Taxonomy of bacteria in the genomic era

by

Fabini Delmo Orata

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

Doctor of Philosophy

in

Microbiology and Biotechnology

Department of Biological Sciences
University of Alberta

ABSTRACT

Identification of bacterial isolates is important for taxonomic purposes, as well as to predict behaviour and properties of organisms in evolutionary, ecological, industrial, or medical contexts. This process embraces a polyphasic approach (phenotypic, genotypic, and phylogenetic), but has been ultimately based on comparisons of 16S rRNA gene sequences and genomic DNA–DNA hybridization (DDH). It was only in the past decade that modern taxonomy has begun the process of incorporating whole-genome sequencing (WGS) data, thanks to significant advancements in next-generation sequencing (NGS) technologies, to perform taxonomic assignments of bacteria. As such, genome-based pairwise similarity indexes were introduced as part of the taxonomic polyphasic approach. This includes multilocus sequence analysis of the core genome, *in silico* (digital) DDH (dDDH) and average nucleotide identity (ANI) for species delineations, and average amino acid identity (AAI) and the percentage of conserved proteins (POCP) for delineations within higher taxonomic ranks.

For this thesis, WGS was employed in characterizing a novel species of *Vibrio*, *Vibrio cidicii*, which was only distinguishable from its closest known relative, *Vibrio navarrensis*, in one phenotypic test. The former, but not the latter, was able to utilize L-rhamnose as a sole carbon source. Pairwise ANI was calculated from the genomes of four isolates of each species, which ranged from 95.4–95.8% (same-species cutoff for ANI of 95–96%). Although ANI showed borderline results, pairwise dDDH results were more conclusive and ranged from 61.9–64.3% (same-species cutoff for DDH is 70%). This was further supported by phylogenetic analysis of the core genome, which placed the *V. cidicii* isolates into their own monophyletic clade, distinct from *V. navarrensis*. Whereas dDDH and ANI are becoming more commonly

used in species characterizations and specifically demonstrated here to be useful in the characterization of a novel species, higher taxonomic ranks are still mostly circumscribed based on 16S rRNA gene sequence phylogeny.

The use of 16S rRNA gene sequences was predominant in identifying organisms within the family *Rhodobacteraceae*, with the first isolates discovered in the early 1990s, the heyday of sequencing for this marker gene. Within this family is a monophyletic group of organisms called the roseobacter clade, plagued with inconsistencies in nomenclature. In another study described in this thesis, pairwise AAI and POCP were calculated from 290 high-quality whole-genome sequences with the aim of properly delineating genera within this family. Whereas the POCP metric did not show distinct delineations for genus, a cutoff for genus (70%) is proposed based on AAI. Using this cutoff, strains were reclassified to remove polyphyly and paraphyly in some genera. Additionally, we were able to determine 60% AAI as the newly proposed cutoff for family, reclassifying some members into a new family with the proposed name *Stappiaceae* fam. nov.

WGS is useful beyond identification purposes. In a third study described in this thesis, the extent of interspecies horizontal gene transfer (HGT) was determined between two very closely related organisms, *Vibrio cholerae* and *Vibrio metoecus*, to demonstrate the role HGT plays in diversification and speciation. *Vibrio metoecus* is a recently described species that showed high 16S rRNA gene sequence identity with *V. cholerae* (98%), the causative agent of the potent diarrheal disease cholera. It was previously thought to be an atypical variant of *V. cholerae*. However, genome comparisons resulted in only 86–87% ANI between species, indicating that *V. metoecus* is a different species. Both species were co-isolated in a brackish coastal pond in the United States East Coast, which suggests that they are likely in constant

interaction with each other. This presented an opportunity to study interspecies gene exchange in natural populations. Comparative genomic analysis showed a bias in interspecies recombination, where *V. metoecus* was a recipient of up to three times more genes from *V. cholerae* as it was the donor. Interestingly, it was determined by qPCR that *V. cholerae* is three times more abundant in the environment than *V. metoecus*, and the former was present throughout the summer but the latter was only detectable at the end of the season. This difference in abundance and seasonality could be major contributors in the HGT bias from *V. cholerae* to *V. metoecus*.

WGS has also shown promise in epidemiological investigations and was proven to be a powerful tool in investigating the origin of the cholera outbreak in Haiti. After the devastating 2010 earthquake in Haiti, cholera spread quickly around the country. Initial epidemiological reports linked the outbreak to the United Nations peacekeeping troops deployed to Haiti from Nepal after news reports showed improper sewage treatment in their camp. Traditional epidemiological investigations were enhanced by WGS and phylogenetic analysis and allowed the determination of the exact source of the outbreak. The clinical *V. cholerae* isolates from Haiti showed very high genomic resemblance to clinical isolates from Nepal, with only one to two nucleotide differences in their core genome, supporting the hypothesis that pathogenic *V. cholerae* was imported to Haiti from an external source through human transmission.

The various studies presented herein demonstrate the usefulness of WGS in taxonomy, population genetics, and microbial epidemiology. This is made possible by significant advances in NGS technologies as well as sequence analysis methods.

PREFACE

Some of the research conducted for this thesis are collaborative works, and individual author contributions are listed below for each chapter. Chapters 2, 4, and 6 have been published in peer-reviewed journals and are included in this thesis in practically unaltered forms (except for minor changes and formatting).

1. A version of Chapter 2 has been published as:

"Orata, F.D., Xu, Y., Gladney, L.M., Rishishwar, L., Case, R.J., Boucher, Y., Jordan, I.K., Tarr, C.L. (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(10), 4148-4155."

- F.D.O. and Y.B. designed the study; L.M.G. and C.L.T. provided the *V. cidicii* and *V. navarrensis* isolates and genome sequences; F.D.O. and L.R. performed bioinformatic analyses; Y.X., L.M.G., and R.J.C. performed biochemical tests; F.D.O. and Y.B. wrote the manuscript; and Y.B., I.K.J., and C.L.T. supervised the project.

2. A version of Chapter 3 will be submitted for publication as:

"Orata, F.D., Liang, K., Boucher, Y., and Case, R.J. Whole genome-based taxonomy of *Rhodobacteraceae* and *Stappiaceae* fam. nov. in the order *Rhodobacterales*."

- F.D.O., Y.B., and R.J.C. designed the study and wrote the manuscript; F.D.O. and K.L. performed bioinformatic analyses; and Y.B. and R.J.C. supervised the project.

3. A version of Chapter 4 has been published as:

"Orata, F.D., Kirchberger, P.C., Méheust, R., Barlow, E.J., Tarr, C.L., and 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(10), 2941-2954."

- F.D.O. and Y.B. designed the study and wrote the manuscript; Y.B. performed sampling; C.L.T. provided the clinical *V. metoecus* genome sequences; F.D.O. performed whole-genome sequencing and analyses; P.C.K. performed phylogenetic analysis; R.M. quantified interspecies gene transfers; E.J.B. provided multiple scripts used in data analysis; and Y.B. supervised the project.

4. A version of Chapter 5 will be submitted for publication as:

"Orata, F.D., Liang, K., Nasreen, T., Hussain, N.A.S., and Boucher, Y. Differences in abundance and seasonal patterns of *Vibrio cholerae* and *Vibrio metoecus* lead to a directional bias in interspecies horizontal gene transfer."

- F.D.O. and Y.B. designed the study and wrote the manuscript; Y.B. performed sampling; F.D.O. performed whole-genome sequencing and assembly; F.D.O. and K.L. performed bioinformatic analyses; T.N. performed qPCR; F.D.O. and N.A.S.H. performed the natural transformation assay; and Y.B. supervised the project.

5. A version of Chapter 6 has been published as:

"Orata, F.D., Keim, P.S., and Boucher, Y. (2014). The 2010 cholera outbreak in Haiti: how science solved a controversy. *PLoS Pathogens* 10(4), e1003967."

- F.D.O. and Y.B. wrote the manuscript, with revisions from P.S.K.; and F.D.O. created the figures.

Beyond the research conducted for this thesis, I was also involved in other collaborative works that led to several publications. Appendix A provides a complete list of authored and co-authored works in the course of my doctorate program.

"I look at the term **species**, as one arbitrarily given for the sake of convenience to a set of individuals closely resembling each other, and that it does not essentially differ from the term variety, which is given to less distinct and more fluctuating forms. The term variety, again, in comparison with mere individual differences, is also applied arbitrarily, and for mere convenience sake."

– Charles Darwin*

*Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* London, UK: John Murray.

ACKNOWLEDGMENTS

This accomplishment would not have been possible without the help and support of all my colleagues, family, and friends.

Firstly, I would like to thank my supervisor, Dr. Yan Boucher, for taking a chance on me despite my little-to-no background in bioinformatics prior to joining your lab. You have been a great supervisor and mentor, and I could not have asked for a better supervisor. Thank you for providing a conducive environment for me to learn and grow as a scientist, not to mention the free food and coffee. You are one of the reasons why graduate school was such a wonderful experience for me.

I would also like to thank Dr. Rebecca Case, who in many ways became my unofficial co-supervisor. I enjoyed collaborating with you on several projects. Thank you for all the valuable work and life advice you have given me.

I thank my supervisory committee, Dr. David Wishart, Dr. Felix Sperling, and Dr. Stefan Pukatzki for all your great insights and suggested improvements for my thesis. Additionally, I am grateful for all the collaborative works that helped shape this thesis and served as significant learning opportunities, namely those from Dr. Cheryl Tarr and Lori Gladney (Centers for Disease Control and Prevention), Dr. I. King Jordan and Dr. Lavanya Rishishwar (Georgia Institute of Technology), Dr. Raphaël Méheust (Université Pierre et Marie Curie), and Dr. Paul Keim (Northern Arizona University).

