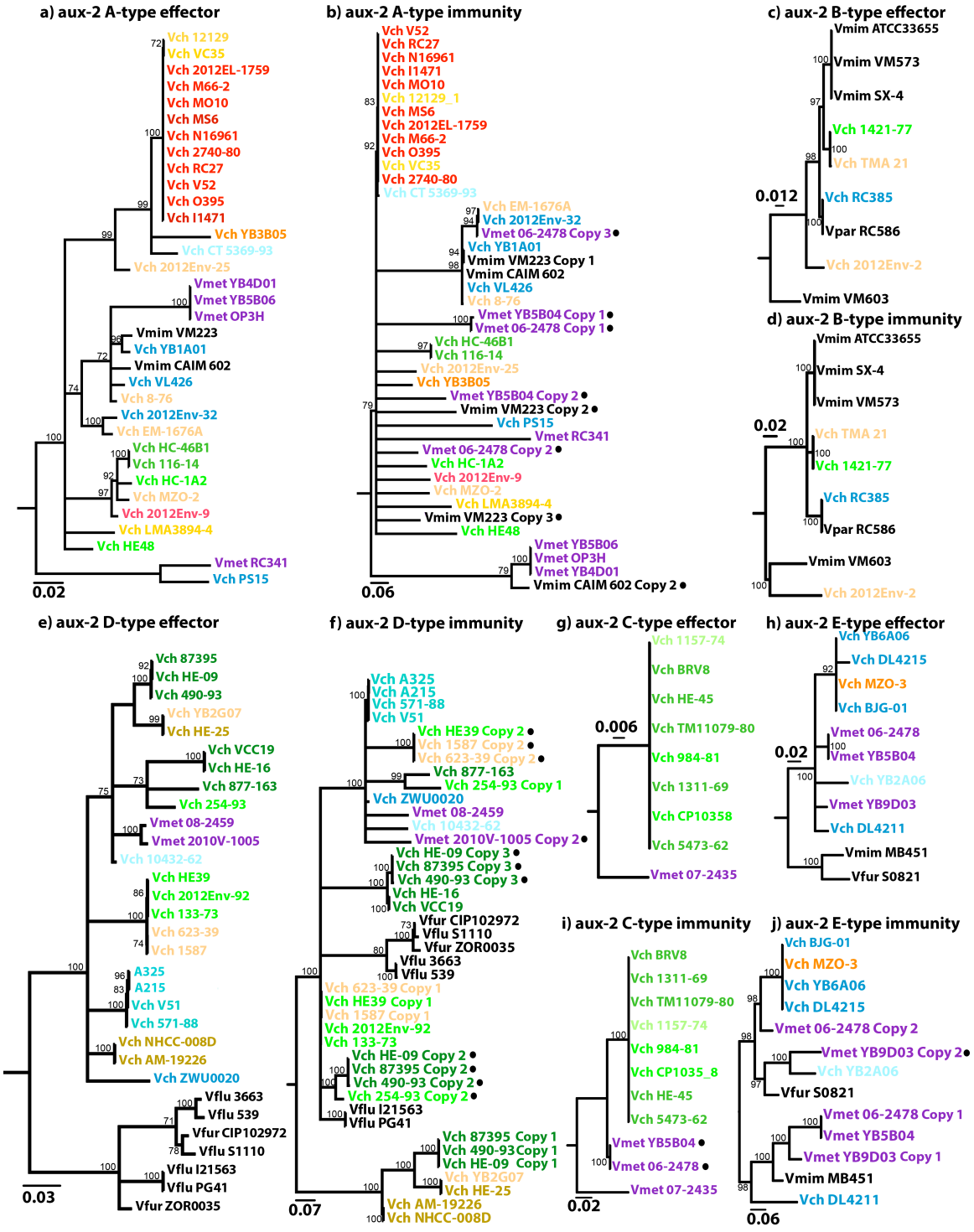


Figure 4.9: Single gene phylogenies of aux-2 effector and immunity genes. a) A-type effector b) A-type immunity c) B-type effector d) B-type immunity e) D-type effector f) D-type immunity g) C-type effector h) E-type effector i) C-type immunity j) E-type immunity. Colouring of

strains corresponds to whole genome tree of figure 4.7. Phylogenetic trees were calculated using the GTR+Gamma Maximum Likelihood model implemented in RAxML based on single gene alignments (aux-2 A-type effector: 3,268 bp; aux-2 B-type effector: 3,261 bp; aux-2 C-type effector: 3,531 bp; aux-2 D-type effector: 3,726 bp; aux-2 E-type effector: 3,594 bp; aux-2 A-type immunity: 741 bp; aux-2 B-type immunity: 1,119 bp; aux-2 C-type immunity: 951bp; aux-2 D-type immunity: 1,162 bp; aux-2 E-type immunity: 1,018 bp. Black dots indicate orphan immunity genes (i.e. genes not directly downstream of an effector). Statistical branch support was obtained from 100 bootstrap repeats. Branches with support <70 were collapsed, as such every branch has a bootstrap support of >70. Scale bar indicates substitutions/site. Vch = *Vibrio cholerae*, Vmet = *V. metoecus*, Vmim = *V. mimicus*, Vfur = *V. furnissii*, Vflu = *V. fluvialis*.

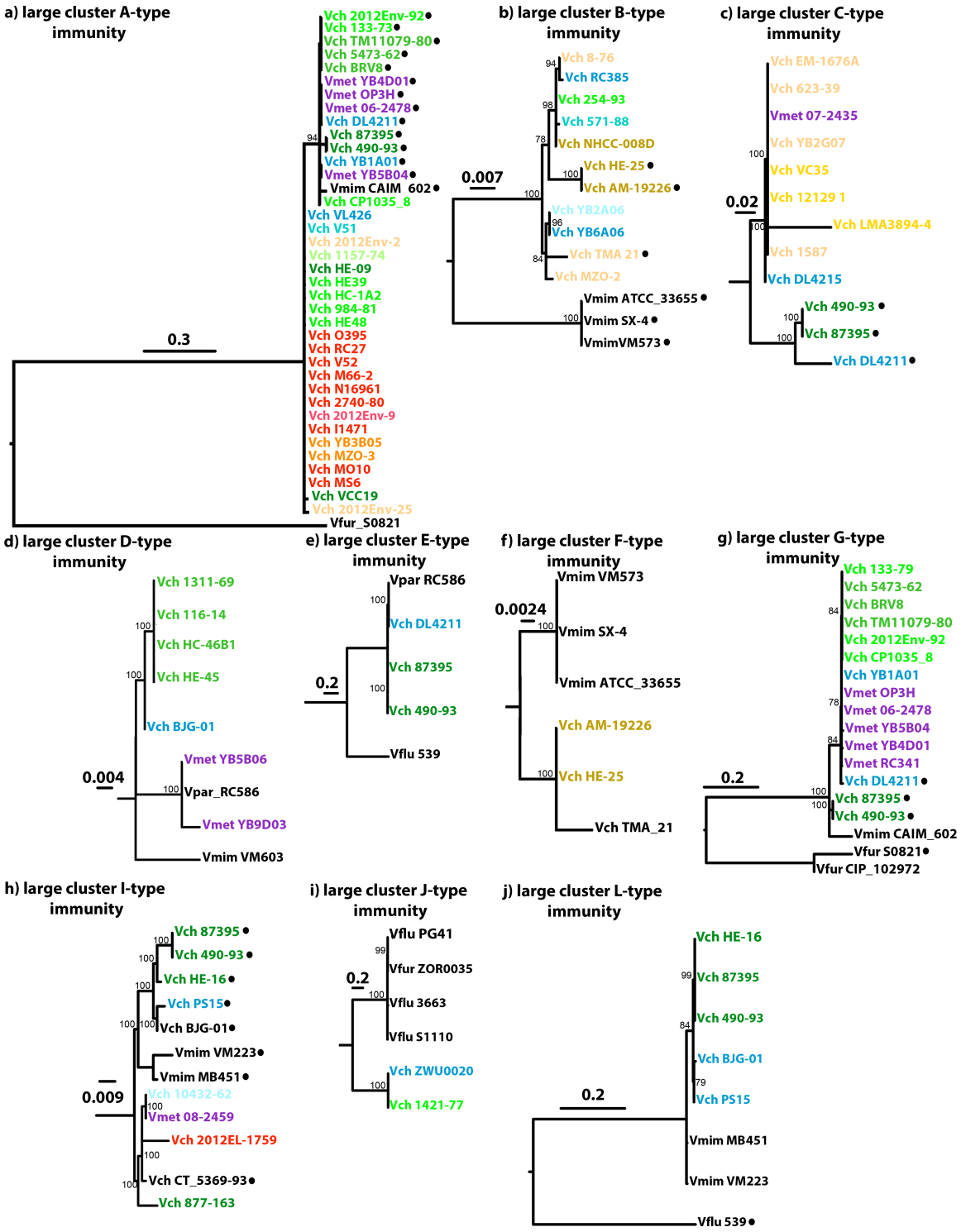


Figure 4.10: Single gene phylogenies of large cluster immunity genes. Colouring of strains corresponds to whole genome tree of figure 4.7. Phylogenetic trees were calculated using the GTR+Gamma Maximum Likelihood model implemented in RAxML based on single gene alignments (main A-type immunity: 402 bp; main B-type immunity: 966 bp; main C-type immunity: 384bp; main D-type immunity: 399 bp; main E-type immunity: 346 bp; main F-type immunity: 409 bp; main G-type immunity: 585; main I-type immunity: 445 bp; main J-type immunity: 384 bp; main L-type immunity: 402 bp. Black dots indicate orphan immunity genes

(i.e. genes not directly downstream of an effector). Statistical branch support was obtained from 100 bootstrap repeats. Branches with support <70 were collapsed, as such every branch has a bootstrap support of >70. Scale bar indicates substitutions/site. Vch = *Vibrio cholerae*, Vmet = *V. metoecus*, Vmim = *V. mimicus*, Vfur = *V. furnissii*, Vflu = *V. fluvialis*.

4.4.4 Homologous recombination without specific integration sites leads to the mosaic structure of T6SS clusters

Distribution patterns of EI modules and phylogenies of effector and immunity genes suggest that genes within T6SS clusters are frequently transferred between *Vibrio*. These transfer could occur at random sites, or at specific recombination sites like the integration of the *Vibrio* pathogenicity island VPI-1 (189). In the latter scenario, the size of recombined regions as well as the location of recombination breakpoints should not vary greatly between events, as each recombinant region would be integrated at a specific location. Such recombination hotspots have been predicted to exist in the adaptor protein encoding gene *tap-1* and *vgrG-1* of the aux-1 cluster (174). We performed a scan for recombinant regions in T6SS clusters based on phylogenetic discordance and patterns of single nucleotide polymorphisms (126). Contrary to expectations of recombination being more frequent at a specific site, the size of recombined regions and the associated positions of recombination breakpoints varied greatly in all three T6SS clusters (Figure 4.11). We could detect a number of recombination breakpoints located around the hypothesized recombination hotspots. However, we also found numerous recombination breakpoints inside effector and immunity genes, structural genes and noncoding regions. Interestingly, we detected very few recombination events within the actin-crosslinking domain at the 3'-end of *vgrG*-1, but found a relatively large number of breakpoints at the 5'-end of this gene. As hypothesized for bacteriocin genes, this difference in the number of detected recombination events could simply be a result of the availability of specific genes or gene regions as substrate for recombination(190). The actin crosslinking domain, present in only a small subset of *Vibrio,* could thus only rarely recombine compared to the ubiquitous 5'-end of *vgrG*-1 or the highly conserved central region of *tap*-1.

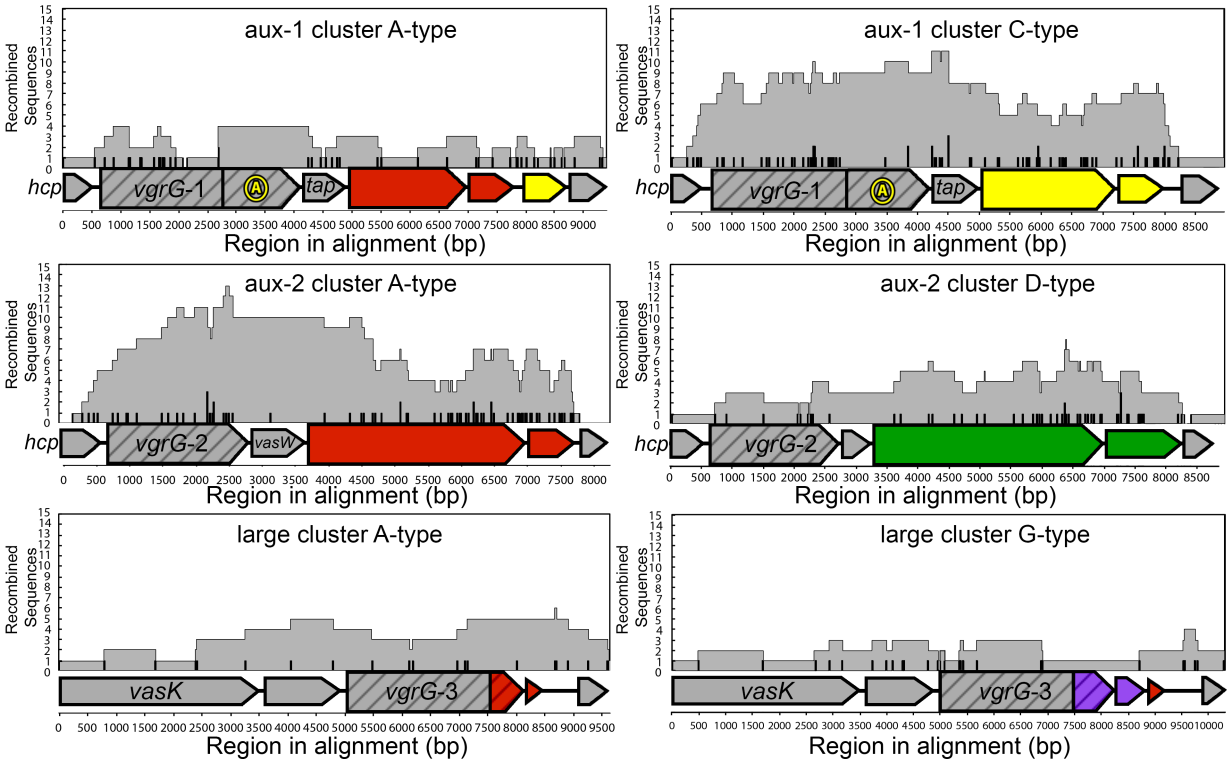


Figure 4.11: Location of recombination tracts and breakpoint on T6SS clusters. X-axis indicates region in the alignment starting at hcp-1 and 2 for aux-1 and aux-2 cluster and vasK for large cluster. Grey areas indicate presence of a recombined region detected by >3 algorithms implemented in RDP4, vertical hash marks indicate presence of corresponding recombination breakpoints in that region of the alignment. Multiple recombination regions/breakpoints are stacked on top of each other. Horizontal block arrows indicate coding sequences, horizontal lines noncoding regions. Striped arrows denote genes encoding VgrG effectors, non-striped arrows variable effector (large arrows) or immunity genes (small arrows). Colours indicate homology between either effector or immunity genes. Identically coloured effector and immunity genes are part of the same EI module. Grey arrows indicate conserved genes.

Overall, recombination appears to not be confined to specific regions within the T6SS clusters (Figure 4.11). Horizontally transferred DNA integrates in any sufficiently homologous site and thus essentially everywhere along the T6SS region, potentially incorporating non-homologous regions between the integration sites and making each individual region a mosaic composed of DNA from multiple different origins. This mosaic structure is also apparent in phylogenies of individual T6SS genes, with sequences of genes from distantly related strains clustering together, and weak overall bootstrap support due to the existence of genes composed of DNA from multiple distantly related bacteria (Figures 4.7-4.10). This is particularly

apparent in widespread genes such as those encoding C-type effector and immunity proteins (Figures 4.7 and 4.8).

4.4.5 A model for the establishment of immunity gene arrays through displacement of effector genes by EI modules

Homologous recombination exclusively cannot explain the existence of numerous arrays of multiple immunity genes, as their formation involves the addition of novel genes with no homology to sequences present in the recipient strain. We propose an event where a horizontally transferred EI module is inserted into a T6SS cluster, displacing the ancestral effector gene or effector gene domain but conserving the ancestral immunity gene, which is shifted downstream of the new EI module.

A putative mechanism to account for these observations could be akin to previously described homology-facilitated illegitimate recombination: a conserved stretch of an incoming DNA element, in this case represented by the upstream region of the EI module, serves as an "anchor" by forming a heteroduplex with a homologous sequence in the target region, thereby facilitating the integration of the non-homologous end through illegitimate recombination (191, 192). Multiple successive EI insertions could then give rise to longer arrays of a single effector with multiple immunity genes as those reported here for numerous strains. The mechanism can be illustrated most accurately by examining the structure of the large cluster (L-type) found in *V. cholerae* strains 87395 and 490-93 (Figure 4.5b). This particular gene assembly evolved by successive transitions through various simpler intermediate forms, whose structure is present in other strains included in this study. (I) First, an ancestral A-type EI module is replaced by a G-type EI module, replacing the stretch of DNA encoding the A-type *vgrG*-3 effector region with a G-type effector and immunity gene while shifting the A-type immunity gene to the back of the array; (II) This G-type EI module is then replaced by a C-type EI module, shifting the G and A immunity genes further back; (III) The new C-type EI module is then replaced by an E-type EI module, giving rise to an E-type EI module followed by C, G and A immunity genes; (IV) A final insertion event of an L-type EI module (containing an I-type immunity gene as a remnant of an earlier replacement of an I-type EI module) occurs, shifting back the 3'-end of the E-type effector and creating an array containing six different immunity genes.

Two theoretical alternatives to the successive replacement of an effector gene by a novel EI module could also explain the presence of immunity gene arrays. The first is the complete replacement of an EI-module by a DNA fragment encoding a different module containing a 3'-

extended region with additional immunity genes. Although we found several examples of this type of event, they do not explain how the more complex immunity arrays initially formed, only how they were introduced in a new strain. Another alternative explanation for the existence of immunity gene arrays is that they are created by successive gene duplications. If duplication of immunity genes occurred, multiple homologous genes found in one strain would be more closely related to each other than to those found in other strains. However, as the phylogenies of immunity genes in Figure 4.7-4.10 show, multiple alleles found in the a single genome do not cluster together and thus likely originate from different *Vibrio* strains rather than from duplication events. Furthermore, duplication events cannot explain heterogeneous arrays consisting of different immunity type genes.

Integration of DNA elements into chromosomal T6SS clusters of *Vibrio* thus likely occurs through at least two different mechanisms: I) normal homologous recombination leading to the replacement of a region in the T6SS cluster (as shown *in-vivo* by Koskiniemi et *al*. (193) in *Salmonella*, where an orphan EI module replaces the variable 3' region and immunity protein of a contact dependent toxin) ; and II) replacement of an effector by a novel EI module (perhaps) through homology-facilitated illegitimate recombination with conservation of the ancestral immunity gene. The latter was also hypothesized for the diversification of recombination hot spot (RHS) protein coding loci (which includes T6SS regions) with constant 5' and variable 3' regions to explain the frequent observation of strings of "orphaned" 3' regions (194-196).

It appears likely that the first mechanism would occur more often, as closely related strains with different EI modules mostly do not contain orphan immunity genes indicative of illegitimate displacement events. For example, strains V52 and 2012EL-1759 are closely related and contain complete A- and I-type EI modules in the main cluster, respectively, with no orphan immunity genes in either of them (Figure 4.2). Furthermore, experimental evidence in other organisms shows that homology-facilitated illegitimate recombination is quite rare (several orders of magnitude less frequent than regular homologous recombination in *Pseudomonas* (192), *Streptococcus*(191) and *Acinetobacter*(197)). While no comparisons of these two processes have been done for *Vibrio*, high rates of homologous recombination are commonly observed in multi-locus sequence typing or whole genome studies of this genus(63, 124, 135, 140, 198). The insertion of T6SS EI pairs is reminiscent of site-specific recombination in the *Vibrio* chromosomal integron region(109, 199). In integrons, gene cassettes are added at an insertion site downstream of an integrase gene, whose gene product facilitates this process. Addition of a new gene cassette leads to the displacement of old gene cassettes to the back of

the array(200), which parallels our observation of immunity gene displacement. However, so far there exists no evidence that EI insertion is facilitated by an integrase in a similar manner. The relative uniformity of the O1/O139 T6SS arrays also stands in contrast with the variability in integron cassette content of that clade(201), indicating that EI change, especially through illegitimate recombination, proceeds much less rapidly.

4.4.6 The T6SS effector-immunity gene combination of pandemic *V. cholerae* strains evolved and spread through a series of horizontal gene transfer events

Pandemic *V. cholerae* strains not only possess the ability to cause lethal disease in humans, but also a unique composition of T6SS modules that gives them (at least *in-vivo*) unmatched competitive abilities in interactions with conspecific strains (141). All sequenced genomes of strains from the lineage containing pandemic *V. cholerae* (with the single known exception being the aforementioned 2012EL-1759) harbour the same T6SS A-type aux-1 module accompanied by an actin-crosslinking VgrG-1 and an extended 3'-region with a C-type immunity protein; A-type EI modules in aux-2 and the large cluster; and a TseH/TsiH module encoded close to the chromosomal integron region (Fig. 4.2). The specific combination of EI modules found at the three T6SS clusters of pandemic *V. cholerae* strains appears to have assembled progressively through a combination of both previously proposed integration mechanisms of horizontally transferred T6SS elements (Fig. 4.12a). (I) The CAA module combination, which is found in strains basal to the lineage that gave rise to pandemic *V. cholerae*, was likely the starting point for the evolution of the modern pandemic T6SS structure. (II) The antibacterial TseH-TseI module was likely acquired by the CAA common ancestor of the lineage containing pandemic strains and its sister group (exemplified by 2012-Env9), potentially providing a competitive advantage over strains lacking this genetic element (176). (III) The A-type EI module subsequently replaced the ancestral C-type effector while displacing the C-type immunity gene into the extended 3'-region in the ancestor of the lineage that gave rise to modern pandemic strains (Figure 4.12b). A single gene phylogeny of the C-type immunity gene provides some evidence for this hypothesis, as the alleles found in the pandemic strains lineage and its sister group strains remain comparatively similar, in congruence with the common ancestry of these groups (Figure 4.7). Acquisition of a new, rare A-type EI module while retaining the C-type immunity gene (whose expression is up-regulated nearly 2 folds during infection of the host (171)) likely enhanced the competitive advantage of strains in the pandemic lineage. Whether the VgrG-1 actin crosslinking domain shared by all *V. cholerae* in the lineage

containing pandemic strains was included in that recombination event or inserted at a separate point before or after remains unclear, although parsimony would imply integration in a single event.

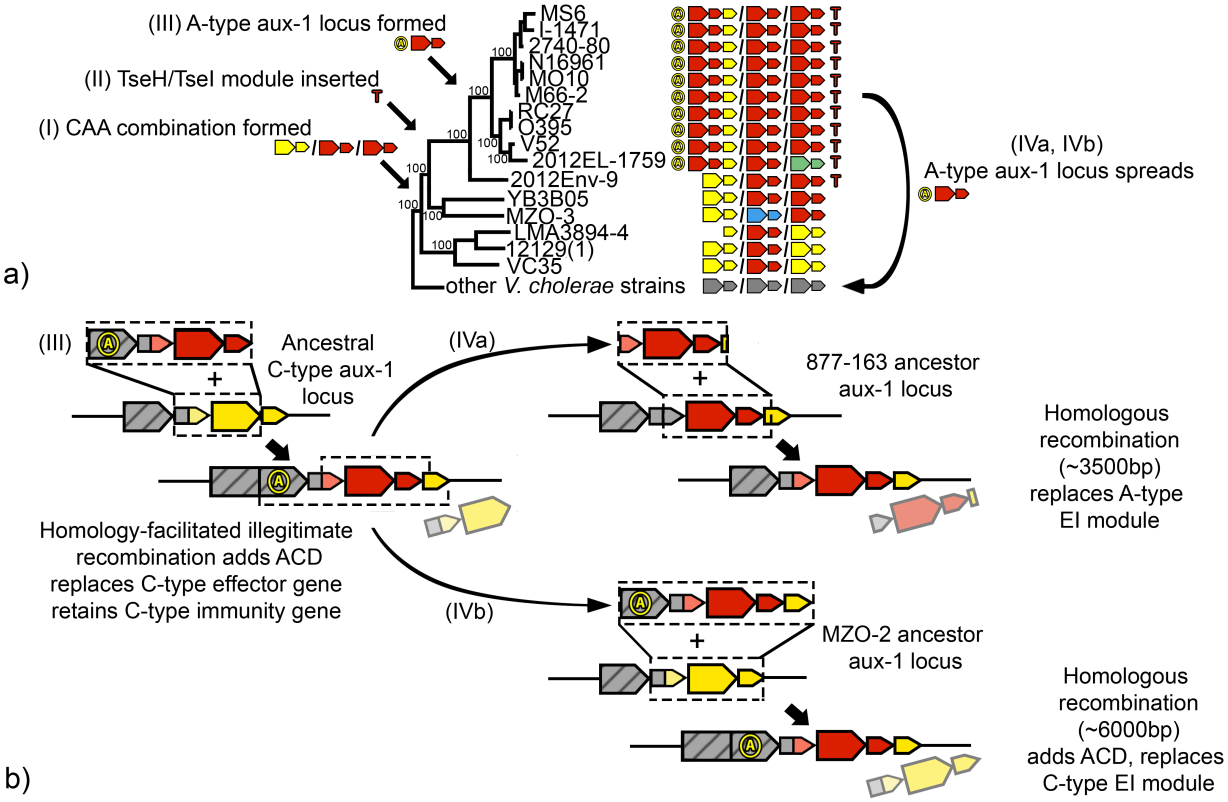


Figure 4.12: Evolution of the *Vibrio cholerae* aux-1 cluster. a) Evolutionary events along the phylogeny of the lineage containing modern pandemic group V. cholerae. b) Schematic representation of HGT events in the aux-1 cluster. (I) CAA module combination forms through homologous replacement of other EI modules in the ancestor of pandemic group V. cholerae and related strains. (II) TseH-TseI module is inserted close to the chromosomal integron region of the strain ancestral to the pandemic and pandemic sister group (represented by 2012-Env9). (III) Original C-type EI module is replaced by an A-type EI module in the ancestor of lineage containing pandemic V. cholerae, retaining the C-type immunity gene. Parts of the modern aux-1 cluster of pandemic V. cholerae are subsequently inserted into 877-163 (IVa) and MZO-2 (IVb) (see also Fig. 4.13). Arrows indicate coding sequences, lines noncoding regions. Striped arrows denote VgrG effectors, non-striped arrows variable effector (large arrows) or immunity genes (small arrows). Colours indicate homology between either effector or immunity genes. Identically coloured effector and immunity genes are part of the same EI module. Grey arrows indicate conserved genes. Regions of recombination are indicated by striped boxes.

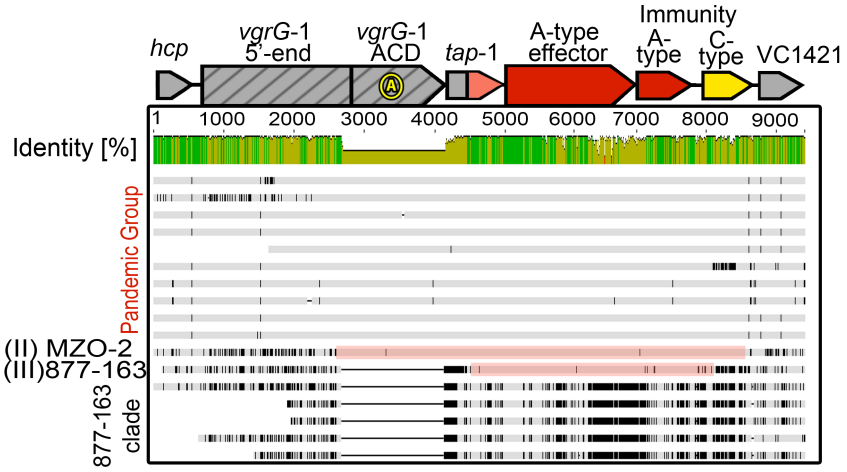


Figure 4.13: Nucleotide alignment of aux-1 clusters of the lineage containing pandemic *V. cholerae* and putative recombinant regions. Grey regions indicated conserved sites while black lines indicate divergence from the consensus sequence. Red bars highlight regions in MZO-2 and 877-163 with high identity to the pandemic group aux-1 cluster, indicating putative horizontal gene transfer.

Interestingly, parts of the lethal T6SS structure have proceeded to spread from the lineage containing pandemic strains into distantly related, non-pandemic strains. Both strains MZO-2 and 877-163 possess aux-1 clusters containing effector and immunity genes nearly identical to their homologs in pandemic strains (Figures 4.7 and 4.8), making it likely that the latter was the donor. In accordance to our finding that integration of recombinant DNA into T6SS clusters does not occur at specific sites, the size of recombined regions differs for both strains (Fig 4.12b and 4.13). (IVa) The size of the region received by 877-163 is around 3.5 Kbp and includes the 3'-end of *tap-1* upstream of the EI module and a small part of the C-type immunity gene downstream. (IVb) In contrast, the region received by MZO-2 extends beyond the adaptor *tap-1* gene and the actin-crosslinking domain of *vgrG-1* upstream of the EI module and beyond the C-type immunity gene downstream (more than 6 Kbp). We interpret this as evidence that, similar to *rhs* elements in other Gram negative bacteria (202), complex E-I arrays created by successive displacements of effector genes by E-I modules could form a pool of structurally stable elements that can be transferred between diverse strains through frequent homologous recombination.

4.5 **Conclusions**

Our observation of mosaic arrays of immunity genes in the T6SS clusters of *V. cholerae* and closely related species suggests a selective advantage for the presence of multiple immunity genes. The retention of immunity genes by shifting them into the extended 3'-regions of the *Vibrio* T6SS clusters provides a single cell within an otherwise homogeneous population a mechanism to successfully acquire a EI module without being killed by surrounding kin bacteria. *V. cholerae* becomes naturally competent when reaching high cell densities on chitinous surfaces (107), conditions that also lead to up-regulation of T6SS gene expression (93).Therefore, any cell acquiring a new EI module is probably surrounded by now incompatible sister cells. Since even more effective EI module combinations succumb to less effective ones when greatly outnumbered (141), a newly acquired EI module combination would likely be rapidly overwhelmed by sister cells. Our observation that the extended 3'-region of T6SS clusters retains ORFs coding for additional immunity proteins provides an explanation for the EI diversity reported previously and validated here. EI modules could be successfully acquired in a numerically superior population of incompatible cells by only replacing the effector but retaining the immunity gene. Due to the aforementioned simultaneous upregulation of T6SS activity and natural competence, incoming non-compatible cells killed by the T6SS of resident Vibrios would represent an easily available source of DNA encoding new EI modules and other potentially beneficial genetic elements (93). Larger amounts of DNA freed from subsequently killed former sister cells could then provide a readily available additional food source (203).

After acquisition of a novel EI module by a lineage of bacteria, expression of additional immunity proteins (such as the C-type immunity protein VC1420 in pandemic strains(171)) could also confer protection against more distantly related strains with different effectors. Additional immunity proteins encoded in the 3'-extended region belonging to the same type as the main immunity protein present in the EI module but displaying some sequence divergence could protect strains against similarly divergent effectors that cannot be effectively bound by the main immunity protein (187). This would be particularly beneficial in competition involving widespread and diverse EI modules such as the aux-1 C-type.

Relaxed selective pressure on a redundant immunity gene could furthermore give cells a significant edge in a T6SS mediated arms race. In colicin E-I modules, mutations in an immunity gene conferring additional resistance to foreign effector types are thought to be followed by mutations in effector genes that enable it to avoid immunity of other strains, leading to the emergence of a competitively superior strain(190). A second copy of an immunity gene would

allow one of the genes to diverge without having to retain the immunity function against the cell's own effector and could considerably speed up the evolution of novel functionalities (195). Furthermore, it would appear mechanistically easier for a single effector to mutate to overcome binding by immunity proteins of other strains than for a single immunity protein being able to bind all potential effector variants. A larger repertoire of immunity proteins, even of the same type, could thus confer an advantage in an effector-rich environment.

Thus, understanding compatibility of various *Vibrio* strains might require taking into account not only the effectors and immunity proteins encoded in EI modules, but also those found in the extended 3'-regions of T6SS clusters, as well as sequence divergence within effector and immunity proteins of the same type.

In summary, we provide a comprehensive overview of the *V. cholerae* T6SS EI module diversity in a phylogenetic context, expand the repertoire of the *Vibrio* T6SS by multiple novel putative effectors, extend T6SS clusters to include an additional 3' region and put forth a hypothetical model for the evolution of mosaic immunity gene arrays in this 3' extended region. Furthermore, our analysis makes it possible to trace the genesis of possibly the most effective module combination, found in the *V. cholerae* lineage containing pandemic strains, through stepwise acquisition of singular elements along their pathway from harmless environmental bacteria to deadly human pathogens.

Table S4.1: NCBI Accession number of all genomes.