A big thank you goes out to my wonderful labmates, Dr. Paul Kirchberger, Tania Nasreen, Tareq Islam, Jed Barlow, Nora Hussain, and Kevin Liang for all your significant contribution to my thesis. I apologize for my very loud singing in the lab. To Paul: I started this

journey with you, and you have become one of my closest friends. I appreciate our friendly competition that constantly pushed me to work harder. Also, thank you to Dr. Leen Labeeuw, Dr. Anna Bramucci, Yue Xu, and Teaghan Mayers for all your help. I consider all of you not just colleagues but lifelong friends. I hope that our paths cross again soon.

Finally, thank you to my family for your unconditional support. I am sure you are very relieved to know that I am finally done with school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction – Application of Whole-Genome Sequencing to Prokaryotic Identification and Taxonomy

---

CHAPTER 1

1.1. The phylo-phenetic species concept for prokaryotes

A species is the basic unit of biological classification and taxonomic rank. The 1942 species definition by Ernst Mayr, used in most modern textbooks, is that a species constitutes of "groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups" (Mayr, 1942; De Queiroz, 2005). This "biological species concept" implies that two members of the same species can produce fertile offspring, typically by sexual reproduction (Mayr, 1942; Bernstein et al., 1985). And while this simple definition fits well for most multicellular organisms, this does not apply to all organisms, such as those that reproduce asexually, as in most unicellular organisms and parthenogenetic or apomictic multicellular organisms (De Meeûs et al., 2007). In addition, horizontal gene transfer (HGT) adds a whole level of complexity to the issue, since organisms such as prokaryotes are capable of acquiring genetic material that is not originally from their parent (De la Cruz and Davies, 2000; Soucy et al., 2015). HGT also occurs in eukaryotes (Andersson, 2005; Keeling and Palmer, 2008) and between prokaryotes and eukaryotes (Dunning Hotopp, 2011). Thus, the issue of what actually constitutes a species remains controversial, as it is difficult to develop a definition of a species that applies to all organisms. This is called the "species problem" (De Queiroz, 2005; 2007; Fraser et al., 2009; Ereshefsky, 2010).

The prokaryotic species concept evolved in parallel to the design of laboratory techniques and bioinformatic tools that permitted the retrieval of useful information (Rosselló-Móra and Amann, 2001; Cohan, 2002). It is circumscribed based on three major points: a demonstration of genomic boundaries of the taxon (Mallet, 1995), the exhaustive phenotypic description of the

taxon (Tindall et al., 2010), and the recognition of the monophyletic nature of members of the taxon (Wayne et al., 1987; Baum, 2008). This polyphasic approach to taxonomy (Vandamme et al., 1996) is discussed in more detail below. This "phylo-phenetic concept" defines a prokaryotic species as a "monophyletic and genomically coherent cluster of individual organisms that show a high degree of overall similarity with respect to many independent characteristics, and is diagnosable by a discriminative phenotypic property" (Rosselló-Móra and Amann, 2001; Gevers et al., 2005). This definition is generally agreed among microbiologists to be useful, pragmatic, and universally applicable among prokaryotes. Furthermore, it covers the primary goals of taxonomy such as rapid and reliable identification of strains.

1.2. The polyphasic approach in microbial identification

The characterization and identification of a strain is a key element in taxonomy. However, prokaryotic taxonomy is largely a matter of scientific judgment since there is no "official" classification of *Bacteria* and *Archaea* (Sneath and Brenner, 1992; Garrity, 2016). The system described in the *Bergey's Manual of Systematics of Archaea and Bacteria* (or the *Bergey's Manual of Systematic Bacteriology*) is the most widely accepted classification system by microbiologists (Whitman, 2015). The *Bergey's Manual* contains a compendium of systematic information useful for identification purposes and is the most complete and authoritative description of bacterial and archaeal diversity (Krieg and Garrity, 2005; Whitman, 2015).

When a new microorganism is isolated from nature and is thought to be unique, a comprehensive description of the organism is pursued, which should sufficiently demonstrate that the isolate is distinct from already described taxa. Descriptions should use the polyphasic

approach in describing microorganisms (Vandamme et al., 1996) to categorize an organism based on similarities with closely related taxa. This includes phenotypic analysis, which examines the morphological, metabolic, physiological, and chemical characteristics. Meanwhile, genotypic analysis considers characteristics of the genome and phylogenetic analysis seeks to place organisms within an evolutionary framework through phylogenetic reconstruction (Tindall et al., 2010), where more closely related organisms should cluster together in a phylogenetic tree than more distantly related ones (Baum, 2008).

1.2.1. Phenotypic approaches

The phenotype refers to observable characteristics, traits that can be used to differentiate one species from another. Usual phenotypic characteristics taken into account, such as colony morphology, cell morphology, biochemical tests, optimum pH and temperature, etc., are too simplistic and do not reflect the true scope of an organism phenotype (Tindall et al., 2010). Phenotypic tests should be as comprehensive as possible (Table 1.1). Instead of doing tests individually, various biochemical tests are available as a single kit, such as the Analytical Profile Index (API) kit from bioMérieux. The API kit includes strips that contain 20 miniature biochemical tests (Holmes et al., 1978; Shayegani et al., 1978; Overman et al., 1985). There are also high-throughput identification assays available, such as the Microbial Identification System from Biolog. This system can test for up to 96 different characteristics in one 96-well plate, with 20 different plates available from their catalogue representing various tests (e.g., carbon, nitrogen, phosphorus, and sulfur utilization; biosynthetic pathway and nutrient stimulation; osmotic, ionic, and pH responses; and sensitivity to chemicals) for a total of 1,920 tests (Bochner et al., 2001; Bochner, 2003; 2009; Shea et al., 2012). Both kits are relatively quick and easy to

perform, as they mainly involve the inoculation of strips/wells with an overnight culture of the unknown organism plus a period of incubation. On the other hand, the growth requirement means that these kits may not be applicable to all organisms that do not grow under optimized temperatures for these kits to work correctly. Individually, phenotypic characteristics may not be sufficient to determine genetic relatedness, yet as a whole, they provide descriptive information enabling the recognition of taxa (Rosselló-Móra and Amann, 2001).

Table 1.1. Some phenotypic characteristics determined in prokaryotic identification.

The list is adapted from Tindall et al. (2010) and Madigan et al. (2014).

Category	Characteristics
Morphology	Colony morphology (size, shape, edge, opacity, color, fast or slow spreading); cell morphology (size, shape, arrangement); Gram reaction; presence and location of flagella; pattern of flagellation; presence stalks or appendages; spore formation; fruiting-body formation; presenc of intracellular structures (gas vacuoles, sulphur granules, or polyhydroxybutyrate granules)
Motility	Rapid, slow, or nonmotile; swimming, gliding, or swarming motility
Metabolism	Fermentation of sugars; utilization of carbon, nitrogen, or sulfur compounds; nitrogen fixation; growth factors
Physiology	Range of temperature, pH, and salinity for growth; response to presence/absence of oxygen (aerobic, anaerobic, facultative); presence of catalase or oxidase; production of extracellular enzymes
Cell lipid chemistry	Respiratory lipoquinones; fatty acid composition; ether-linked lipids (isoprenoid- and non-isoprenoid based); hydrophobic side chains of lipids; polar lipids; extracellular constitutents (lipopolysaccharide, mycolic acid)
Cell wall chemistry	Peptidoglycan type; diamino acid composition of crosslinks; mode of crosslinkage; complete amino acid composition
Others	Pigments; luminescence; antibiotic sensitivity; serotype; production of secondary metabolites

1.2.2. Genotypic and phylogenetic approaches

Modern prokaryotic taxonomy has been strongly influenced by developments in genotypic methods. The pragmatic species concept for prokaryotes is ultimately based on these traits, similarity based on 16S rRNA sequences and genomic similarity based on DNA–DNA hybridization (DDH; Stackebrandt and Goebel, 1994). These are considered most important for genomic identification of species, although other traits are also examined. The G+C content of

DNA was recognized as having a direct link to codon usage (De Ley, 1968) and is still a useful parameter to classify prokaryotes (Tindall et al., 2010). A more comprehensive list of genotypic tests is discussed by Tindall et al. (2010) in their guidelines on a proper species description.

1.2.2.1. 16S rRNA sequencing

One of the most prevalent gene-based techniques for identification of unknown organisms in the species and higher levels is the comparison of 16S rRNA gene sequences with known organisms. The use of the gene is justified, as it is universally present in all prokaryotes and it is highly conserved (Kitahara and Miyazaki, 2013). Evidence has shown that two organisms with 16S rRNA genes sharing at least 97% sequence identity based on full-length sequences belong to the same species (Martinez-Murcia and Collins, 1990; Collins et al., 1991; Amann et al., 1992; Fox et al., 1992; Martinez-Murcia et al., 1992), but more stringent thresholds of 98.2–99% have been suggested (Stackebrandt and Ebers, 2006; Meier-Kolthoff et al., 2013). Those that typically share above 95% sequence identity are suggested to belong to the same genus (Yarza et al., 2008; Tindall et al., 2010). However, the use of the 16S rRNA gene is also debatable since it provides limited phylogenetic resolution in the species level and poor discriminatory power for some genera due to its high sequence conservation (Bosshard et al., 2006; Mignard and Flandrois, 2006; Janda and Abbott, 2007). Some examples would be comparisons within the genus *Vibrio* (Thompson et al., 2005; Gladney and Tarr, 2014) and *Bacillus* (Janda and Abbott, 2007; Wang and Ash, 2015). Also, the 16S rRNA gene can exist in various copy numbers of up to 15 or more copies (Klappenbach et al., 2001; Acinas et al., 2004) and can be horizontally transferred (Schouls et al., 2003; Acinas et al., 2004; Dewhirst et al., 2005; Kitahara and Miyazaki, 2013). Thus, it is important to note that the use of 16S rRNA gene

sequences alone should not describe a species, but may provide the first indication that a novel species has been isolated (Tindall et al., 2010). There is justification for using 23S rRNA gene sequences (Ludwig and Schleifer, 1994; Bavykin et al., 2004; Dewhirst et al., 2005; Konstantinidis and Tiedje, 2005b; Hunt et al., 2006), although the dataset is currently much smaller for this gene than the 16S rRNA gene. It is highly recommended to perform other tests such as nucleic acid hybridization.