Species	Strain	Accession	Accession 2
V. cholerae	MS6	AP014524.1	AP014525.1
V. cholerae	I-1471	CM003111.1	CM003112.1
V. cholerae	2740-80	AAUT00000000.1	
V. cholerae	N16961	AE003852.1	AE003853.1
V. cholerae	MO10	AAKF00000000.3	
V. cholerae	M66-2	CP001233.1	CP001234.1
V. cholerae	RC27	ADAI00000000.1	
V. cholerae	O395	CP001235.1	CP001236.1
V. cholerae	V52	AAKJ00000000.2	
V. cholerae	2012EL-1759	JNEW00000000.1	
V. cholerae	2012Env-9	JSTH00000000.1	
V. cholerae	CP1037(10)	ALDB00000000.1	
V. cholerae	YB3B05	LBGB00000000.1	
V. cholerae	MZO-3	AAUU00000000.1	
V. cholerae	LMA3984-4	CP002555.1	CP002556.1

V. cholerae	12129(1)	ACFQ00000000.1
V. cholerae	VC35	AMBR00000000.1
V. cholerae	HE-25	ALEC00000000.1
V. cholerae	AM-19226	AATY00000000.1
V. cholerae	NHCC-008D	APGC00000000.1
V. cholerae	1587	AAUR00000000.1
V. cholerae	MZO-2	AAWF00000000.1
V. cholerae	8-76	JIDN00000000.1
V. cholerae	2012Env-2	JSTD00000000.1
V. cholerae	623-39	AAWG00000000.1
V. cholerae	2012Env-25	JSTE00000000.1
V. cholerae	EM-1676A	APFY00000000.1
V. cholerae	YB2G07	LBGA00000000.1
V. cholerae	TMA21	ACHY00000000.1
V. cholerae	1421-77	JMBL00000000.1
V. cholerae	254-93	JMBP00000000.1
V. cholerae	133-73	JIDK00000000.1
V. cholerae	2012Env-92	JSTJ00000000.1
V. cholerae	HE-39	AFOQ00000000.1
V. cholerae	HC-1A2	AJRO00000000.1
V. cholerae	HE-48	AFOR00000000.1
V. cholerae	984-81	JMBM00000000.1
V. cholerae	CP1035(8)	AJRM00000000.1
V. cholerae	TMM11079-80	ACHW00000000.1
V. cholerae	1311-69	JIDJ00000000.1
V. cholerae	BRV8	CTBD00000000.1
V. cholerae	5473-62	JIDI00000000.1
V. cholerae	HE-45	ALED00000000.1
V. cholerae	HC-46B1	AJSL00000000.1
V. cholerae	116-14	CGHE00000000.1
V. cholerae	1157-74	JIDL00000000.1
V. cholerae	490-93	JIDQ00000000.1
V. cholerae	HE-09	AFOP00000000.1
V. cholerae	87395	APFL00000000.1
V. cholerae	HE-16	ALEB00000000.1
V. cholerae	VCC19	ATEV00000000.2
V. cholerae	877-163	LBNV00000000.1
V. cholerae	RC385	AAKH00000000.3
V. cholerae	YB1A01	LBCL00000000.1
V. cholerae	YB6A06	LBGK00000000.1
V. cholerae	BJGO1	AFOU00000000.1
V. cholerae	DL4211	MOLL00000000.1
V. cholerae	PS15	AIJR00000000.1

V. cholerae	2012Env-32	JSTF00000000.1	
V. cholerae	DL4215	MOLM00000000.1	
V. cholerae	ZWU0020	JRJX00000000.1	
V. cholerae	VL426	ACHV00000000.1	
V. cholerae	571-88	JIDO00000000.1	
V. cholerae	V51	AAKI00000000.2	
V. cholerae	A325	CWSO00000000.1	
V. cholerae	A215	CWSL00000000.1	
V. cholerae	10432-62	GCA_000969265.1	
V. cholerae	YB2A06	LBFX00000000.1	
V. cholerae	CT5369-93	ADAL00000000.1	
V. metoecus	OP3H	JJMN00000000.1	
V. metoecus	YB4D01	LBGO00000000.1	
V. metoecus	YB5B06	LBGQ00000000.1	
V. metoecus	2010V-1005	LCUG00000000.1	
V. metoecus	RC341	ACZT00000000.1	
V. metoecus	06-2478	LCUD00000000.1	
V. metoecus	YB5B04	LBGP00000000.1	
V. metoecus	08-2459	LCUF00000000.1	
V. metoecus	YB9D03	LBGR00000000.1	
V. metoecus	07-2435	LCUE00000000.1	
V. mimicus	VM223	ADAJ00000000.1	
V. mimicus	CAIM602	AOMO00000000.1	
V. mimicus	VM603	ACYU00000000.1	
V. mimicus	MB451	ADAF00000000.1	
V. mimicus	SX-4	ADOO00000000.1	
V. mimicus	VM573-73	ACYV00000000.1	
V. mimicus	ATCC33655	NZ_CP014042.1	NZ_CP014043.1
V. furnissii	CIP102972	ACZP00000000.1	
V. furnissii	SO821	LKHS00000000.1	
V. fluvialis	3663	JXXQ00000000.1	
V. fluvialis	539	JQHX00000000.1	
V. fluvialis	S1110	LKHR00000000.1	
V. furnissii	ZOR0035	JTLJ00000000.1	
V. fluvialis	I21563	ASXT00000000.1	
V. fluvialis	PG41	ASXS00000000.1	

Chapter 5

High-throughput sequencing of a protein-coding gene allows detailed tracking of *Vibrio cholerae* population dynamics and confirms the presence of pandemic-related O1 strains in a cholera-free region

---

5.1 Abstract

Vibrio cholerae is a common but low abundance member of coastal microbial communities worldwide. Currently, a single lineage of this highly diverse bacterium is capable of causing pandemics of the lethal diarrheal disease cholera. However, horizontal gene transfer events of virulence factors have the potential of enabling other strains to cause local epidemics as well. As such, studying the population dynamics of *V. cholerae* could play a key role in the discovery of novel pathogenic variants as well as provide insights into the biology of the organism. However, current culture- or culture-independent techniques are either too laborious or do not offer a high enough resolution to accurately assess the population structure of this species. In this study, we develop an amplicon sequencing strategy based on a highly-diverse *V. cholerae* core-genome protein-coding gene. With the use of clustering-free approaches to variant-calling, this provides strain-level resolution of the dynamics of a *V. cholerae* population in the coastal eastern United States. Considerable variation is observed in the relative abundance of various strains as well as extensive mosaic sympatry over the course of two years. This method allowed the detection of pandemic-related O1 serogroup *V. cholerae* in this cholera-free region of the world. This suggests that the phylogenetic group which gave rise to pandemic *V. cholerae* is much more diverse and widespread than previously believed.

5.2 Introduction

Throughout the course of history, seven pandemics of the deadly diarrheal disease cholera have been recorded (204). The first six pandemics were caused by a lineage of the bacterium *Vibrio cholerae* belonging to the O1 serogroup, one of the species' over 200 serogroups, as

defined by antigenic properties of their lipopolysaccharide layer (205). The 7th and currently ongoing pandemic is caused by a different, closely related group of strains that have undergone a number of horizontal gene transfer events that lead to a switch in biotype (of different biochemical properties and phage sensitivity) from “classical” to “El Tor” (103). Recent genomic analyses have revealed extensive additional diversity within this so-called pandemic group (PG) of *V. cholerae*, with several waves of divergent El Tor lineages spreading across continents from the Bay of Bengal, where the PG is endemic (206). Despite the constant presence of El Tor *V. cholerae* in this region, cholera outbreaks only occur seasonally due to changes in abundance of that lineage in the local waters (76). Similar bloom and bust dynamics occur in regions free of cholera (for example (207)), where numerous non PG-*V. cholerae* lineages exist as innocuous members of bacterial communities in coastal and brackish waters (95). However, some non-PG lineages have independently acquired virulence factors enabling them to cause enteric infections when present in large enough abundances (99). As such, tracking the population structure of *V. cholerae* is important both in regions where cholera is endemic as well as in regions currently free of cholera, especially with the expected increase in range of the cholera and general *Vibrio*-associated diseases due to global warming (208).

Currently, an abundance of tools and methods are available to track the population structure of bacteria in general and *V. cholerae* specifically (209). Isolation of pure cultures and their analysis using phenotypic assays, marker genes, multi-locus sequence typing or whole genome sequencing is the most informative. It is also exceedingly labour-intensive and expensive, making it unfeasible for long-term monitoring, as a large number of isolates is needed to adequately represent the population structure of even a single location (162). Another popular technique to measure the abundance of *V. cholerae*, both specifically for toxigenic *V. cholerae* or for the species overall, is qPCR (210). However, qPCR can not provide any information on population structure and might be susceptible to false positives in the detection of specific strains, as no sequence information is obtained.

The gold standard in the investigation of bacterial population structure is the amplification of various regions of the bacterial 16S rRNA gene and sequencing using high-throughput methods such as Illumina-Miseq or IonTorrent. Early limits in sequencing technology have led to the standardized method of clustering sequence reads into operational taxonomic units of 97% sequence identity (211). Using this method, it is possible to easily track a large number of *Vibrio* (and other) associated OTUs (212); however, the process of clustering severely limits the resolution of this technique, as clusters of 97% identity often represent multiple species of *Vibrio*

(for example (118, 213). Improvements in sequencing technology (lowering of error rates) and novel clustering-free methods of differentiating sequence reads have in recent years accelerated the obsolescence of this practice, as the differentiation of reads differing by as little as a single base pair has moved into the realm of possibility (214-216). Thus, the only limiting factor is the diversity of the 16s rRNA gene itself, which in the case of *V. cholerae* provides insufficient resolution for high-throughput sequencing methods (156).

In this study we develop a primer pair based on the highly variable region of the *viuB* gene specific to *V. cholerae* to drastically improve the resolution of amplicon-based high-throughput sequencing in monitoring *V. cholerae* at the subspecies level. I then demonstrate the efficacy of this method by observing the structure of a *Vibrio* population in a cholera-free location in the eastern United States over the course of two consecutive summers. I successfully replicate the results of previous multi-locus sequence typing based studies (162) but also discovered a considerable isolation bias for specific strains. Most importantly, I detected the presence of PG *V. cholerae* in this cholera-free region. Fractionated sampling allowed us to furthermore define potential niches for different strains of *V. cholerae*, lending support to the hypothesis that PG *V. cholerae* differs in its environmental niche from harmless environmental strains. Additionally, we observe extensive variation in strain diversity between samples located only several meters apart from each other, as well as between different particle fraction sizes and points in time. This mosaic sympatry could play a key role in the divergence of *V. cholerae* strains into a number of different ecological niches.

5.3 Methods

5.3.1 Finding marker genes suitable for differentiation of *Vibrio cholerae* clonal complexes

In order to find a primer set suitable for the amplification of products capable of differentiating *V. cholerae* strains, protein coding genes from a dataset of 68 *Vibrio* were analysed. This included 20 V. cholerae genomes from 5 clonal complexes found in the Oyster Pond (US east coast) sequenced in chapter 2 (162) , 22 additional *V. cholerae* genomes from public databases, 10 genomes of its closest relative *V. metoecus* and 16 other *Vibrio* species. Their genes were clustered based on 30% amino acid sequence identity of their protein products (131) using OrthoMCL 2.0 (132). From the 946 gene families found only in *V. cholerae*, multi-

copy genes were removed. Alignments of alleles from single-copy gene families present in all *V. cholerae* were created using ClustalW 2.0 (155) and maximum likelihood phylogenetic trees constructed using the GTR+GAMMA substitution model implemented in RAxML 8.0 (130). All single gene trees were then inspected for the ability to differentiate the 5 major clonal complexes described in Chapter 2 (162). The alignments producing these trees were then manually inspected for variable regions of less than 300bp that differentiated clonal complexes from each other and from other *V. cholerae* by at least 2bp and were also flanked by conserved sites of around 20bp. Candidate primer sets were synthesized by Integrated DNA Technologies (Iowa, USA), and tested for optimal amplification results. Optimal results were ultimately obtained using a primer set derived from *viuB* (involved in iron acquisition via vibriobactin utilization (217)).

5.3.2 Sampling and DNA extraction

We collected two different types of water samples in consecutive years. In the first year (2008), samples were collected from three different sites at Oyster Pond, MA, USA – the pond itself, and adjacent lagoon that connects the to the ocean, and the ocean outflow. Three samples were taken from the aforementioned locations at 0.5m depth at a distance of 5m each month from June to September. 50ml of water was pushed through a 4.5cm Millipore Durapore filter (size 0.22µm) using a polypropylene syringe. A second sample collection was conducted in 2009 in parallel with isolation described in (162), from June to October. 100 liters of water were first filtered through a 63µm nylon mesh net to capture large particles such as zooplankton. Large particles were crushed in a 50ml tissue grinder after transfer using 20ml of sterile filtered local water. 2ml of the crushed material was diluted 200-fold (to an equivalent of 50ml of water) and pushed through a 4.5cm Millipore Durapore filters (size 0.22 µm) using a polypropylene syringe. Similarly, 50ml of water passed through the mesh net was pushed through a series of in-line 4.5cm Millipore Durapore filters (sizes 5µm, 1µm and 0.22µm) using a polypropylene syringe. DNA extraction from the filters using a Qiagen DNEasy Blood and Tissue Kit was performed as follows: 0.25 g of sterile zirconium beads were added to cut-up filter pieces in a 1.5ml screw cap tube and 360ul Cell Lysis Buffer ATL and bead beating performed for 30 seconds at maximum speed. 40µl Proteinase K was added and tubes vortexed for several seconds. Further steps followed the instructions of the manufacturer.

To assess bias in amplification, a mock community containing known concentrations of DNA from 13 *V. cholerae* strains containing different *viuB*-alleles was created. Strains were grown

separately over night in LB medium at 37°C and extracted identically to the filters as mentioned above. Genomic DNA concentrations were measured using a Qubit Fluorometer (ThermoFisher) with a Qubit dsDNA HS Assay Kit (ThermoFisher). Equal concentrations of extracted genomic DNA from 12 strains, and double concentration for the 13th strain were then pipetted together. This mock community was then treated as a normal sample, described below.

5.3.3 PCR and sequencing

To amplify *viuB*-alleles from extracted DNA from fractionated water samples and the mock community, a touchdown PCR was performed using 0.5µl each of 10pmol forward and reverse primer (for *viuB:* *viuB*2f CCGTTAGACAATACCGAGCAC and *viuB*5r TTAGGATCGCGCACTAACCAC)

0.4µl 10mM dNTP-Mix (ThermoFisher), 0.4 µl Phire Hot Start II DNA Polymerase (ThermoFisher), 0.5µl of Molecular Biology Grade Bovine Serum Albumin (20mg/ml, New England Biolabs), 5µl of 5x Phire Buffer and 2 µl of template DNA. The PCR reaction was performed as follows: Initial denaturation 98°C for 4 min, followed by 10 cycles of denaturation 98°C for 10 sec, annealing 60°C for 6 sec (reduced by 1°C per cycle) and extension 72°C for 1 sec, followed by 23 cycles of denaturation 98°C for 10 sec, annealing 50°C for 6 sec (reduced by 1°C per cycle) and extension 72°C for 1 sec and a final extension of 72°C for 1 min.

In preparation for sequencing, dual-indexed sequences using indices developed by Kozich et al. (218) were created as follows: 2µl of preceding *viuB*-PCR reaction were used as template for a second PCR reaction of just 2 cycles with the same reagents as mentioned above. Forward and reverse primers used in this reaction consisted of appropriate Illumina-adapters, a sample-specific 8 nucleotide index sequence, a 10 nucleotide pad, 2 nucleotide linker and the gene specific sequence described above. (see (218)). This tagging PCR reaction was performed as follows: Initial denaturation 98°C for 30 sec, followed by 2 cycles of denaturation 98°C for 10 sec, annealing 55°C for 6 sec and extension 72°C for 1 sec and final extension 72°C for 1 min. This strategy of amplification using gene-specific primers and subsequent tagging to create dual-indexed PCR products was performed both to improve yield of PCR reactions (as direct amplification using indexed primers often did not succeed or resulted in only very faint bands) and to prevent biased amplification due to unexpected interaction of non-primer sequences with template.

In order to obtain a sufficient concentration of amplicons for subsequent sequencing, 8 tagging reactions were performed for each sample. These reactions were pooled and run on a 2% agarose gel in 1x Tris-Acetate-EDTA Buffer. This reaction created a product consisting of two bands of very similar size, a smaller band representing only half-tagged PCR products, and a slightly bigger band of fully tagged product. The larger bands were cut out of the gel. PCR products were then purified using Wizard SV Gel and PCR Clean-Up System (Promega) according to the instructions by the manufacturer. Concentration of cleaned up PCR products was then measured using a Qubit Fluorometer (ThermoFisher) with a Qubit dsDNA HS Assay Kit (ThermoFisher) and pooled together in equal concentrations (>10ng/ul). The pooled samples were then concentrated using a Wizard SV Gel and PCR Clean-Up System (Promega) according to the instructions by the manufacturer. Quality control of the pooled and concentrated sample was performed using an Agilent 2100 Bioanalyzer. Sequencing was performed using Illumina Miseq technology with a V3 600cycle reagent kit.

5.3.4 Sequence analysis

De-multiplexed raw reads were processed in *R* using the DADA2 pipeline, as described in (214). DADA2 uses an error rate model calculated from the both the abundance of reads as well as their quality scores to achieve single base pair resolution of amplicons. Forward and reverse reads were trimmed due to a drop-off in read quality in the first 10 bp as well as after 240bp and 160 bp for forward and reverse read respectively. Furthermore, reads with a maximum expected error rate > 1 were discarded. After this procedure, 4938 unique sequences remained in the dataset. Chimera detection implemented in DADA2 was then performed on pooled samples, leaving a total of 2278 unique sequences. To account for the possibility of real chimeras of protein coding genes in closely related organisms (due to recombination or homoplasic mutations), chimeras were compared with a reference dataset of *viuB*-alleles found in 782 sequenced *V. cholerae* genomes. 6 “real” chimeras were then included back into the dataset. Among the remaining unique sequences, 129 of these proved to be alleles of *viuB*. Only 43 *viuB*-alleles were composed of more than 1000 reads (with an average of 100,000 reads per sample) and considered for further analysis. 24 unique allleles were found both in multiple samples and also corresponded to a *viuB* allele found in a dataset of 782 *V. cholerae* genomes (obtained from NCBI) and could thus be considered unambiguously “real”. The remaining alleles fell into three categories: (I) present in multiple samples but differing in only 1 bp from “real” alleles, (II) present in multiple samples and differing in more than 1bp from other alleles, and (III)

present in only a single sample. Since our mock community containing only known *viuB*-alleles showed that occasional alleles 1bp away from real sequences remained after error correction, we remapped alleles of category I onto the confirmed real alleles. 2 unique alleles fell in category II and were considered unambiguously novel *viuB*-alleles. 16 alleles of category III were discarded for a conservative estimate of 26 *viuB*-alleles present in our dataset. Further analysis of population structure was performed using the PRIMER-E Software Suite (Quest Research Ltd).

5.2.4 Whole genome sequencing and analysis

73 genomes corresponding to 17 different *V. cholerae* clonal complexes and 7 singletons isolated from the Oyster Pond and lagoon, as previously identified by multi-locus sequence typing (162) were chosen for whole genome sequencing by Fabini Orata as described in (140). De-novo assembly of reads into contiguous sequences was conducted using CLC Genomics Workbench 5.0 (CLC Bio, Aarhus, Denmark). Assembled genomes were aligned using mugsy, standard settings (128). Short locally collinear blocks (LCBs) smaller than 500bp were removed and alignments converted into FASTA format using the Galaxy Web Server (129). Geneious 6.1.7 was then used to visually check the quality of alignments and to remove all positions containing one or more gaps (219). This was performed for all members of each CC separately to assess the variation within a CC, and with representative members of each CC and ST to assess the variation between them. A maximum likelihood phylogenetic trees constructed using the GTR+GAMMA substitution model implemented in RAxML 8.0 (130).

Additionally, a dataset of all available *V. cholerae* was downloaded from NCBI and a core-genome phylogenetic tree constructed using Parsnp (220). The phylogenetic tree was then visualized and annotated in iTOL (221) according to *viuB* types found using local blastn searches against all genomes. Typing of type VI secretion system genes was performed as described previously (222).

5.3.5 qPCR

Real-time quantitative PCR was performed by Tania Nasreen for the enumeration of all *V. cholerae* using a *viuB*-based primer set, and for PG *V. cholerae* using a primer set amplifying the *rfb*O1 region characteristic for this group (Nasreen et al., in preparation). The *viuB*-probe

was 5'-/56- FAM/TCATTTGGC/ZEN/CAGAGCATAAACCGGT/3IABkFQ/-3', fw-primer 5'-TCGGTATTGTCTAACGGTAT- 3' and rev-primer *rev: 5'- CGATTCGTGAGGGTGATA-3'.* The *rfb*O1 probe was 5'-/5HEX/AGAAGTGTG/ZEN/TGGGCCAGGTAAAGT/3IABkFQ/-3', fw-primer 5'-GTAAAGCAGGATGGAAACATATTC-3' and rev-primer 5'-TGGGCTTACAAACTCAAGTAAG-3'. Reaction conditions in an Illumina Eco Real-Time PCR system were as follows: Activation at 95°C for 2 min followed by 40 cycles of 95°C for 15 s, 60°C for 1 min. The volume of the reaction was 10µl, including 5 µl of master mix*, 1 µl of primer-probe, 2 µl of molecular grade water and 2 µl of template. The assay included standards of known copy number and no template control for each master mix to assure the reaction is contamination free. The limit of detection determined as 3 copies per reaction. The 2X QPCR Mastermix (*Dynamite*) used in this study is a proprietary mix developed, and distributed by the Molecular Biology Service Unit (MBSU), in the department of Biological Science at the University of Alberta, Edmonton, Alberta, Canada. It contains Tris (pH 8.3), KCl, MgCl 2 , Glycerol, Tween 20, DMSO, dNTPs, ROX as a normalizing dye, and an antibody inhibited Taq polymerase.

5.4 Results and Discussion

5.4.1 A small region of *viuB* offers base-pair exact differentiation of *V. cholerae* strains

In a previous study, we categorized *V. cholerae* isolates gathered from Oyster Pond, MA, USA, through a multi-locus sequence typing strategy using 7 partially sequenced housekeeping genes (162). Here, we initially aimed to find a single gene region smaller than 300bp in the genome of *V. cholerae*, allowing us to differentiate five clonal complexes (CCs) dominating the population in our isolate-based study. The size restriction of 300bp was imposed to allow for the amplification of this gene region using 300bp overlapping reads in a Illumina Miseq V3 technology kit. Screening a dataset of single-copy genes found in *V. cholerae* but not in closely related species, we identified a hypervariable stretch of 292bp flanked by conserved regions in the *viuB* gene and developed a strategy to successfully amplify this gene from DNA extracted from environmental samples. Using a high-resolution sample inference algorithm (214) and careful curation of the resulting reads to differentiate real *viuB*-alleles from PCR-artifacts (see methods), we were able to differentiate the majority of CCs previously identified using 7 partially sequenced housekeeping genes (162) using a stretch of 272bp of *viuB* (20bp shorter than the variable region of 292bp to account for reduced read quality in the last 10bp of Illumina reads). In three instances (CC2+15, CC6+8 and CC9+12), an identical *viuB* sequence can be attributed

to the close phylogenetic relationship between those strains (Figures 5.2, 5.7). This method was then applied on time-series data of DNA extracted from water samples taken over the course of multiple months, overlapping with our previous isolation-based approach.

To benchmark the ability of this technique to detect *viuB*-alleles corresponding to specific *V. cholerae* strains in a semi-quantitative way, we compared the proportion of isolates of specific CCs isolated in a single sampling effort in the Oyster Pond and adjacent lagoon (162) with the proportion of reads of corresponding to *viuB*-alleles obtained from the same samples (Fig. 5.1a).

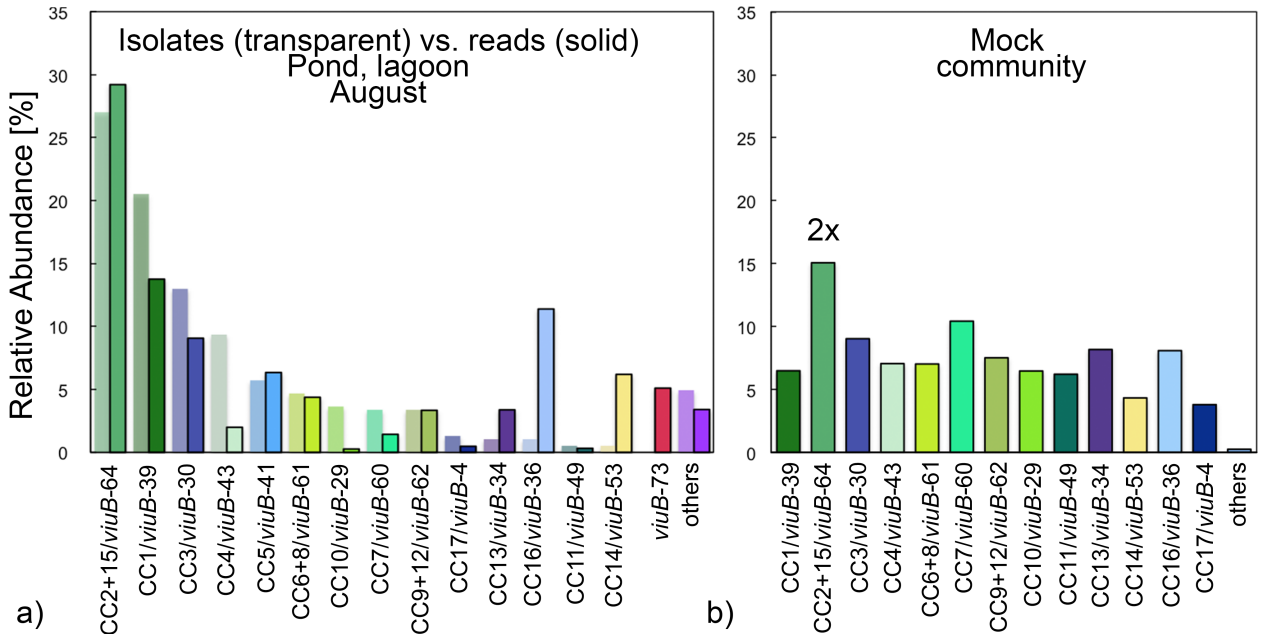


Figure 5.1: Comparison of *viuB*-amplicon sequencing with isolation in the detection of *V. cholerae* strains. a) Relative abundance of isolates sampled from Oyster Pond and adjacent lagoon from four different size fractions (162) is compared to relative abundance of *viuB* reads corresponding to these isolates in the same month. Transparent bars represent isolates, solid bars represent reads. To account for different read depths, subsampling to the sample with the lowest read number was performed. b) Performance of *viuB*-sequencing on a mock community of 12 genomes with different viuB-alleles in equal concentration, and one genome in double concentration. Other reads due to PCR errors or undetected chimeras correspond to <0.25% of the total mock community.

Additionally, to assess potential bias of our technique in the amplification of specific *viuB*-alleles, we performed *viuB*-amplification on a mock communities consisting of 13 different *viuB*-

alleles in equal concentration (one allele in double concentration, Fig. 5.1b). Fig. 5.1a shows that the relative abundance of many *viuB*-alleles (CC1, CC2+15, CC3, CC5, CC6+8 and CC9+12) corresponds rather closely with the previously assessed relative abundance of isolates. In some cases (CC4, CC7, CC10, CC11 and CC17), a larger number of isolates was obtained than corresponding *viuB*-reads would indicate. Furthermore, multiple CCs (CC13, C14 and CC16) were found in much lower relative abundance than corresponding *viuB*-reads. However, as the mock community in Fig. 5.1b indicates, biases in amplification are rather minimal, with the *viuB*-reads that are overrepresented compared to isolates CC14 and 17 actually amplified at a slightly lower rate than others. Thus, it seems reasonable to assume that the differences in culture and culture-independent methods are not necessarily due to biased amplification. Overrepresentation of isolates compared to reads can be attributed to variation due to the relatively low number of strains obtained (385 in a single month). On the other hand, overrepresentation of reads compared to isolates can have multiple reasons. One is the possibility that the same *viuB*-allele is found in multiple strains. The other is the reduced culturability of some *V. cholerae* strains compared to others. Under conditions of stress, *V. cholerae* and other bacteria are known to enter a viable but non culturable state that still allows growth, yet prevents isolation using standard methods (223). Indication for this is found in the prevalence of *viuB*-73, completely lacking in Oyster Pond isolates possessing this allele.

5.4.2 Most *viuB*-alleles are specific to closely related strains of *V. cholerae*

To assess the capability of *viuB* in differentiating not just strains corresponding to Oyster Pond isolates, but *V. cholerae* strains in general, we assembled a dataset of all 782 *V. cholerae* genomes available to date, including genomes of all CCs found in our sampling site that were sequenced after the development of our primer set (Fig. 5.3). This dataset contained a total of 70 unique *viuB*-alleles of 272bp. In pairwise comparisons between unique *viuB* regions, the majority of allele pairs (>99%) differed by at least 2bp from each other (Fig. 5.3), largely eliminating the possibility of misidentifying alleles due to single base pair errors resulting from PCR errors or false base-calling.

To identify *viuB*-alleles shared between strains, we constructed a phylogenetic tree of all *V. cholerae* genomes using the core-genome algorithm implemented by Parsnp (220) and mapped the *viuB*-alleles of all CCs onto the resulting cladogram depicted in Figure 5.3 (as indicated by coloured squares denoting CC identity and numbered *viuB* allele). We furthermore mapped all additional *viuB*-alleles (a total of 26, including CC alleles) that we unambiguously identified as

representing “real” environmental sequences rather than PCR artifacts (see methods) onto this cladogram (Fig. 5.3, colored squares of numbered *viuB*-alleles without CC indication).

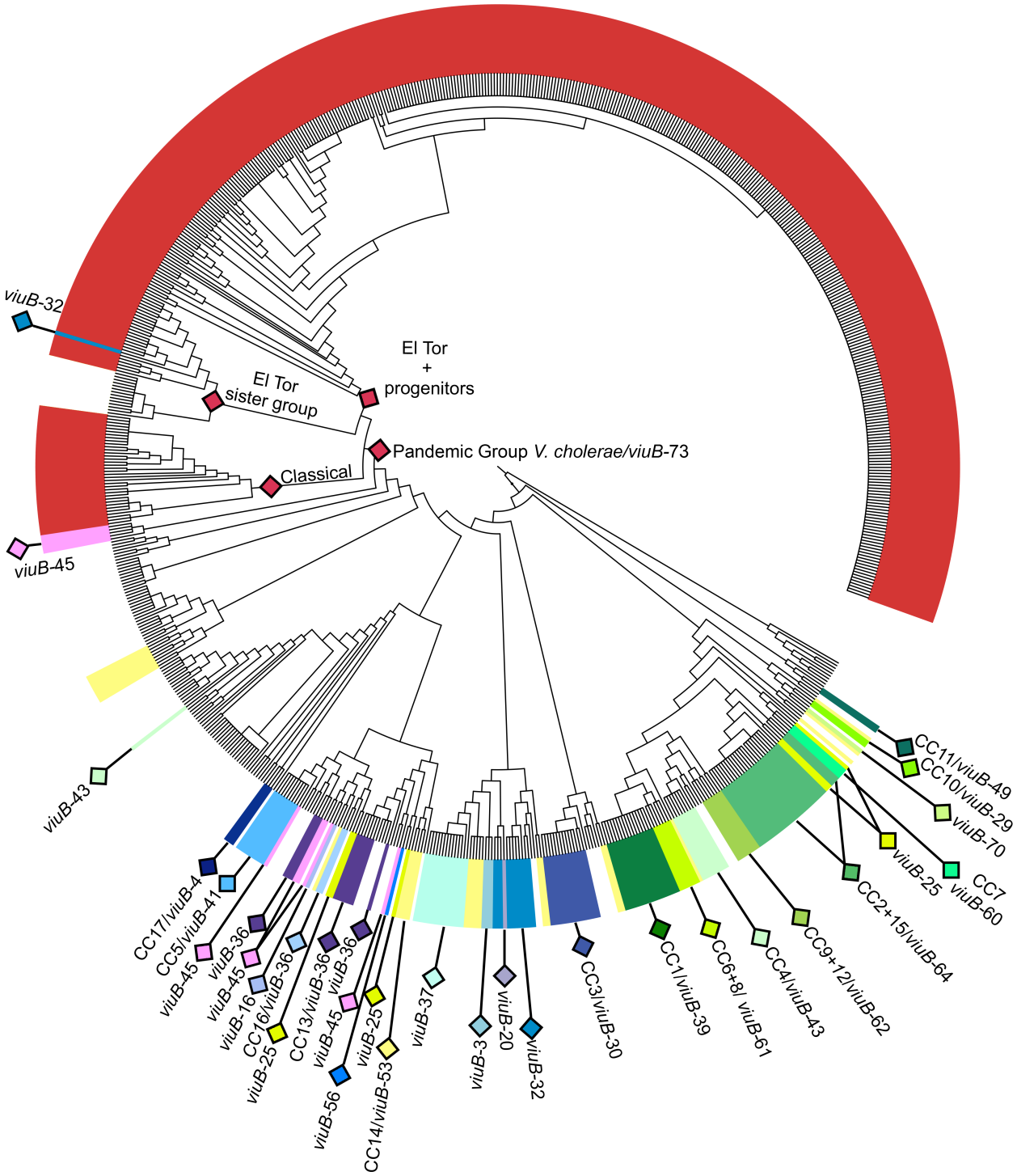


Figure 5.2: Highly variable *viuB* alleles offer strain level differentiation of *V. cholerae* sequences. a) Unique *viuB* alleles from this study mapped on core-genome phylogeny of 782 *V. cholerae* genomes. Phylogeny was created using Parsnp (220) based on a reference genome of N16961. Branch lengths are ignored for ease of viewing. Leaves of the tree are coloured according to the *viuB*-allele found in that particular genome, indicated by coloured rectangles. Alleles corresponding to previously described *V. cholerae* clonal complexes from the Oyster Pond sampling site are denoted with CC and a number, other *viuB*-alleles just as *viuB* and a number. Unmarked bars coloured in witch haze yellow correspond to *viuB*-53, an allele found in multiple phylogenetically distant genomes.

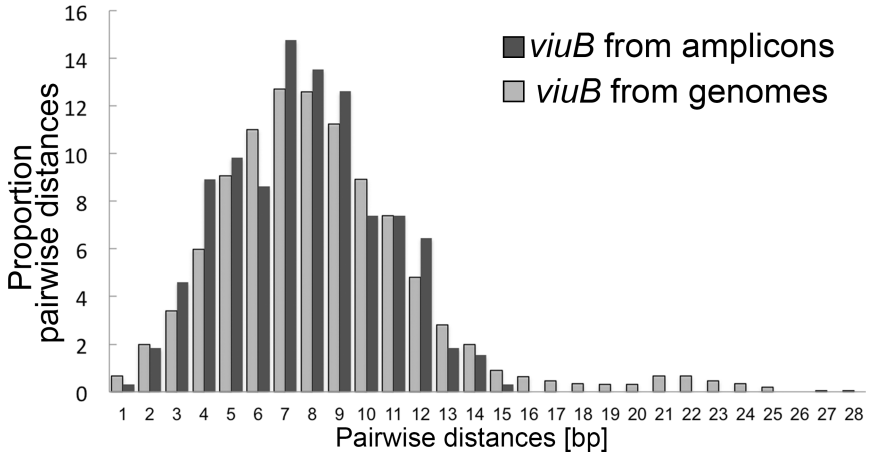


Figure 5.3: Pairwise distance comparisons between all *viuB* alleles amplified from water samples and between *viuB* alleles from sequenced genomes. X-axis denoted the pairwise distance in basepairs, Y-axis the proportion of pairwise comparisons with a specific difference in basepairs.

As Figure 5.2 shows, *viuB*-alleles are generally identical between closely related strains, and only a minority of CCs share their *viuB* allele with other genomes: *viuB*-53 of CC14 is widespread in a number of distantly related *V. cholerae* genomes and can thus not be relied on to identify a single strain of the species. The presence of multiple strains possessing *viuB*-53 could be the explanation for the strong overrepresentation of this allele compared to corresponding isolates of CC14. *viuB*-36 of CC13 is also found in two closely related groups of strains from Bangladesh and might not necessarily represent a single CC. The *viuB*-43 allele of CC4 is furthermore found in a single strain from India, 1157-74. In addition to these overlaps, a few rare *viuB*-alleles, *viuB*-45 and *viuB*-25 are found in multiple, distantly related lineages. Overall however, we observe close correspondence of *viuB*-alleles with phylogeny, allowing the use of this gene as an approximate marker for strain specific detection of *V. cholerae*.

5.4.3 Detection of *viuB*-alleles unique to the pandemic-group *V. cholerae* in a cholerae-free region

Most notably in Figure 5.2, almost all 441 *V. cholerae* isolates belonging to the so-called pandemic group (including all strains of the 6th and 7th cholera pandemic and non-pandemic relatives (163)) possess a single *viuB* allele that is not found in any strains outside of this clade – *viuB*-73. The only exceptions are the non-pandemic Russian strain I-1471 that shares the allele *viuB*-32 with a group of non-pandemic isolates, and several non-pandemic members of the so-called El Tor sister group (163) (not coloured in Figure 5.2). As Figure 5.1a. shows, around 5% of all *viuB*-reads amplified from water samples of the Oyster Pond and lagoon in August 2009 were *viuB*-73 and potentially represent *V. cholerae* of the pandemic group. As we performed *viuB*-amplification from fractionated water samples over the course of five months (June-October 2009) at three locations (pond, lagoon and ocean), we were able to identify *viuB*-73 sporadically in all three locations (Fig. 5.4). To discount the possibility of PCR-contamination with PG-type *viuB* amplicons, we performed *viuB*-probe based qPCR enumerating the total abundance of *V. cholerae*, and specific enumeration of the *rfB*O1 gene exclusive to O1 serogroup strains on all eight *viuB*-73 positive samples. (Table 5.1.).

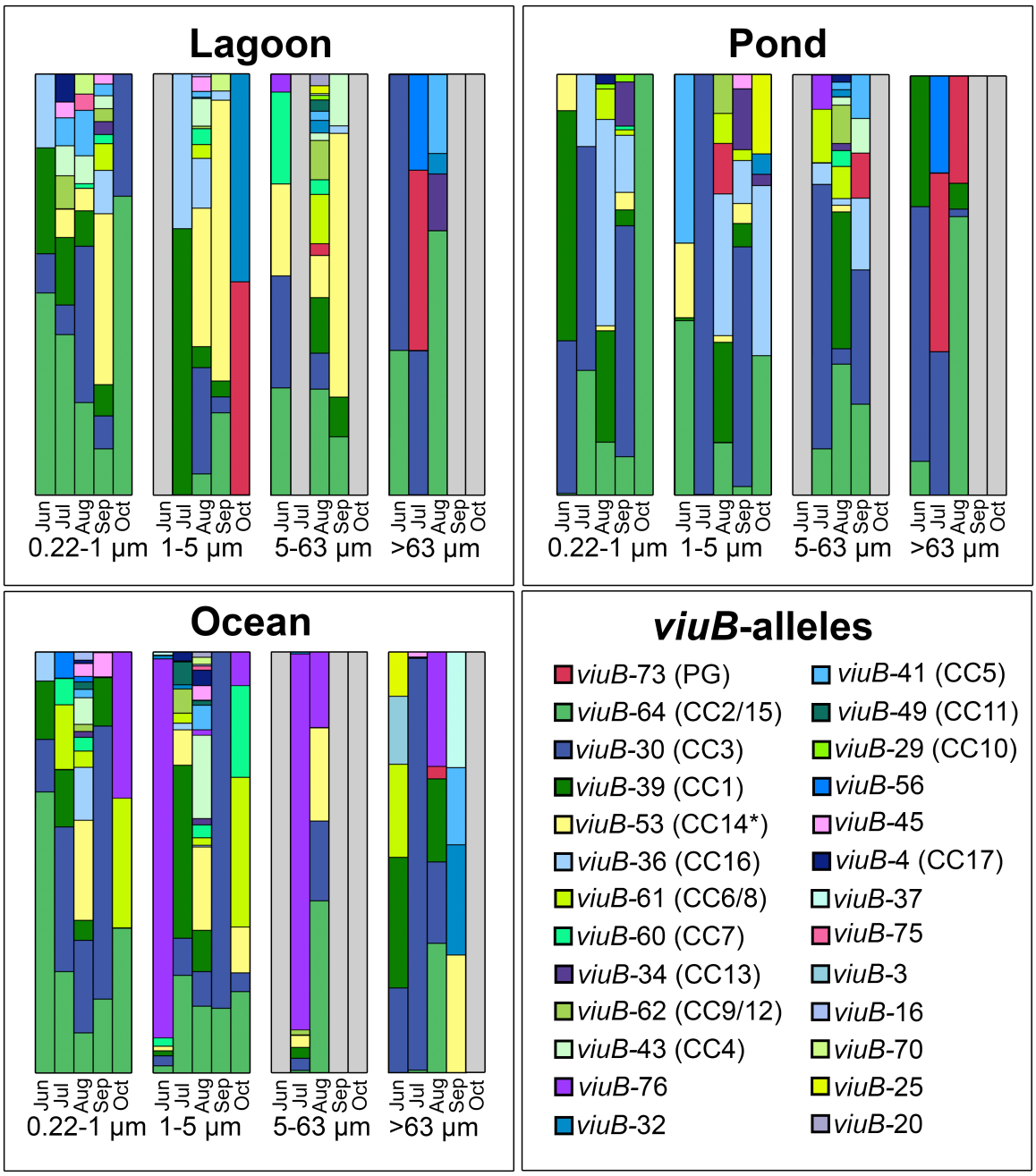


Figure 5.4: Relative abundance of *viuB*-reads in a time-series of fractionated water samples. Read numbers were subsampled to the sample with lowest reads. Full grey bars indicate missing data due to lack of amplification. Asterisk in viuB-53/CC14 indicates strong polyphyletic signal for this allele, which should not be considered strain specific.

Table 5.1: Enumeration of total *V. cholerae* and PG *V. cholerae* through qPCR

Sample	*viuB*/ml	*rfb*/ml	*rfB*O1/*viuB*	*viuB*-73/*viuB*
Lagoon Jul >63 µm	468	188	40.19	42.81
Lagoon Aug5-63 µm	5	2	37.17	2.73
Lagoon Oct1-5 µm	54	12	23.36	50.46
Pond Jul >63 µm	810	265	32.72	42.41
Pond Aug >63 µm	209	82	39.29	29.2
Pond Sep 5-63 µm	8	0	0.00	10.69
Pond Aug 1-5 µm	n/d	n/d	n/d	11.96
Ocean Aug 63 µm	121	85	70.37	2.95

One sample had degraded to the point where qPCR could no longer be performed, but qPCR succeeded in the seven other samples. In six of these samples, we could successfully enumerate *rfb*O1, indicating the presence of DNA from O1 serogroup strains and discounting the possibility of amplicon-contamination. While we can not expect 1:1 correlation of a quantitative method to our amplicon-based approach, the proportion of total *V. cholerae* to PG *V. cholerae* assessed via qPCR more often than not corresponded rather closely to the proportion of total *viuB* to *viub*-73. In two cases (Lagoon August 1-5µm and Ocean August >63µm), the proportion of PG *V. cholerae* assessed by qPCR considerably exceeded that of total *viuB* to *viuB*-73. A possible explanation for this could be the sporadic presence of *rfB*O1 in non PG-strains (103). However, all other *viuB*-alleles detected in these particular samples correspond to known strains that do not contain *rfb*O1, so the cause of this discrepancy is unclear. Another sample fell very close to the reliable detection limit of qPCR, potentially giving unreliable results for one of the two methods. Another sample (Pond September 1-5µm) was positive for *viuB*-73 yet did not show any *rfB*O1 amplification, again potentially because of the very low cell number in that sample. Additional samples that did not contain *viuB*-73 were also found to be *rfB*O1-negative.

Nonetheless, the combination of amplicon sequencing of *viuB* and qPCR provides substantial evidence that PG *V. cholerae* strains are part of local populations on the of the United States east coast.

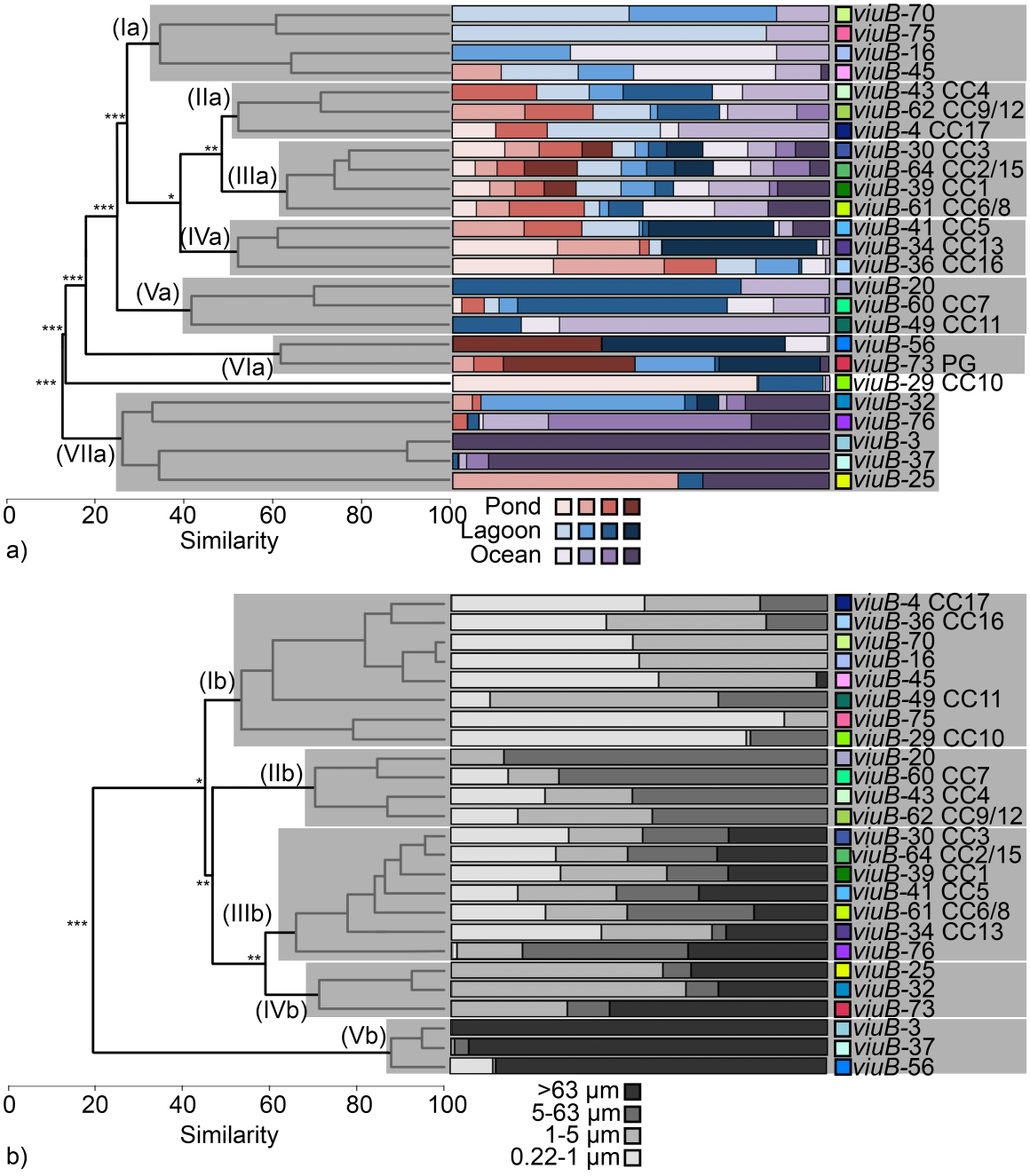
5.4.4 Bloom and bust cycles and specific habitats for different strains of *V. cholerae*

In a previous culture-based study, we noted that the *V. cholerae* population in Oyster Pond is dominated by a small number of phylogenetically distinct CCs, with substantial fluctuations over the course of a month (162). In this study, we covered a larger timeframe consisting of an entire

summer/fall season of the months of June-October with DNA extracted from fractionated water (filtered through 63-,5-,1- and 0.22um filters). As Fig. 5.4 shows, similar to our previous culture-based effort, few *viuB*-alleles (*viuB*-64, *viuB*-30 and *viuB*-39, corresponding to the most numerously isolated CC2+15, CC3 and CC1) are present in the majority of most samples, yet their relative abundance changes considerably from June to October in all locations and size fractions. A much larger number of alleles are detected only sporadically, occasionally increasing in relative abundance but generally not stably persisting in any particular location or fraction size. This again confirms our previous culture-based results that showed just a small number of CCs persisting between two consecutive samplings (162).

As previously noted, filter sizes serve as proxies for different bacterial habitats, from pelagic bacteria in the smallest size fraction to smaller particles and ultimately zoo- and phytoplankton (65). Based on the genomic divergence of sympatric isolates, we previously hypothesized that various strains of V. cholerae might be in the process of evolutionary divergence from each other and not necessarily inhabit the same niche. However, due to the relatively low number of isolates from each potential particle-size niche, we were only able to differentiate between a lagoon and pond location, with most strains showing slight tendencies towards one or the other habitat (162).

However, our amplicon-based method offers much greater statistical power. To infer potential habitats of different *V. cholerae* strains we performed similarity profile analysis (122) of the occurrence of specific alleles in the size fractions of the three locations. Figure 5.5 shows statistically significant clustering both based on the combination of location and size fraction as well as on size fraction alone. 7 different statistical clusters exist on the level of location-size fraction: (Ia) Mostly confined to pond and lagoon and either pelagic or confined to small particles <5μm, (IIa) mostly found pond and lagoon but absent from the largest particles, (IIIa+IVa) found in all habitats and size fractions, (Va) found mostly in lagoon and ocean, (VIa)found mostly in pond and lagoon with no pelagic cells, and (VIIa) mostly lagoon and ocean, with a large number of zooplankton/phytoplankton associated. Similarly, clustering based on fraction size alone shows multiple potential habitats: (Ib) Mostly pelagic and small particles, (IIb) absent on zooplankton, (IIIb) present on all size fractions, (IVb) found on particles but not pelagic and (Vb) mostly found on zooplankton. Clustering based on location alone only significantly differentiates allleles predominantly found in the ocean versus more generalist or pond/lagoon associated alleles (Figure 5.3c).


(Ia)
(IIa)
(IIIa)
(IVa)
(Va)
(VIa)
(VIIa)
viuB-70
viuB-75
viuB-16
viuB-45
viuB-43 CC4
viuB-62 CC9/12
viuB-4 CC17
viuB-30 CC3
viuB-64 CC2/15
viuB-39 CC1
viuB-61 CC6/8
viuB-41 CC5
viuB-34 CC13
viuB-36 CC16
viuB-20
viuB-60 CC7
viuB-49 CC11
viuB-56
viuB-73 PG
viuB-29 CC10
viuB-32
viuB-76
viuB-3
viuB-37
viuB-25
0 20 40 60 80 100
Similarity
Pond
Lagoon
Ocean
a)
(Ib)
(IIb)
(IIIb)
(IVb)
(Vb)
viuB-4 CC17
viuB-36 CC16
viuB-70
viuB-16
viuB-45
viuB-49 CC11
viuB-75
viuB-29 CC10
viuB-20
viuB-60 CC7
viuB-43 CC4
viuB-62 CC9/12
viuB-30 CC3
viuB-64 CC2/15
viuB-39 CC1
viuB-41 CC5
viuB-61 CC6/8
viuB-34 CC13
viuB-76
viuB-25
viuB-32
viuB-73
viuB-3
viuB-37
viuB-56
0 20 40 60 80 100
Similarity
>63 µm
5-63 µm
1-5 µm
0.22-1 µm
b)

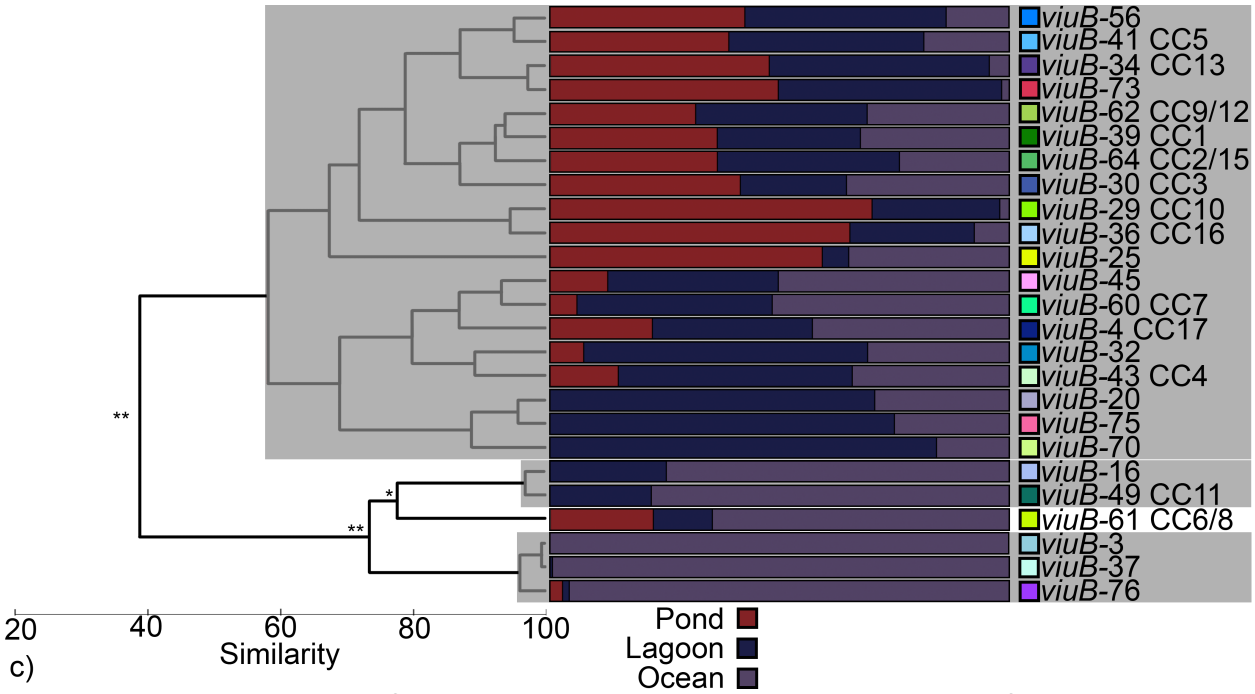


Figure 5.5: Putative habitats of *V. cholerae* strains based on similarity profile analysis. a) Statistically significant clustering of *viuB*-alleles based on their occurrence in samples from various different fraction sizes and locations. b) Statistically significant clustering of *viuB*-alleles based on their occurrence in samples from various size fractions only. c) Statistically significant clustering of *viuB*-alleles based on their occurrence in ocean, pond or lagoon sampling sites. All graphs were created by UPGMA clustering of *viuB*-alleles based on Bray-Curtis similarity as calculated by the relative contribution of samples to the total number of reads for each allele. Subsampling was performed to the lowest sample size. Similarity profile analysis (122) to differentiate statistically significant differences in inferred bacterial habitats from random. Greyed out clusters represent cluster whose members do not differ from each other in their distribution, but from those of other clusters. *** $p < 0.001$, ** $p < 0.01$, * $p<0.05$).

It should be noted that many strains are found only in a small number of samples (see Fig. 5.4), and thus their absence in certain sample types might not necessarily be biologically meaningful. *V. cholerae* strains present in most samples (represented by alleles *viuB*-64/CC2+15, *viuB*-30/CC3 and *viuB*-39/CC1) can be considered generalists in regard to size fractions, although displaying some differences in locational preference. A few strains found both in a large number of samples and display a distribution significantly different from others can also be identified. For example, the *viuB*-76 allele, not corresponding to any sequenced *Vibrio* genome, appears to be mostly excluded from the pond-lagoon ecosystem and is predominantly found in the ocean. Similarly, alleles *viuB*-61 and *viuB*-49 (corresponding to isolates CC6/8 and CC1 respectively) are predominantly found in the ocean and lagoon samples, perhaps indicating a higher salinity tolerance in these lineages. Conversely, *viuB*-56, corresponding to a singleton sequence type which we previously isolated, and *viuB*-73,

corresponding to PG *V. cholerae*, are rarely found in ocean samples. Allele *viuB*-73 is furthermore absent from the pelagic size fraction. Considering the large number of different other *viuB*-alleles found in the pelagic size fraction, the complete absence of this allele is indicative of a more particle/host associated niche for strains harbouring that particular allele. Zooplankton and other particles have in fact long been suggested as the primary niche of the *V. cholerae* species *(101, 102).* In areas where cholera is endemic, filtering of drinking water through folded cloth removes zooplankton and significantly reduces the incidence of cholera (224). Paradoxically, considerable numbers of *V. cholerae* are found in pelagic forms. A specific niche for *V. cholerae* of the pandemic group could reconcile these two conflicting findings. Divergence of ecological niches on the level of particle association is not an unprecedented phenomenon in *Vibrio*. The marine bacterium *V. cyclitrophicus* is separated in two lineages: One steadily attached to particles by biofilm formation, and one only loosely associated with particles and capable of rapid exploitation of newly emerging nutrient hotspots (225). Curiously, this divergence in ecology is mediated by a few 'ecological islands', horizontally transferred regions that have swept through a diverse population and homogenize strains of similar ecology only at certain genomic regions. In contrast, extensive diversity is still observed in regions of the genome that are not relevant for the adaptation to the two specific niches (66). A similar situation could be occurring in *V. cholerae* populations, and further genomic analysis might shed light on the potential genomic basis of ecological niches in that species.

5.4.5 Spatially proximal samples harbour different strains of *V. cholerae*

It is important to note the samples from the dataset depicted in Figure 5.4 do not represent a single location, but rather the pooled DNA from three replicate samples taken 5 meters apart. Therefore, these samples present an "average" of three slightly separate samplings. While this is not necessarily a disadvantage in the inference of general habitats, potential differences between locations in close proximity to each other are not considered. A second dataset consisted of DNA extracted from unfractionated water samples collected in the previous year in the months of June-September with 2-3 samples taken 5 meters apart from each other. Importantly, these samples were not pooled but sequenced separately, serving as biological replicates. In 16S amplicon-based studies employing OTU-clustering methods to group similar sequences into species-level groups, it is common that replicate samples taken at the same location are highly similar to each other (212). However, the low resolution of 16S rRNA obscures the picture of within-species diversity. In contrast, this might not necessarily be the

case in this instance, when a more diverse marker gene and methods offering base-pair-level resolution are used. As Figure 5.6a shows, at the subspecies level these biological replicates vary considerably between each other. In the most drastic example, three samples taken 5 meters apart at the Lagoon site in August 2008 do not share any *viuB*-alleles with each other (Fig 5.6b). Thus, it appears that quite often, different strains of *V. cholerae* do not coexist in true sympatry, but rather in what is termed mosaic sympatry (226). In mosaic sympatry, species (or in this case strains) are distributed seemingly at random, co-occuring in some spots or niches and existing separately in others. Conditions of mosaic sympatry are readily provided by an aquatic environment consisting of numerous small patches of resources (organic particles) that undergo frequent turnover, mixing and separating organisms stochastically (227). Mosaic sympatry has been implicated to play an important role in the divergence of bacteria with largely overlapping niches (58). Strains of *V. cholerae* might evolutionarily diverge in exactly this kind of environment. As the whole-genome phylogeny in Figure 5.7 shows, all CCs and additional singleton sequence types previously isolated from the Oyster Pond show a considerable degree of divergence. We calculated that in most cases, members of a single CC differ from each other in only a few hundred bp at most, but members of different CCs generally show distances of 40-60 kbp from each others. As we have previously shown, different lineages of *V. cholerae* (CC1, CC2, CC3, CC5 and CC13) show very similar, but not completely overlapping carbon use profiles (162). Thus, despite using a mostly similar set of resources and sharing a generalist niche that allows both a pelagic and particle associated existence, a considerable amount of genetic divergence has accumulated between CCs. In a fully sympatric environment, niche exclusion would likely lead to the extinction of all but the most well adapted lineage and not allow for this level of genetic divergence between CCs (although phage-mediated kill-the-winner dynamics might prevent purging of diversity (77)).

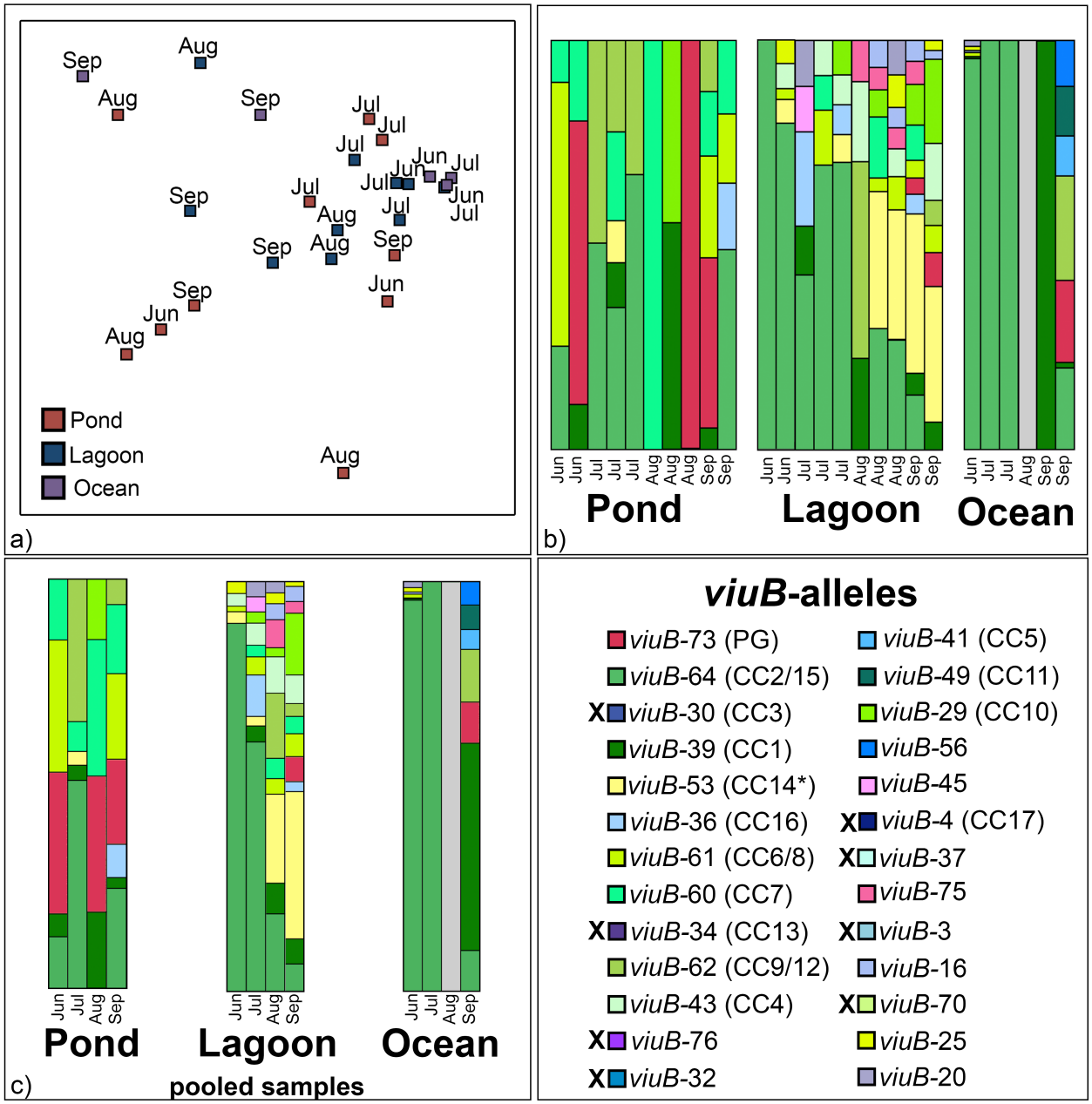


Figure 5.6: No clustering of locations based on proximity of unfractionated samples or month/location. a) MDS graph based on Bray-Curtis similarity of different unfractionated samples. b) Relative abundance of *viuB*-alleles in different samples. c) Relative abundance of *viuB*-alleles, duplicate samples pooled after subsampling to the lowest sample size. Asterisk in *viuB*-53/CC14 indicates strong polyphyletic signal for this allele, which should not be considered strain specific**.** X in front of alleles indicate absence from 2008 sampling.

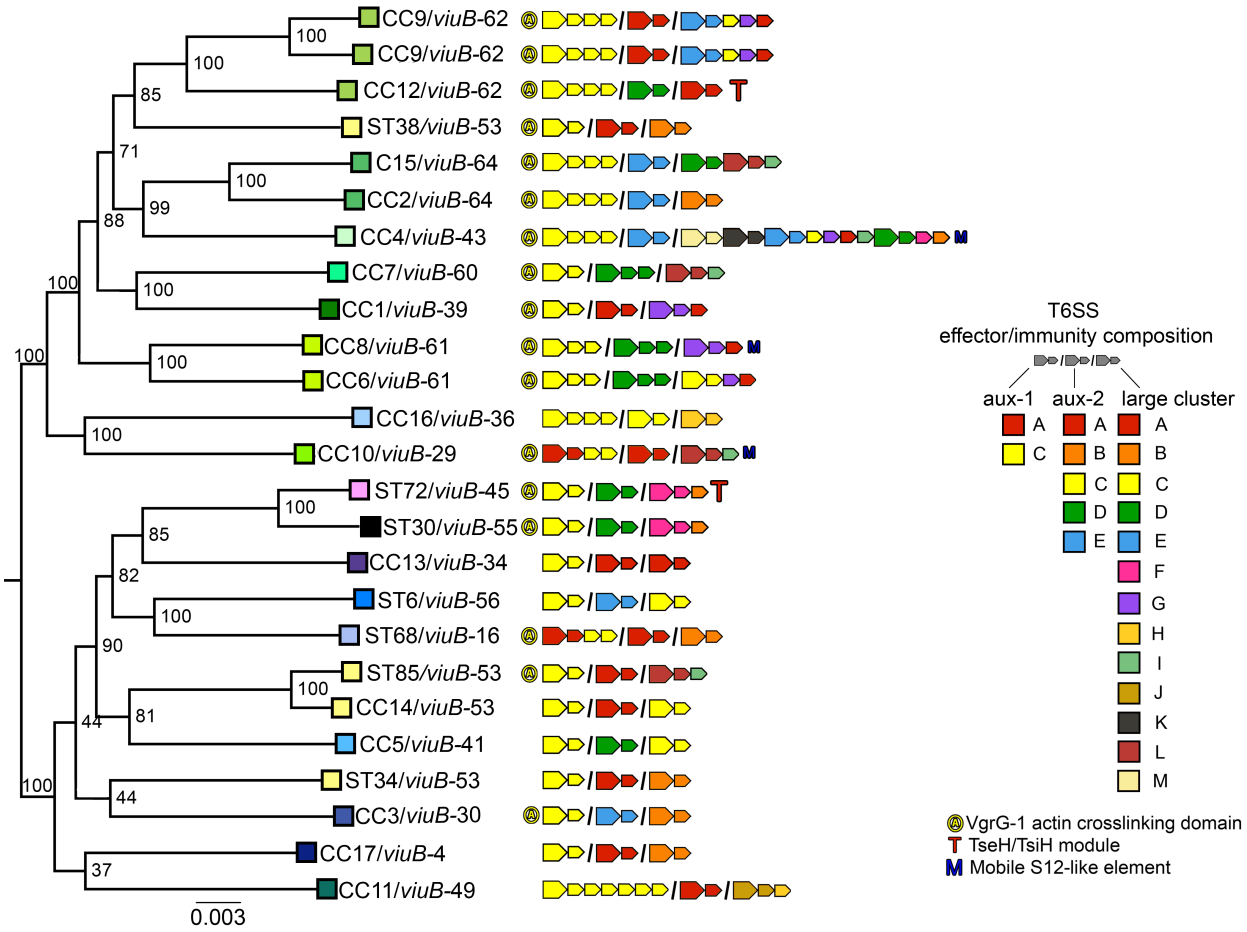


Figure 5.7: Type VI secretion system effector and immunity genes from Oyster Pond isolates. Phylogeny built using the GTR+Gamma substitution model implemented in RAxML on a 2,948,969 bp whole-genome alignment of sequenced Oyster Pond genomes. Large arrows next to strain names indicate effector, small arrows immunity genes, colour of arrows indicate different effector/immunity gene families. Auxiliary cluster 1, 2 and the large cluster are separated by slashes. Statistical branch support was obtained from 100 bootstrap repeats, and bootstrap support for relevant bipartitions is indicated. Scale bar indicates substitutions/site.

Conditions of mosaic sympatry might be provided by *V. cholerae* itself: each *V. cholerae* genome encodes three loci of bactericidal effectors that are injected into other cells by their Type VI secretion system (141). The presence of cognate immunity proteins prevents self-intoxication and killing of kin-bacteria (141), but as we have recently shown, even closely related strains of Vibrios encode different combinations of effector-immunity pairs in each genomic locus, potentially serving as an effective system for discriminating close relatives and competing foreign strains (222). As Figure 5.7 shows, almost every single CC and ST from Oyster Pond encodes a unique combination of effector and immunity proteins. An active Type VI secretion

system could thus play an important role in monopolizing a patch of resources for a single strain of *Vibrio* (228). It could also play an important role in structuring the population: As long as an environment remains stable, the first colonizers will achieve numerical superiority and prevent other strains from successfully gaining a foothold, effectively ensuring the dominance of one or few lineages with different effector-immunity combinations. This could explain the population structure we observe in our sampling: strains that are present in large numbers in one month generally retain their dominance to the next, while many other strains occur sporadically but fail to become established. Thus, initial conditions of a population could have far reaching consequences in the future.

5.4.6 Change in population structure of *V. cholerae* in the span of a year

Comparing Figure 5.4 and 5.6, it is notable that samples from 2009 display a noticeably higher number of *viuB*-alleles than most samples from 2008. Pooling of data from separately amplified replicate samples does not considerably change that picture (Fig. 5.6c). Of course, care must be taken in interpreting this discrepancy - due to separation of different size fractions and the pooling of extracted DNA from replicate samples before PCR in the second, but not in the first year of sampling, the datasets are not directly comparable. By separating a water sample into size fractions, it seems reasonable to assume that rarer alleles that would normally be dwarfed by more numerous alleles found in other size fractions can be detected in size fractions with lower cell number. Nonetheless, some general trends remain obvious: eight *viuB*-alleles that are found in 2009 are not present in any of the locations in the preceding year. This comes to no great surprise for alleles and corresponding strains like *viuB*-4/CC17, which are both rarely found in culture and culture independent methods in the following year. In contrast, *viuB*-30/CC3 is one of the most commonly found alleles and strains from 2009, and completely absent from the preceding year. This could represent a prime example of “first-come-first-serve” dynamics (88), or “survival of the common (229)”, where competitive interactions prevent the establishment of new strains. Only after a strong disturbance in population structure (such as lowered temperatures in the winter months leading to a reduction in *Vibrio* numbers) can new strains become stable members of a population. This appears to have been the case with *viuB*-30/CC3 and potentially other strains carrying different alleles. It is interesting to note that with the exception of *viuB*-30/CC3 and *viuB*-34/CC13, the habitat of most newcomers seems to be mostly limited to the ocean and lagoon location, which might represent more transient habitats, with oceanic currents or human transport serving as vectors for novel strains(230). A further,

perhaps coincidental observation pertains the phylogenetic placement of these newcomer strains: As Figure 5.1 and 5.6 show, the Oyster Pond isolates from 2009 fall into two broad clades: One comprising almost exclusively of strains isolated in the Oyster Pond, with the few other isolates mostly found in the American Gulf Coast (coloured in green tones in all figures), and a second clade consisting of a wide range of international isolates, including all newcomers (coloured in blue tones). One might consider this indication of certain biogeographical factors influencing the diversity of *V. cholerae.*

5.5 **Conclusions**

As with 16S rRNA OTUs or oligotypes, care must be taken to not overinterpret data gained from sequencing of protein coding genes. A *viuB*-allele does not necessarily correspond to a specific group of *V. cholerae* strains, and strains of *V. cholerae* do not necessarily correspond to ecologically meaningful groups. Horizontal gene transfer or convergent evolution can and does erase uniqueness of alleles. Furthermore, horizontal gene transfer can create natural chimeras that are detected by chimera removal algorithms, and PCR errors can appear as real alleles, confounding the true population structure. Only with comparison to available genomic data from sympatric isolates, confirmation of findings through qPCR and careful curation of data can we begin to deduce information from this novel method. This approach severely under-uses the potential power of this method, as frequently occurring one-off alleles are removed even though corresponding reads occasionally reach numbers comparable to dominant alleles in the population – the population structure could thus be much more dynamic than what we assess in our conservative approach. Furthermore, the diversity of the population itself is probably vastly underestimated – rare alleles can not be considered real until independently verified, and single alleles obviously offer lower resolution than MLST or genomic data. With these caveats in mind, our method nonetheless offers considerable insights into the population structure and dynamics of *V. cholerae*. The dynamics observed over the course of months and between two years make it clear that populations of *V. cholerae* undergo frequent, drastic changes in strain composition. Thus, given the large genetic variability both of potentially ecologically and clinically relevant genes observed in members of this species (103), simply enumerating the total amount of *V. cholerae* cells does not necessarily suffice in giving an accurate insight into *V. cholerae* ecology or disease risk.