1.2.2.2. DNA–DNA hybridization

Nucleic acid hybridization techniques (DNA–DNA and DNA–RNA) have allowed for the indirect comparison of genomes of two organisms and the measurement of their genomic similarity (Madigan et al., 2014). These techniques were introduced into prokaryotic systematics in the 1960s (McCarthy and Bolton, 1963; Brenner et al., 1967; Johnson and Ordal, 1968). 16S rRNA gene sequencing eventually replaced DNA–RNA hybridization (Tindall et al., 2010), although DDH is still widely used and is considered the "gold standard" for circumscribing a bacterial species (Figure 1.1; Wayne et al., 1987; Tindall et al., 2010). The major advantage of DDH is that it often produces more sharply defined clusters of strains than those solely circumscribed by phenotypic traits (Krieg, 1988; Rosselló-Móra and Amann, 2001). A DDH value of at least 70% is recommended as the suitable threshold for species (Brenner, 1973; Johnson, 1973; Wayne et al., 1987), which corresponds to at least 97% 16S rRNA sequence identity (Figure 1.2; Stackebrandt and Goebel, 1994). Also, DDH provides absolute resolution where the 16S rRNA gene does not. A good example would be the comparison between the type strains of *Bacillus globisprus* and *Bacillus psychrophilus* which share > 99.5% 16S rRNA sequence identity, yet exhibit only 50% DDH (Fox et al., 1992). Three type strains of

Edwardsiella species exhibit 99.4–99.8% 16S rRNA sequence identity but only 28–50% DDH to each other (Janda and Abbott, 2007).

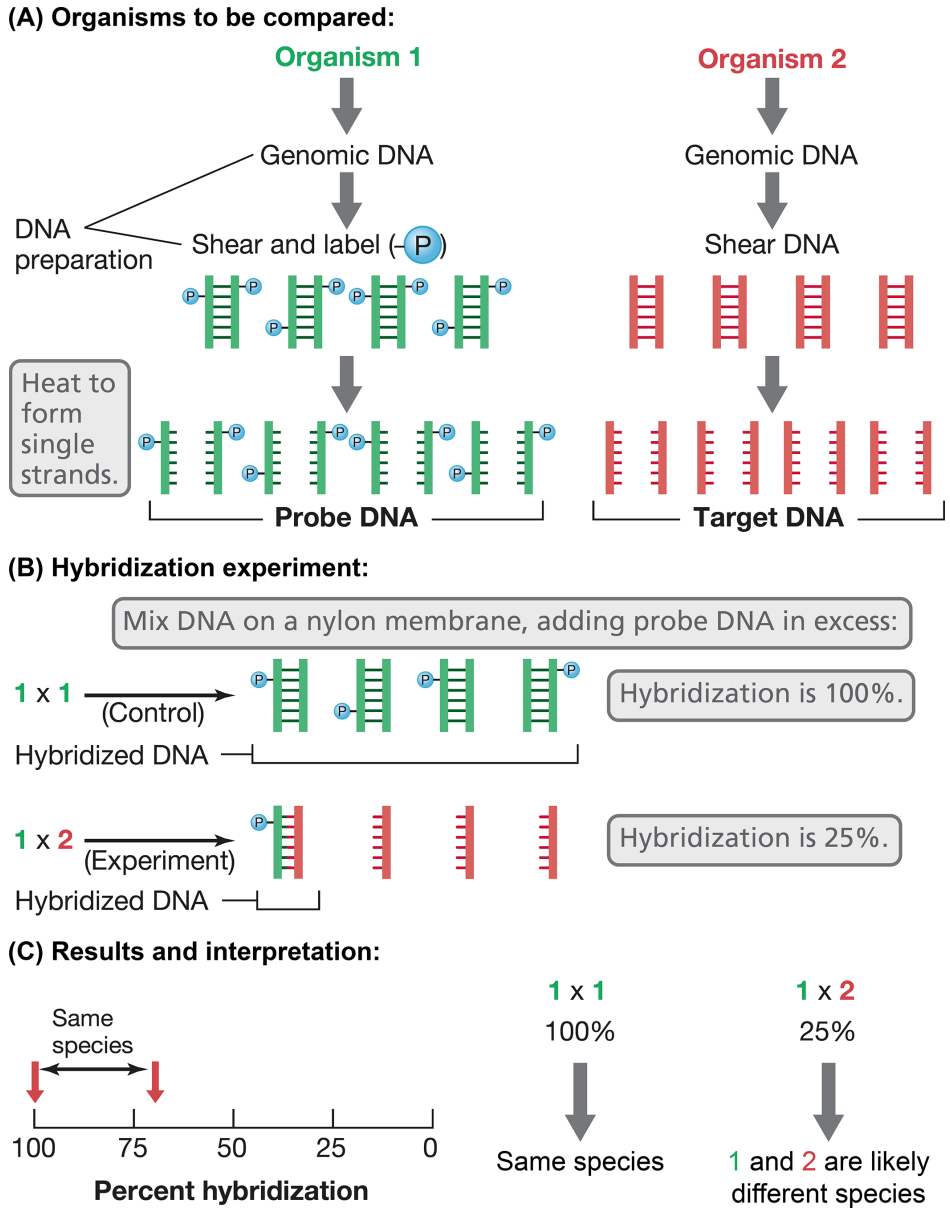

Figure 1.1. The general DDH experiment.

(A) Genomic DNA is isolated from the two organisms to be compared. They are then sheared into small fragments and denatured to produce single strands. DNA fragments from organism 1 are labeled (shown here as a radioactive phosphate) and will serve as probe. (B) Sheared single-stranded DNA fragments from each organism are immobilized on a membrane and then hybridized with the labeled probe from organism 1. Radioactivity in the hybridized DNA is measured and compared with the control (organism 1 DNA hybridizing with itself). (C) Radioactivity in the control is taken as 100% DDH and radioactivity with organism 2 is compared against the control. Figure adapted from Madigan et al. (2014).

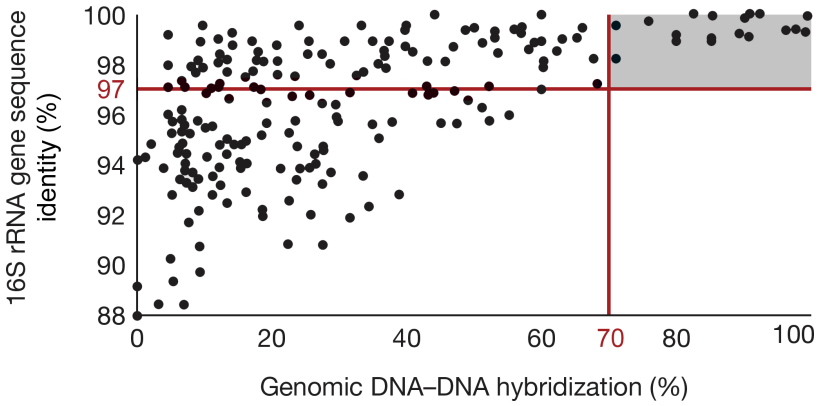

Figure 1.2. Relationship between 16S rRNA gene sequence identity and genomic DDH.

The graph shows pairwise comparisons of microorganisms on the basis of their 16S rRNA sequence identity plotted against DDH values. Two organisms are likely to belong to the same species if they exhibit at least 97% 16S rRNA sequence identity and 70% DDH (highlighted in gray). Figure adapted from Stackebrandt and Ebers (2006), Rosselló-Móra and Amann (2001), and Madigan et al. (2014).

1.3. Use of whole-genome sequencing in taxonomy

The significant decline in price of DNA sequencing and the advances in next-generation sequencing (NGS) technologies have provided a huge dataset of whole-genome sequences. Genomes provide valuable insights into microbial physiology and evolution and can be utilized in many ways in the characterization of strains. A greater selection of genes other than 16S rRNA that provide different degrees of resolution have become available with whole-genome sequences, such as single-copy, protein-coding housekeeping genes (Zeigler, 2003; Santos and Ochman, 2004). Unlike the 16S rRNA gene, protein-coding genes have a better resolution power because of the reduced selection at the third codon position (Koonin and Novozhilov, 2017). Core housekeeping genes commonly used in phylogenetic analysis include *recA* (Eisen, 1995; Thompson et al., 2005), *gyrB* (Bavykin et al., 2004; Wang et al., 2007) and *rpoB* (Tarr et al., 2007), but hundreds to thousands of core genes, genes that are present in all members of a taxon, can be used (Segata and Huttenhower, 2011; Vernikos et al., 2015). The use of these genes could complement or resolve issues from 16S rRNA gene sequence analysis. Common methods where

these genes are used include multilocus sequence typing of five to seven genes, used for intraspecies discrimination (Maiden et al., 1998), and multilocus sequence analysis (MLSA), more commonly used to determine relationships between organisms at the genus or family level (Stackebrandt et al., 2002; Gevers et al., 2005; Glaeser and Kämpfer, 2015). In MLSA, the individual gene sets are first aligned and then concatenated to reconstruct a phylogeny that provides better resolution than the 16S rRNA phylogeny. Genomes can also be compared to determine gene content (presence or absence) and synteny, the order of genes in the genome. Also, the determination of G+C content has become trivial with the availability of whole-genome sequences (Meier-Kolthoff et al., 2014). Lastly, there are several indexes that are obtained by pairwise comparisons of genomes, which are discussed in detail below. All these provide significant insights into genomic relationships between strains.