While *V. cholerae* is a globally spread species, the distribution of strains might be mosaic both on the meter scale to larger geographic regions. Specific communities of strains appear to

dominate certain locations and prevent the invasion of strains that might be more prevalent in other regions of the world. Only after significant disturbances by outside forces might new strains be able to fill the niches left open by perhaps ecologically equivalent, yet phylogenetically differentiated strains. The invasion of *viuB*-30/CC3 into the Oyster Pond population after a decrease in *Vibrio* numbers in the winter is one example. A more famous example could perhaps be seen with the invasion of earthquake-stricken Haiti by PG *V. cholerae* from Nepal (231). The ecological disturbance caused by climatic factors and the earthquake itself could have, as some scientists have suggested (208), provided the necessary conditions for a change in *Vibrio* population structure, enabling the establishment of foreign strains. In addition, our results show that local PG-*V.cholerae* strains can already reside in cholera-free regions as a non-dominant member of the population and perhaps rise to potentially dangerous dominance after changes in population structure (232).

Chapter 6

General discussion, future experiments and random thoughts

6.1 A tale of two vibrios

The working hypothesis throughout the course of my thesis was that the species of *V. cholerae* is divided into many ecologically differentiated lineages. The discovery and description of *V. metoecus* as in independent species at least partially validates this hypothesis. It also demonstrates a weakness in the previously described phenetic species concept – while *V. metoecus* is clearly delineated genomically and phylogenetically (albeit not using the 16S rRNA marker gene), the differentiation by phenotypic characteristic is problematic. Very few biochemical properties differentiate the species, and singular groups within *V. cholerae* possess traits that were considered characteristic for *V. metoecus* only – in at least one case due to a direct horizontal gene transfer event mentioned in chapter 2. Furthermore, closer investigation of the different anti-protozoan activity of the T6SS of the two species in light of genomic data from chapter 4 might also erode another major phenotypic difference between the species.

In chapter 4 we observe that not only numerous strains of *V. cholerae* (174), but also all closely related species harbour a truncated version of the VgrG-1 protein. This protein still contains the domains homologous to the phage base plate and tail spike, yet lacks the C-terminal actin-crosslinking domain that is thought to be involved in cytotoxicity against human cells and predatory *Dictyostelium* slime molds (169). The inability of *V. metoecus* to kill *Dictyostelium* has previously been described as a distinguishing characteristic of *V. metoecus* from *V. cholerae* (where conserved RTX-toxins enable actin-crosslinking independent of VgrG-1 (233)). A truncated VgrG-1 as well as a lack of *rtx* homologs (140) could explain this difference. However, T6SS regulation is complex (234), and details about the environmental conditions of expression remain largely unknown. Furthermore, as chapters 4 and 5 show, there is a large within-species variability in the T6SS-mediated predator defence repertoire of *V. cholerae* and *V. metoecus*. Some strains of V. cholerae contain a full repertoire consisting of a VgrG-1 actin-crosslinking domain, RTX and the anti-eukaryotic lipase VasX (effector A in the aux-2 locus) (235). Others do not contain any, and many strains contain different combinations of these

effectors, not to mention the possibility of additional T6SS dependent and independent factors. One would therefore expect a certain hierarchy in the efficacy of killing – at the top strains containing all three or more anti-predatory effectors, and at the bottom *V. metoecus* containing none. However, given the large number of strain in-between these extremes, one would also expect to see no clear separation in the ability of *V. cholerae* and *V. metoecus* to kill eukaryotic predators. A more systematic investigation of this topic using strains described in this thesis is currently underway and I expect it to erase this phenotypic difference between the two species.

Regardless of the lack of clear phenotypic differentiation, comparative genomic studies of these two species have proven to be a valuable source of information in investigating the evolution of recently diverged bacterial lineages. Before its formal description, *V. metoecus* was subject to several studies under the provisional name *V. metecus*. Haley et al. noted that the first sequenced genome of *V. met(o)ecus*, RC341, contained a complete pathway responsible for the chitin-mediated initiation of natural competence known from *V. cholerae (152)*. This pathway presumably facilitates levels of horizontal gene transfer similar to *V. cholerae*. Haley et al. noted the presence of multiple horizontally transferred virulence factors associated with pathogenic *V. cholerae* – the virulence regulators *toxR*/*toxS*, hemolysins, lipases as well as incomplete *Vibrio* Seventh Pandemic (VSP) islands I and II. Interestingly, Haley et al. also noted that while the cholera toxin genes *ctxAB* and *tcp*-island genes are missing in *V. metoecus*, the genomic prerequisites for infection by CTX-Phage are presence (152). As *V. meteocus* thus possesses the theoretical potential to acquire additional virulence genes from *V. cholerae* or the same environmental sources as *V. cholerae*, the degree of horizontal gene transfer between those two species has been the main topic of further publications. Boucher et al. showed that sympatrically occurring V. *metoecus* and *V. cholerae* resemble each other more closely based on the content of their chromosomal integron than *V. cholerae* from two different continents (109). In a later whole-genome based analysis of the two *Vibrio* species from Oyster Pond, we showed that the integron regions of those two species are virtually indistinguishable ((140), see appendix). We also observed highly directional horizontal transfer of core genes from *V. cholerae* to *V. metoecus*. We hypothesized this to be due to the unequal numbers of the two species present in Oyster Pond – since *V. cholerae* appears much more abundant than V. metoecus, more DNA of that species should be available for horizontal gene transfer. This hypothesis however is based only on the number of isolated bacteria, which might differ from the actual number present in the environment. Such a situation could severely skew the enumeration of *V. cholerae* by qPCR methods that do not differentiate the two species. The development of a high-throughput sequencing method to directly target *V. cholerae* but exclude

V. metoecus described in chapter 4 has led to an interesting spin-off by my colleague T. Nasreen: The development of a new qPCR primer set based on the *V. cholerae*-specific *viuB* gene to allow for a more accurate enumeration of the abundance of this species in the environment (Nasreen et al., in preparation).

It is interesting to note that not just in the case of the Oyster Pond sampling, *V. metoecus* is isolated in much rarer occasions than the relatively ubiquitous *V. cholerae*. Our own sampling efforts in Bangladesh have to this day not resulted in the isolation of *V. metoecus*, and *V. metoecus*-specific qPCR has failed to detect this species in water samples from this country. To date, only two strains have been isolated from locations outside of the United States East Coast – one in Germany (26) and one in Italy (210), although the lack of mention in the literature could be due to the decreased pathogenic and thus publication potential compared to its sister species. Another factor could be the aforementioned possibility of a severe isolation bias towards *V. cholerae.* Nevertheless, the rarity of isolation of *V. metoecus* warrants discussion, especially in light of the near complete overlap of these two species in major niche dimensions that differentiate *V. cholerae* from congeneric species (137).

We can speculate that biotic factors might be the cause of our observation. A likely influence in the different prevalence of these species is the composition of the prokaryotic microbiota of the Oyster Pond and lagoon, where *V. metoecus* is predominantly found in the lagoon). As abundances of bacterial OTUs correlate more strongly with each other than with abiotic factors or eukaryotes in marine environments (138). However, protozoan feeding has also been shown to be a significant element in control of *V. cholerae* numbers (236), and *V. cholerae* employs its type 6 secretion system (T6SS) to prevent eukaryotic predation (84). The lack of effective defences against protozoans perhaps prevents *V. metoecus* from gaining a foothold in the pond, where predatory protists might preferentially feed on them (237).

6.2 **Population structure through time and space**

The population structure of *V. cholerae* in Oyster Pond proved to be considerably more clonal than expected. As mentioned before, these expectations were influenced by dozens of previous studies assembling historically and spatially separate isolates into “populations”, perhaps misrepresenting the actual structure. However, it is also possible that our study is simply an outlier. A recent survey of *V. cholerae* diversity based on MLST of 472 isolates from the Austrian saline lake Neusiedlersee represents the only comparable study sampling a large

number of *V.cholerae* from a single location (207). The authors claim that the population structure of this lake varies considerably from the one in Oyster Pond. They isolated *V. cholerae* from 3 habitats: The vast reed habitat on the shore of the lake, “open” waters (the maximum depth of the lake is less than 2m) and an intermediate region between the two. While they did not cluster their isolates into clonal complexes, their concatenated phylogenies of 4 partially sequenced housekeeping genes found 95 sequence types falling within 41 major clades. 18/19 sequence types were found in the open water and intermediate habitat, while the reed habitat contained 84 sequence types, 71 of those exclusively found there. This was interpreted as evidence for a considerably higher diversity and rate of recombination for *V. cholerae* in the Neusiedlersee reed habitat compared to Oyster Pond. However, as far as the datasets are comparable, these differences become less striking upon closer inspection. Nearly 40% of their isolates stem from the highly diverse reed habitat that was sampled at a low depth on 38 different occasions between 2011 and 2012, while the rest of their isolates stems from 3 simultaneous samplings at all three habitats. These co-temporal samplings all display *Vibrio* population structure similar to the Oyster Pond, while multiple shallow samplings in the reed habitat over a long time period resemble those found in collections of geographically and temporally widespread isolates. Since the reed habitat serves as the nesting grounds numerous migratory bird species, which have been hypothesized to act as carriers of *V. cholerae* (238), their population might thus not only resemble, but actually represent an international strain collection. This at first glance contradictory study can thus be considered as actually confirming some of the main findings of chapters 2 and 5: The dominance of few *V. cholerae* clonal complexes over a short time period, with strong fluctuations over longer time.

6.3 **Clonal complex ecology**

It is interesting to ponder the emergence of *Vibrio* clonal complexes in light of the models of ecotype theory and other explanatory frameworks for microbial diversity. Are all clonal complexes likely to represent ecotypes with specific niche adaptations? Both the genomic divergence and the presence of genes unique to each clonal complex might point into that direction, yet it should be noted that much of this variation might be neutral (239). As chapter 5 shows, the most dominant members of our *V. cholerae* population occur in both the same location and size fractions, indicating a lack of habitat preference. Similar carbon use profile (see chapter 2) also shed doubt on the assumption that different clonal complexes would represent considerably ecologically divergent lineages. On the other hand, the nucleotide

differences between members of a single clonal complex fall well within the low variation expected within a single ecotype. Furthermore, most clonal complexes isolated from the Oyster Pond are phylogenetically approximately equidistant, their genomes differing in approximately 40-60k bp from each other. As mentioned in chapter 2, this distance appears too wide to be accounted for purely by drift within a single ecotype between periods of selection. Specific lineages are likely to have truly evolved to conquer new niches such as the human gut in the case of PG *V.cholerae.* PG *V. cholerae* has accumulated a considerable number of virulence factors over the course of its evolution (103), including a highly specific T6SS (240) and appears to be more particle-associated than conspecifics. Other lineages such as the bioluminescent CC1 strains described in chapter 2 are also strong candidates for possessing ecologically distinct survival strategies. However, many differences observed between *Vibrio* lineages could have evolved neutrally in allopatry, and given the opportunity, an ecologically equivalent strain can readily migrate into a new environment, as observed with CC3 in chapter 5 – corresponding to the geotype plus Boeing model of ecotype theory (241). The population sizes of *V. cholerae* could also simply be too small and their structures too mosaic to allow for purging of diversity upon the emergence of a superior genotype. Given prolonged existence at relatively low cell numbers in a diverse community of microorganisms, different lineages of *V. cholerae* could remain effectively isolated from one another and potentially diverge to the observed level of nucleotide differences without necessarily having to change their fundamental niche.

6.4 **T6SS as a barrier to gene flow**

T6SS mediated interactions and the extensive horizontal gene transfer of EI modules described in chapter 4 could have a significant influence on the population and evolutionary dynamics of *V. cholerae*. This could be especially true in the initial emergence of diversity among *V. cholerae* and other species with similar systems. The uptake of a novel EI module with the retention of the old immunity gene could initiate a genome-wide selective sweep, as the recipient strain would enjoy a clear advantage over its ancestral lineage and actively eradicate its progenitors. However, the advantage of protection from an intermediate neighbour could be short lived, as ancestral and derived strains would soon become separated and perhaps not come into contact with each other too often. If, as shown in *Myxococcus (242)*, contact dependent competition serves as a barrier to horizontal gene transfer, transfer of an EI module could therefore serve as a starting point of diversification or enhance already existing diversification.

The phylogenetic tree in Figure 5.7 could provide an insight into the occurrence of coupled genomic and T6SS diversification. The most closely related Oyster Pond isolates, belonging to CC14 and ST85, already vary in the large cluster EI modules (albeit not in a pattern indicative of direct acquisition of an EI with retention of the ancestral immunity gene). Conversely, CC6 shows an EI module combination that appears directly derived from its relative CC8, with a C-type EI module in the large cluster replacing the G-type effector yet retaining the G and A-type immunity genes. More distantly related isolates belonging to CC9 based on the MLST scheme from chapter 2 as well as relatively closely related ST72 and ST30 retain the same EI module combinations. However, ST30 possesses a GIVchS12-type pathogenicity island (177) containing an additional T6SS locus (see appendix) that could prevent the peaceful coexistence and horizontal gene transfer between these two strains.

If T6SS serve as a barrier to recombination, levels of horizontal gene transfer should be consistently higher between strains possessing the same EI module combinations than between strains with different combinations. The prerequisite for testing this hypothesis this would be a large enough dataset of sympatrically occurring *Vibrio* lineages isolated at the same time. The Oyster Pond dataset assembled during this thesis is unfortunately not ideal for this type of analysis, since all major clonal complexes differ in their module composition. Only CC17, ST34 and ST38 (with module combination CAB) would be suitable candidates, hardly enough to answer this question conclusively. Furthermore, given the rapid change of EI module composition between strains, periods of EI compatibility could be too short to allow for measureable periods of increased horizontal gene transfer between momentarily compatible strains.

6.5 **On the diversity of T6SS**

With the currently known repertoire of T6SS effectors in *V. cholerae* and close relatives, the species can harbour a theoretical 130 different module combinations (not taking into account immunity genes). 46 out of 130 EI module combinations are observed in the dataset from chapter 4 (consisting of all unique EI combination found in GenBank at the time of publication) and additional Oyster Pond genomes. Since only a relatively small number of *Vibrio* strains has been sequenced, it is likely that more combinations will be observed as more data becomes available. However, it should be noted that some combinations could result in suboptimal loading of the T6SS spear and strong selection against this particular combination due to physical incompatibility of some effector proteins.

While it seems obvious that any one EI module combination would be beneficial due to its bactericidal effect, is there a benefit in possessing a *specific* combination? Different effectors could show different efficacies against various target organisms (243), but it appears unlikely that different lineages of *V. cholerae* would encounter such radically different bacterial adversaries as to warrant fine-scale selection for specific EI module combinations. It seems therefore more likely that the diversity is an effect of intraspecific competition. Hypothetically, there could be a clear hierarchy of EI module combinations, starting with the (at least in-vitro) competitively superior AAA combination of PG *V. cholerae* and ending in a single most feeble EI combination. Alternatively "rock-paper-scissors"-type interactions could lead to situations any single strain can outcompete some other strains, but is in turn outcompeted by others. These competitions could also underlie similar dynamics as those exhibited by different species of *Streptomyces* (229): In a series of pairwise competitions in-vitro, neither a clear hierarchical nor a rock-paper-scissors interaction was observed. Rather, it was found that species fell into different tiers of competitive ability, where species of higher tier would generally outcompete strains of lower tiers, yet be equally matched with those belonging to the same tier. However, numerous instances of lower tier competitors outcompeting specific higher tier strains were also observed. Importantly, in pairwise competitions where unequal numbers of both species were pitted against each other, only very few species were able to establish dominance when outnumbered.

As mentioned in the preceding chapters, a similar positive frequency dependent selection could be at play in *Vibrio* strains: What matters in competition between strains could be less the specific combination of EI modules (which might vary in 1:1 competitive ability) but rather the initial number of bacteria possessing any one combination. Our own efforts to find a hierarchy in T6SS mediated competitive ability of Oyster Pond *Vibrio* isolates have shown that competitions of equal number of non-compatible strains end in stalemates. Only when outnumbering a strain 1:10 or more is a considerable reduction in the cell number of the minority strain observable. As such, simply having *any* EI module combination could be beneficial, as long as it differs from those of competitors and thus prevents them from invading the space occupied by a specific *Vibrio* strain.

Evidence for the hypothesis that the uniqueness of the EI module combination rather than any specific combination is biologically important comes from a simple summation of all observed effector combinations found in chapter 4 and the additional Oyster Pond strains described in chapter 5. Half of all EI combinations are only found in a single strain or group of

closely related strains, and more than 75% in just one or two. If specific EI modules were universally beneficial, one would expect that the combination of frequent horizontal gene transfer and strong selection would spread them throughout a population. Rather, the prevalent situation seems to be that of extreme diversity. A few combinations such as CAA do appear multiple times in distantly related strains. As mentioned before, AAA strains enjoy a considerable competitive advantage over other strains *in-vitro* (141) , and CAA might also be competitively favoured (albeit to a lesser degree). Why would AAA then not be more widespread? Aux-1 C-type as well as aux-2 and large cluster A-type EI modules are “common”, and a fortuitous CAA combination could become independently fixed in multiple strains, while other combinations are subject to frequent turnover. The comparative lack of “superior” AAA strains could simply be explained by the relative rarity of the aux-1 A-type EI module in the environmental gene pool, perhaps due to a relatively recent introduction (the origin of pandemic-group *V. cholerae* has been estimated to have occurred 10,000 to 500 years ago (244)). The same could apply to the many other EI module combinations containing rare EI modules.

A recent study highlighting the recycling of T6SS toxins expands the potential value of orphan immunity genes and could be the reason for their retention in some strains (245). As T6SS activity in *V. cholerae* is unspecific, i.e. targeting of cells is random (246), many T6SS attacks would affect neighbouring cells with the same EI module set. Binding of effectors by cognate immunity proteins then makes them available for reattachment to the T6SS spear. A cell expressing multiple orphan immunity proteins could thus effectively steal effectors from foreign strains and use them against other bacteria. Furthermore, under the assumption that EI incompatibility leads to a reduction in horizontal gene transfer between different lineages of *Vibrio*, a strain expressing a variety of immunity proteins could kill foreign strains and take up their DNA with impunity, rapidly changing its genetic makeup. A recent study by Shapiro et al. showed that pandemic group *V. cholerae* possess a unique combination of alleles of different genes thought to increase virulence, assembled through horizontal gene transfers from environmental strains, which possess only some of these alleles (247). The acquisition of a novel aux-1 A-type EI module while retaining the immunity gene corresponding to the common C-type effector could have accelerated this process. Despite the frequent occurrence of horizontal gene transfer of EI modules and the obvious benefit of such an event, the described dynamics could ultimately prevent gene-specific selective sweeps of superior EI-modules, as their utility would decrease with increased frequency in a population.

Any putative influence of T6SS on the evolution of *V. cholerae* ultimately depends on the population structure and dynamics of this species. T6SS incompatibility could ensure that different lineages of *V. cholerae* are physically separated in situations when they are capable of DNA uptake. If single particles are colonized by just a single bacterial cell before others can settle as well, then genetic exchange is limited to mostly clones and the occasional invaders, effectively homogenizing strains possessing the same EI module combination. If multiple different strains colonize a single particle, the situation would be different. It has been shown both through modelling and also experimentally that a mix of strains displaying contact-dependent incompatibility through T6SS segregates by killing of surrounding bacteria until only well defined assemblages of compatible bacteria remain (228). In such situations, characterized by rampant initial killing between well mixed groups of strains, and continuous killing at the bacterial "frontlines" after segregation of strains, large amount of DNA from other lineages would be available for uptake. It is unclear which scenario predominates under natural conditions. As this work has shown, considerable proportions of the total *V. cholerae* population can be found both on particles and free swimming. Depending on the density of uncolonized particles and free swimming bacteria ready to colonize them, Vibrios either could readily recombine their DNA during co-colonization, blurring established lineages, or remain separated long enough to evolve the differences we observe. Another factor to take into account would be the presence of bacteria other than Vibrios, making it unlikely that any single *Vibrio* strain would be a first colonizer or that two strains would meet on the same particle.

6.6 In the future: Fine-scale and global-scale sampling and experimental microcosms to study the population dynamics of *V. cholerae*

The major drawback of the sampling effort in the Oyster Pond ecosystem is the lack of metadata. Without metadata, the causes for the changes in population structure remain unknown. More recent time-series sampling efforts in Bangladesh have incorporated the measurement of a large amount of environmental parameters from pH to phytoplankton concentrations that can be connected to changes in population structure, perhaps even giving insight into the ecology of certain strains. The question of scale in the abundance of specific strains (often dramatic at a distance of few meters) will have to be addressed by fine-scale sampling of transects in a single location and comparison of populations from distant geographic locales. Amplifying *viuB*-alleles from different volumes of water or perhaps

specifically from single particles could furthermore provide great insights into the micro-scale variation within a bacterial population.

Advances in synthetic microcosms have also made it possible to experimentally answer many questions brought up in this thesis. Datta et al. recently devised a chitin bead incubator that allows the observation of bacterial population dynamics in a naturalized experimental environment (248). Future experiments planned in our lab will use this system in conjunction with the *viuB* marker gene to track the dynamics of *V. cholerae* and the influence of T6SS mediated competition. Multiple different strains can be competed in pairs or groups, both in the presence of chitin beads, enabling contact dependent killing, or in their absence, and outcomes compared. The role of stochasticity or determinism in *Vibrio* populations can easily be tested by running multiple experiments in parallel or by disturbing established populations and seeing if their structure remains similar after recovery. Strains can also be introduced into already established populations and their rise or demise tracked using the *viuB* marker.

6.7 **Concluding remarks**

Questions about the extent, causes and consequences of bacterial diversity pervade the fields of microbial ecology and evolution on all taxonomic levels. The study of *Vibrio cholerae*, perhaps the most infamous member of a large genus of aquatic heterotrophic bacteria, has mostly remained at the periphery of this endeavour. While hundreds of studies have dealt with the evolution and diversity of the PG lineage of *V. cholerae*, the idiosyncrasies of this clade as a recently emerged human pathogen perhaps make it a special case, and many findings from studying that particular groups might have minimal cross-applicability. By focusing on the much more numerous non-clinical lineages of the species, I believe that I have laid valuable groundwork for moving the study of *V. cholerae* from a mostly clinical application standpoint to a more broad perspective that can be incorporated into a greater framework of basic research in microbial evolution. Environmental, ecologically differentiated non-pathogenic strains of *Vibrio cholerae* could play an important role in preventing the spread of pandemic *V.cholerae* across the globe. The continued presence of pandemic *V. cholerae* in the previously cholera-free Haiti after the introduction by UN peacekeepers (231) indicates that environmental factors alone are likely not responsible for the patchy distribution of this pathogen. Global ocean currents and other means of transportation are likely to have spread occasional pandemic *V. cholerae* bacteria across the globe, but local *V. cholerae* (249) might have prevented them from becoming a permanent part of the local flora. Only when overwhelming numbers of pandemic *V.*

cholerae are introduced via a human vector (numbering in the trillions released by a cholera victim), might they be able to gain a foothold against a diverse and locally adapted pre-existing community of *V. cholerae*.

References

1. **Mayr E.** 1942. Systematics and the origin of species, from the viewpoint of a zoologist. Harvard University Press.
2. **Doolittle WF.** 2012. Population genomics: how bacterial species form and why they don't exist. Current Biology **22:**R451-R453.
3. **Shapiro BJ, Polz MF.** 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol **22:**235-247.
4. **Doolittle WF, Zhaxybayeva O.** 2009. On the origin of prokaryotic species. Genome Res **19:**744-756.
5. **Huq A, Haley BJ, Taviani E, Chen A, Hasan NA, Colwell RR.** 2012. Detection, isolation, and identification of Vibrio cholerae from the environment. Curr Protoc Microbiol **Chapter 6:**Unit6A 5.
6. **Coyne JA, Orr HA.** 2004. Speciation. Sunderland, MA. Sinauer Associates, Inc.
7. **Sneath PH, Sokal RR.** 1973. Numerical taxonomy. The principles and practice of numerical classification.
8. **Avise JC.** 2000. Phylogeography: the history and formation of species. Harvard university press.
9. **Mayr E.** 1998. Two empires or three? Proceedings of the National Academy of Sciences **95:**9720-9723.
10. **Vandamme P, Pot B, Gillis M, De Vos P, Kersters K, Swings J.** 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiological reviews **60:**407-438.
11. **Hada H, Stemmler J, Grossbard M, West P, Potrikus C, Hastings J, Colwell R.** 1985. Characterization of non-O1 serovar Vibrio cholerae (Vibrio albensis). Systematic and applied microbiology **6:**203-209.
12. **Chatterjee B.** 1993. Numerical taxonomy of vibrio & allied organisms. The Indian journal of medical research **97:**162-167.
13. **Mackie T.** 1922. The Serological Relationships of the Paracholera Vibrios to Vibrio choleræ, and the Serological Races of the Paracholera Group. British journal of experimental pathology **3:**231.
14. **Woese CR, Fox GE.** 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences **74:**5088-5090.
15. **Fox Gc-a, Stackebrandt E, Hespell R, Gibson J, Maniloff J, Dyer T, Wolfe R, Balch W, Tanner R, Magrum L.** 1980. The phylogeny of prokaryotes. Science **209:**457-463.
16. **Palys T, Nakamura L, Cohan FM.** 1997. Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. International Journal of Systematic and Evolutionary Microbiology **47:**1145-1156.
17. **Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF.** 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature **430:**551-554.
18. **Rosselló-Mora R, Amann R.** 2001. The species concept for prokaryotes. FEMS microbiology reviews **25:**39-67.
19. **Polz MF, Hanage WP.** 2013. Quantitative and Theoretical Microbial Population Biology. doi:10.1007/978-3-642-30123-0_35**:**31-42.
20. **Lawson DJ.** 2012. Populations in statistical genetic modelling and inference.
21. **Mayr E.** 1963. Animal species and evolution. Animal species and evolution.
22. **Lewontin RC.** 1970. The units of selection. Annual Review of Ecology and Systematics **1:**1-18.

23. **O'Malley MA.** 2008. 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. Stud Hist Philos Biol Biomed Sci **39:**314-325.
24. **Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK, Fierer N, Townsend AR, Cleveland CC, Stanish L.** 2011. Global patterns in the biogeography of bacterial taxa. Environmental microbiology **13:**135-144.
25. **Vos M, Wolf AB, Jennings SJ, Kowalchuk GA.** 2013. Micro-scale determinants of bacterial diversity in soil. FEMS microbiology reviews **37:**936-954.
26. **Octavia S, Salim A, Kurniawan J, Lam C, Leung Q, Ahsan S, Reeves PR, Nair GB, Lan R.** 2013. Population structure and evolution of non-O1/non-O139 Vibrio cholerae by multilocus sequence typing. PLoS One **8:**e65342.
27. **Didelot X, Falush D.** 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics **175:**1251-1266.
28. **Wollenberg M, Ruby E.** 2009. Population structure of Vibrio fischeri within the light organs of Euprymna scolopes squid from two Oahu (Hawaii) populations. Applied and environmental microbiology **75:**193-202.
29. **Turner JW, Paranjpye RN, Landis ED, Biryukov SV, González-Escalona N, Nilsson WB, Strom MS.** 2013. Population structure of clinical and environmental Vibrio parahaemolyticus from the Pacific Northwest coast of the United States. PLoS One **8:**e55726.
30. **Rossello-Mora R, Amann R.** 2015. Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol doi:10.1016/j.syapm.2015.02.001.
31. **Gevers D, Dawyndt P, Vandamme P, Willems A, Vancanneyt M, Swings J, De Vos P.** 2006. Stepping stones towards a new prokaryotic taxonomy. Philos Trans R Soc Lond B Biol Sci **361:**1911-1916.
32. **Stackebrandt E, Goebel B.** 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic Bacteriology **44:**846-849.
33. **Thompson FL, Iida T, Swings J.** 2004. Biodiversity of vibrios. Microbiol Mol Biol Rev **68:**403-431, table of contents.
34. **Thompson CC, Amaral GR, Campeao M, Edwards RA, Polz MF, Dutilh BE, Ussery DW, Sawabe T, Swings J, Thompson FL.** 2014. Microbial taxonomy in the post-genomic era: Rebuilding from scratch? Arch Microbiol doi:10.1007/s00203-014-1071-2.
35. **Stackebrandt E, Ebers J.** 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiology today **33:**152.
36. **Richter M, Rossello-Mora R.** 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A **106:**19126-19131.
37. **Auch AF, Jan M, Klenk H-P, Göker M.** 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Standards in Genomic Sciences **2:**117.
38. **Sentausa E, Fournier PE.** 2013. Advantages and limitations of genomics in prokaryotic taxonomy. Clinical Microbiology and Infection **19:**790-795.
39. **Kersters K, Vancanneyt M.** 2005. Bergey's manual of systematic bacteriology. Springer Verlag.
40. **Jain R, Rivera MC, Lake JA.** 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Sciences **96:**3801-3806.
41. **LAMBERT MA, Hickman-Brenner F, FARMER III J, MOSS CW.** 1983. Differentiation of Vibrionaceae species by their cellular fatty acid composition. International Journal of Systematic and Evolutionary Microbiology **33:**777-792.

42. **Erler R, Wichels A, Heinemeyer E-A, Hauk G, Hippelein M, Reyes NT, Gerdts G.** 2015. VibrioBase: a MALDI-TOF MS database for fast identification of Vibrio spp. that are potentially pathogenic in humans. Systematic and applied microbiology **38:**16-25.
43. **Staley JT.** 2006. The bacterial species dilemma and the genomic–phylogenetic species concept. Philosophical Transactions of the Royal Society of London B: Biological Sciences **361:**1899-1909.
44. **Koeppel AF, Wu M.** 2013. Species matter: the role of competition in the assembly of congeneric bacteria. ISME J doi:10.1038/ismej.2013.180.
45. **Konstantinidis KT, Rossello-Mora R.** 2015. Classifying the uncultivated microbial majority: A place for metagenomic data in the Candidatus proposal. Syst Appl Microbiol doi:10.1016/j.syapm.2015.01.001.
46. **Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K.** 2016. A new view of the tree of life. Nature Microbiology **1:**16048.
47. **Cohan FM.** 2001. Bacterial species and speciation. Systematic biology **50:**513-524.
48. **Atwood K, Schneider LK, Ryan FJ.** 1951. Periodic selection in Escherichia coli. Proceedings of the National Academy of Sciences **37:**146-155.
49. **Koeppel AF, Wertheim JO, Barone L, Gentile N, Krizanc D, Cohan FM.** 2013. Speedy speciation in a bacterial microcosm: new species can arise as frequently as adaptations within a species. The ISME journal **7:**1080-1091.
50. **Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL.** 2005. Re-evaluating prokaryotic species. Nature Reviews Microbiology **3:**733-739.
51. **Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF.** 2005. Genotypic diversity within a natural coastal bacterioplankton population. Science **307:**1311-1313.
52. **Polz MF, Alm EJ, Hanage WP.** 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics **29:**170-175.
53. **Majewski J, Cohan FM.** 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics **152:**1459-1474.
54. **Papke RT, Ramsing NB, Bateson MM, Ward DM.** 2003. Geographical isolation in hot spring cyanobacteria. Environmental Microbiology **5:**650-659.
55. **Fraser C, Hanage WP, Spratt BG.** 2005. Neutral microepidemic evolution of bacterial pathogens. Proceedings of the National academy of Sciences of the United States of America **102:**1968-1973.
56. **Cohan FM, Perry EB.** 2007. A systematics for discovering the fundamental units of bacterial diversity. Curr Biol **17:**R373-386.
57. **Cohan FM.** 2005. Periodic selection and ecological diversity in bacteria, p 78-93, Selective sweep. Springer.
58. **Shapiro BJ, Polz MF.** 2015. Microbial Speciation. Cold Spring Harb Perspect Biol **7:**a018143.
59. **Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW.** 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science **344:**416-420.
60. **Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, Mira A.** 2009. Explaining microbial population genomics through phage predation. Nat Rev Microbiol **7:**828-836.
61. **Wiedenbeck J, Cohan FM.** 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev **35:**957-976.

62. **Melendrez MC, Becraft ED, Wood JM, Olsen MT, Bryant DA, Heidelberg JF, Rusch DB, Cohan FM, Ward DM.** 2015. Recombination Does Not Hinder Formation or Detection of Ecological Species of Synechococcus Inhabiting a Hot Spring Cyanobacterial Mat. Front Microbiol **6:**1540.
63. **Vos M, Didelot X.** 2009. A comparison of homologous recombination rates in bacteria and archaea. The ISME journal **3:**199-208.
64. **Vulić M, Dionisio F, Taddei F, Radman M.** 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proceedings of the National Academy of Sciences **94:**9763-9767.
65. **Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF.** 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science **320:**1081-1085.
66. **Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ.** 2012. Population genomics of early events in the ecological differentiation of bacteria. Science **336:**48-51.
67. **Cordero OX, Polz MF.** 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol **12:**263-273.
68. **Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR.** 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J doi:10.1038/ismej.2015.241.
69. **Cohan FM.** 2016. Bacterial speciation: genetic sweeps in bacterial species. Current Biology **26:**R112-R115.
70. **Fraser C, Hanage WP, Spratt BG.** 2007. Recombination and the nature of bacterial speciation. Science **315:**476-480.
71. **Sheppard SK, McCarthy ND, Falush D, Maiden MC.** 2008. Convergence of Campylobacter species: implications for bacterial evolution. Science **320:**237-239.
72. **Denef VJ, Banfield JF.** 2012. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. Science **336:**462-466.
73. **Thingstad TF, Vage S, Storesund JE, Sandaa RA, Giske J.** 2014. A theoretical analysis of how strain-specific viruses can control microbial species diversity. Proc Natl Acad Sci U S A **111:**7813-7818.
74. **Suttle CA.** 2005. Viruses in the sea. Nature **437:**356-361.
75. **Winter C, Bouvier T, Weinbauer MG, Thingstad TF.** 2010. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. Microbiol Mol Biol Rev **74:**42-57.
76. **Faruque SM, Naser IB, Islam MJ, Faruque AS, Ghosh AN, Nair GB, Sack DA, Mekalanos JJ.** 2005. Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages. Proc Natl Acad Sci U S A **102:**1702-1707.
77. **Thingstad TF, Pree B, Giske J, Vage S.** 2015. What difference does it make if viruses are strain-, rather than species-specific? Front Microbiol **6:**320.
78. **Hibbing ME, Fuqua C, Parsek MR, Peterson SB.** 2010. Bacterial competition: surviving and thriving in the microbial jungle. Nat Rev Microbiol **8:**15-25.
79. **Foster KR, Bell T.** 2012. Competition, not cooperation, dominates interactions among culturable microbial species. Current biology **22:**1845-1850.
80. **Pfennig DW, Pfennig KS.** 2012. Evolution's wedge: competition and the origins of diversity. Univ of California Press.
81. **Kümmerer K.** 2009. Antibiotics in the aquatic environment–a review–part I. Chemosphere **75:**417-434.