1.3.1. General principles of next-generation sequencing

Beginning in 2005, the traditional Sanger DNA sequencing method (Sanger et al., 1977; Swerdlow and Gesteland, 1990; Swerdlow et al., 1990; Hunkapiller et al., 1991) has experienced revolutionary changes, leading to the era of NGS (Reuter et al., 2015; Goodwin et al., 2016; Mardis, 2017). From sequencing a single to few hundred fragments (Figure 1.3A), NGS technologies have evolved into massively parallel sequencing of millions of DNA fragments of whole genomes. This is done through a cyclic-array strategy (Figure 1.3B), which involves the sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection (Mitra and Church, 1999). The general workflow for NGS is as follows: first, library preparation is accomplished by random fragmentation of DNA followed by ligation of adapters to both ends of these fragments. These adapters are universal sequences that

can be used to polymerase-amplify the fragments. Second, the fragments are immobilized on a surface, either on a bead or a flat glass microfluidic channel covered with adapter sequences complementary to those on the fragments. Third, upon immobilization of fragments, amplification can then be achieved using several approaches, including *in situ* PCR colonies (polonies; Mitra and Church, 1999), emulsion PCR (Dressman et al., 2003), or bridge PCR (Adessi et al., 2000; Fedurco et al., 2006). This results in amplified fragments spatially clustered on the surface. Lastly, the sequencing process itself consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition.

The cyclic-array strategy offers several advantages. Whereas Sanger sequencing requires the cloning of a fragment into a vector (i.e., transformation of *Escherichia coli* and colony picking), NGS does not and is therefore not bottlenecked by this process, allowing for massive parallelism. Also, because fragments are immobilized on a surface, they can be enzymatically manipulated by a single reagent volume, translating into lower overall costs for sequencing.

Several reviews have tackled in detail the differences between specific NGS technologies (Shendure and Ji, 2008; Kircher and Kelso, 2010; Metzker, 2010; Glenn, 2011; Shendure et al., 2011; Mardis, 2013; Reuter et al., 2015; Goodwin et al., 2016; Mardis, 2017), and these will not be discussed in detail here. A recent review presents in great detail all currently available sequencing technologies and their specifications (Table 1 of Goodwin et al., 2016). Here, I will only highlight the two sequencing technologies that were used in my work: the Illumina and Pacific Biosciences platforms.

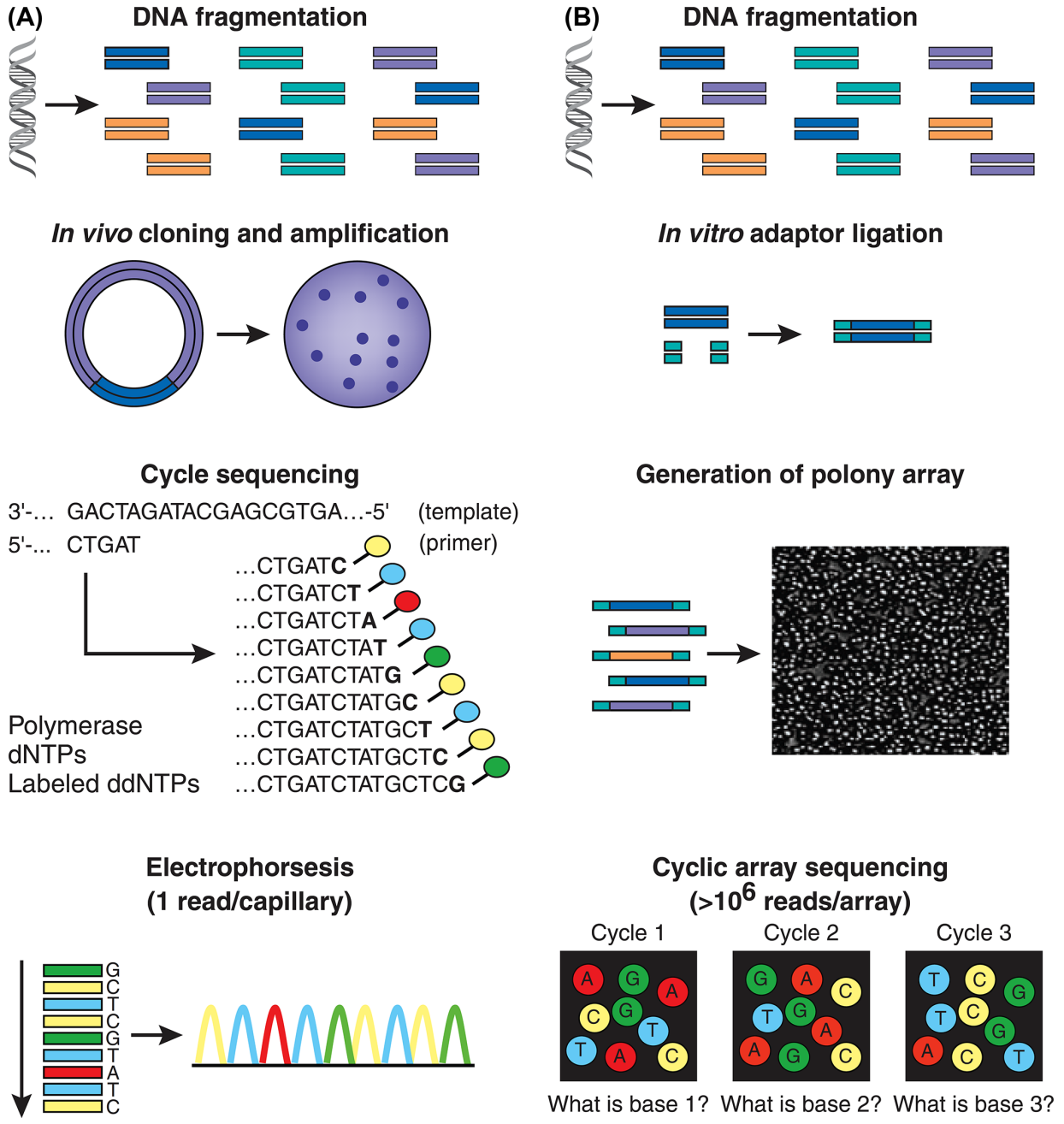

Figure 1.3. Work flow of conventional Sanger sequencing versus NGS.

(A) In Sanger sequencing, genomic DNA is fragmented, cloned into a plasmid vector, and used to transform *E. coli*. For each sequencing reaction, a single bacterial colony is picked and the plasmid DNA is isolated. During sequencing, a ladder of ddNTP-terminated, dye-labeled products is generated and subjected to high-resolution electrophoretic separation within a capillary. A sequence is generated as fluorescently labeled fragments of discrete sizes pass a detector. (B) In NGS with cyclic-array methods, genomic DNA is fragmented and ligated with adapters, which is then subjected to amplification resulting in an array of millions of spatially immobilized PCR colonies (polonies). Each polony contains many copies of a single DNA fragment. Since polonies are tethered to a surface, the polonies can be manipulated in parallel during sequencing using a single reagent volume. Similarly, imaging-based detection during each extension cycle can also be done in parallel. Successive iterative cycles of enzymatic manipulation and imaging-based detection are used to build up a contiguous sequencing read for each array feature. Figure obtained from Shendure and Ji (2008).

1.3.1.1. Illumina sequencing

The Illumina platform has its origins in work by Turcatti et al. (Fedurco et al., 2006; Turcatti et al., 2008) and arguably the most popular platform because of its simplicity of producing the library and amplifying DNA fragments. Its error rate of just 0.1% per read makes it an attractive platform over everything else (Goodwin et al., 2016). Briefly, genomic DNA is fragmented into several hundred base pairs in length, and adaptors are ligated to the ends of each fragment. Additional motifs are added into the adapters by reduced cycle amplification, including primer binding sites, unique indexes, and regions complementary to short strand oligonucleotides (oligos) on the flow cell. Fragments are then allowed to bind into the flow cell oligos to immobilize them, and they are amplified by bridge PCR (Adessi et al., 2000; Fedurco et al., 2006). Each fragment eventually forms a cluster of about 1,000 copies, and hundreds of millions of clusters are simultaneously produced throughout the flow cell. Sequencing is initiated with primers that anneal into the amplified fragments, and DNA polymerase and a mixture of four deoxynucleotides, each labelled with a different fluorophore, are pumped through the flow cell. After the addition of a nucleotide, the cluster is excited by a light source, which emits a specific fluorescent signal depending on the type of nucleotide that was previously incorporated, a process called sequencing by synthesis (Chen, 2014). All clusters are read simultaneously in a massively parallel process. After the completion of the first read, the indexes are sequenced in the same manner. Sequencing by synthesis is repeated for the complementary reverse strand in paired-end sequencing. The pooled sequences are sorted based on their unique indexes, and forward and reverse reads are paired during assembly creating contiguous sequences.

1.3.1.2. Pacific Biosciences sequencing

Pacific Biosciences (PacBio) sequencing uses Single Molecular Real Time (SMRT) technology, harnessing the power of the polymerase by monitoring it in real time as it synthesizes DNA (Eid et al., 2009; Korlach et al., 2010). Sequencing is performed in a SMRT cell, containing 80,000 visualization chambers called zero-mode waveguides (ZMW). A ZMW is a cylindrical "hole" about 70 nm in diameter with a detection volume of only 20 zeptoliters (20×10^{-21} liters). Here, a single DNA template–polymerase complex is immobilized at the bottom of the ZMW. Nucleotides, each labelled with a different fluorophore on the phosphate group, are introduced into the chamber. As the polymerase incorporates the correct nucleotide, it is held within the detection volume of the chamber for tens of milliseconds. During this time, a fluorescent signal is emitted and detected at the bottom of the well. Then, as part of the natural incorporation cycle, the polymerase cleaves the bond holding the fluorophore in place, and the fluorescent signal diffuses out of the detection volume, during which the signal in the ZMW immediately returns to base line until the next nucleotide is incorporated and a new signal is detected. This process occurs in parallel throughout all the 80,000 ZMWs of a SMRT cell. The power of SMRT sequencing lies in its long read lengths, where a run can produce up to 20 kb reads, as opposed to reads of only 75–300 bp from Illumina (Goodwin et al., 2016). Thus, PacBio sequencing is ideal for sequencing a genome to completion. A notable weakness of this technology is the high error rate (11–14% per read; Roberts et al., 2013). This error rate can be reduced by generating circular consensus sequence reads through sufficient sequencing passes of the template, where a coverage of 15 passes yields greater than 99% accuracy (Eid et al., 2009). However, coverage is limited by the life time of the polymerase, resulting to longer templates yielding fewer passes (Detter et al., 2014). Additionally, PacBio sequences may be

complemented with sequences produced from other technologies (e.g., Illumina) to account for sequencing errors when attempting to close a genome.

1.3.2. Genome-based similarity indexes for taxon delineation

Several similarity indexes obtained by pairwise comparisons of whole genomes are now available for use in identification. The two most commonly used metrics are *in silico* DNA–DNA hybridization (Auch et al., 2010; Meier-Kolthoff et al., 2013) and average nucleotide identity (Konstantinidis and Tiedje, 2005a; Goris et al., 2007; Richter and Rosselló-Móra, 2009) for species delineations. For ranks higher than species, however, genome-based metrics are not as utilized, and 16S rRNA phylogeny or sequence identity is still the standard (Krieg and Garrity, 2005). Pairwise comparisons of all protein-coding genes in a genome and calculating average amino acid identity (Konstantinidis and Tiedje, 2005b; Luo et al., 2014) and the percentage of conserved proteins (Qin et al., 2014) have been proposed as methods to determine higher rank classification.

1.3.2.1. *In silico* DNA–DNA hybridization

The availability of whole-genome sequences has now made it possible to simulate DDH *in silico* and replace experimental DDH (Auch et al., 2010; Meier-Kolthoff et al., 2013). *In silico* (digital) DDH (dDDH) avoids the major pitfalls of experimental DDH, since the latter is a tedious process, which may vary depending on specific methods employed by each laboratory, and values obtained can be difficult to reproduce between laboratories (Gevers et al., 2005). dDDH can easily be calculated using the Genome-to-Genome Distance Calculator (GGDC; http://ggdc.dsmz.de/ggdc.php; Auch et al., 2010; Meier-Kolthoff et al., 2013), the only *in silico*

tool currently available for this analysis. GGDC employs the Genome BLAST Distance Phylogeny (GBDP) approach, originally devised to infer phylogenetic distances from whole genomes (Henz et al., 2005). In GGDC, the two genomes being compared are aligned by the Basic Local Alignment Search Tool (BLAST; Altschul et al., 1990) algorithm without the need to artificially cut the genomes into pieces (Meier-Kolthoff et al., 2013). Regions of significant matches called high-scoring segment pairs (HSPs) are considered statistically significant if the associated expected value (*E*-value) is sufficiently low (10^{-2} or less; Meier-Kolthoff et al., 2013). These HSPs would represent regions in the genomes that would hybridize in a wet lab-based DDH. Values are eventually reported as percent DDH, as with experimental DDH values. The major limiting factor of dDDH is the availability of whole-genome sequences, which is not required by traditional DDH, just the genomic DNA extract. The advantage of genome sequence availability is that the sequences can be used for subsequent analyses to confirm identity, as well as further downstream work beyond species descriptions. The DDH standard of at least 70% for strains to be considered to belong to the same species is also used for dDDH (Brenner, 1973; Johnson, 1973; Wayne et al., 1987; Auch et al., 2010; Meier-Kolthoff et al., 2013).

1.3.2.2. Average nucleotide identity

Average nucleotide identity (ANI) employs genome sequences in a manner slightly different from DDH (Figure 1.1). In the method employed by Goris et al. (2007), a genome is artificially cut (genome 1) into 1,020-bp fragments (as is the desired length in traditional DDH) and used to search against the other genome sequence (genome 2) of interest by also using the BLAST algorithm (Altschul et al., 1990) to determine best BLAST hits between genomes (at least 30% sequence identity and 70% coverage). The method is then reversed (i.e., genome 2 is

fragmented and compared with the full sequence of genome 1). The bidirectional best BLAST hits are reported as the average percent identity from all comparisons or ANI (Konstantinidis and Tiedje, 2005a; Goris et al., 2007). This was correlated with traditional DDH, and it was determined that 70% DDH corresponded to 95% ANI for two genomes to belong to the same species (Figure 1.4; Goris et al., 2007). Interestingly, a correlation study of 16S rRNA sequence identity and ANI based on pairwise comparisons of 6,787 whole-genome sequences revealed that the 95% ANI threshold corresponds to 98.65% 16S rRNA sequence identity (Kim et al., 2014a), which falls within the higher threshold for 16S rRNA sequence identity (98.2–99%) proposed previously for species definition (Stackebrandt and Ebers, 2006; Meier-Kolthoff et al., 2013).

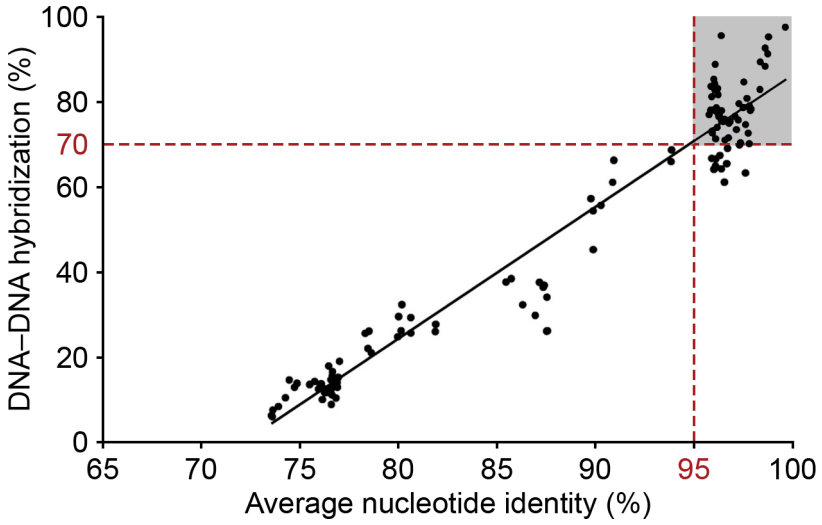

Figure 1.4. Relationship between DDH and ANI.

The graphs show pairwise comparisons of microorganisms on the basis of their experimental DDH values plotted against ANI. Two organisms are likely to belong to the same species if they exhibit at least 70% DDH and 95% ANI (highlighted in gray). Figure adapted from Goris et al., 2007.

Unlike dDDH, several tools are currently available to calculate ANI, such as the standalone tool JSpecies (Richter and Rosselló-Móra, 2009) or online ANI Calculators by Konstantinos Konstantinidis et al. (http://enve-omics.ce.gatech.edu/ani; Rodriguez-R and

Konstantinidis, 2005) and EzBioCloud (https://www.ezbiocloud.net/tools/ani; Yoon et al., 2017). Both dDDH and ANI are not the same metric and use different algorithms to obtain the desired values (Goris et al., 2007; Meier-Kolthoff et al., 2013). dDDH measures gene content because it only considers highly significant BLAST hits (e.g., potential homologous regions) from whole-genome alignments, whereas ANI measures sequence identity from random fragmentation of the genome. Some argue that dDDH is a more significant index than ANI, since the former yielded higher correlations with traditional DDH than the latter (Meier-Kolthoff et al., 2013). However, with access to already available genome sequences, both indexes can be calculated and reported in a species description.

1.3.2.3. Average amino acid identity

Whereas gene- or genome-based delineations for species are well established between organisms (16S rRNA sequence identity, DDH, and ANI), 16S rRNA phylogeny or sequence identity is still the widely used standard for delineations of higher taxonomic ranks (Krieg and Garrity, 2005). A proposed boundary for genus and family based on 16S rRNA sequence identity are 95% and 88%, respectively (Yarza et al., 2008; Tindall et al., 2010).

Whole-genome comparisons in the nucleotide level, such as with ANI, is not suitable for delineation of higher ranks since approximately 90% of sequence data is abandoned (i.e., falls below the best BLAST hit cutoffs; Qin et al., 2014). For higher ranks, comparison with amino acid sequences is more suitable since protein sequences evolve slower than DNA sequences (Drummond et al., 2005; Zhang and Yang, 2015). Average amino acid identity (AAI) is one of the proposed indexes for delineations higher than the species rank. AAI applies the same concept as ANI but using amino acid sequences of shared protein-coding sequences (CDSs). Here, all

CDSs from one genome are compared against the full genome sequence of another genome by TBLASTN (Altschul et al., 1990; Gertz et al., 2006) and vice versa to determine bidirectional best BLAST hits (at least 30% sequence identity over at least 70% coverage) and reported as the AAI (Konstantinidis and Tiedje, 2005b). An AAI boundary for genus would fall in the 60–80% AAI range, according to the pairwise comparisons of 410 genomes performed by Luo et al. (2014). Definite AAI cutoff values for higher ranks are currently not available. In fact, there is a significant overlap in AAI values between ranks (Luo et al., 2014) and, surprisingly, even between domains (Konstantinidis and Tiedje, 2005b), implying a continuum of genetic diversity in prokaryotes. A few tools that are easy to use for AAI calculation are available, such as the online AAI Calculator (http://enve-omics.ce.gatech.edu/aai; Luo et al., 2014) and the standalone tool CompareM (https://github.com/dparks1134/CompareM; Donovan Parks, Australian Centre for Ecogenomics).