82. **Riley MA, Wertz JE.** 2002. Bacteriocins: evolution, ecology, and application. Annu Rev Microbiol **56:**117-137.
83. **Ruhe ZC, Low DA, Hayes CS.** 2013. Bacterial contact-dependent growth inhibition. Trends Microbiol **21:**230-237.
84. **Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, Heidelberg JF, Mekalanos JJ.** 2006. Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. Proc Natl Acad Sci U S A **103:**1528-1533.
85. **Strassmann JE, Gilbert OM, Queller DC.** 2011. Kin discrimination and cooperation in microbes. Annual review of microbiology **65:**349-367.
86. **Velicer GJ, Plucain J.** 2016. Evolution: bacterial territoriality as a byproduct of kin discriminatory warfare. Current Biology **26:**R364-R366.
87. **Lyons NA, Kraigher B, Stefanic P, Mandic-Mulec I, Kolter R.** 2016. A Combinatorial Kin Discrimination System in Bacillus subtilis. Curr Biol **26:**733-742.
88. **Stefanic P, Kraigher B, Lyons NA, Kolter R, Mandic-Mulec I.** 2015. Kin discrimination between sympatric Bacillus subtilis isolates. Proc Natl Acad Sci U S A **112:**14042-14047.
89. **Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, Hussain F, Le Roux F, Mincer T, Polz MF.** 2012. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. Science **337:**1228-1231.
90. **Czaran TL, Hoekstra RF, Pagie L.** 2002. Chemical warfare between microbes promotes biodiversity. Proc Natl Acad Sci U S A **99:**786-790.
91. **Kerr B, Riley MA, Feldman MW, Bohannan BJ.** 2002. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. Nature **418:**171-174.
92. **Bari SM, Roky MK, Mohiuddin M, Kamruzzaman M, Mekalanos JJ, Faruque SM.** 2013. Quorum-sensing autoinducers resuscitate dormant Vibrio cholerae in environmental water samples. Proc Natl Acad Sci U S A **110:**9926-9931.
93. **Borgeaud S, Metzger LC, Scrignari T, Blokesch M.** 2015. The type VI secretion system of Vibrio cholerae fosters horizontal gene transfer. Science **347:**63-67.
94. **Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L.** 2000. DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature **406:**477-483.
95. **Vezzulli L, Pruzzo C, Huq A, Colwell RR.** 2010. Environmental reservoirs of Vibrio cholerae and their role in cholera. Environ Microbiol Rep **2:**27-33.
96. **Hazen TH, Pan L, Gu J-D, Sobecky PA.** 2010. The contribution of mobile genetic elements to the evolution and ecology of Vibrios. FEMS microbiology ecology **74:**485-499.
97. **Gangarosa E, Mosley W.** 1974. Epidemiology and surveillance of cholera. In: Barua D, Burrows W (Editors) Cholera Philadelphia: WB Saunders**:**381 - 403.
98. **Colwell R, Kaper J, Joseph S.** 1977. Vibrio cholerae, Vibrio parahaemolyticus, and other vibrios: occurrence and distribution in Chesapeake Bay. Science **198:**394-396.
99. **Islam A, Labbate M, Djordjevic SP, Alam M, Darling A, Melvold J, Holmes AJ, Johura FT, Cravioto A, Charles IG, Stokes HW.** 2013. Indigenous Vibrio cholerae strains from a non-endemic region are pathogenic. Open Biol **3:**120181.
100. **Haley BJ, Chen A, Grim CJ, Clark P, Diaz CM, Taviani E, Hasan NA, Sancomb E, Elnemr WM, Islam MA, Huq A, Colwell RR, Benediktsdottir E.** 2012. Vibrio cholerae in an Historically Cholera-Free Country. Environ Microbiol Rep **4:**381-389.
101. **Nahar S, Sultana M, Naser MN, Nair GB, Watanabe H, Ohnishi M, Yamamoto S, Endtz H, Cravioto A, Sack RB, Hasan NA, Sadique A, Huq A, Colwell RR, Alam M.** 2011. Role of Shrimp Chitin in the Ecology of Toxigenic Vibrio cholerae and Cholera Transmission. Front Microbiol **2:**260.

102. **Pruzzo C, Vezzulli L, Colwell RR.** 2008. Global impact of Vibrio cholerae interactions with chitin. Environmental microbiology **10:**1400-1410.
103. **Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley BJ, Taviani E, Jeon YS, Kim DW, Lee JH, Brettin TS, Bruce DC, Challacombe JF, Detter JC, Han CS, Munk AC, Chertkov O, Meincke L, Saunders E, Walters RA, Huq A, Nair GB, Colwell RR.** 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae. Proc Natl Acad Sci U S A **106:**15442-15447.
104. **Haley BJ, Choi SY, Grim CJ, Onifade TJ, Cinar HN, Tall BD, Taviani E, Hasan NA, Abdullah AH, Carter L, Sahu SN, Kothary MH, Chen A, Baker R, Hutchinson R, Blackmore C, Cebula TA, Huq A, Colwell RR.** 2014. Genomic and phenotypic characterization of Vibrio cholerae non-O1 isolates from a US Gulf Coast cholera outbreak. PLoS One **9:**e86264.
105. **Schuster BM, Tyzik AL, Donner RA, Striplin MJ, Almagro-Moreno S, Jones SH, Cooper VS, Whistler CA.** 2011. Ecology and genetic structure of a northern temperate Vibrio cholerae population related to toxigenic isolates. Appl Environ Microbiol **77:**7568-7575.
106. **Faruque SM, Mekalanos JJ.** 2012. Phage-bacterial interactions in the evolution of toxigenic Vibrio cholerae. Virulence **3:**556-565.
107. **Meibom KL, Blokesch M, Dolganov NA, Wu C-Y, Schoolnik GK.** 2005. Chitin induces natural competence in Vibrio cholerae. Science **310:**1824-1827.
108. **Waldor MK, Mekalanos JJ.** 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science **272:**1910-1914.
109. **Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF.** 2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global Vibrio cholerae populations. MBio **2**.
110. **Keymer DP, Miller MC, Schoolnik GK, Boehm AB.** 2007. Genomic and phenotypic diversity of coastal Vibrio cholerae strains is linked to environmental factors. Appl Environ Microbiol **73:**3705-3714.
111. **Zo YG, Chokesajjawatee N, Arakawa E, Watanabe H, Huq A, Colwell RR.** 2008. Covariability of Vibrio cholerae microdiversity and environmental parameters. Appl Environ Microbiol **74:**2915-2920.
112. **Zo YG, Chokesajjawatee N, Grim C, Arakawa E, Watanabe H, Colwell RR.** 2009. Diversity and seasonality of bioluminescent Vibrio cholerae populations in Chesapeake Bay. Appl Environ Microbiol **75:**135-146.
113. **Boucher Y, Orata FD, Alam M.** 2015. The out-of-the-delta hypothesis: dense human populations in low-lying river deltas served as agents for the evolution of a deadly pathogen. Front Microbiol **6:**1120.
114. **Vezzulli L, Guzmán CA, Colwell RR, Pruzzo C.** 2008. Dual role colonization factors connecting Vibrio cholerae's lifestyles in human and aquatic environments open new perspectives for combating infectious diseases. Current opinion in biotechnology **19:**254-259.
115. **Preheim SP, Boucher Y, Wildschutte H, David LA, Veneziano D, Alm EJ, Polz MF.** 2011. Metapopulation structure of Vibrionaceae among coastal marine invertebrates. Environ Microbiol **13:**265-275.
116. **Szabo G, Preheim SP, Kauffman KM, David LA, Shapiro J, Alm EJ, Polz MF.** 2013. Reproducibility of Vibrionaceae population structure in coastal bacterioplankton. ISME J **7:**509-519.
117. **Shapiro BJ, Polz MF.** 2015. Microbial Speciation. Cold Spring Harbor perspectives in biology **7:**a018143.
118. **Kirchberger PC, Turnsek M, Hunt DE, Haley BJ, Colwell RR, Polz MF, Tarr CL, Boucher Y.** 2014. Vibrio metoecus sp. nov., a close relative of Vibrio cholerae isolated

from coastal brackish ponds and clinical specimens. Int J Syst Evol Microbiol **64:**3208-3214.
119. **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C.** 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28:**1647-1649.
120. **Spratt BG, Hanage WP, Li B, Aanensen DM, Feil EJ.** 2004. Displaying the relatedness among isolates of bacterial species–the eBURST approach. FEMS Microbiology Letters **241:**129-134.
121. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology **75:**7537-7541.
122. **Clarke KR, Somerfield PJ, Gorley RN.** 2008. Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage. Journal of Experimental Marine Biology and Ecology **366:**56-69.
123. **Feil EJ, Holmes EC, Bessen DE, Chan M-S, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE.** 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proceedings of the National Academy of Sciences **98:**182-187.
124. **Keymer DP, Boehm AB.** 2011. Recombination shapes the structure of an environmental Vibrio cholerae population. Appl Environ Microbiol **77:**537-544.
125. **McVean G, Awadalla P, Fearnhead P.** 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:**1231-1241.
126. **Martin DP, Murrell B, Golden M, Khoosal A, Muhire B.** 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evolution **1:**vev003.
127. **Thompson JD, Gibson T, Higgins DG.** 2002. Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics doi:10.1002/0471250953.bi0203s00**:**2.3. 1-2.3. 22.
128. **Angiuoli SV, Salzberg SL.** 2011. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics **27:**334-342.
129. **Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J.** 2010. Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology**:**19.10. 11-19.10. 21.
130. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30:**1312-1313.
131. **Rost B.** 1999. Twilight zone of protein sequence alignments. Protein engineering **12:**85-94.
132. **Li L, Stoeckert CJ, Roos DS.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research **13:**2178-2189.
133. **Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research **25:**3389-3402.
134. **Tatusov RL, Galperin MY, Natale DA, Koonin EV.** 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic acids research **28:**33-36.
135. **Esteves K, Mosser T, Aujoulat F, Hervio-Heath D, Monfort P, Jumas-Bilak E.** 2015. Highly diverse recombining populations of Vibrio cholerae and Vibrio parahaemolyticus in French Mediterranean coastal lagoons. Frontiers in Microbiology **6:**708.
136. **Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, Nevo E, Cohan FM.** 2008. Identifying the

fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci U S A **105:**2504-2509.
137. **Materna AC, Friedman J, Bauer C, David C, Chen S, Huang IB, Gillens A, Clarke SA, Polz MF, Alm EJ.** 2012. Shape and evolution of the fundamental niche in marine Vibrio. ISME J **6:**2168-2177.
138. **Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I.** 2012. Defining seasonal marine microbial community dynamics. The ISME journal **6:**298-308.
139. **Faruque SM, Naser IB, Islam MJ, Faruque A, Ghosh A, Nair GB, Sack DA, Mekalanos JJ.** 2005. Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages. Proceedings of the National Academy of Sciences of the United States of America **102:**1702-1707.
140. **Orata FD, Kirchberger PC, Méheust R, Barlow EJ, Tarr CL, Boucher Y.** 2015. The Dynamics of Genetic Interactions between Vibrio metoecus and Vibrio cholerae, Two Close Relatives Co-Occurring in the Environment. Genome biology and evolution **7:**2941-2954.
141. **Unterweger D, Miyata ST, Bachmann V, Brooks TM, Mullins T, Kostiuk B, Provenzano D, Pukatzki S.** 2014. The Vibrio cholerae type VI secretion system employs diverse effector modules for intraspecific competition. Nat Commun **5:**3549.
142. **Grim CJ, Taviani E, Alam M, Huq A, Sack RB, Colwell RR.** 2008. Occurrence and expression of luminescence in Vibrio cholerae. Appl Environ Microbiol **74:**708-715.
143. **Zarubin M, Belkin S, Ionescu M, Genin A.** 2012. Bacterial bioluminescence as a lure for marine zooplankton and fish. Proc Natl Acad Sci U S A **109:**853-857.
144. **Wood JM.** 2015. Bacterial responses to osmotic challenges. The Journal of general physiology **145:**381-388.
145. **Dutton G.** 2012. Glucuronic Acid Free and Combined: Chemistry, Biochemistry, Pharmacology, and Medicine. Elsevier.
146. **Fieldhouse RJ, Turgeon Z, White D, Merrill AR.** 2010. Cholera-and anthrax-like toxins are among several new ADP-ribosyltransferases.
147. **Isidean S, Riddle M, Savarino S, Porter C.** 2011. A systematic review of ETEC epidemiology focusing on colonization factor and toxin expression. Vaccine **29:**6167-6178.
148. **Beltrán P, Delgado G, Navarro A, Trujillo F, Selander RK, Cravioto A.** 1999. Genetic Diversity and Population Structure ofVibrio cholerae. Journal of clinical microbiology **37:**581-590.
149. **Våge S, Storesund JE, Thingstad TF.** 2013. Adding a cost of resistance description extends the ability of virus–host model to explain observed patterns in structure and function of pelagic microbial communities. Environmental microbiology **15:**1842-1852.
150. **Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP.** 2009. The bacterial species challenge: making sense of genetic and ecological diversity. science **323:**741-746.
151. **Choopun N.** 2004. The population structure of Vibrio cholerae in Chesapeake Bay.
152. **Haley BJ, Grim CJ, Hasan NA, Choi SY, Chun J, Brettin TS, Bruce DC, Challacombe JF, Detter JC, Han CS, Huq A, Colwell RR.** 2010. Comparative genomic analysis reveals evidence of two novel Vibrio species closely related to V. cholerae. BMC Microbiol **10:**154.
153. **Nishiguchi MK, Nair VS.** 2003. Evolution of symbiosis in the Vibrionaceae: a combined approach using molecules and physiology. International Journal of Systematic and Evolutionary Microbiology **53:**2019-2026.
154. **Lane D.** 1991. 16S/23S rRNA sequencing. In 'Nucleic acid techniques in bacterial systematics'.(Eds E Stackebrandt, M Goodfellow) pp. 115–175. John Wiley and Sons: Chichester, UK.

155. **Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R.** 2007. Clustal W and Clustal X version 2.0. Bioinformatics **23:**2947-2948.
156. **Chun J, Huq A, Colwell RR.** 1999. Analysis of 16S-23S rRNA intergenic spacer regions of Vibrio cholerae and Vibrio mimicus. Applied and environmental microbiology **65:**2202-2208.
157. **Nandi B, Nandy RK, Mukhopadhyay S, Nair GB, Shimada T, Ghose AC.** 2000. Rapid method for species-specific identification ofVibrio cholerae using primers targeted to the gene of outer membrane protein OmpW. Journal of clinical microbiology **38:**4145-4151.
158. **Huq A, Haley BJ, Taviani E, Chen A, Hasan NA, Colwell RR.** 2006. Detection, isolation, and identification of Vibrio cholerae from the environment. Current protocols in microbiology**:**6A. 5.1-6A. 5.51.
159. **Pascual J, Macian MC, Arahal DR, Garay E, Pujalte MJ.** 2010. Multilocus sequence analysis of the central clade of the genus Vibrio by using the 16S rRNA, recA, pyrH, rpoD, gyrB, rctB and toxR genes. Int J Syst Evol Microbiol **60:**154-165.
160. **Preheim SP, Timberlake S, Polz MF.** 2011. Merging taxonomy with ecological population prediction in a case study of Vibrionaceae. Appl Environ Microbiol **77:**7195-7206.
161. **Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM.** 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol **57:**81-91.
162. **Kirchberger PC, Orata FD, Barlow EJ, Kauffman KM, Case RJ, Polz MF, Boucher Y.** 2016. A Small Number of Phylogenetically Distinct Clonal Complexes Dominate a Coastal Vibrio cholerae Population. Appl Environ Microbiol **82:**5576-5586.
163. **Boucher Y.** 2016. Sustained Local Diversity of Vibrio cholerae O1 Biotypes in a Previously Cholera-Free Country. mBio **7:**e00570-00516.
164. **Nelson EJ, Harris JB, Morris JG, Jr., Calderwood SB, Camilli A.** 2009. Cholera transmission: the host, pathogen and bacteriophage dynamic. Nat Rev Microbiol **7:**693-702.
165. **Bachmann V, Kostiuk B, Unterweger D, Diaz-Satizabal L, Ogg S, Pukatzki S.** 2015. Bile salts modulate the mucin-activated type VI secretion system of pandemic Vibrio cholerae. PLoS Negl Trop Dis **9:**e0004031.
166. **Durand E, Nguyen VS, Zoued A, Logger L, Pehau-Arnaudet G, Aschtgen MS, Spinelli S, Desmyter A, Bardiaux B, Dujeancourt A, Roussel A, Cambillau C, Cascales E, Fronzes R.** 2015. Biogenesis and structure of a type VI secretion membrane core complex. Nature **523:**555-560.
167. **Zheng J, Ho B, Mekalanos JJ.** 2011. Genetic analysis of anti-amoebae and anti-bacterial activities of the type VI secretion system in Vibrio cholerae. PloS one **6:**e23876.
168. **Cianfanelli FR, Monlezun L, Coulthurst SJ.** 2016. Aim, Load, Fire: The Type VI Secretion System, a Bacterial Nanoweapon. Trends Microbiol **24:**51-62.
169. **Pukatzki S, Ma AT, Revel AT, Sturtevant D, Mekalanos JJ.** 2007. Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. Proc Natl Acad Sci U S A **104:**15508-15513.
170. **Heisler DB, Kudryashova E, Grinevich DO, Suarez C, Winkelman JD, Birukov KG, Kotha SR, Parinandi NL, Vavylonis D, Kovar DR.** 2015. ACD toxin–produced actin oligomers poison formin-controlled actin polymerization. Science **349:**535-539.
171. **Fu Y, Waldor MK, Mekalanos JJ.** 2013. Tn-Seq analysis of Vibrio cholerae intestinal colonization reveals a role for T6SS-mediated antibacterial activity in the host. Cell host & microbe **14:**652-663.

172. **Brooks TM, Unterweger D, Bachmann V, Kostiuk B, Pukatzki S.** 2013. Lytic activity of the Vibrio cholerae type VI secretion toxin VgrG-3 is inhibited by the antitoxin TsaB. Journal of Biological Chemistry **288:**7618-7625.
173. **Durand E, Cambillau C, Cascales E, Journet L.** 2014. VgrG, Tae, Tle, and beyond: the versatile arsenal of Type VI secretion effectors. Trends in microbiology **22:**498-507.
174. **Unterweger D, Kostiuk B, Otjengerdes R, Wilton A, Diaz-Satizabal L, Pukatzki S.** 2015. Chimeric adaptor proteins translocate diverse type VI secretion system effectors in Vibrio cholerae. EMBO J **34:**2198-2210.
175. **Unterweger D, Kitaoka M, Miyata ST, Bachmann V, Brooks TM, Moloney J, Sosa O, Silva D, Duran-Gonzalez J, Provenzano D.** 2012. Constitutive type VI secretion system expression gives Vibrio cholerae intra-and interspecific competitive advantages. PLoS One **7:**e48320.
176. **Altindis E, Dong T, Catalano C, Mekalanos J.** 2015. Secretome Analysis of Vibrio cholerae Type VI Secretion System Reveals a New Effector-Immunity Pair. MBio **6**.
177. **Labbate M, Orata FD, Petty NK, Jayatilleke ND, King W, Kirchberger PC, Allen C, Mann G, Mutreja A, Thomson NR, Boucher Y, Charles IG.** 2016. A genomic island in Vibrio cholerae with VPI-1 site-specific recombination characteristics contains CRISPR-Cas and type VI secretion modules. . Scientific Reports **6**.
178. **Gardner SN, Slezak T, Hall BG.** 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics **31:**2877-2878.
179. **Zhaxybayeva O, Lapierre P, Gogarten JP.** 2004. Genome mosaicism and organismal lineages. TRENDS in Genetics **20:**254-260.
180. **Sawyer S.** 2000. GENECONV: Statistical tests for detecting gene conversion (version 1.81). Department of Mathematics, Washington University, St Louis, Mo doi:10.1093/oxfordjournals.molbev.a040567.
181. **Smith JM.** 1992. Analyzing the mosaic structure of genes. Journal of molecular evolution **34:**126-129.
182. **Posada D, Crandall KA.** 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences **98:**13757-13762.
183. **Davis B, Fanning G, Madden J, Steigerwalt A, Bradford H, Smith H, Brenner D.** 1981. Characterization of biochemically atypical Vibrio cholerae strains and designation of a new pathogenic species, Vibrio mimicus. Journal of Clinical Microbiology **14:**631-639.
184. **Zhang J, Zhang H, Liu Y, Zhan L, She Z, Dong C, Dong Y.** 2014. Crystallization and preliminary X-ray study of TsiV3 from Vibrio cholerae. Acta Crystallographica Section F: Structural Biology Communications **70:**335-338.
185. **Pukatzki S, McAuley SB, Miyata ST.** 2009. The type VI secretion system: translocation of effectors and effector-domains. Curr Opin Microbiol **12:**11-17.
186. **Miyata ST, Unterweger D, Rudko SP, Pukatzki S.** 2013. Dual expression profile of type VI secretion system immunity genes protects pandemic Vibrio cholerae. PLoS Pathog **9:**e1003752.
187. **Russell AB, Singh P, Brittnacher M, Bui NK, Hood RD, Carl MA, Agnello DM, Schwarz S, Goodlett DR, Vollmer W, Mougous JD.** 2012. A widespread bacterial type VI secretion effector superfamily identified using a heuristic approach. Cell Host Microbe **11:**538-549.
188. **Karaolis DK, Somara S, Maneval DR, Johnson JA, Kaper JB.** 1999. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. Nature **399:**375-379.

189. **Rajanna C, Wang J, Zhang D, Xu Z, Ali A, Hou Y-M, Karaolis D.** 2003. The vibrio pathogenicity island of epidemic Vibrio cholerae forms precise extrachromosomal circular excision products. Journal of bacteriology **185:**6893-6901.
190. **Riley MA.** 1998. Molecular mechanisms of bacteriocin evolution. Annual review of genetics **32:**255-278.
191. **Prudhomme M, Libante V, Claverys JP.** 2002. Homologous recombination at the border: insertion-deletions and the trapping of foreign DNA in Streptococcus pneumoniae. Proc Natl Acad Sci U S A **99:**2100-2105.
192. **Meier P, Wackernagel W.** 2003. Mechanisms of homology-facilitated illegitimate recombination for foreign DNA acquisition in transformable Pseudomonas stutzeri. Molecular microbiology **48:**1107-1118.
193. **Koskiniemi S, Garza-Sánchez F, Sandegren L, Webb JS, Braaten BA, Poole SJ, Andersson DI, Hayes CS, Low DA.** 2014. Selection of orphan Rhs toxin expression in evolved Salmonella enterica serovar Typhimurium. PLoS Genet **10:**e1004255.
194. **Jackson AP, Thomas GH, Parkhill J, Thomson NR.** 2009. Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. BMC Genomics **10:**584.
195. **Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L.** 2012. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol Direct **7:**18.
196. **Poole SJ, Diner EJ, Aoki SK, Braaten BA, t'Kint de Roodenbeke C, Low DA, Hayes CS.** 2011. Identification of functional toxin/immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems. PLoS Genet **7:**e1002217.
197. **de Vries J, Wackernagel W.** 2002. Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. Proceedings of the National Academy of Sciences **99:**2094-2099.
198. **Pretzer C, Druzhinina IS, Amaro C, Benediktsdóttir E, Hedenström I, Hervio‐Heath D, Huhulescu S, Schets FM, Farnleitner AH, Kirschner AK.** 2016. High genetic diversity of Vibrio cholerae in the European lake Neusiedler See is associated with intensive recombination in the reed habitat and the long-distance transfer of strains. Environmental Microbiology doi:10.1111/1462-2920.13612.
199. **Koenig JE, Bourne DG, Curtis B, Dlutek M, Stokes HW, Doolittle WF, Boucher Y.** 2011. Coral-mucus-associated Vibrio integrons in the Great Barrier Reef: genomic hotspots for environmental adaptation. ISME J **5:**962-972.
200. **Mazel D.** 2006. Integrons: agents of bacterial evolution. Nat Rev Microbiol **4:**608-620.
201. **Labbate M, Boucher Y, Joss MJ, Michael CA, Gillings MR, Stokes HW.** 2007. Use of chromosomal integron arrays as a phylogenetic typing system for Vibrio cholerae pandemic strains. Microbiology **153:**1488-1498.
202. **Jackson AP, Thomas GH, Parkhill J, Thomson NR.** 2009. Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. BMC genomics **10:**1.
203. **Redfield RJ.** 1993. Genes for Breakfast: The Have-Your-Cake and-Eat-It-Too of Bacterial Transformation. Journal of Heredity **84:**400-404.
204. **Rosenberg CE.** 2009. The cholera years: The United States in 1832, 1849, and 1866. University of Chicago Press.
205. **Chatterjee S, Chaudhuri K.** 2003. Lipopolysaccharides of Vibrio cholerae: I. Physical and chemical characterization. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease **1639:**65-79.

206. **Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M.** 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature **477:**462-465.
207. **Schauer S, Jakwerth S, Bliem R, Baudart J, Lebaron P, Huhulescu S, Kundi M, Herzig A, Farnleitner AH, Sommer R, Kirschner A.** 2015. Dynamics of Vibrio cholerae abundance in Austrian saline lakes, assessed with quantitative solid-phase cytometry. Environ Microbiol doi:10.1111/1462-2920.12861.
208. **Chowdhury FR, Nur Z, Hassan N, Seidlein L, Dunachie S.** 2017. Pandemics, pathogenicity and changing molecular epidemiology of cholera in the era of global warming. Annals of Clinical Microbiology and Antimicrobials **16:**10.
209. **Rahaman M, Islam T, Colwell RR, Alam M.** 2015. Molecular tools in understanding the evolution of Vibrio cholerae. Frontiers in microbiology **6:**1040.
210. **Vezzulli L, Stauder M, Grande C, Pezzati E, Verheye HM, Owens NJ, Pruzzo C.** 2015. gbpA as a novel qPCR target for the species-specific detection of Vibrio cholerae O1, O139, non-O1/non-O139 in Environmental, Stool, and Historical Continuous Plankton Recorder Samples. PloS one **10:**e0123983.
211. **Huse SM, Welch DM, Morrison HG, Sogin ML.** 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental microbiology **12:**1889-1898.
212. **Siboni N, Balaraju V, Carney R, Labbate M, Seymour JR.** 2016. Spatiotemporal Dynamics of Vibrio spp. within the Sydney Harbour Estuary. Front Microbiol **7:**460.
213. **Orata FD, Xu Y, Gladney LM, Rishishwar L, Case RJ, Boucher Y, Jordan IK, Tarr CL.** 2016. Characterization of clinical and environmental isolates of Vibrio cidicii sp. nov., a close relative of Vibrio navarrensis. International Journal of Systematic and Evolutionary Microbiology **66:**4148-4155.
214. **Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP.** 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nature methods.
215. **Tikhonov M, Leach RW, Wingreen NS.** 2015. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. ISME J **9:**68-80.
216. **Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML.** 2015. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J **9:**968-979.
217. **Wyckoff EE, Mey AR, Payne SM.** 2007. Iron acquisition in Vibrio cholerae. Biometals **20:**405.
218. **Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD.** 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol **79:**5112-5120.
219. **Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, Heled J, Kearse M.** 2010. Geneious v6. 1.3. Available at: ht tp.
220. **Treangen TJ, Ondov BD, Koren S, Phillippy AM.** 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome biology **15:**524.
221. **Letunic I, Bork P.** 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23:**127-128.
222. **Kirchberger PC, Unterweger D, Provenzano D, Pukatzki S, Boucher Y.** 2017. Sequential displacement of Type VI Secretion System effector genes leads to evolution of diverse immunity gene arrays in Vibrio cholerae. Scientific Reports **7:**45133.
223. **Halpern M, Landsberg O, Raats D, Rosenberg E.** 2007. Culturable and VBNC Vibrio cholerae: interactions with chironomid egg masses and their bacterial population. Microb Ecol **53:**285-293.

224. **Colwell RR, Huq A, Islam MS, Aziz KM, Yunus M, Khan NH, Mahmud A, Sack RB, Nair GB, Chakraborty J, Sack DA, Russek-Cohen E.** 2003. Reduction of cholera in Bangladeshi villages by simple filtration. Proc Natl Acad Sci U S A **100:**1051-1055.
225. **Yawata Y, Cordero OX, Menolascina F, Hehemann JH, Polz MF, Stocker R.** 2014. Competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. Proc Natl Acad Sci U S A **111:**5622-5627.
226. **Mallet J.** 2008. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos Trans R Soc Lond B Biol Sci **363:**2971-2986.
227. **Polz MF, Hunt DE, Preheim SP, Weinreich DM.** 2006. Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. Philos Trans R Soc Lond B Biol Sci **361:**2009-2021.
228. **McNally L, Bernardy E, Thomas J, Kalziqi A, Pentz J, Brown SP, Hammer BK, Yunker PJ, Ratcliff WC.** 2017. Killing by Type VI secretion drives genetic phase separation and correlates with increased cooperation. Nature Communications **8**.
229. **Wright ES, Vetsigian KH.** 2016. Inhibitory interactions promote frequent bistability among competing bacteria. Nature communications **7**.
230. **Colwell RR.** 1996. Global climate and infectious disease: the cholera paradigm. Science **274:**2025.
231. **Orata FD, Keim PS, Boucher Y.** 2014. The 2010 cholera outbreak in Haiti: how science solved a controversy. PLoS Pathog **10:**e1003967.
232. **Jutla A, Whitcombe E, Hasan N, Haley B, Akanda A, Huq A, Alam M, Sack RB, Colwell R.** 2013. Environmental factors influencing epidemic cholera. The American journal of tropical medicine and hygiene **89:**597-607.
233. **Fullner KJ, Mekalanos JJ.** 2000. In vivo covalent cross-linking of cellular actin by the Vibrio cholerae RTX toxin. The EMBO journal **19:**5315-5323.
234. **Miyata ST, Bachmann V, Pukatzki S.** 2013. Type VI secretion system regulation as a consequence of evolutionary pressure. Journal of medical microbiology **62:**663-676.
235. **Miyata ST, Kitaoka M, Brooks TM, McAuley SB, Pukatzki S.** 2011. Vibrio cholerae requires the type VI secretion system virulence factor VasX to kill Dictyostelium discoideum. Infection and immunity **79:**2941-2949.
236. **Worden AZ, Seidel M, Smriga S, Wick A, Malfatti F, Bartlett D, Azam F.** 2006. Trophic regulation of Vibrio cholerae in coastal marine waters. Environmental Microbiology **8:**21-29.
237. **Bell T, Bonsall MB, Buckling A, Whiteley AS, Goodall T, Griffiths RI.** 2010. Protists have divergent effects on bacterial diversity along a productivity gradient. Biology letters **6:**639-642.
238. **Halpern M, Senderovich Y, Izhaki I.** 2008. Waterfowl—the missing link in epidemic and pandemic cholera dissemination? PLoS Pathog **4:**e1000173.
239. **Gogarten JP, Townsend JP.** 2005. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol **3:**679-687.
240. **Pukatzki S, Provenzano D.** 2013. Vibrio cholerae as a predator: lessons from evolutionary principles. Front Microbiol **4:**384.
241. **Cohan FM, Koeppel AF.** 2008. The origins of ecological diversity in prokaryotes. Curr Biol **18:**R1024-1034.
242. **Wielgoss S, Didelot X, Chaudhuri RR, Liu X, Weedall GD, Velicer GJ, Vos M.** 2016. A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium Myxococcus xanthus. ISME J **10:**2468-2477.
243. **Bernardy EE, Turnsek MA, Wilson SK, Tarr CL, Hammer BK.** 2016. Diversity of Clinical and Environmental Isolates of Vibrio cholerae in Natural Transformation and Contact-Dependent Bacterial Killing Indicative of Type VI Secretion System Activity. Appl Environ Microbiol doi:10.1128/AEM.00351-16.

244. **Devault AM, Golding GB, Waglechner N, Enk JM, Kuch M, Tien JH, Shi M, Fisman DN, Dhody AN, Forrest S.** 2014. Second-pandemic strain of Vibrio cholerae from the Philadelphia cholera outbreak of 1849. New England Journal of Medicine **370:**334-340.
245. **Vettiger A, Basler M.** 2016. Type VI Secretion System Substrates Are Transferred and Reused among Sister Cells. Cell **167:**99-110 e112.
246. **Ho BT, Dong TG, Mekalanos JJ.** 2014. A view to a kill: the bacterial type VI secretion system. Cell Host Microbe **15:**9-21.
247. **Shapiro BJ, Levade I, Kovacikova G, Taylor RK, Almagro-Moreno S.** 2016. Origins of pandemic Vibrio cholerae from environmental gene pools. Nature Microbiology **2:**16240.
248. **Datta MS, Sliwerska E, Gore J, Polz MF, Cordero OX.** 2016. Microbial interactions lead to rapid micro-scale successions on model marine particles. Nat Commun **7:**11965.
249. **Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, Alam M, Haley BJ, Taviani E, Hine E, Su Q.** 2012. Genomic diversity of 2010 Haitian cholera outbreak strains. Proceedings of the National Academy of Sciences **109:**E2010-E2017.

Appendix 1

The Dynamics of Genetic Interactions between *Vibrio metoecus* and *Vibrio cholerae*, Two Close Relatives Co-Occurring in the Environment

---

The Dynamics of Genetic Interactions between *Vibrio metoecus* and *Vibrio cholerae*, Two Close Relatives Co-Occurring in the Environment

Fabini D. Orata[1], Paul C. Kirchberger[1], Raphaël Méheust[2], E. Jed Barlow[3], Cheryl L. Tarr[4], and Yan Boucher[1,*]

[1]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

[2]Unité Mixte de Recherche 7138, Evolution Paris-Seine, Institut de Biologie Paris-Seine, Université Pierre et Marie Curie, Paris, France

[3]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

[4]Enteric Diseases Laboratory Branch, Division of Foodborne, Waterborne, and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA

*Corresponding author: E-mail: yboucher@ualberta.ca.