1.3.2.4. Percentage of conserved proteins

A more recently proposed metric is by calculating the percentage of conserved proteins (POCP) between organisms (Qin et al., 2014). The number of genes two genomes have in common depends on their evolutionary distance, rather than phenotype (Huynen and Bork, 1998; Snel et al., 1999; Konstantinidis and Tiedje, 2005b). However, the size of the genomes being compared also needs to be taken into account, since larger genomes, despite relatedness, may also have many genes in common (Snel et al., 1999; Qin et al., 2014). The POCP between two genomes is calculated as:

$$\frac{C1 + C2}{T1 + T2} \times 100\%$$

Where C1 and C2 represent the conserved number of proteins between the two genomes being compared, and T1 and T2 represent the total number of proteins in those genomes, respectively (Qin et al., 2014). The proposed genus boundary based on POCP is 50% (Qin et al., 2014).

1.4. Thesis objectives and outline

The overarching objective of this research is the application of whole-genome sequencing (WGS) in bacterial identification, population genetics, and epidemiology in the study of various marine microorganisms. Whole genome-based similarity indexes for species identification, dDDH and ANI, were used in distinguishing *Vibrio cidicii*, a recently described species, from its closest known relatives, *Vibrio navarrensis* and *Vibrio vulnificus*. Additionally, dDDH, ANI, AAI, and POCP were applied in ironing out taxonomic inconsistencies in the family, genus, and even species levels that have plagued a huge and well-studied group of marine microorganisms, the *Rhodobacteraceae*. Furthermore, WGS was used in investigating the role of interspecies HGT between two closely related vibrios, *Vibrio cholerae* and *Vibrio metoecus*, that may have contributed into their diversification and speciation. Lastly, the role of WGS in epidemiology is discussed, specifically in tracking the source of the 2010 cholera outbreak in Haiti.

1.4.1. Characterization of the novel species *Vibrio cidicii* (Chapter 2)

Vibrio cidicii is a novel species of the genus *Vibrio* that I recently described (Chapter 2). Its closest known relative is *V. navarrensis*, with isolates obtained from sewage in Navarra, Spain (Urdaci et al., 1991) and the German Baltic Sea (Jores et al., 2007). Both species are closely related to *V. vulnificus*, responsible for a large proportion of deaths related to seafood consumption with over 50% fatality rates in patients with septicemia (Jones and Oliver, 2009).

On the other hand, *V. navarrensis* was only recently associated with human illness with the isolation of strains from blood (Gladney et al., 2014; Gladney and Tarr, 2014). More isolates were identified through routine characterization of phenotypically similar historical isolates by the Centers for Disease Control and Prevention (CDC, Atlanta, Georgia, USA). However, four of the isolates the CDC characterized clustered separately on a phylogenetic tree after performing MLSA of three housekeeping genes (*pyrH*, *recA*, and *rpoA*; Gladney and Tarr, 2014). Three of these isolates were recovered from human clinical specimens (blood).

I hypothesize that the four isolates potentially belong to a novel species of the genus *Vibrio*. The isolates were further characterized phenotypically to determine more distinguishing characteristics. Using available whole-genome sequences (Gladney et al., 2014), the core genome of *V. navarrensis* and these four isolates were determined, as well as genes unique for each species. Core-genome MLSA was performed, and ANI and dDDH were calculated to further demonstrate that the four isolates belong to a novel species distinct from *V. navarrensis*.

1.4.2. Family, genus, and species delimitations in *Rhodobacteraceae* (Chapter 3)

The family *Rhodobacteraceae* belongs to the order *Rhodobacterales*, class *Alphaproteobacteria*, and phylum *Proteobacteria*, circumscribed on the basis of phylogenetic analysis of 16S rRNA gene sequences (Garrity et al., 2005). Members of this family are key players of biogeochemical cycling (Wagner-Döbler and Biebl, 2006), and comprise up to 20% of bacterial communities in pelagic environments and are often mutualists of eukaryotes (Moran et al., 2007). Most members of the family *Rhodobacteraceae* are assigned to the roseobacter group (Giovanni and Rappé, 2000). Strains within this group have at least 89% 16S rRNA sequence identity and exist as monophyletic in a phylogenetic tree (Buchan et al., 2005; Luo and Moran,

2014). Thus, this group is also referred to as the roseobacter "clade." However, despite the phylogenetic coherence of this clade within *Rhodobacteraceae*, phylogeny within the clade is not. Several genera within the group, such as *Leisingera*, *Phaeobacter*, *Rosoebacter*, and *Ruegeria*, are not monophyletic (Buchan et al., 2005; Wagner-Döbler and Biebl, 2006; Luo and Moran, 2014; Simon et al., 2017). Delimitation standards should be set to resolve this issue, which was the main objective in Chapter 3. Additionally, whereas the majority of previous studies have performed comprehensive phylogenetic analysis of just the rosoebacter clade (Buchan et al., 2005; Wagner-Döbler and Biebl, 2006; Moran et al., 2007; Luo and Moran, 2014), an analysis of the phylogenetic affiliation of this group as part of the family *Rhodobacteraceae* has only been carried out recently (Simon et al., 2017).

The *Bergey's Manual* only lists 25 genera under this family as of its publication in 2005 (Garrity et al., 2005). On the other hand, NCBI Taxonomy currently lists, as of October 5, 2017, a total of 163 identified genera, 110 of which have genome sequences available. This constitutes 254 identified species with available genome sequences. The majority of the strains were classified on the basis of their 16S rRNA gene sequences, and whole-genome phylogeny has only been recently performed but with a smaller dataset (Luo and Moran, 2014; Simon et al., 2017). I employed comparative genomic analysis using all available high-quality *Rhodobacteraceae* genomes (290 in this study) to set genome-based standards for family and genus delineations. Also, some strains were reclassified into their proper species. It is important that the incoherent assignment of scientific names in this family needs to be resolved first even before ecological and evolutionary interpretations, which can be made confusing due to the current taxonomic inconsistencies.

1.4.3. Role of interspecies horizontal gene transfer between *Vibrio cholerae* and *Vibrio metoecus* (Chapter 4 and Chapter 5)

Chapters 4 and 5 present the dynamics of genetic interactions by gene exchange between *V. cholerae* and the recently described *V. metoecus*. *Vibrio cholerae* thrives in brackish estuarine water regions, either in planktonic state or associated with the surfaces of blue-green algae, oysters, crustaceans, fish, or copepods (Colwell, 1996; Boucher et al., 2015). It is the etiological agent of the potent diarrheal disease cholera, responsible for 1.2–4.3 million cases and 28,000–142,000 deaths worldwide every year (Ali et al., 2012). However, disease causing *V. cholerae* is part of just a single genetic lineage, the phylocore genome/pandemic-generating group, from an extremely diverse species harbouring more than 200 serogroups (Chun et al., 2009; Boucher et al., 2015; Boucher, 2016). The majority of environmental isolates are actually non-toxigenic (Faruque and Mekalanos, 2012). The closest known relative of *V. cholerae* is *V. metoecus*, with clinical and environmental strains isolated, so far, from Europe (Italy and Spain) and mostly from the USA (Haley et al., 2010; Boucher et al., 2011; Kirchberger et al., 2014; Vezzulli et al., 2015; Carda Diéguez, 2016).

Vibrio metoecus, unofficially named *Vibrio metecus* (Haley et al., 2010; Boucher et al., 2011), was initially described as an atypical variant of *V. cholerae* (Choopun, 2004; Haley et al., 2010; Boucher et al., 2011; Kirchberger et al., 2014). It appears as yellow *V. cholerae*-like cells on thiosulfate citrate bile salts sucrose agar and shares a high 16S rRNA sequence identity (≥ 98%) with *V. cholerae.* It also resembles *V. cholerae* in the majority of biochemical characteristics, but is negative for acetoin (Voges-Proskauer assay), amylase, and lipase production (Choopun, 2004; Kirchberger et al., 2014). More phenotypic differences between *V.*

cholerae and *V. metoecus* were determined by Kirchberger et al. (2014), where the latter is able to grow using N-acetyl-D-galactosamine or D-glucuronic acid as a sole carbon source.

In their environmental reservoir, bacteria can acquire genetic material from other organisms as a result of HGT. This process plays an important role in the evolution, adaptation, maintenance, and transmission of virulence in bacteria (De la Cruz and Davies, 2000). Boucher et al. (2011) and Kirchberger et al. (2016) co-isolated *V. cholerae* and *V. metoecus* from Oyster Pond (Falmouth, MA, USA), a cholera-free, brackish coastal pond in the United States East Coast. Their co-isolation presents an opportunity to study the dynamics of interspecies interactions of natural populations without the interference of clinical cases. To confirm the identity of these isolates, the genomes of representative strains from both species were sequenced and compared by ANI. Genome sequence data were analyzed to determine the extent of their genetic interactions. I hypothesize that, due to their co-isolation, both species may likely be in constant interaction with each other and exchange a significant region of their genomes through HGT (Chapter 4).

From all 480 strains that were successfully isolated from Oyster Pond, only 41 (8.5%) of them were *V. metoecus* (Kirchberger et al., 2016), which led me to hypothesize that the significant difference in quantity may lead to a significant bias in HGT. To investigate this, more genomes of *V. cholerae* and *V. metoecus* were sequenced to provide a bigger picture of the extent of genetic interactions in the same environment. Also, culture-based isolation may be biased (i.e., towards the isolation of *V. cholerae* in this case), so quantitative PCR, a culture-independent approach to quantify the species from Oyster Pond to confirm this significant difference in abundance (Chapter 5).

1.4.4. Determining the source of the cholera outbreak in Haiti (Chapter 6)

Chapter 6 discusses the application of WGS in epidemiology, specifically in determining the origin of the cholera outbreak that plagued Haiti since 2010. Shortly after the onset of the outbreak, initial news reports broke out about the United Nations troops showing improper sewage waste disposal in their camp (Al Jazeera English, 2010; Katz, 2010), which angered the Haitians and caused riots in the country (BBC News, 2010b; Desvarieux, 2010). It was through WGS that the source of the outbreak was traced back to Nepal, where the troops trained prior to deployment to Haiti. This chapter discuses several independent studies involving WGS and comparative genomics (Chin et al., 2011; Hendriksen et al., 2011; Reimer et al., 2011; Katz et al., 2013; Eppinger et al., 2014) that led to the identification of the closest relatives of the Haitian *V. cholerae* strains (i.e., the Nepalese strains).

1.5. References

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *Journal of Bacteriology* 186(9)**,** 2629-2635.

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P., et al. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research* 28(20)**,** e87.

Al Jazeera English (2010). *UN investigates Haiti outbreak* [Online]. Doha, Qatar: Al Jazeera Satellite Network. Available: http://www.aljazeera.com/news/africa/2010/10/201010284141214196 7.html [Accessed 25 March 2014].

Ali, M., Lopez, A.L., You, Y.A., Kim, Y.E., Sah, B., Maskery, B., et al. (2012). The global burden of cholera. *Bulletin of the World Health Organization* 90(3)**,** 209-218A.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215(3)**,** 403-410.

Amann, R.I., Lin, C., Key, R., Montgomery, L., and Stahl, D.A. (1992). Diversity among *Fibrobacter* isolates: towards a phylogenetic classification. *Systematic and Applied Microbiology* 15(1)**,** 23-31.

Andersson, J.O. (2005). Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences* 62(11)**,** 1182-1197.

Auch, A.F., Klenk, H.P., and Göker, M. (2010). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Standards in Genomic Sciences* 2(1)**,** 142-148.

Baum, D. (2008). Reading a phylogenetic tree: the meaning of monophyletic groups. *Nature Education* 1(1)**,** 190.

Bavykin, S.G., Lysov, Y.P., Zakhariev, V., Kelly, J.J., Jackman, J., Stahl, D.A., et al. (2004). Use of 16S rRNA, 23S rRNA, and *gyrB* gene sequence analysis to determine phylogenetic relationships of *Bacillus cereus* group microorganisms. *Journal of Clinical Microbiology* 42(8)**,** 3711-3730.

BBC News (2010b). *UN appeals for calm after cholera riots in Haiti* [Online]. London, UK: British Broadcasting Corporation. Available: http://www.bbc.co.uk/news/world-latin-america-11772283 [Accessed 25 March 2014].

Bernstein, H., Byerly, H.C., Hopf, F.A., and Michod, R.E. (1985). Sex and the emergence of species. *Journal of Theoretical Biology* 117(4)**,** 665-690.

Bochner, B.R. (2003). New technologies to assess genotype-phenotype relationships. *Nature Reviews Genetics* 4(4)**,** 309-314.

Bochner, B.R. (2009). Global phenotypic characterization of bacteria. *FEMS Microbiology Reviews* 33(1)**,** 191-205.

Bochner, B.R., Gadzinski, P., and Panomitros, E. (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Research* 11(7)**,** 1246-1255.

Bosshard, P.P., Zbinden, R., Abels, S., Boddinghaus, B., Altwegg, M., and Bottger, E.C. (2006). 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. *Journal of Clinical Microbiology* 44(4)**,** 1359-1366.

Boucher, Y. (2016). Sustained local diversity of *Vibrio cholerae* O1 biotypes in a previously cholera-free country. *MBio* 7(3)**,** e00570-16.

Boucher, Y., Cordero, O.X., Takemura, A., Hunt, D.E., Schliep, K., Bapteste, E., et al. (2011). Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *MBio* 2(2)**,** e00335-10.

Boucher, Y., Orata, F.D., and 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. *Frontiers in Microbiology* 6**,** 1120.

Brenner, D.J. (1973). Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria. *International Journal of Systematic Bacteriology* 23(4)**,** 298-307.

Brenner, D.J., Martin, M.A., and Hoyer, B.H. (1967). Deoxyribonucleic acid homologies among some bacteria. *Journal of Bacteriology* 94(2)**,** 486-487.

Buchan, A., González, J.M., and Moran, M.A. (2005). Overview of the marine *Roseobacter* lineage. *Applied and Environmental Microbiology* 71(10)**,** 5665-5677.

Carda Diéguez, M. (2016). *Vibrio vulnificus: from water to host.* Doctor of Philosophy Thesis, University of Valencia.

Chen, C.Y. (2014). DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Frontiers in Microbiology* 5**,** 305.

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., et al. (2011). The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 364(1)**,** 33-42.

Choopun, N. (2004). *The population structure of Vibrio cholerae in Chesapeake Bay.* Doctor of Philosophy Thesis, University of Maryland.

Chun, J., Grim, C.J., Hasan, N.A., Lee, J.H., Choi, S.Y., Haley, B.J., et al. (2009). Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 106(36)**,** 15442-15447.

Cohan, F.M. (2002). What are bacterial species? *Annual Review of Microbiology* 56(1)**,** 457-487.

Collins, M.D., Rodrigues, U., Ash, C., Aguirre, M., Farrow, J.A.E., Martinez-Murcia, A., et al. (1991). Phylogenetic analysis of the genus *Lactobacillus* and related lactic acid bacteria as determined by reverse transcriptase sequencing of 16S rRNA. *FEMS Microbiology Letters* 77(1)**,** 5-12.

Colwell, R.R. (1996). Global climate and infectious disease: the cholera paradigm. *Science* 274(5295)**,** 2025-2031.

De la Cruz, F., and Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in Microbiology* 8(3)**,** 128-133.

De Ley, J. (1968). "Molecular biology and bacterial phylogeny," in *Evolutionary Biology,* eds. T. Dobzhansky, M.K. Hecht & W.C. Steere. (Boston, MA, USA: Springer), 103-156.

De Meeûs, T., Prugnolle, F., and Agnew, P. (2007). Asexual reproduction: genetics and evolutionary aspects. *Cellular and Molecular Life Sciences* 64(11)**,** 1355-1372.

De Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences of the United States of America* 102(Suppl. 1)**,** 6600-6607.

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology* 56(6)**,** 879-886.

Desvarieux, J. (2010). *At the heart of Haiti's cholera riots, anger at the U.N.* [Online]. New York, NY, USA: Time, Inc. Available: http://content.time.com/time/world/article/0,8599,2032437,00.html [Accessed 25 March 2014].

Detter, J.C., Johnson, S.L., Bishop-Lilly, K.A., Chain, P.S., Gibbons, H.S., Minogue, T.D., et al. (2014). "Nucleic acid sequencing for characterizing infectious and/or novel agents in complex samples," in *Biological Identification: DNA Amplification and Sequencing, Optical Sensing, Lab-On-Chip and Portable Systems,* ed. R.P. Schaudies. (Sawston, UK: Woodhead Publishing), 3-53.

Dewhirst, F.E., Shen, Z., Scimeca, M.S., Stokes, L.N., Boumenna, T., Chen, T., et al. (2005). Discordant 16S and 23S rRNA gene phylogenies for the genus *Helicobacter*: implications for phylogenetic inference and systematics. *Journal of Bacteriology* 187(17)**,** 6106-6118.

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., and Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of

genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* 100(15)**,** 8817-8822.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* 102(40)**,** 14338-14343.

Dunning Hotopp, J.C. (2011). Horizontal gene transfer between bacteria and animals. *Trends in Genetics* 27(4)**,** 157-163.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910)**,** 133-138.

Eisen, J.A. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *Journal of Molecular Evolution* 41(6)**,** 1105-1123.

Eppinger, M., Pearson, T., Koenig, S.S., Pearson, O., Hicks, N., Agrawal, S., et al. (2014). Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *MBio* 5(6)**,** e01721-14.

Ereshefsky, M. (2010). Microbiology and the species problem. *Biology & Philosophy* 25(4)**,** 553-568.

Faruque, S.M., and Mekalanos, J.J. (2012). "Epidemiology and molecular biology of *Vibrio cholerae*," in *Foodborne and Waterborne Bacterial Pathogens: Epidemiology, Evolution and Molecular Biology,* ed. S.M. Faruque. (Norfolk, UK: Caister Academic Press), 211-230.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* 34(3)**,** e22.

Fox, G.E., Wisotzkey, J.D., and Jurtshuk, P., Jr. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic Bacteriology* 42(1)**,** 166-170.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323(5915)**,** 741-746.

Garrity, G.M. (2016). A new genomics-driven taxonomy of *Bacteria* and *Archaea*: are we there yet? *Journal of Clinical Microbiology* 54(8)**,** 1956-1963.

Garrity, G.M., Bell, J.A., and Lilburn, T. (2005). "Order III. *Rhodobacterales* ord. nov.," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part C: The Alpha-, Beta-, Delta-, and Epsilonproteobacteria,* eds. D.J. Brenner, N.R. Krieg & J.T. Staley. 2nd ed. (New York, NY, USA: Springer), 161-229.

Gertz, E.M., Yu, Y.K., Agarwala, R., Schaffer, A.A., and Altschul, S.F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology* 4**,** 41.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., et al. (2005). Opinion: re-evaluating prokaryotic species. *Nature Reviews Microbiology* 3(9)**,** 733-739.

Giovanni, S.J., and Rappé, M.S. (2000). "Evolution, diversity, and molecular ecology of marine prokaryotes," in *Microbial Ecology of the Oceans,* ed. D.L. Kichman. (New York, NY, USA: John Wiley & Sons, Inc.), 47-84.

Gladney, L.M., Katz, L.S., Knipe, K.M., Rowe, L.A., Conley, A.B., Rishishwar, L., et al. (2014). Genome sequences of *Vibrio navarrensis*, a potential human pathogen. *Genome Announcements* 2(6)**,** e01188-14.

Gladney, L.M., and Tarr, C.L. (2014). Molecular and phenotypic characterization of *Vibrio navarrensis* isolates associated with human illness. *Journal of Clinical Microbiology* 52(11)**,** 4070-4074.

Glaeser, S.P., and Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology* 38(4)**,** 237-245.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11(5)**,** 759-769.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17(6)**,** 333-351.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57(1)**,** 81-91.