Accepted: October 2, 2015

Data deposition: The whole-genome sequences generated in this study have been deposited in the DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL), and GenBank databases under the BioProject accession PRJNA281423. The individual genome accession numbers are listed in supplementary table S1, Supplementary Material online.

Abstract

Vibrio metoecus is the closest relative of *Vibrio cholerae*, the causative agent of the potent diarrheal disease cholera. Although the pathogenic potential of this new species is yet to be studied in depth, it has been co-isolated with *V. cholerae* in coastal waters and found in clinical specimens in the United States. We used these two organisms to investigate the genetic interaction between closely related species in their natural environment. The genomes of 20 *V. cholerae* and 4 *V. metoecus* strains isolated from a brackish coastal pond on the US east coast, as well as 4 clinical *V. metoecus* strains were sequenced and compared with reference strains. Whole genome comparison shows 86–87% average nucleotide identity (ANI) in their core genes between the two species. On the other hand, the chromosomal integron, which occupies approximately 3% of their genomes, shows higher conservation in ANI between species than any other region of their genomes. The ANI of 93–94% observed in this region is not significantly greater within than between species, meaning that it does not follow species boundaries. *Vibrio metoecus* does not encode toxigenic *V. cholerae* major virulence factors, the cholera toxin and toxin-coregulated pilus. However, some of the pathogenicity islands found in pandemic *V. cholerae* were either present in the common ancestor it shares with *V. metoecus*, or acquired by clinical and environmental *V. metoecus* in partial fragments. The virulence factors of *V. cholerae* are therefore both more ancient and more widespread than previously believed. There is high interspecies recombination in the core genome, which has been detected in 24% of the single-copy core genes, including genes involved in pathogenicity. *Vibrio metoecus* was six times more often the recipient of DNA from *V. cholerae* as it was the donor, indicating a strong bias in the direction of gene transfer in the environment.

Key words: *Vibrio metoecus*, *Vibrio cholerae*, horizontal gene transfer, genomic islands, integron, comparative genomics.

Introduction

The genus *Vibrio* constitutes a diverse group of gammaproteobacteria ubiquitous in marine, brackish, and fresh waters. There are currently over 100 species of vibrios that have been described (Gomez-Gil et al. 2014). This includes clinically significant pathogens such as *Vibrio cholerae*, *Vibrio parahaemolyticus*, and *Vibrio vulnificus* among many others. *Vibrio cholerae*, the causative agent of the potent diarrheal disease cholera, is the most notorious of these human pathogens. Cholera remains a major public health concern, with an estimated 1.2–4.3 million cases and 28,000–142,000 deaths every year worldwide (Ali et al. 2012).

A novel *Vibrio* isolate, initially identified as a nonpathogenic environmental variant of *V. cholerae* (Choopun 2004), was

recently revealed to be a distinct species based on comparative genomic analysis (Haley et al. 2010). Additional environmental strains of this species have been isolated since then (Boucher et al. 2011). Also, since 2006, several clinical strains have been recovered from a range of specimen types (blood, stool, ear, and leg wound) and characterized by the Centers for Disease Control and Prevention (CDC, Atlanta, GA). This recently described species, now officially called *V. metoecus* (Kirchberger et al. 2014), is even more closely related to *V. cholerae* than any other known *Vibrio* species based on biochemical and genotypic tests (Boucher et al. 2011; Kirchberger et al. 2014). Previously, the closest known relative of *V. cholerae* was *Vibrio mimicus*, which was first described as a biochemically atypical strain of *V. cholerae* and named after the fact that it "mimicked *V. cholerae*" phenotypically (Davis et al. 1981).

The discovery of a closely related but distinct species which co-occurs with *V. cholerae* in the environment (Boucher et al. 2011) presents a unique opportunity to investigate the dynamics of interspecies interactions at the genetic level. In their environmental reservoir, bacteria can acquire genetic material from other organisms as a result of horizontal gene transfer (HGT; De la Cruz and Davies 2000). HGT plays an important role in the evolution, adaptation, maintenance, and transmission of virulence in bacteria. It can launch nonpathogenic environmental strains into new pathogenic lifestyles if they obtain the right virulence factors. The two major virulence factors that have led to the evolution from nonpathogenic to toxigenic *V. cholerae* are the cholera toxin (CTX), which is responsible for the cholera symptoms (Waldor and Mekalanos 1996), and the toxin-coregulated pilus (TCP), which is necessary for the colonization of the small intestine in the human host (Taylor et al. 1987). These elements are encoded in genomic islands, specifically called pathogenicity islands, and have been acquired horizontally by phage infections (Waldor and Mekalanos 1996; Karaolis et al. 1999). Another genomic island, the integron, is used to capture and disseminate gene cassettes, such as antibiotic resistance genes (Stokes and Hall 1989). Integrons have been identified in a diverse range of bacterial taxa, and are known to play a major role in genome evolution (Mazel 2006; Boucher et al. 2007). As evidenced by multiple HGT events across a wide range of phylogenetic distances, integrons themselves, not only the cassettes they carry, may have been mobilized within and between species throughout their evolutionary history (Boucher et al. 2007). Integrons are ubiquitous among vibrios, but in some species, such as *V. cholerae*, it can occupy up to 3% of the genome and can contain over a hundred gene cassettes with a wide range of biochemical functions (Mazel et al. 1998; Heidelberg et al. 2000).

Here, we investigate the extent of genetic interaction between *V. metoecus* and *V. cholerae* through comparative genomic analysis, with the focus on the genomic islands, known hotspots for HGT (Dobrindt et al. 2004). The co-isolation of both species in the same environment (Boucher et al. 2011) indicates that *V. metoecus* is likely in constant interaction with *V. cholerae*. Our results show that there is a high rate of gene exchange between species, so rapid in the chromosomal integron that this region is indistinguishable between species. Multiple HGT events were also inferred in the core genome, including genes implicated in pathogenicity, with the majority with *V. metoecus* as a recipient of *V. cholerae* genes, suggesting a directional bias in interspecies gene transfer.

Materials and Methods

Bacterial Strains Used

The *V. metoecus* and *V. cholerae* isolates sequenced in this study as well as genome sequences of additional isolates for comparison are listed in supplementary table S1, Supplementary Material online. Environmental strains of *V. metoecus* and *V. cholerae* were isolated from Oyster Pond (Falmouth, MA) on August and September 2009 using previously described methods (Boucher et al. 2011). Isolates were grown overnight at 37 °C in tryptic soy broth (Becton Dickinson, Sparks, MD) with 1% NaCl (BDH, Toronto, ON, Canada). The sequences of the clinical *V. metoecus* strains were determined by the CDC. Additional sequences were obtained from the National Center for Biotechnology Information (Bethesda, MD) GenBank database.

Genomic DNA Extraction and Quantitation

Genomic DNA was extracted from overnight bacterial cultures with the DNeasy Blood and Tissue Kit (QIAGEN, Hilden, Germany). The concentration for each extract was determined using the Quant-iT PicoGreen double-stranded DNA Assay Kit (Molecular Probes, Eugene, OR) and the Synergy H1 microplate reader (BioTek, Winooski, VT).

Genome Sequencing and Assembly

The genomic DNA extracts were sent to the McGill University and Génome Québec Innovation Centre (Montréal, QC, Canada) for sequencing, which was performed using the TrueSeq library preparation kit and the HiSeq PE100 sequencing technology (Illumina, San Diego, CA). The contiguous sequences were assembled de novo with the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark). Functional annotations of the draft genomes were done in RAST v2.0 (Rapid Annotation using Subsystem Technology; Aziz et al. 2008).

Whole Genome Alignment

A circular BLAST (Basic Local Alignment Search Tool) atlas was constructed to visually compare whole genomes. The annotated genome sequences of *V. metoecus* and *V. cholerae* were aligned by BLASTN (Altschul et al. 1990) against a reference, *V. cholerae* N16961 (Heidelberg et al. 2000), using the CGView Comparison Tool (Grant et al. 2012).

Determination of Orthologous Gene Families and Pan-Genome Analysis

Orthologous groups of open-reading frames (ORFs) from all strains of *V. metoecus* and *V. cholerae* were determined by pairwise bidirectional BLASTP using the OrthoMCL pipeline v2.0 (Li et al. 2003) with 30% match cutoff, as proteins sharing at least 30% identity are predicted to fold similarly (Rost 1999). The gene families were assigned into functional categories based on the Clusters of Orthologous Groups of proteins (COG) database (Tatusov et al. 2000). The pan- and core genome profiles for each species were determined with PanGP v1.0.1 (Zhao et al. 2014) using the distance guide algorithm, repeated 100 times. Sample size and amplification coefficient were set to 1,000 and 100, respectively.

Determination of Genomic Islands

The major genomic islands of *V. cholerae* N16961 were identified using IslandViewer (Langille and Brinkman 2009) and confirmed with previously published data (Heidelberg et al. 2000; Chun et al. 2009). To determine whether a putative homolog is present, ORFs in these genomic islands were compared against the ORFs of *V. metoecus* and *V. cholerae* by calculating the BLAST score ratio (BSR) between reference and query ORF (Rasko et al. 2005) using a custom-developed Perl script (National Microbiology Laboratory, Winnipeg, MB, Canada). Only BSR values of at least 0.3 (for 30% amino acid identity) were considered (Rost 1999).

Determination of the Integron Regions

The chromosomal integron regions of *V. metoecus* and *V. cholerae* were recovered by finding the locations of the integron integrase gene *intI4* and the *attI* and *attC* recombination sites, identified with the ISAAC software (Improved Structural Annotation of *attC*; Szamosi 2012). The *intI4* and gene cassette sequences were used to calculate the ANI (Konstantinidis and Tiedje 2005; Goris et al. 2007) between strains (intra- and interspecies) in JSpecies v1.2.1 (Richter and Rosselló-Móra 2009), using the bidirectional best BLAST hits between nucleotides. The ANI of the integron region was compared with the ANI of 1,560 single-copy core ORFs ($\approx$1.42 Mb).

Phylogenetic Analyses

Using the PhyloPhlAn pipeline v0.99 (Segata et al. 2013), 3,978 amino acid positions based on 400 universally conserved bacterial and archaeal proteins were determined. The concatenated alignment was used to construct a core genome maximum-likelihood (ML) phylogenetic tree, with a BLOSUM45 similarity matrix using the Jones-Taylor-Thorton (JTT) + category (CAT) amino acid evolution model optimized for topology/length/rate using the nearest neighbor interchange (NNI) topology search. Robustness of branching was estimated with Shimodaira-Hasegawa-like (SH-like) support values from 1,000 replicates.

Nucleotide sequences within a gene family were aligned with ClustalW v2.1 (Larkin et al. 2007), and an ML tree was constructed using RAxML v8.1.17 (Stamatakis 2014) using the general time reversible (GTR) nucleotide substitution model and gamma distribution pattern. Robustness of branching was estimated with 100 bootstrap replicates. Interspecies gene transfer events were determined and quantified by comparison of tree topologies using the Phangorn package v1.99-11 (Schliep 2011) in R v3.1.2 (R Development Core Team 2014). A tree was partitioned into clades and determined whether the clades were perfect or not. Following the definition by Schliep et al. (2011), we defined a perfect clade as a partition that is both complete and homogeneous for a given taxonomic category (e.g., a clade with all *V. metoecus*, and only *V. metoecus*). At least one gene transfer event was hypothesized if a tree did not show perfect clades for neither *V. metoecus* nor *V. cholerae* (i.e., in a rooted tree, *V. metoecus* and *V. cholerae* are both polyphyletics).

Resulting alignments of the 1,184 single-copy core gene families not exhibiting HGT were concatenated, and alignment columns with at least one gap were removed using Geneious (Kearse et al. 2012). A final alignment with a total length of 771,455 bp was obtained and used to construct a core genome ML phylogenetic tree with RAxML v8.1.17 (Stamatakis 2014), as described above.

Results and Discussion

Vibrio cholerae is widely studied, and the genomes of globally diverse clinical and environmental isolates are available (supplementary table S1, Supplementary Material online). On the other hand, there are currently only two *V. metoecus* genomes available. Strain RC341 was isolated from Chesapeake Bay (MD) in 1998. It was presumptively identified as a variant *V. cholerae* based on 16 S ribosomal RNA gene similarity to *V. cholerae* (Choopun 2004), but was later reclassified into its current species (Haley et al. 2010; Kirchberger et al. 2014). Strain OP3H was isolated in 2006 from Oyster Pond, a brackish pond in Cape Cod, MA. OP3H is considered the type strain of *V. metoecus*, which was recently officially described as a species (Kirchberger et al. 2014). A screen was performed for atypical *V. cholerae* isolates from a historical collection of clinical isolates at the CDC and identified that several of them were, in fact, *V. metoecus* (Boucher et al. 2011). Additional environmental *V. metoecus* strains were isolated in 2009 from Oyster Pond. While examining the population structure and surveying the mobile gene pool of environmental *V. cholerae* in Oyster Pond, Boucher et al. (2011) discovered that both *V. metoecus* and *V. cholerae* co-occur in this location. To gain a better understanding of

the *V. metoecus* species, we sequenced the genomes of four clinical *V. metoecus* strains originating from patients in the United States and an additional four from Oyster Pond. To be able to evaluate genetic interactions between strains of two different species from the same environment, we sequenced an additional 20 genomes of *V. cholerae* isolates from the same Oyster Pond samples (fig. 1).

Vibrio metoecus: The Closest Relative of *V. cholerae*

To obtain a visual comparison of the genomes, provide an overall impression of genome architecture and identify highly conserved and divergent regions, a circular BLAST atlas was constructed (Grant et al. 2012). *Vibrio metoecus* and representative *V. cholerae* genomes were compared by BLASTN alignment of coding sequences (Altschul et al. 1990) against the reference *V. cholerae* N16961, a pandemic strain from Bangladesh isolated in 1971 whose entire genome was sequenced to completion and carefully annotated (Heidelberg et al. 2000). The BLAST atlas shows a clear distinction between species, as sequence identity is higher within a species than between different species for most genes (fig. 2).

On average, *V. metoecus* shares 84% of its ORFs with *V. cholerae*, whereas 89–91% ORFs are shared between strains of the same species (supplementary table S2, Supplementary Material online). In contrast, *V. mimicus*, previously the closest known relative of *V. cholerae*, shares only 64–69% of ORFs with *V. cholerae* (Hasan et al. 2010). It was determined previously that the recommended cutoff point for prokaryotic species delineation by DNA–DNA hybridization (DDH) is 70%, which corresponds to 85% of conserved protein-coding genes for a pair of strains (Goris et al. 2007). These results show clear distinction between the three closely related species based on conserved genes, and *V. metoecus* is a much closer relative to *V. cholerae* than *V. mimicus*.

Another fundamental measure of relatedness between bacterial strains is ANI. This measure was proposed as a modern replacement to the traditional DDH method to determine relatedness of organisms, but still provide equivalent information (i.e., DNA–DNA similarity; Konstantinidis and Tiedje 2005; Goris et al. 2007). The ANI of the core genome is 86–87% between species and 98–100% within species (fig. 3*a*), showing a clear distinction between *V. metoecus* and *V. cholerae*. Two organisms belonging to the same species will have an ANI of at least 95%, corresponding to 70% DDH (Goris et al. 2007), although earlier studies have proposed a 94% cutoff (Konstantinidis and Tiedje 2005). For this reason, we have currently classified the clinical strain 07-2435 as *V. metoecus* as it shows 94% ANI with other *V. metoecus* strains but only 87% ANI with *V. cholerae* (fig. 3*a*).

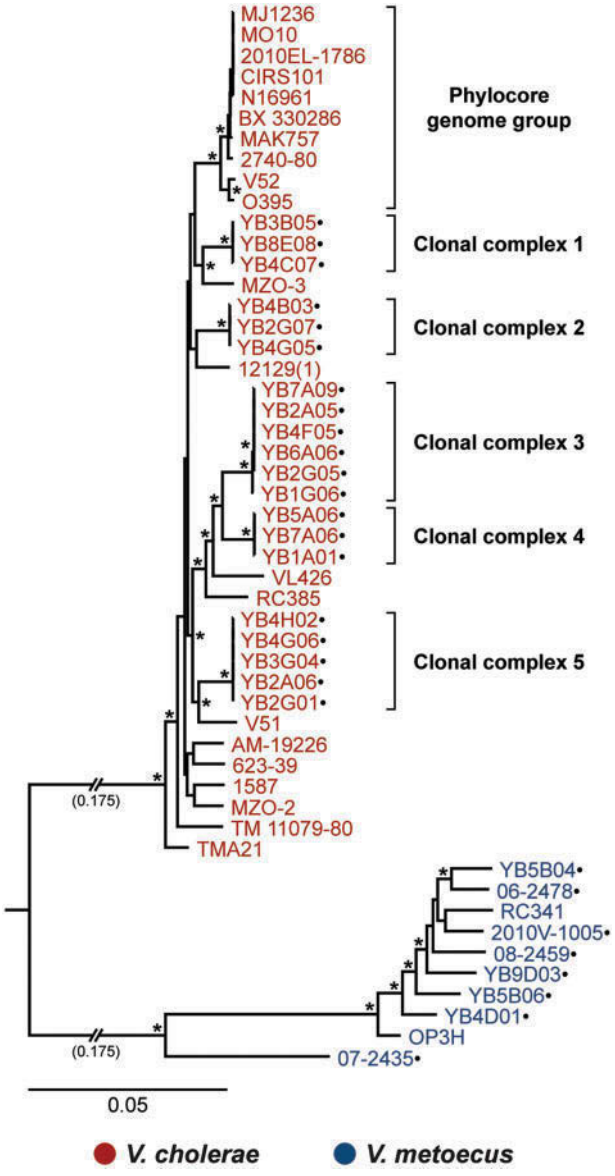


Fig. 1.—The phylogenetic relationship of the *V. metoecus* and *V. cholerae* strains. The ML phylogenetic tree was constructed from the concatenated sequence alignment of single-copy core gene families (771,455 bp). All reliable bootstrap support values are indicated with * and are at least 97% for this tree. The scale bar represents nucleotide substitutions per site. Shortened branch lengths, approximately 3.5× the scale bar (0.175), are indicated. Strains with their genomes sequenced in this study are indicated by dots. Multiple *V. cholerae* strains from Oyster Pond (MA) belong to the same clonal complex.

A Portion of the Genome Escapes the Species Boundary between *V. metoecus* and *V. cholerae*

The BLAST atlas allows for the clear distinction between strains belonging to the *V. cholerae* species and those belonging to the *V. metoecus* species. However, there is a clear and visible exception in one genomic region: The integron. Sequence

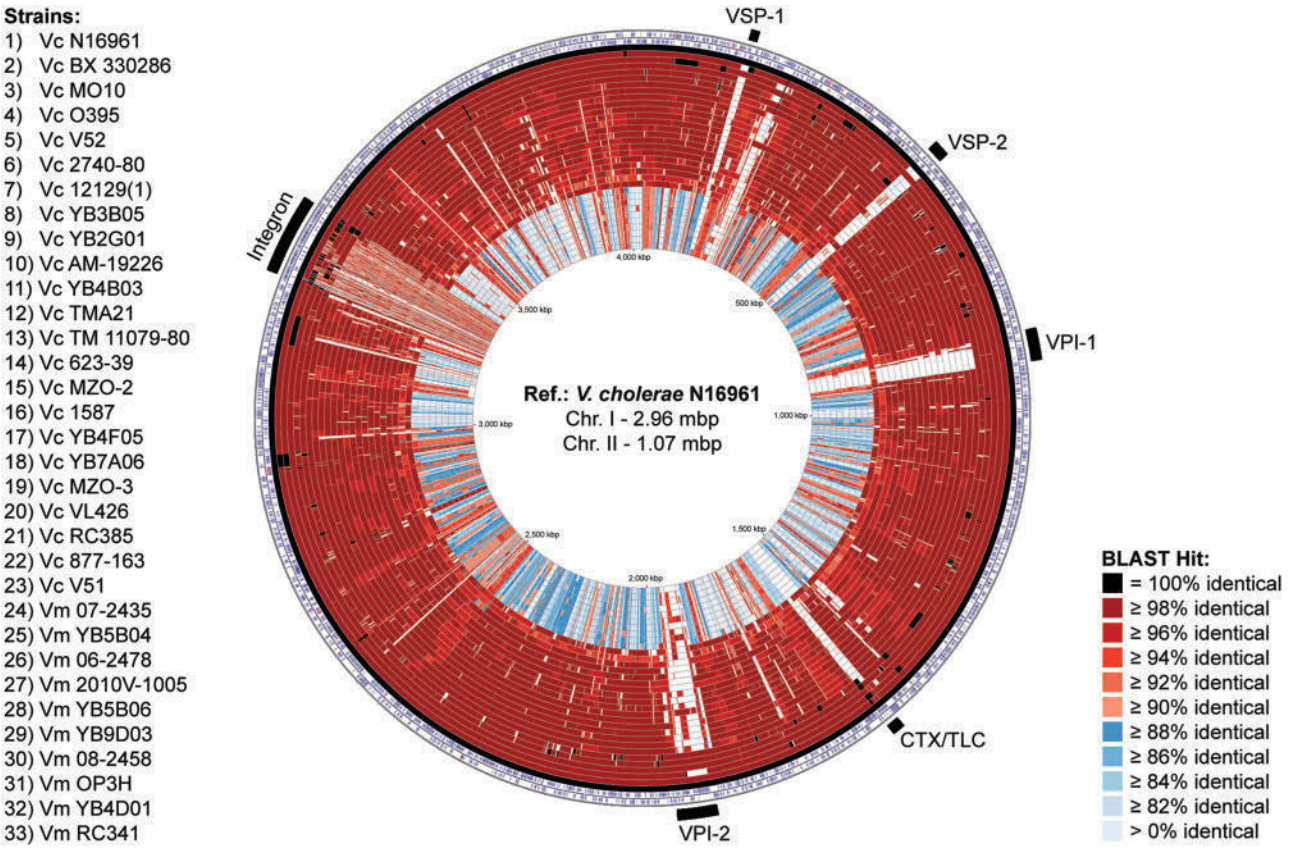


Fig. 2.—The *V. metoecus* (Vm) and *V. cholerae* (Vc) BLAST atlas. The map compares sequenced genomes against the reference (ref.), *V. cholerae* N16961. The two outermost rings show the forward and reverse strand sequence features of the reference. The next 33 rings show regions of sequence similarity detected by BLASTN comparisons between genes of the reference and query genomes. White regions indicate the absence of genes. Outermost black bars indicate the location of the major genomic islands. VSP, *Vibrio* seventh pandemic island; VPI, *Vibrio* pathogenicity island; CTX/TLC, cholera toxin/toxin-linked cryptic; chr., chromosome.

identity of genes found in the integron region does not seem to differ within and between species (fig. 2).

The integron is a region of the genome capable of gene capture and excision (Stokes and Hall 1989) and can occupy up to 3% of the genome in *V. cholerae* (Heidelberg et al. 2000). Although the size of the chromosomal integron region varies between isolates, there is no significant difference in length and number of ORFs between species and between clinical and environmental isolates (supplementary table S3, Supplementary Material online). The ANI of the integron region was determined between pairs of strains and compared with the ANI of the core genome (fig. 3). Although ANI is 86–87% between species and 98–100% within species for the core genome (fig. 3*a*), the integron region displays an average pairwise ANI of 93–94%, both within and between species (fig. 3*b*). Gene cassettes from the 10 *V. metoecus* and 11 *V. cholerae* integron regions were grouped into orthologous gene families, and the occurrence of HGT was quantified for gene families with at least two *V. metoecus* and *V. cholerae* members by the construction of phylogenetic trees. Of the 116 gene families considered, 109 or 94% do not show distinct separation between the two species in a phylogenetic tree. The high number of genes shared between species and their high nucleotide identity are likely the result of frequent interspecies HGT (figs. 2 and 3*b*). A previous study by Boucher et al. (2011) showed that there is indeed a high frequency of gene exchange in the integron region between *V. cholerae* and *V. metoecus*, specifically from the same geographic location (i.e., *V. cholerae* and *V. metoecus* in Oyster Pond) as compared with the same species in different locations (i.e., *V. cholerae* from Bangladesh and the United States). Here, we show that not only is the frequency of interspecies HGT high in the integron, but that its level is such that this region becomes indistinguishable between species.

Although the functions of the majority of integron gene cassettes are unknown (Boucher et al. 2007), many of the known genes are antibiotic resistance genes and are implicated in the evolution of bacteria highly resistant to antibiotics (Collis and Hall 1995; Rowe-Magnus and Mazel 2002). Looking into the predicted functions of the 116 gene families

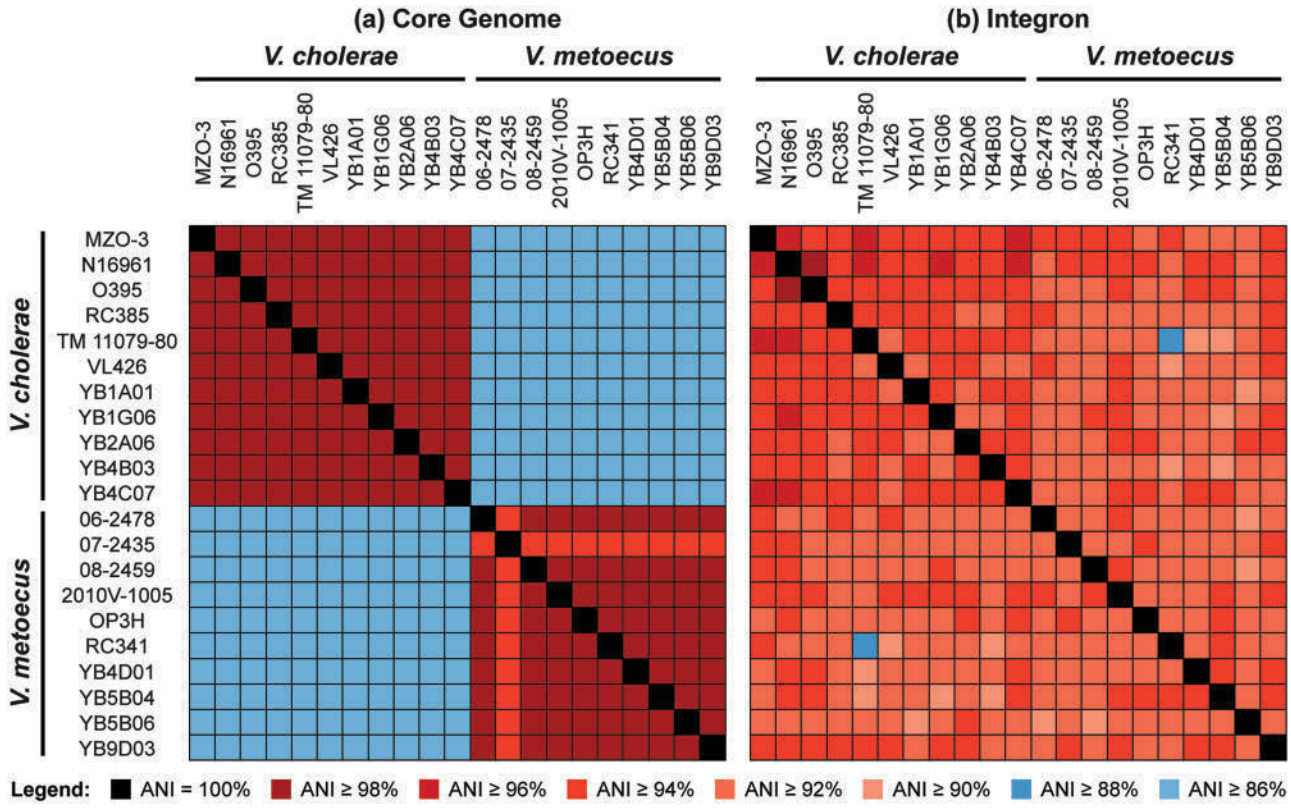


Fig. 3.—ANI of the core genome versus chromosomal integron region of *V. metoecus* and *V. cholerae*. (*a*) Intra- and interspecies pairwise comparison of the 1,560 single-copy core genes (≈1.42 Mb). (*b*) Intra- and interspecies pairwise comparison of the integron gene cassettes.

comprising 1,452 gene cassettes, the majority of which are shared between *V. metoecus* and *V. cholerae*, reveals genes that encode proteins involved in transport and metabolism of various molecules (supplementary fig. S1, Supplementary Material online), suggesting a major contributing function of the integron for host acquisition and distribution of important resources in the environment by bacteria (Koenig et al. 2008). Gene cassettes encoding nicotinamidase-related amidases are present in multiple copies. Nicotinamidase catalyzes the deamination of nicotinamide to produce ammonia and nicotinic acid (Petrack et al. 1965). A key enzyme in many organisms, nicotinamidase has been shown to be important in the proliferation of bacteria pathogenic to mammalian hosts including humans (Purser et al. 2003; Kim et al. 2004). Other genes present are involved in basic cellular functions such as acetyltransferases, involved in posttranslational modifications of ribosomal proteins, the functional significance of which remains unclear but may have regulatory roles (Nesterchuk et al. 2011). Some genes are part of the plasmid stabilization systems, which include the toxin–antitoxin (TA) systems. TA systems are frequently found in gene cassette arrays for the stabilization and prevention of loss of gene cassettes. They also play additional roles in stress response, bacterial persistence, and phage defense (Iqbal et al. 2015).

A Lack of Reciprocity: Directional Gene Flow from *V. cholerae* to *V. metoecus*

To get a quantitative estimate of the amount of HGT between *V. cholerae* and *V. metoecus*, we investigated the amount of interspecies recombination taking place in their core genomes. An ML tree was constructed for each of the 1,947 gene families comprising the *V. metoecus*–*V. cholerae* core genome (fig. 4). The trees were then analyzed for gene transfer events by partitioning them into clades (Schliep 2011). In our analysis, following the definition by Schliep et al. (2011), a gene transfer is hypothesized if a member of one species clusters with members of the other species in a clade, and the tree cannot be partitioned into perfect clades, which must consist of all members from the same species and only of that species. Considering only the single-copy core genes, we have inferred interspecies HGT in 376 of 1,560 genes (24%; supplementary table S4, Supplementary Material online). Our analysis excluded 387 core genes that have duplicates in at least one of the genomes, as it is difficult to reliably assess HGT in genes from large paralogous families (Ge et al. 2005). Using this method, it was possible to determine directionality of HGT, whether from *V. cholerae* to *V. metoecus* or vice versa. HGT was qualified by examining the individual gene trees, and only reliable clustering with at

least 70% bootstrap support was considered (Hillis and Bull 1993). A total of 655 interspecies gene transfer events were detected, with the majority (489 or 75%; $P = 0.0053$) with *V. metoecus* as the recipient (i.e., *V. metoecus* members clustering within the *V. cholerae* clade). On the other hand, we detected 166 (25%) of gene transfer events with *V. cholerae* as the recipient (supplementary table S5, Supplementary Material online).

To investigate whether this bias in directionality of HGT was due to differences in the origin or ecology of strains from one species or the other, we performed the analysis using only environmental strains from Oyster Pond. To ensure equal genetic diversity for both species, we compared the same number of isolates from each species. The 20 *V. cholerae* isolates we sequenced for this study can be grouped into five clonal complexes as determined by multilocus sequence typing of seven housekeeping genes. All the isolates from the same clonal complex cluster together in a core genome phylogenetic tree (fig. 1). They also exhibit 100% ANI only with each other but not with isolates from other clonal complexes (supplementary table S6, Supplementary Material online). Indeed, members of the same clonal complex always cluster together in all the individual gene trees examined (fig. 4). We therefore randomly chose one isolate from each *V. cholerae* clonal complex from Oyster Pond, yielding a final data set of five genomes from each species. A total of 224 interspecies gene transfer events were detected in this environment-specific data set, where 192 (86%; $P = 0.0012$) involved *V. metoecus* as the recipient and only 32 (14%) with *V. cholerae* as the recipient (table 1). One possibility to explain this bias could be that *V. cholerae* genes are more abundant in the environment and therefore more accessible to *V. metoecus*. Indeed, using culture-based methods, *V. cholerae* was ten times more abundant than *V. metoecus* in Oyster Pond. Another possibility is that *V. cholerae* is more refractory to HGT as they contain more barriers to gene uptake, such as restriction-modification systems, or that *V. metoecus* is more permissive, containing more DNA uptake systems (conjugative plasmids, natural competence machinery or phages). However, no significant difference could be found in the number or nature of proteins involved in restriction-modification or DNA uptake systems between *V. metoecus* and *V. cholerae* in our study, although poorly transformable *V. cholerae*, despite having an intact and perfectly functioning DNA uptake system, have been reported (Katz et al. 2013). Additionally, nuclease activity by Dns, Xds, and other DNases can inhibit natural transformation (Blokesch and Schoolnik 2008; Gaasbeek et al. 2009). We also surveyed our *V. metoecus* and *V. cholerae* genomes for predicted DNases and found no significant difference between species.

Despite the directional gene transfer from *V. cholerae* to *V. metoecus*, it seems that the latter might have contributed to the virulence of its more famous relative by HGT. Interspecies recombination was detected in four core genes where at least one clinical *V. cholerae* grouped in the same clade with *V. metoecus* (fig. 4). Interestingly, three of these genes are implicated, whether directly or indirectly, in *V. cholerae* pathogenesis. VC2614 encodes a cyclic adenosine monophosphate regulatory protein, a global regulator of gene expression in *V. cholerae* including CTX and TCP (Skorupski and Taylor 1997). It appears that HGT in this case occurred in the ancestor of the phylocore genome (PG) group, which contains all pandemic strains (fig. 1; Chun et al. 2009), with a clinical *V. metoecus* strain as the possible donor. The new version of this cyclic-AMP regulatory protein was eventually lost in the classical O1 strain (O395). VC2545 encodes an inorganic pyrophosphatase, and its expression in *V. cholerae* may play an important role during human and mouse infection (Lombardo et al. 2007). This transfer was only between clinical *V. metoecus* and classical O1. VCA0925 encodes a dihydroorotase essential for pyrimidine biosynthesis. Biosynthesis of nucleotides is the single most critical metabolic function for growth of pathogenic bacteria in the bloodstream because of scarcity of nucleotide precursors but not other nutrients, and the genes involved serve as potential antibiotic targets for treatments of blood infection (Samant et al. 2008). Here, gene transfer involved not just the PG group of *V. cholerae* but also the environmental strains of clonal complex 5 and 623-39.

Although these interspecies recombination events do not represent novel gene acquisitions, gaining a new allele of a gene can often have important consequences in a pathogen, changing its fitness in the host. This has been demonstrated for single-point mutations in *ompU*, *vpvC*, and *ctxB*. The *ompU* gene encodes for the major outer membrane porin OmpU, generally for the transport of hydrophilic solutes, but has been shown to provide *V. cholerae* resistance to bile acids and antimicrobial peptides in the host (Provenzano et al. 2000; Mathur and Waldor 2004). It is suggested that it can also act as a receptor for phage to infect *V. cholerae* (Seed et al. 2014). The *vpvC* gene encodes for diguanylate cyclase, and the mutation results in a switch from the smooth to rugose phenotype in *V. cholerae* (Beyhan and Yildiz 2007). The single-point mutations in these genes result in a *V. cholerae* that is less susceptible to phage infection, contributing to the evolutionary success of the pathogen (Beyhan and Yildiz 2007; Seed et al. 2014). *Vibrio cholerae* responsible for cholera outbreaks in Bangladesh have changing genotypes of *ctxB*, a subunit of CTX (Waldor and Mekalanos 1996), also caused by a single-point mutation (Rashed et al. 2012). The years 2006 and 2007 saw a dominance of *V. cholerae* with the *ctxB* genotype 1 (*ctxB*1). *Vibrio cholerae* with the *ctxB* genotype 7 (*ctxB*7) outcompeted *ctxB*1 from 2008 to 2012. However, there appears to be a shift back to *ctxB*1 since 2013. The changing *ctxB* genotypes were associated with differing levels of severity of cholera. This also suggests CTX phage-mediated evolution, survival, and dominance of *V. cholerae* (Rashed et al. 2012; Rashid et al. 2015).

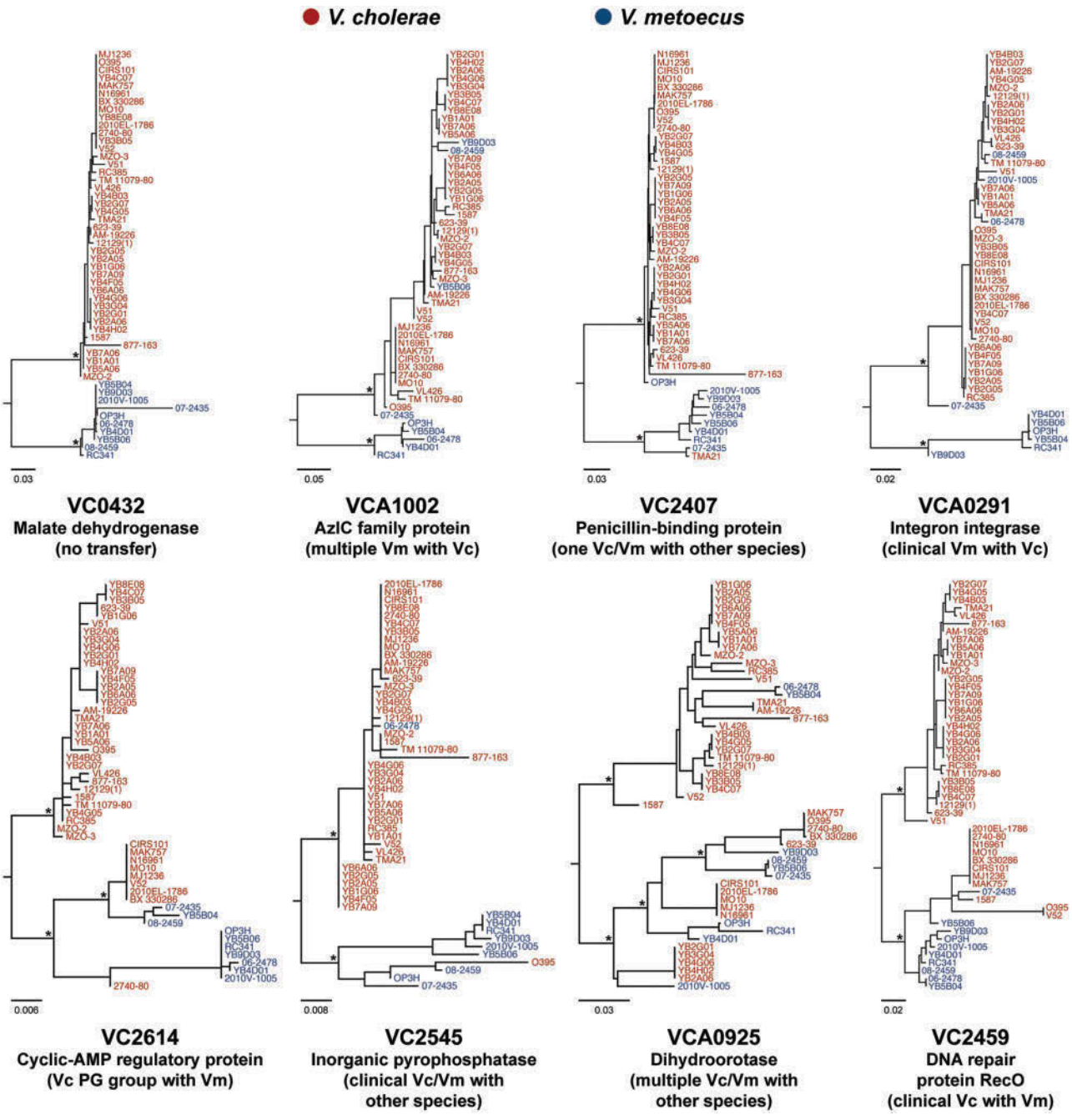


Fig. 4.—Representative HGT between *V. metoecus* (Vm) and *V. cholerae* (Vc). The trees are representative ML phylogenetic trees from 1,560 orthologous families of single-copy core genes showing various examples of transfer events. Bottom trees: Transfers involving at least one clinical *V. cholerae* clustering with *V. metoecus*. Relevant bootstrap support (>70%) is indicated with *. The scale bars represent nucleotide substitutions per site.

Components of Major Pathogenicity Islands Are More Ancient than the *V. cholerae* Species

A BSR map (Rasko et al. 2005) was constructed to show the presence or absence of the genes comprising the major pathogenicity islands in various *V. metoecus* and *V. cholerae* isolates (fig. 5). Using the genes from *V. cholerae* N16961 as reference, BLASTP was used to determine the presence of homologous genes in the other strains (Altschul et al. 1990). The major *V. cholerae* virulence factors, CTX and TCP, which are encoded by pathogenicity islands that have been acquired horizontally by phage infections of the CTXΦ and VPIΦ, respectively (Waldor and Mekalanos 1996; Karaolis et al. 1999), are absent from all clinical and environmental *V. metoecus* (fig. 5*a*). The absence of CTX and TCP in

Table 1

HGT Count for *Vibrio metoecus* and Representative *Vibrio cholerae* Strains from Oyster Pond (MA) Based on 376 Single-Copy Core Genes with Inferred HGT

Species and Strain	HGT Count	Percent of Total
Vibrio metoecus OP3H	55	25
Vibrio metoecus YB4D01	43	19
Vibrio metoecus YB5B06	37	17
Vibrio metoecus YB5B04	30	13
Vibrio metoecus YB9D03	27	12
	192	**86**
Vibrio cholerae YB2G01 (CC 5)	16	7
Vibrio cholerae YB4F05 (CC 3)	9	4
Vibrio cholerae YB4B03 (CC 2)	4	2
Vibrio cholerae YB7A06 (CC 4)	2	1
Vibrio cholerae YB3B05 (CC 1)	1	0
Total	**32**	**14**

NOTE.—Only one strain from each clonal complex (CC) was included. An HGT event was hypothesized when a strain clustered with members of the other species in a phylogenetic tree, with reliable bootstrap support (>70%). Unequal variance *t*-test, $P = 0.0012$.

V. metoecus is consistent with the absence of reports on a toxigenic *V. metoecus*.

Interestingly, our results show some of the other major pathogenicity islands to be present in some *V. metoecus* and nonpandemic *V. cholerae* strains in fragments and not as a complete presence or absence. This is evident in the *Vibrio* pathogenicity island 2 (VPI-2), which can be divided into four subclusters we call "islets," as indicated in figure 5*a*. These four islets match the previous description of Jermyn and Boyd (2002) for VPI-2: 1) A type-1 restriction-modification system for protection against viral infection, 2) a *nan-nag* cluster for sialic acid metabolism, 3) a Mu phage-like region, and 4) a number of ORFs of unknown function. We hypothesize two scenarios as to the fragmentation of these genomic islands 1) that the islands were obtained as a whole and sections were eventually lost, or 2) that the islands were acquired independently in islets and were accreted into the same region in the genome. Evolution would favor the latter hypothesis, as it is more parsimonious for fewer environmental strains to independently acquire certain islets of the islands rather than a majority of the strains acquiring whole islands and losing most regions eventually (Freeman and Herron 2007). Phylogenetic trees were constructed for the gene families that constitute the four putative islets of VPI-2. Gene trees for islet B, the *nan-nag* cluster, show distinct clustering of *V. metoecus* and *V. cholerae*, suggesting the acquisition of this region by a common ancestor, which diverged and evolved independently after speciation, with more recent isolated HGT events between *V. metoecus* and *V. cholerae* (fig. 5*b* and supplementary fig. S2, Supplementary Material online). A similar pattern of distinct clustering of *V. metoecus* and *V. cholerae* is also observed in islet A, but the latter is only present in O1 El Tor *V. cholerae* and two *V. metoecus* strains (fig. 5*a*), suggesting that it was horizontally transferred between the two species and likely absent from their common ancestor. Furthermore, islet C, the putative Mu phage-like region, is only detected in *V. cholerae* of the PG group and TM 11079-80, an O1 El Tor environmental isolate. This islet is absent in *V. metoecus*, which suggests a more recent acquisition of this region only by certain *V. cholerae*. Finally, islet D is prevalent in the majority of the isolates, whether *V. metoecus* or *V. cholerae*, which do not cluster by species in the phylogeny (fig. 5*b*). This suggests frequent interspecies HGT of its component genes. Taken together, these results support that the VPI-2 island emerged by accretion of smaller islets with different evolutionary histories before reaching the form currently found in *V. cholerae* O1 El Tor or classical pandemic strains. The *nan-nag* cluster (islet B) is likely ancestral, being present before speciation of *V. cholerae* and *V. metoecus*, with islets A and D acquired later by the ancestor of pandemic *V. cholerae* through HGT within or between species and islet C added most recently through HGT from an unknown source.

The *Vibrio* seventh pandemic islands 1 and 2 (VSP-1 and VSP-2, respectively) are genomic islands believed to be present and unique only among the seventh pandemic isolates of *V. cholerae* (Dziejman et al. 2002; O'Shea et al. 2004). These VSPs are hypothesized to provide a fitness advantage to these isolates. However, multiple variants of VSP-2 have been detected in *V. cholerae*, including non-O1/O139 strains, by acquisition and loss of genes at specific loci within a conserved core genomic backbone (Taviani et al. 2010). This core VSP-2 is also present in two *V. metoecus* isolates, the clinical 2010 V-1005 and environmental RC341 (fig. 5*a*), and may have been acquired from *V. cholerae*, as indicated by the great similarity of genes in this region to *V. cholerae* and phylogenetic analysis (fig. 5*b* and supplementary fig. S3, Supplementary Material online). This variant of VSP-2 is stable and present in diverse strains isolated from different times and geographic locations and may be the one circulating among non-O1/O139 isolates (Taviani et al. 2010). VSP-1 is present almost in its entirety in environmental *V. cholerae* VL426 and *V. metoecus* RC341 (fig. 5*a*); similar strains in the environment may serve as reservoirs of VSP-1. There is no correlation between the presence of VSP-1 and VSP-2 in non-O1/O139 *V. cholerae*, indicating that both islands were acquired independently in different HGT events by seventh pandemic *V. cholerae* (Taviani et al. 2010). The presence of both of the entire VSP-1 and the core of VSP-2 in *V. metoecus* strains indicate interspecies movement of pathogenicity islands, suggesting that interspecies transfer can contribute to the evolution of pathogenic variants.

Fundamental Genetic Differences between *V. metoecus* and *V. cholerae*

To determine genetic differences between *V. cholerae* and *V. metoecus* and the unique gene content of each species,

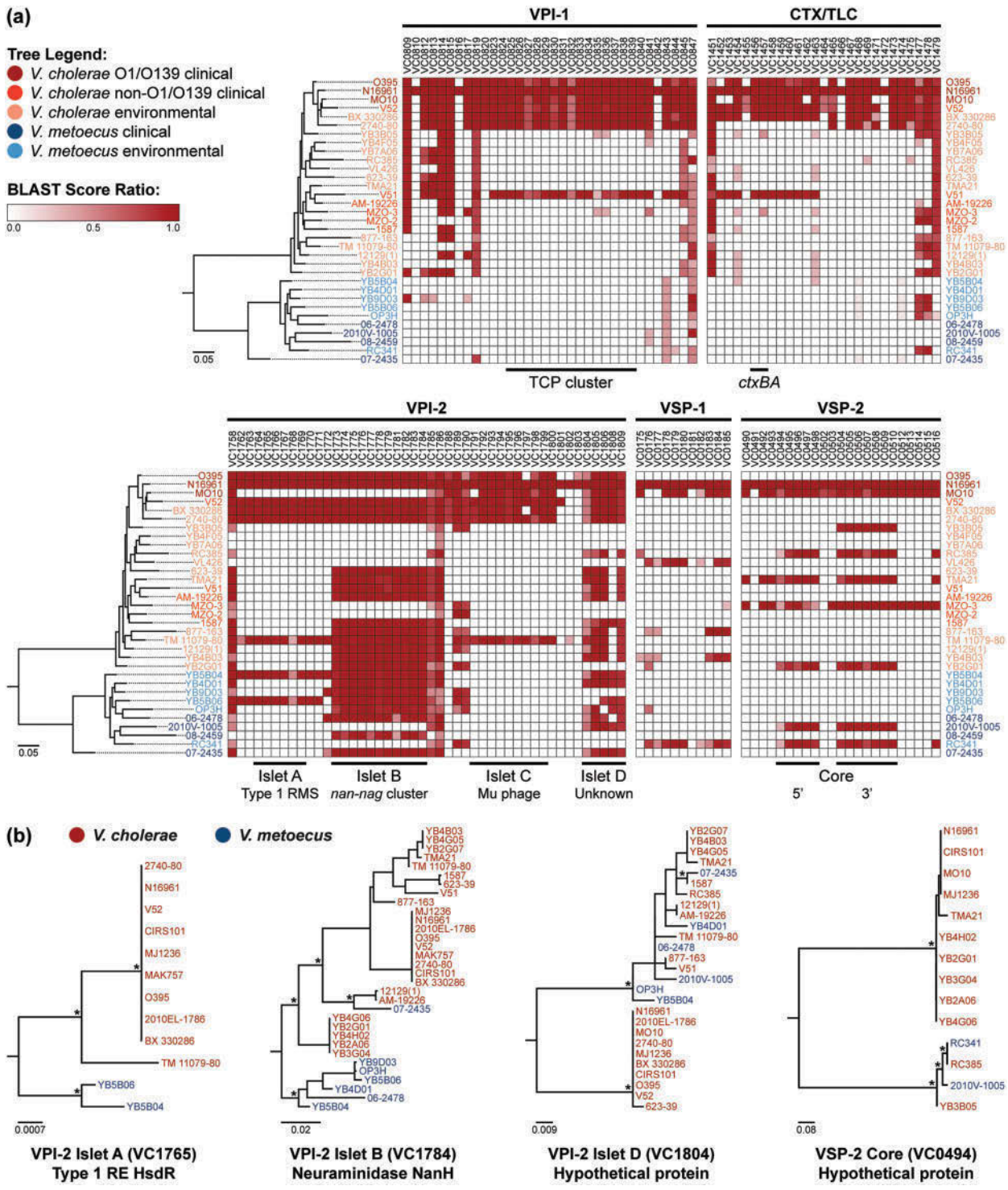


Fig. 5.—Virulence factors present in *V. metoecus* and *V. cholerae*. (*a*) The phylogenetic relationship of the *V. metoecus* and *V. cholerae* strains is shown on the left of each BSR map. The ML phylogenetic tree was constructed using 3,978 amino acid positions based on 400 universally conserved bacterial and archaeal proteins. The scale bars represent amino acid substitutions per site. The columns on the BSR maps show genes (locus tags) from genomic islands VPI-1, CTX/TLC, VPI-2, VSP-1, and VSP-2 of the reference, *V. cholerae* N16961. The black bars at the bottom of the BSR maps indicate the TCP cluster of VPI-1, *ctxAB* of CTX/TLC, islets of VPI-2, and core regions of VSP-2. The gradient bar shows the BSRs and their corresponding colors, with white regions indicating the absence of genes. Only BSR values of at least 0.3 were included. VPI, *Vibrio* pathogenicity island; CTX/TLC, cholera toxin/toxin-linked cryptic; VSP, *Vibrio* seventh pandemic island; RMS, restriction-modification system. (*b*) Representative ML phylogenetic trees of orthologous gene families of the VPI-2 islets and the VSP-2 core. Relevant bootstrap support (>70%) is indicated with *. The scale bars represent nucleotide substitutions per site. RE, restriction endonuclease.

we first compiled their pan- and core genomes (supplementary fig. S4, Supplementary Material online). The pan-genome is the entire gene repertoire of a bacterial species, whereas the core genome comprises genes shared by all the strains (Tettelin et al. 2005, Vernikos et al. 2015). ORFs from both species were assigned to orthologous groups based on sequence similarity, yielding pan- and core genomes containing 5,613 and 2,089 gene families, respectively, based on the 42 *V. cholerae* genomes used in this study (supplementary fig. S4*a*, Supplementary Material online). This differs from the previous estimate of Chun et al. (2009), who determined the *V. cholerae* core genome to contain 2,432 gene families based on 23 strains, a higher core genome size than we obtained from our data set. The reduced core genome size is expected as the number of shared genes decreases with the addition of each new genome (Tettelin et al. 2005). It also depends on the degree of relatedness of the organisms. A study on 32 *Vibrionaceae* genomes, including 18 representative *V. cholerae*, established a core genome of only 1,000 gene families (Vesth et al. 2010). The *V. metoecus* pan- and core genomes constitute 4,298 and 2,872 gene families, respectively, based on the ten genomes currently available (supplementary fig. S4*b*, Supplementary Material online). The difference in pan- and core genome sizes of *V. cholerae* and *V. metoecus* can be explained by the significant difference in the number of genomes used. We expect the pan- and core genomes of *V. metoecus* to ultimately reach sizes similar to that of *V. cholerae* when genomes of additional strains become available.

As a newly described species, very little is currently known about the biology of *V. metoecus* and what sets it apart genetically from *V. cholerae*. From the combined pan-genome of both species, orthologous gene families present in various groups of strains were determined: Families unique to *V. metoecus* and *V. cholerae*, or unique to clinical and environmental strains (supplementary fig. S5, Supplementary Material online). Function was predicted for each gene family based on the COG database (supplementary fig. S6, Supplementary Material online). *Vibrio metoecus* contains more unique gene families than *V. cholerae* that are involved in carbohydrate transport and metabolism (supplementary fig. S6*a*, Supplementary Material online). In the species description study by Kirchberger et al. (2014), it was determined that although the majority of biochemical and growth characteristics of *V. metoecus* resemble *V. cholerae*, the former was mainly differentiated from the latter for its ability to utilize the complex sugars D-glucuronic acid and *N*-acetyl-D-galactosamine. Indeed, multiple β-galactosidase/β-glucuronidase enzymes for the breakdown of D-glucuronic acid (Louis and Doré 2014) were present in our *V. metoecus*-specific COG data set, but not in *V. cholerae*. Multiple hexosaminidases for the hydrolysis of terminal *N*-acetyl-D-hexosamine (Magnelli et al. 2012) were also detected in *V. metoecus*, which supports the phenotype observed by Kirchberger et al. (2014). Additionally, genes unique for clinical *V. metoecus* and clinical *V. cholerae* were identified (supplementary fig. S6*b*, Supplementary Material online). Clinical *V. cholerae* have more genes encoding proteins involved in replication, recombination, and repair (mostly transposases), and signal transduction, such as the GGDEF family protein. Transposases in pathogenicity islands can contribute to the instability and mobilization of virulence genes (Schmidt and Hensel 2004). The GGDEF family protein is critical in biofilm formation (García et al. 2004) and is highly induced in *V. cholerae* during infection in humans and mice (Lombardo et al. 2007). As expected, genes of the CTX and TCP clusters were not found in our clinical *V. cholerae*-specific data set because they are not unique to clinical strains, but are also present in some environmental ones (fig. 5*a*). Among the genes uniquely found in clinical *V. metoecus* is a putative *mdaB* (modulator of drug activity B) gene. The *mdaB* gene has been shown to play an important role in oxidative stress resistance and host colonization in *Helicobacter pylori* (Wang and Maier 2004), and may also contribute to the fitness of clinical *V. metoecus* in the host.

Conclusion

The discovery of *V. metoecus*, the closest known relative of *V. cholerae*, presents an opportunity to study the HGT events between species and the role this might play in the evolution of pathogenesis. In contrast to the core genome, which is distinctly more similar between members of the same species, the chromosomal integron region, occupying approximately 3% of *V. cholerae* and *V. metoecus* genomes, represents a pool of genes which is freely exchanged between these two species. This genomic region displays no greater similarity within than between species. Genomic islands encoding pathogenicity factors, known to play a role in pandemic *V. cholerae* virulence, are also occasionally found in *V. metoecus*, either completely or in part. This includes VPI-2, found in most pandemic *V. cholerae*, as well as the VSP islands, previously believed to be specific to *V. cholerae* strains from the seventh pandemic. VPI-2 and VSP-2 seem to have assembled over time by accretion of smaller units, which we call islets. Some islets, such as the *nan-nag* cluster of the VPI-2 (islet B) for sialic acid metabolism, have been stable over time and were present in the common ancestor of *V. metoecus* and *V. cholerae*. Other islets, such as islet A (restriction-modification system) and islet D (unknown function) of VPI-2, the core of VSP-2, or the entire VSP-1 island seem to move frequently between *V. metoecus* and *V. cholerae* and are not restricted to pandemic strains.

The most striking finding is that even the core genome of *V. cholerae* is susceptible to frequent interspecies recombination with *V. metoecus*. Twenty-four percent of the genes found in all *V. cholerae* and *V. metoecus* had experienced interspecies recombination. There also seems to be a

directional bias to these recombination events. In Oyster Pond, in particular, *V. metoecus* is the recipient of genes six times more than *V. cholerae*. The cause of this bias is unclear, but it does not seem to be restricted to a single environment, as all *V. metoecus* are recipients of more interspecies DNA transfers than any of the *V. cholerae* strains investigated. One possibility is that *V. cholerae* is more abundant in most environments than *V. metoecus* and there is, therefore, simply more of its DNA available for uptake. Indeed, in this study, *V. cholerae* was isolated ten times more frequently than *V. metoecus* from Oyster Pond, which is consistent with the observed HGT bias. However, this explanation is very tentative and requires more evidence, as this study is the first one to isolate *V. cholerae* and *V. metoecus* quantitatively from the same site, and this was done using a culture-based method. This relative abundance would not necessarily be obtained with more accurate culture-free quantitative methods. Also, HGT could be biased because of differences in phage abundance/susceptibility, presence of DNA uptake systems, or restriction-modification systems. Nonetheless, this is, to our knowledge, the first quantitative report of HGT bias for bacteria in the natural environment and has fundamental implications for understanding the evolution of microbial populations.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful for the assistance of Tania Nasreen, Paul Stothard (University of Alberta), Lee Katz, Mike Frace, Maryann Turnsek (Centers for Disease Control and Prevention), Éric Bapteste (Université Pierre et Marie Curie), Gary Van Domselaar, and Aaron Petkau (National Microbiology Laboratory). They appreciate the helpful discussions with Rebecca Case, Stefan Pukatzki, and David Wishart (University of Alberta). This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research, the Canadian Foundation for Innovation (to Y.B.), and the Alberta Innovates—Technology Futures (to F.D.O. and P.C.K.).

Literature Cited

Ali M, et al. 2012. The global burden of cholera. Bull World Health Organ. 90:209–218A.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Aziz RK, et al. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Beyhan S, Yildiz FH. 2007. Smooth to rugose phase variation in *Vibrio cholerae* can be mediated by a single nucleotide change that targets c-di-GMP signalling pathway. Mol Microbiol. 63:995–1007.

Blokesch M, Schoolnik GK. 2008. The extracellular nuclease Dns and its role in natural transformation of *Vibrio cholerae*. J Bacteriol. 190:7232–7240.

Boucher Y, et al. 2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. MBio 2:e00335–10.

Boucher Y, Labbate M, Koenig JE, Stokes HW. 2007. Integrons: mobilizable platforms that promote genetic diversity in bacteria. Trends Microbiol. 15:301–309.

Choopun N. 2004. The population structure of *Vibrio cholerae* in Chesapeake Bay [Ph.D. thesis]. [College Park (MD)]: University of Maryland.

Chun J, et al. 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. Proc Natl Acad Sci U S A. 106:15442–15447.

Collis CM, Hall RM. 1995. Expression of antibiotic resistance genes in the integrated cassettes of integrons. Antimicrob Agents Chemother. 39:155–162.

Davis BR, et al. 1981. Characterization of biochemically atypical *Vibrio cholerae* strains and designation of a new pathogenic species, *Vibrio mimicus*. J Clin Microbiol. 14:631–639.

De la Cruz F, Davies J. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol. 8:128–133.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2:414–424.

Dziejman M, et al. 2002. Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. Proc Natl Acad Sci U S A. 99:1556–1561.

Freeman S, Herron JC. 2007. Evolutionary analysis. San Francisco (CA): Pearson Benjamin Cummings.

Gaasbeek EJ, et al. 2009. A DNase encoded by integrated element CJIE1 inhibits natural transformation of *Campylobacter jejuni*. J Bacteriol. 191:2296–2306.

García B, et al. 2004. Role of the GGDEF protein family in *Salmonella* cellulose biosynthesis and biofilm formation. Mol Microbiol. 54:264–277.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3:e316.

Gomez-Gil B, et al. 2014. The family *Vibrionaceae*. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F, editors. The prokaryotes—gammaproteobacteria. Berlin (Germany): Springer. p. 659–747.

Goris J, et al. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 57:81–91.

Grant JR, Arantes AS, Stothard P. 2012. Comparing thousands of circular genomes using the CGView Comparison Tool. BMC Genomics 13:202.

Haley BJ, et al. 2010. Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. BMC Microbiol. 10:154.

Hasan NA, et al. 2010. Comparative genomics of clinical and environmental *Vibrio mimicus*. Proc Natl Acad Sci U S A. 107:21134–21139.

Heidelberg JF, et al. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477–483.

Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol. 42:182–192.

Iqbal N, Guérout AM, Krin E, Le Roux F, Mazel D. 2015. Comprehensive functional analysis of the 18 *Vibrio cholerae* N16961 toxin-antitoxin

systems substantiates their role in stabilizing the superintegron. J Bacteriol. 197:2150–2159.
Jermyn WS, Boyd EF. 2002. Characterization of a novel *Vibrio* pathogenicity island (VPI-2) encoding neuraminidase (*nanH*) among toxigenic *Vibrio cholerae* isolates. Microbiology 148:3681–3693.
Karaolis DK, Somara S, Maneval DR Jr, Johnson JA, Kaper JB. 1999. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. Nature 399:375–379.
Katz LS, et al. 2013. Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. MBio 4:e00398–13.
Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.
Kim S, et al. 2004. *Brucella abortus* nicotinamidase (PncA) contributes to its intracellular replication and infectivity in mice. FEMS Microbiol Lett. 234:289–295.
Kirchberger PC, et al. 2014. *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. Int J Syst Evol Microbiol. 64:3208–3214.
Koenig JE, et al. 2008. Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. Environ Microbiol. 10:1024–1038.
Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 102:2567–2572.
Langille MG, Brinkman FS. 2009. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25:664–665.
Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.
Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Lombardo MJ, et al. 2007. An *in vivo* expression technology screen for *Vibrio cholerae* genes expressed in human volunteers. Proc Natl Acad Sci U S A. 104:18229–18234.
Louis P, Doré J. 2014. Functional metagenomics of human intestinal microbiome β-glucuronidase activity. In: Nelson KE, editor. Encyclopedia of metagenomics. New York: Springer. p. 1–8.
Magnelli P, Bielik A, Guthrie E. 2012. Identification and characterization of protein glycosylation using specific endo- and exoglycosidases. Methods Mol Biol. 801:189–211.
Mathur J, Waldor MK. 2004. The *Vibrio cholerae* ToxR-regulated porin OmpU confers resistance to antimicrobial peptides. Infect Immun. 72:3577–3583.
Mazel D. 2006. Integrons: agents of bacterial evolution. Nat Rev Microbiol. 4:608–620.
Mazel D, Dychinco B, Webb VA, Davies J. 1998. A distinctive class of integron in the *Vibrio cholerae* genome. Science 280:605–608.
Nesterchuk MV, Sergiev PV, Dontsova OA. 2011. Posttranslational modifications of ribosomal proteins in *Escherichia coli*. Acta Naturae 3:22–33.
O'Shea YA, et al. 2004. The *Vibrio* seventh pandemic island-II is a 26.9 kb genomic island present in *Vibrio cholerae* El Tor and O139 serogroup isolates that shows homology to a 43.4 kb genomic island in *V. vulnificus*. Microbiology 150:4053–4063.
Petrack B, Greengard P, Craston A, Sheppy F. 1965. Nicotinamide deamidase from mammalian liver. J Biol Chem. 240:1725–1730.
Provenzano D, Schuhmacher DA, Barker JL, Klose KE. 2000. The virulence regulatory protein ToxR mediates enhanced bile resistance in *Vibrio cholerae* and other pathogenic *Vibrio* species. Infect Immun. 68:1491–1497.
Purser JE, et al. 2003. A plasmid-encoded nicotinamidase (PncA) is essential for infectivity of *Borrelia burgdorferi* in a mammalian host. Mol Microbiol. 48:753–764.
R Development Core Team. 2014. R: a language and environment for statistical computing. Version 3.1.2. Vienna (Austria): R Foundation for Statistical Computing.
Rashed SM, et al. 2012. Genetic characteristics of drug-resistant *Vibrio cholerae* O1 causing endemic cholera in Dhaka, 2006-2011. J Med Microbiol. 61:1736–1745.
Rashid MU, et al. 2015. *ctxB*1 outcompetes *ctxB*7 in *Vibrio cholerae* O1, Bangladesh. J Med Microbiol. Advance Access published October 19, 2015; doi: 10.1099/jmm.0.000190.
Rasko DA, Myers GS, Ravel J. 2005. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics 6:2.
Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 106:19126–19131.
Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng. 12:85–94.
Rowe-Magnus DA, Mazel D. 2002. The role of integrons in antibiotic resistance gene capture. Int J Med Microbiol. 292:115–125.
Samant S, et al. 2008. Nucleotide biosynthesis is critical for growth of bacteria in human blood. PLoS Pathog. 4:e37.
Schliep K, Lopez P, Lapointe FJ, Bapteste É. 2011. Harvesting evolutionary signals in a forest of prokaryotic gene trees. Mol Biol Evol. 28:1393–1405.
Schliep KP. 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.
Schmidt H, Hensel M. 2004. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev. 17:14–56.
Seed KD, et al. 2014. Evolutionary consequences of intra-patient phage predation on microbial populations. Elife 3:e03497.
Segata N, Börnigen D, Morgan XC, Huttenhower C. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun. 4:2304.
Skorupski K, Taylor RK. 1997. Cyclic AMP and its receptor protein negatively regulate the coordinate expression of cholera toxin and toxin-coregulated pilus in *Vibrio cholerae*. Proc Natl Acad Sci U S A. 94:265–270.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Stokes HW, Hall RM. 1989. A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. Mol Microbiol. 3:1669–1683.
Szamosi JC. 2012. ISAAC: an improved structural annotation of *attC* and an initial application thereof [M.Sc. thesis]. [Hamilton (ON)]: McMaster University.
Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36.
Taviani E, et al. 2010. Discovery of novel *Vibrio cholerae* VSP-II genomic islands using comparative genomic analysis. FEMS Microbiol Lett. 308:130–137.
Taylor RK, Miller VL, Furlong DB, Mekalanos JJ. 1987. Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. Proc Natl Acad Sci U S A. 84:2833–2837.
Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A. 102:13950–13955.

Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. Curr Opin Microbiol. 23:148–154.
Vesth T, et al. 2010. On the origins of a *Vibrio* species. Microb Ecol. 59:1–13.
Waldor MK, Mekalanos JJ. 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 272: 1910–1914.
Wang G, Maier RJ. 2004. An NADPH quinone reductase of *Helicobacter pylori* plays an important role in oxidative stress resistance and host colonization. Infect Immun. 72:1391–1396.
Zhao Y, et al. 2014. PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinformatics 30:1297–1299.

Associate editor: Rachel O'Neill

Appendix 2

A genomic island in *Vibrio cholerae* with VPI-1 site-specific recombination characteristics contains CRISPR-Cas and type VI secretion modules.

SCIENTIFIC REPORTS

OPEN

Received: 18 February 2016
Accepted: 10 October 2016
Published: 15 November 2016

A genomic island in *Vibrio cholerae* with VPI-1 site-specific recombination characteristics contains CRISPR-Cas and type VI secretion modules

Maurizio Labbate[1], Fabini D. Orata[2], Nicola K. Petty[3], Nathasha D. Jayatilleke[1], William L. King[1], Paul C. Kirchberger[2], Chris Allen[1], Gulay Mann[4], Ankur Mutreja[5], Nicholas R. Thomson[5], Yan Boucher[2] & Ian G. Charles[3,†]

Cholera is a devastating diarrhoeal disease caused by certain strains of serogroup O1/O139 *Vibrio cholerae*. Mobile genetic elements such as genomic islands (GIs) have been pivotal in the evolution of O1/O139 *V. cholerae*. Perhaps the most important GI involved in cholera disease is the *V. cholerae* pathogenicity island 1 (VPI-1). This GI contains the toxin-coregulated pilus (TCP) gene cluster that is necessary for colonization of the human intestine as well as being the receptor for infection by the cholera-toxin bearing CTX phage. In this study, we report a GI (designated GI*Vch*S12) from a non-O1/O139 strain of *V. cholerae* that is present in the same chromosomal location as VPI-1, contains an integrase gene with 94% nucleotide and 100% protein identity to the VPI-1 integrase, and attachment (*att*) sites 100% identical to those found in VPI-1. However, instead of TCP and the other accessory genes present in VPI-1, GI*Vch*S12 contains a CRISPR-Cas element and a type VI secretion system (T6SS). GIs similar to GI*Vch*S12 were identified in other *V. cholerae* genomes, also containing CRISPR-Cas elements and/or T6SS's. This study highlights the diversity of GIs circulating in natural *V. cholerae* populations and identifies GIs with VPI-1 recombination characteristics as a propagator of CRISPR-Cas and T6SS modules.

Vibrio cholerae is a species of bacteria commonly found in marine and estuarine waters and the causative agent of the diarrhoeal disease cholera[1]. Although more than 200 serogroups of the bacterium are known, only two, O1 and O139, are responsible for pandemics of the cholera disease[2]. Historically, there have been seven pandemics of cholera, with the current seventh pandemic caused by O1 strains of the El Tor biotype and the sixth and presumably the previous five pandemics caused by O1 strains of the classical biotype[1]. The evolution of pandemic strains has been greatly influenced by lateral gene transfer (LGT), which led to the acquisition of many novel virulence factors[3].

Integrative mobile genetic elements (MGEs) such as transposons and genomic islands (GIs) are particularly important in LGT processes, as they help facilitate integration of non-homologous transferred DNA into replicons such as the chromosome or a resident plasmid(s)[4]. Studies comparing the genomes of *V. cholerae* strains consistently find MGEs as one of the sources, if not the major source, of genome variation[5,6]. For example, the seventh pandemic strains contain two GIs called the *Vibrio* seventh pandemic islands I and II not present in the strains isolated from the previous pandemics[7]. Most important to pathogenicity of the pandemic strains are genes found on two MGEs, the cholera toxin bacteriophage (CTXϕ) and a GI called the *V. cholerae* pathogenicity island

[1]University of Technology Sydney, School of Life Sciences, Sydney, 2007, Australia. [2]University of Alberta, Department of Biological Sciences, Edmonton, T6G 2E9, Canada. [3]University of Technology Sydney, The ithree institute, Sydney, 2007, Australia. [4]Defence Science and Technology Group, Melbourne, 3207, Australia. [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, United Kingdom. [†]Present address: Institute of Food Research, Norwich Research Park, Norwich, NR4 7UA, United Kingdom. Correspondence and requests for materials should be addressed to M.L. (email: maurizio.labbate@uts.edu.au)

1 (VPI-1)[8,9]. Although these MGEs are common in pandemic strains, examples of CTXϕ and VPI-1 in non-O1/O139 strains have been documented and are indicative of their mobility across natural populations of *V. cholerae* and close relatives[10–12]. The CTXϕ contains genes encoding a potent enterotoxin that, when secreted in the human intestinal tract, results in significant loss of fluid that can lead to death within 24 hours if left untreated[1]. VPI-1 contains the genes encoding the toxin-coregulated pilus (TCP; thus, VPI-1 is also known as the TCP pathogenicity island) that is required for adhesion to the intestinal wall as well as genes encoding the accessory colonization factor and the regulatory genes *toxT* and *tcpPH*[13–15]. TCP is also the receptor for CTXϕ and thus a prerequisite for infection and subsequent lysogeny in the emergence of toxigenic *V. cholerae*[8].

Another virulence factor associated with frequent LGT is the type VI secretion system (T6SS) of *V. cholerae* and other Gram-negative bacteria. Bacteria that harbor T6SS produce a membrane-spanning protein complex capable of puncturing eukaryotic or prokaryotic cells and injecting toxic effector proteins into their targets[16]. In *V. cholerae*, the presence of T6SS has previously only been reported as a conserved chromosomal element and displays toxic activity against macrophages[17]. The T6SS is made up of a series of proteins encoded by genes in three different locations of the *V. cholerae* genome: a main cluster on chromosome 2 and two smaller auxiliary clusters on chromosomes 1 and 2[17,18]. The characteristic proteins of the T6SS are Hcp and VgrG. Hcp is encoded by two alleles on the *V. cholerae* chromosome and polymerizes to form the nanotube that protrudes from the bacterial cell surface[19,20]. The tip of Hcp is decorated with VgrG proteins that form a trimer[21]. VgrG proteins are conserved at their N-termini but carry specialist C-termini with enzymatic activity[21]. For example, *V. cholerae* contains three VgrG alleles with two encoding proteins with actin cross-linking (VgrG-1) and peptidoglycan degrading (VgrG-3) activity[22,23]. Additional effector proteins with diverse enzymatic activities can be added to the VgrG spike for delivery into target cells[24].

Acquisition of novel genes through the uptake of MGEs can have obvious beneficial effects for harmless and pathogenic bacteria alike. However, not all LGT is advantageous, and bacteria have evolved a variety of methods to prevent the spread of harmful DNA sequences in their genomes[25]. One recently discovered defense mechanism against unwanted LGT is the CRISPR-Cas system. CRISPR-Cas modules consist of *cas* genes and an array of short direct repeats separated by highly variable spacer sequences that correspond to genetic elements such as bacteriophages or MGEs[26]. Transcription of the CRISPR array with subsequent slicing of the transcript into smaller CRISPR RNAs acts in concert with the Cas proteins to specifically recognize foreign DNA and cleave it[26]. The system acts as an adaptive immune system for bacteria as it allows for the synthesis and incorporation of new spacers into the array following invasion of a foreign DNA molecule, thus providing immunity to the host cell.

Non-O1/O139 strains of *V. cholerae* have been hypothesized to be an important source of new MGEs that could relocate into pandemic strains. Here, we report a GI with VPI-1 recombination characteristics that harbors both a CRISPR-Cas module and an auxiliary T6SS locus in a non-O1/O139 strain of *V. cholerae* from Sydney, Australia. This GI likely provides recipient cells not only with a defense mechanism against maladaptive LGT, but also with a potential competitive advantage over bacteria lacking this GI and perhaps a novel virulence factor. We also show that similar GIs are present in other non-O1/O139 strains from around the globe.

Methods

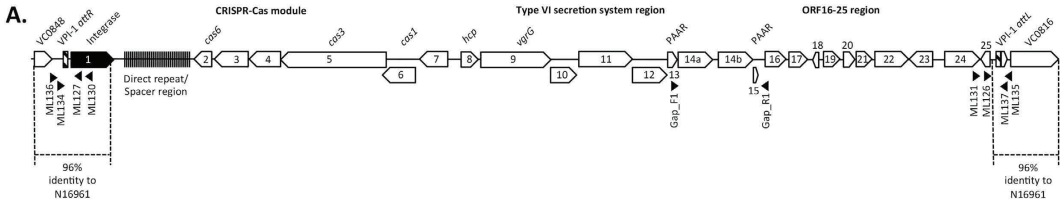
Bacterial strain, growth conditions, and molecular biology methods. The non-O1/O139 *V. cholerae* S12 strain was isolated from the Georges River (Sydney, Australia) in 2009[27] and routinely cultured on Luria-Bertani medium at 37 °C. The whole genome of *V. cholerae* S12 was paired-end sequenced at the Wellcome Trust Sanger Institute (Hinxton, UK) using Illumina HiSeq 2000 (San Diego, CA, USA) from DNA extracted using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI, USA). For extraction of GI*Vch*S12 circles, a plasmid extraction was carried out on S12 using the PureYield Plasmid Miniprep Kit (Promega). PCR was performed using 2X MangoMix (Bioline, London, UK) that consists of DNA polymerase, dNTPs, Mg^{2+} and an orange dye with 30 cycles of denaturation at 93 °C for 30 sec, the appropriate annealing temperature for 30 sec and an extension of 72 °C (1 min/kb). All primers were acquired from Integrated DNA Technologies (Coralville, IA, USA) (Table 1) and used at a final concentration of 10 μM. PCR amplicons intended for sequencing were excised from 1% agarose gels and purified using the Wizard SV Gel and PCR Clean-Up Kit (Promega) and sequenced using the Sanger method at Macrogen (Seoul, South Korea).

Bioinformatic analyses. The Illumina HiSeq whole genome sequencing reads for *V. cholerae* S12 were filtered to remove low quality reads with average read quality less than Q20 and low quality trailing ends with base quality less than Q20 using Prinseq-lite v0.20.4[28]. Reads were then *de novo* assembled into contiguous sequences (contigs) using Velvet v1.2.10[29] and the assemblies were improved by scaffolding using SSPACE v2.0[30], gap filling using GapFiller v1.10[31], reordering of contigs against the *V. cholerae* N16961 reference genome using Mauve v2.4.0[32] and removal of contigs shorter than 300 bp. The final improved draft genome assembled into 83 scaffolds from 4,624,354 read pairs with an average read length of 75 bp to give a total genome size of 4,061,577 bp with average depth of coverage of 171 reads.

The genomic region encoding GI*Vch*S12 was identified on contig 000009 by pairwise comparison of the *V. cholerae* S12 draft genome to the complete reference genome of *V. cholerae* N16961 using the program Mauve v2.4.0[32]. Analysis of GI*Vch*S12 on contig 000009 identified three assembly gaps between ORFs 13 and 15 due to a putative repeat of ORF14 (annotated as ORF14a and ORF14b). To confirm this repeat, a PCR with primers Gap_F1 and Gap_R1 that anneal outside of this repeat region (see Fig. 1) resulted in an expected ~2.5 kb product (as opposed to a ~1.3–1.5 kb product if only one copy of ORF14 was present). Two of the assembly gaps were closed through sequencing the ends of the amplicon creating the GI*Vch*S12 sequence in GenBank file KU722393. GI*Vch*S12 was annotated using Prokka[33] and the automated annotation was manually curated with the aid of BLAST against the non-redundant NCBI protein database (Bethesda, MD, USA) using BLASTP

Primer	Sequence (5′-3′)	Target
ML126	ACTTCTCGAAAGCGGATCAA	*attL* end of GI*Vch*S12
ML127	AAGCCATCACCATCGAAAAG	*attR* end of GI*Vch*S12
ML130	GCTACCTTTGGCTTCAATCG	*attR* end of GI*Vch*S12
ML131	TGGCAACAAGATGACTTTATCG	*attL* end of GI*Vch*S12
ML134	TCCCTAGCTTCCGCTTGTAA	Between VC0848 and *attR* of GI*Vch*S12
ML135	TCAGTGATGCAGGTTGTTCA	Within VC0816
ML136	GGGAATTTTGCAGTCTGAGG	Between VC0848 and *attR* of GI*Vch*S12
ML137	ATAGGGAGTGGGGCGTTAAT	Within VC0816
Gap_F1	GCGTTTTTATCAATGGCAAACC	Within ORF 13 of GI*Vch*S12
Gap_R1	ACACAGGGCTACCTCTAGATGG	Within and just past ORF 15 of GI*Vch*S12

Table 1. Primers used in this study.



B.

	Protein identification and other relevant information		Protein identification and other relevant information
VC0848	SmpB; SsrA-binding protein. 99% nucleotide identity to VC0848 in N16961.	ORF 14a	Uncharacterised protein. Contains a signal peptide.
ORF 1	Phage family integrase.	ORF 14b	Uncharacterised protein. Contains a signal peptide.
ORF 2	CRISPR-associated protein Cas6/Csy4.	ORF 15	Zn-binding Pro-Ala-Ala-Arg (PAAR) domain protein.
ORF 3	Hypothetical protein.	ORF 16	Hypothetical protein. Contains transmembrane regions.
ORF 4	Hypothetical protein.	ORF 17	Hypothetical protein. Contains transmembrane regions.
ORF 5	CRISPR-associated protein Cas3/Cas2.	ORF 18	Transcriptional regulatory protein, MerR superfamily.
ORF 6	CRISPR-associated protein Cas1.	ORF 19	Uncharacterised protein with RadC domain.
ORF 7	Transcriptional regulatory protein.	ORF 20	Uncharacterised protein with DUF2787 domain.
ORF 8	Type VI secretion system protein, Hcp1 family.	ORF 21	Uncharacterised protein with DUF2787 domain.
ORF 9	Type VI secretion system protein, Vgr family.	ORF 22	GTPase protein.
ORF 10	Uncharacterised protein with a DUF4123 domain.	ORF 23	Uncharacterised protein with DUF3296 domain.
ORF 11	Hypothetical protein.	ORF 24	Hypothetical protein.
ORF 12	Uncharacterised protein. Contains a signal peptide and a glycosyl-hydrolase family 26 domain.	ORF 25	Transcriptional regulatory protein, MerR superfamily.
ORF 13	Zn-binding Pro-Ala-Ala-Arg (PAAR) domain protein.	VC0816	Uncharacterised protein. 96% nucleotide identity to VC0816 in N16961.

Figure 1. The ~28-kb genomic island, GI*Vch*S12, from *Vibrio cholerae* S12 containing a CRISPR-Cas module and type VI secretion auxiliary locus. Schematic representation of GI*Vch*S12 is shown in (**A**) with the VPI-1 *att* sites given as hatched boxes and the VPI-1 integrase as a black block arrow. Regions of nucleotide identity to VPI-1 and surrounding regions in *V. cholerae* N16961 are shown. All ORFs and their orientation are shown as block arrows. Numbers shown in, above, or below the block arrows correlate to the putative identification shown in (**B**). Primer binding sites are shown by black triangles with their direction of extension indicated by the direction of the triangle.

providing putative identification[34]. The CRISPR-Cas module in GI*Vch*S12 was identified using the online tool CRISPRFinder (http://crispr.i2bc.paris-saclay.fr)[35].

In order to determine if GI*Vch*S12 or similar islands were present in other *V. cholerae*, additional *V. cholerae* genomes were obtained from GenBank. The list of genomes used is provided in Supplementary Material. The genomes were annotated with RAST v2.0 (Rapid Annotation Using Subsystem Technology)[36]. The GI*Vch*S12 ORFs were compared against the ORFs of the *V. cholerae* genomes to determine presence/absence by calculating the BLAST score ratio (BSR)[37]. Significant hits were considered as putative homologues if the BSR values were at least 0.3 (for 30% amino acid identity)[38]. Furthermore, the whole genomes were aligned using Mugsy v1.2.3[39], and the core alignment (2,539,853 bp in total length) was extracted from the Mugsy output using Galaxy v16.04[40] and Geneious v8.1.8[41]. From this alignment, a maximum likelihood phylogenetic tree was constructed using RAxML v8.1.17[42] using the general time reversible (GTR) nucleotide substitution model and gamma distribution pattern. Robustness of branching was estimated with 100 bootstrap replicates. *Vibrio metoecus*, the closest relative of *V. cholerae,* was used as an outgroup[43].

GenBank accession numbers. The full sequence of the GI*Vch*S12 including flanking sequences and the sequenced *attP* and *attB* sites are available in GenBank/ENA/DDBJ and have the accession numbers KU722393, KU722394, and KU722395, respectively. The raw Illumina HiSeq sequencing reads are available under accession number ERR063652 and the improved draft genome assembly of *V. cholerae* S12 can be accessed at MDST00000000.

Results and Discussion

A novel variant of the *Vibrio* pathogenicity island. In order to identify regions of interest in the genome of *V. cholerae* S12, contigs were compared to the closed genome of *V. cholerae* N16961. A novel GI in the same respective location as VPI-1 (between VC0816 and VC0848 on chromosome 1) was identified on contig 000009. This GI of ~28-kb has been given the designation GI*Vch*S12 and contains an integrase with 94% nucleotide and 100% amino acid identity to the VPI-1 integrase gene and protein, respectively, and characteristic VPI-1 *attL* and *attR* sites abutting the GI[44]. Given GI*Vch*S12's location, with identical integrase protein and *att* sites, this GI is likely to have recombination functions identical to VPI-1. Previous studies have observed variations in VPI-1 through PCR analysis of *V. cholerae* strains or BLASTN analysis of *V. cholerae* genomes for VPI-1 genes. For the most part, the variations identified represent minor gene gain/loss events or sequence changes to known ORFs in VPI-1[10–12]. GI*Vch*S12 is different in that it shares practically no gene content with VPI-1.

A CRISPR-Cas module for self-preservation. Bioinformatic analysis of GI*Vch*S12 identified a CRISPR-Cas module and a T6SS auxiliary locus at the *attR* and *attL* ends, respectively (Fig. 1). The CRISPR-Cas module in GI*Vch*S12 contains genes encoding homologues of Cas1, Cas3, and Cas6. Based on the protein sequences and their organization, this CRISPR-Cas system is most likely similar to those of type 1-F[26]. Several spacers displayed 100% identity to various bacteriophage genomes consistent with the module having a role in acting against invading foreign DNA. The association of a CRISPR-Cas module within a GI is intriguing for two reasons. From an ecological perspective, the mobilization of a CRISPR-Cas system benefits the host not only with an adaptive immune system but also by the instant addition of the immunity that comes with the various spacer sequences it already carries. Thus, a host would immediately gain protection from various bacteriophages and other invading foreign DNA within that ecosystem. Secondly, an intriguing study identified a CRISPR-Cas system within a bacteriophage genome that was able to counteract an inhibitory GI present in the bacterial host genome, thus improving the bacteriophages' capacity to successfully infect the bacterial host[45]. As a result, this raises the possibility that the CRISPR-Cas system within GI*Vch*S12 might improve integration efficiency in recipient cells that contain other genetic elements interfering with the GI's integrity and/or integration. Furthermore, once integrated into the host, the CRISPR-Cas system could prevent the replacement of GI*Vch*S12 by VPI-1 or other GIs competing for the same integration sites. The CRISPR-Cas system found on GI*Vch*S12 could therefore promote direct self-preservation or self-preservation by protecting its host.

A novel T6SS auxiliary locus. Also present on GI*Vch*S12 are genes normally associated with three T6SS loci found in all *V. cholerae* genomes, known as the main locus and auxiliary loci 1 and 2. The GI*Vch*S12 locus structurally resembles the two T6SS auxiliary loci, with the presence of an *hcp* gene, a copy of *vgrG,* a gene encoding a protein with a DUF4123 domain, and putative cargo effectors and immunity proteins further downstream. The lack of proteins making up the T6SS machinery indicates that the proteins on this additional auxiliary locus are dependent on the structural proteins encoded on the main chromosomal T6SS locus for effective translocation into target cells that in S12 is present on contig 00022. The auxiliary loci 1 and 2 are present on contigs 00011 and 00021, respectively, although in contig 00021 the sequence breaks before *hcp*, presumably due to the difficulty of assembling repeat regions. At the end of contig 00011, the first 392-bp of *hcp* from auxiliary locus 1 is present before the sequence breaks. In the small contigs of 00079 and 00082 are the first 275-bp and last 108-bp of an *hcp* homologue, respectively, with *hcp* from contig 00079 presumably from auxiliary locus 2. The auxiliary 1 and 2 *hcp* genes in S12 share 99% nucleotide identity with those in *V. cholerae* N16961 and V52. The *hcp* from GI*Vch*S12 is clearly different to both the auxiliary 1 and 2 *hcp* loci sharing 88% nucleotide identity.

The VgrG protein encoded in GI*Vch*S12 is dissimilar to the chromosomal VgrG proteins but, like VgrG-2, lacks a C-terminal effector domain[46]. The DUF4123 domain found in the protein encoded downstream of the *vgrG* gene indicates a function as an accessory loading proteins like *tap*-1 (VC1417 in the auxiliary 1 locus of *V. cholerae* V52), which is responsible for the loading of cargo effectors with antibacterial activity onto VgrG proteins[47,48]. Due to the structural similarity of this locus with chromosomal auxiliary loci, it is likely that the hypothetical protein encoded by ORF 11 (Fig. 1) represents such a cargo effector. Antibacterial T6SS effectors are always accompanied by immunity proteins that provide protection against self-intoxication, making it likely that the homologous proteins encoded by ORFs 12, 14a and 14b act as such. ORF 12 is 61% and 62% identical at the nucleotide level to ORFs 14a and 14b, respectively. ORFs 14a and 14b share 89% nucleotide identity. Expression of effectors from the GI*Vch*S12 T6SS locus is likely to increase the range of T6SS-mediated microbial toxicity as evidenced by the divergent VgrG protein and one or more putative other effectors encoded on GI*Vch*S12. It is therefore likely that this increase in effector repertoire gives cells harboring GI*Vch*S12 an advantage over other cells in T6SS-mediated competition. A series of other genes encoding hypothetical proteins are in close proximity and may have a role in either forming the T6SS apparatus or act as effector proteins. For example, ORFs 13 and 15 (also homologues of each other) encode Zn-binding Pro-Ala-Ala-Arg (PAAR) proteins. Zn-binding PAAR proteins form a conical extension on the VgrG tip and also function to attach effector proteins to the spike[24]. More research is required to gain insight on how *V. cholerae* S12 regulates expression of the different Hcp and effector proteins and to determine the enzymatic activity of such effector proteins.

A successful and globally distributed genomic island. In order to determine whether GIs similar to GI*Vch*S12 were present in the VPI-1 site of other *V. cholerae* genomes, we compared GI*Vch*S12 to 28 other *V. cholerae* strains for which genome sequences are available in public databases. The non-O1/O139 strain from Haiti (2012Env-2) contained a complete GI*Vch*S12-like island with similar CRISPR-Cas and T6SS modules and a complete set of ORFs 16–25 (Fig. 2). RC385 has an almost complete GI*Vch*S12 but lacks ORF14b. A further three GI*Vch*S12-like islands had minor variations. The non-O1/O139 strain 2012Env-32 from Haiti lacks ORF24 and

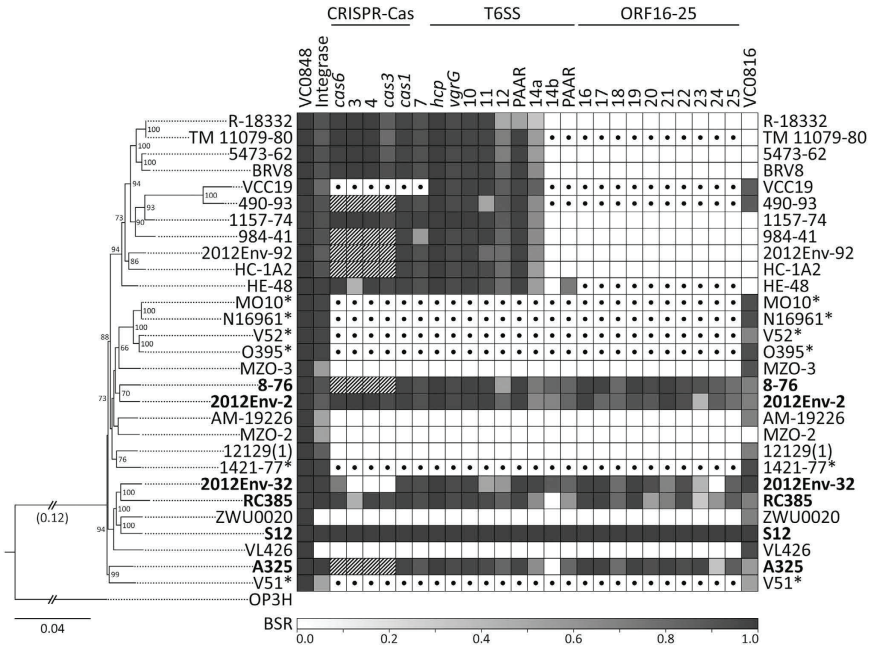


Figure 2. Presence/absence of the GI*Vch*S12 ORFs in various *Vibrio cholerae* strains. The heat map shows the BLAST score ratio (BSR) against the GI*Vch*S12 reference (each column is an ORF). The gradient bar shows the BSR values and their corresponding colours; white indicates the absence of the ORF. Only BSR values of at least 0.3 were included. Strains with similar or complete GI*Vch*S12 are indicated in bold. Striped boxes indicate the presence of a CRISPR-Cas module different from GI*Vch*S12; dotted boxes indicate the presence of other ORFs in those regions different from GI*Vch*S12. Strains with the VPI-1 similar to N16961 are indicated with *. The phylogenetic relationship of the *V. cholerae* strains is shown on the left of the heat map with *Vibrio metoecus* OP3H as outgroup. The maximum likelihood phylogenetic tree was constructed from the core alignment of whole genomes (≈2.5 mb) and using the GTR gamma substitution model with 100 bootstrap replicates (indicated on the tree nodes). The scale bar represents nucleotide substitutions per site. Shortened branch lengths, approximately three times the scale bar (0.12), are indicated.

three other ORFs in the CRISPR-Cas module and the non-O1/O139 strain 8–76 and O1 strain A325 from India and Argentina vary in their CRISPR-Cas modules. A325 also lacks ORF14b. Fifteen other GIs were identified containing a CRISPR-Cas module and/or a T6SS, although for many (in the upper most clade; see Fig. 2), we were unable to locate a contig containing VC0816 so it is unclear if the GI continues beyond ORF14a. One GI (in strain VCC19) harboured a T6SS but contained other genes instead of the CRISPR-Cas module and ORFs 16–25. Two other GIs in strains 490–93 and HE-48 contained a divergent CRISPR-Cas modules and other genes instead of ORFs16-25. Given that all these strains have been isolated from different geographic locations (Supplementary Material) including Europe, South America, North America, and Asia, this data indicates that GIs with VPI-1 recombination characteristics are active in disseminating CRISPR-Cas and T6SS modules across the globe.

GI*Vch*S12 and associated islands can be divided into sub-clusters or islets, as has been previously done for *V. cholerae* pathogenicity island 2 (VPI-2)[12], consisting of the CRISPR-Cas cluster, the T6SS cluster, and the ORFs 16–25 cluster. Given that multiple GIs (Fig. 2) contain different genes in the ORF 16–25 region and in the case of VCC19, in the CRISPR-Cas region, other islets have clearly been acquired by relatives of GI*Vch*S12. We hypothesize that evolution of GI*Vch*S12 and its relatives proceeded through acquisition of these sub-clusters through homologous recombination processes as supported by the observation that the first 293 bp of the GI*Vch*S12 integrase shares 100% identity to the VPI-1 integrase before dropping to 90% for the remainder of the gene.

A genomic island that readily excises as a circle. To confirm that the GI could excise as a circle, a nested PCR strategy was employed using primers annealing within the GI and facing outward toward the *attL* and *attR* ends. First, primers ML130 and ML131 were used in a PCR reaction with template derived from a plasmid preparation of *V. cholerae* S12. Next, 1 μl from the ML130/ML131 PCR was used as template for a fresh PCR reaction employing primers ML126 and ML127 (relative primer binding sites are shown in Fig. 1A) giving an expected fragment of ~580-bp (Fig. 3). As expected, sequence of the PCR product showed that excision occurs at the *att* sites abutting the GI producing an *attP* site identical to what is observed when VPI-1 excises from the genome[44].

PCR of the empty chromosomal site was also conducted using a nested PCR strategy with primers annealing outside the GI and facing in toward the *attL* and *attR* ends. Primers ML134 and ML135 were used in a PCR reaction with template derived from a PCR reaction with ML136 and ML137 using genomic DNA as template. The resulting product (Fig. 3) was sequenced and showed precise excision of GI*Vch*S12 leaving an identical *attB* site as

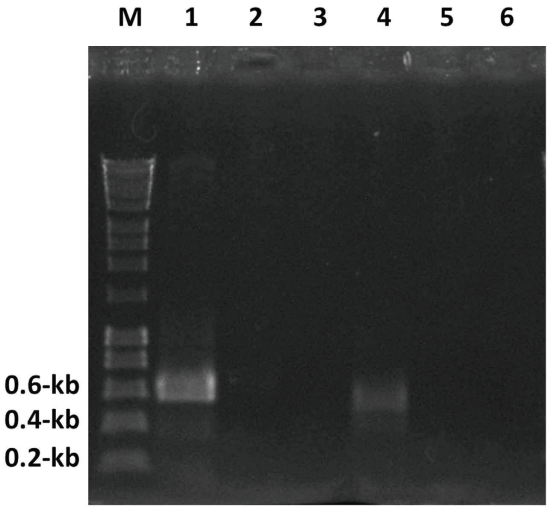


Figure 3. Cropped DNA agarose gel showing the amplicon of a two-stage nested inverse PCR of the excised GI*Vch*S12 circle (lane 1) and the amplicon of a two-stage PCR of the "empty" chromosomal GI*Vch*S12 site (lane 4). Controls for each PCR are given in the adjacent lanes that include nested PCR of the negative control from the first-stage PCR and dH_2O negative control (lanes 2 and 3 for GI*Vch*S12 circle PCR and lanes 5 and 6 for "empty" chromosomal GI*Vch*S12 site PCR) respectively. Lane M is the DNA marker with relevant band sizes shown on the left of the gel.

previously seen with excision of VPI-1[44]. VPI-1 uses both a phage-like integrase and a transposase protein called VpiT to facilitate excision of the GI[44]. Genes encoding these proteins are present within VPI-1.

However, in some *V. cholerae* pandemic strains, *vpiT* is in a different location[44]. VpiT or a homologue of VpiT was not found in GI*Vch*S12 or in the S12 genome.

In conclusion, we report an interesting GI with VPI-1 recombination characteristics housing a CRISPR-Cas element and a T6SS auxiliary locus. This GI is likely to provide its host with a competitive advantage by protecting from bacteriophages as well as adding T6SS-associated bactericidal effectors proteins. Furthermore, this study shows that GIs with VPI-1 recombination characteristics carrying CRISPR-Cas and T6SS modules are circulating in natural *V. cholerae* populations globally.

References

1. Kaper, J., Morris, J. Jr & Levine, M. Cholera. *Clin Microbiol Rev* **8,** 48–86 (1995).
2. Nelson, E. J., Harris, J. B., Glenn Morris, J., Calderwood, S. B. & Camilli, A. Cholera transmission: the host, pathogen and bacteriophage dynamic. *Nat Rev Micro* **7,** 693–702 (2009).
3. Faruque, S. M. & Mekalanos, J. J. Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends in Microbiol* **11,** 505–510 (2003).
4. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2,** 414–424 (2004).
5. Chun, J. *et al.* Comparative genomics reveals mechanisms for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc Natl Acad Sci USA* **106,** 15442–15447 (2009).
6. Grim, C. J. *et al.* Genome sequence of hybrid *Vibrio cholerae* O1 MJ-1236, B-33, and CIRS101 and comparative genomics with *V. cholerae*. *J Bacteriol* **192,** 3524–3533 (2010).
7. Dziejman, M. *et al.* Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proc Natl Acad Sci USA* **99,** 1556–1561 (2002).
8. Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272,** 1910–1914 (1996).
9. Karaolis, D. K. R. *et al.* A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. *Proc Natl Acad Sci USA* **95,** 3134–3139 (1998).
10. Hasan, N. A. *et al.* Nontoxigenic *Vibrio cholerae* non-O1/O139 isolate from a case of human gastroenteritis in the U.S. Gulf Coast. *J Clin Microbiol* **53,** 9–13 (2015).
11. Li, M., Kotetishvili, M., Chen, Y. & Sozhamannan, S. Comparative genomic analyses of the *Vibrio* Pathogencity Island and cholera toxin prophage regions in nonepidemic serogroup strains of *Vibrio cholerae*. *Appl Environ Microbiol* **69,** 1728–1738 (2003).
12. Orata, F. D. *et al.* The dynamics of genetic interactions between *Vibrio metoecus* and *Vibrio cholerae*, two close relatives co-occurring in the environment. *Genome Biol Evol* **7,** 2941–2954 (2015).
13. Herrington, D. A. *et al.* Toxin, toxin-coregulated pili, and the *toxR* regulon are essential for *Vibrio cholerae* pathogenesis in humans. *J Exp Med* **168** (1988).
14. Taylor, R. K., Miller, V. L., Furlong, D. B. & Mekalanos, J. J. Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proc Natl Acad Sci USA* **84,** 2833–2837 (1987).
15. Kovach, M. E., Shaffer, M. D. & Peterson, K. M. A putative integrase gene defines the distal end of a large cluster of ToxR-regulated colonization genes in *Vibrio cholerae*. *Microbiology* **142,** 2165–2174 (1996).
16. Russell, A. B., Peterson, S. B. & Mougous, J. D. Type VI secretion system effectors: poisons with a purpose. *Nat Rev Microbiol* **12,** 137–148 (2014).
17. Pukatzki, S. *et al.* Identification of a conserved bacteria protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proc Natl Acad Sci USA* **103,** 1528–1533 (2006).
18. Das, S. & Chaudhuri, K. Identification of a unique IAHP (IcmF associated homologous proteins cluster in *Vibrio cholerae* and other proteobacteria through in silico analysis. *In Silico Biol* **3,** 287–300 (2003).

19. Williams, S. G., Varcoe, L. T., Attridge, S. R. & Manning, P. A. *Vibrio cholerae* Hcp, a secreted protein coregulated with HlyA. *Infect Immun* **64,** 283–289 (1996).
20. Ballister, E. R., Lai, A. H., Zuckermann, R. N., Cheng, Y. & Mougous, J. D. *In vitro* self-assembly of tailborable nanotubes from a simple protein building block. *Proc Natl Acad Sci USA* **105,** 3733–3738 (2008).
21. Silverman, J. M., Brunet, Y. R., Cascales, E. & Mougous, J. D. Structure and regulation of the type VI secretion system. *Ann Rev Microbiol* **66,** 453–472 (2012).
22. Brooks, T. M., Unterweger, D., Bachmann, V., Kostiuk, B. & Pukatzki, S. Lytic activity of the *Vibrio cholerae* Type VI secretion toxin VgrG-3 is inhibited by the antitoxin TsaB. *J Biol Chem* **288,** 7618–7625 (2013).
23. Dong, T. G., Ho, B. T., Yoder-Himes, D. R. & Mekalanos, J. J. Identification of T6SS-dependent effector and immunity proteins by Tn-seq in *Vibrio cholerae*. *Proc Natl Acad Sci USA* **110,** 2623–2628 (2013).
24. Shneider, M. M. *et al.* PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. *Nature* **500,** 350–353 (2013).
25. Croucher, N. J. *et al.* Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol* **14,** e1002394, doi: 10.1371/journal.pbio.1002394 (2016).
26. Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nat Rev Microbiol* **9,** 467–477 (2011).
27. Islam, A. *et al.* Indigenous *Vibrio cholerae* strains from a non-endemic region are pathogenic. *Open Biol* **3,** 120181 (2013).
28. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* (2011).
29. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short reads assembly using de Bruijn graphs. *Genome Res* **18,** 821–829 (2008).
30. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).
31. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol* (2012).
32. Darling, A. E., Mau, B. & Perna, N. T. progressiveMAUVE: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5,** e11147 (2010).
33. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30** 2068–2069 (2014).
34. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215,** 403–410 (1990).
35. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucl Acids Res* **35,** W52–W57 (2007).
36. Aziz, R. K. *et al.* The RAST server: rapid annotation using subsystems technology. *BMC Genomics* **9,** 75 (2008).
37. Rasko, D. A., Myers, G. S. A. & Ravel, J. Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics* **6,** 1–7, doi: 10.1186/1471-2105-6-2 (2005).
38. Rost, B. Twilight zone of protein sequence alignments. *Protein Engineering* **12,** 85–94 (1999).
39. Angiuoli, S. V. & Salzberg, S. L. Mugsy: Fast multiple alignment of closely related whole genomes. *Bioinformatics* **27,** 334–342 (2010).
40. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11,** 1–13 (2010).
41. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–1649 (2012).
42. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
43. Kirchberger, P. C. *et al. Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. *Int J Syst Evol Microbiol* **64,** 3208–3214 (2014).
44. Rajanna, C. *et al.* The *Vibrio* pathogenicity island of epidemic *Vibrio cholerae* forms precise extrachromosomal circular excision products. *J Bacteriol* **185,** 6893–6901 (2003).
45. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494,** 489–491 (2013).
46. Pukatzki, S., Ma, A. T., Revel, A. T., Sturtevant, D. & Mekalanos, J. J. Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. *Proc Natl Acad Sci USA* **104,** 15508–15513 (2007).
47. Unterweger, D. *et al.* Chimeric adaptor proteins translocate diverse type VI secretion system effectors in *Vibrio cholerae*. *EMBO* **34,** 2198–2210 (2015).
48. Liang, X. *et al.* Identification of divergent type VI secretion effectors using a conserved chaperone domain. *Proc Natl Acad Sci USA* **112,** 9106–9111 (2015).

Acknowledgements

NRT is supported by the Wellcome Trust, grant #098051 to Wellcome Trust Sanger Institute.

Author Contributions

M.L. conceived the experiments and wrote the manuscript. A.M. and N.R.T. sequenced the *V. cholerae* S12 genome. N.D.J., W.L.K., and C.A. conducted the experiments. M.L., F.D.O., N.K.P., P.C.K., and Y.B. performed bioinformatic analyses. G.M. and I.G.C. helped supervise the project. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Labbate, M. *et al.* A genomic island in *Vibrio cholerae* with VPI-1 site-specific recombination characteristics contains CRISPR-Cas and type VI secretion modules. *Sci. Rep.* **6**, 36891; doi: 10.1038/srep36891 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.