Haley, B.J., Grim, C.J., Hasan, N.A., Choi, S.Y., Chun, J., Brettin, T.S., et al. (2010). Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. *BMC Microbiology* 10**,** 154.

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., et al. (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* 2(4)**,** e00157-11.

Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K., and Schuster, S.C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics* 21(10)**,** 2329-2335.

Holmes, B., Willcox, W.R., and Lapage, S.P. (1978). Identification of *Enterobacteriaceae* by the API 20E system. *Journal of Clinical Pathology* 31(1)**,** 22-30.

Hunkapiller, T., Kaiser, R.J., Koop, B.F., and Hood, L. (1991). Large-scale and automated DNA sequence determination. *Science* 254(5028)**,** 59-67.

Hunt, D.E., Klepac-Ceraj, V., Acinas, S.G., Gautier, C., Bertilsson, S., and Polz, M.F. (2006). Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Applied and Environmental Microbiology* 72(3)**,** 2221-2225.

Huynen, M.A., and Bork, P. (1998). Measuring genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* 95(11)**,** 5849-5856.

Janda, J.M., and Abbott, S.L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology* 45(9)**,** 2761-2764.

Johnson, J.L. (1973). Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. *International Journal of Systematic Bacteriology* 23(4)**,** 308-315.

Johnson, J.L., and Ordal, E.J. (1968). Deoxyribonucleic acid homology in bacterial taxonomy: effect of incubation temperature on reaction specificity. *Journal of Bacteriology* 95(3)**,** 893-900.

Jones, M.K., and Oliver, J.D. (2009). *Vibrio vulnificus*: disease and pathogenesis. *Infection and Immunity* 77(5)**,** 1723-1733.

Jores, J., Appel, B., and Lewin, A. (2007). *Vibrio navarrensis* biotype *pommerensis*: a new biotype of *V. navarrensis* isolated in the German Baltic Sea. *Systematic and Applied Microbiology* 30(1)**,** 27-30.

Katz, J.M. (2010). *UN probes base as source of Haiti cholera outbreak* [Online]. Washington, DC, USA: The Washington Times, LLC. Available: http://www.washingtontimes.com/news/2010/oct/27/un-probes-base-as-source-of-haiti-cholera-outbreak [Accessed 25 March 2014].

Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., et al. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4(4)**,** e00398-13.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9(8)**,** 605-618.

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

Kirchberger, P.C., Orata, F.D., Barlow, E.J., Kauffman, K.M., Case, R.J., Polz, M.F., et al. (2016). A small number of phylogenetically distinct clonal complexes dominate a coastal *Vibrio cholerae* population. *Applied and Environmental Microbiology* 82(18)**,** 5576-5586.

Kirchberger, P.C., Turnsek, M., Hunt, D.E., Haley, B.J., Colwell, R.R., Polz, M.F., et al. (2014). *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. *International Journal of Systematic and Evolutionary Microbiology* 64(9)**,** 3208-3214.

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing – concepts and limitations. *Bioessays* 32(6)**,** 524-536.

Kitahara, K., and Miyazaki, K. (2013). Revisiting bacterial phylogeny: natural and experimental evidence for horizontal gene transfer of 16S rRNA. *Mobile Genetic Elements* 3(1)**,** e24210.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Research* 29(1)**,** 181-184.

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

Konstantinidis, K.T., and Tiedje, J.M. (2005b). Towards a genome-based taxonomy for prokaryotes. *Journal of Bacteriology* 187(18)**,** 6258-6264.

Koonin, E.V., and Novozhilov, A.S. (2017). Origin and evolution of the universal genetic code. *Annual Review of Genetics* 51(1)**,** 45-62.

Korlach, J., Bjornson, K.P., Chaudhuri, B.P., Cicero, R.L., Flusberg, B.A., Gray, J.J., et al. (2010). Real-time DNA sequencing from single polymerase molecules. *Methods in Enzymology* 472**,** 431-455.

Krieg, N.R. (1988). Bacterial classification: an overview. *Canadian Journal of Microbiology* 34(4)**,** 536-540.

Krieg, N.R., and Garrity, G.M. (2005). "On using the *Manual*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part A: Introductory Essays,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley & G.M. Garrity. 2nd ed. (New York, NY, USA: Springer), 15-19.

Ludwig, W., and Schleifer, K.H. (1994). Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiology Reviews* 15(2-3)**,** 155-173.

Luo, C., Rodriguez-R, L.M., and Konstantinidis, K.T. (2014). MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research* 42(8)**,** e73.

Luo, H., and Moran, M.A. (2014). Evolutionary ecology of the marine *Roseobacter* clade. *Microbiology and Molecular Biology Reviews* 78(4)**,** 573-587.

Madigan, M.T., Martinko, J.M., Bender, K.S., Buckley, D.H., and Stahl, D.A. (2014). *Brock Biology of Microorganisms.* San Francisco, CA, USA: Benjamin Cummings.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* 95(6)**,** 3140-3145.

Mallet, J. (1995). A species definition for the modern synthesis. *Trends in Ecology & Evolution* 10(7)**,** 294-299.

Mardis, E.R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* 6(1)**,** 287-303.

Mardis, E.R. (2017). DNA sequencing technologies: 2006-2016. *Nature Protocols* 12(2)**,** 213-218.

Martinez-Murcia, A.J., Benlloch, S., and Collins, M.D. (1992). Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. *International Journal of Systematic Bacteriology* 42(3)**,** 412-421.

Martinez-Murcia, A.J., and Collins, M.D. (1990). A phylogenetic analysis of the genus *Leuconostoc* based on reverse transcriptase sequencing of 16S rRNA. *FEMS Microbiology Letters* 58(1)**,** 73-83.

Mayr, E. (1942). *Systematics and the Origin of Species*. New York, NY, USA: Columbia University Press.

McCarthy, B.J., and Bolton, E.T. (1963). An approach to the measurement of genetic relatedness among organisms. *Proceedings of the National Academy of Sciences of the United States of America* 50**,** 156-164.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14**,** 60.

Meier-Kolthoff, J.P., Klenk, H.P., and Göker, M. (2014). Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. *International Journal of Systematic and Evolutionary Microbiology* 64(2)**,** 352-356.

Metzker, M.L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* 11(1)**,** 31-46.

Mignard, S., and Flandrois, J.P. (2006). 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *Journal of Microbiological Methods* 67(3)**,** 574-581.

Mitra, R.D., and Church, G.M. (1999). *In situ* localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research* 27(24)**,** e34.

Moran, M.A., Belas, R., Schell, M.A., Gonzalez, J.M., Sun, F., Sun, S., et al. (2007). Ecological genomics of marine roseobacters. *Applied and Environmental Microbiology* 73(14)**,** 4559-4569.

Overman, T.L., Kessler, J.F., and Seabolt, J.P. (1985). Comparison of API 20E, API rapid E, and API rapid NFT for identification of members of the family *Vibrionaceae*. *Journal of Clinical Microbiology* 22(5)**,** 778-781.

Qin, Q.L., Xie, B.B., Zhang, X.Y., Chen, X.L., Zhou, B.C., Zhou, J., et al. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of Bacteriology* 196(12)**,** 2210-2215.

Reimer, A.R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., et al. (2011). Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerging Infectious Diseases* 17(11)**,** 2113-2121.

Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015). High-throughput sequencing technologies. *Molecular Cell* 58(4)**,** 586-597.

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

Roberts, R.J., Carneiro, M.O., and Schatz, M.C. (2013). The advantages of SMRT sequencing. *Genome Biology* 14(7)**,** 405.

Rodriguez-R, L.M., and Konstantinidis, K.T. (2005). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints* 4(1)**,** e1900v1.

Rosselló-Móra, R., and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiology Reviews* 25(1)**,** 39-67.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., et al. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265(5596)**,** 687-695.

Santos, S.R., and Ochman, H. (2004). Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environmental Microbiology* 6(7)**,** 754-759.

Schouls, L.M., Schot, C.S., and Jacobs, J.A. (2003). Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. *Journal of Bacteriology* 185(24)**,** 7241-7246.

Segata, N., and Huttenhower, C. (2011). Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS One* 6(9)**,** e24704.

Shayegani, M., Maupin, P.S., and McGlynn, D.M. (1978). Evaluation of the API 20E system for identification of nonfermentative Gram-negative bacteria. *Journal of Clinical Microbiology* 7(6)**,** 539-545.

Shea, A., Wolcott, M., Daefler, S., and Rozak, D.A. (2012). "Biolog phenotye microarrays," in *Microbial Systems Biology. Methods in Molecular Biology (Methods and Protocols),* ed. A. Navid. (Totowa, NJ, USA: Humana Press), 331-373.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology* 26(10)**,** 1135-1145.

Shendure, J.A., Porreca, G.J., Church, G.M., Gardner, A.F., Hendrickson, C.L., Kieleczawa, J., et al. (2011). Overview of DNA sequencing strategies. *Current Protocols in Molecular Biology* 96(7.1)**,** 7.1.1-7.1.23.

Simon, M., Scheuner, C., Meier-Kolthoff, J.P., Brinkhoff, T., Wagner-Döbler, I., Ulbrich, M., et al. (2017). Phylogenomics of *Rhodobacteraceae* reveals evolutionary adaptation to marine and non-marine habitats. *ISME Journal* 11(6)**,** 1483-1499.

Sneath, P.H.A., and Brenner, D.J. (1992). “Official” nomenclature lists. *ASM News* 58(1)**,** 175.

Snel, B., Bork, P., and Huynen, M.A. (1999). Genome phylogeny based on gene content. *Nature Genetics* 21(1)**,** 108-110.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* 16(8)**,** 472-482.

Stackebrandt, E., and Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today* 6(4)**,** 152-155.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A., Kampfer, P., Maiden, M.C., et al. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* 52(3)**,** 1043-1047.

Stackebrandt, E., and Goebel, B.M. (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 and Evolutionary Microbiology* 44(4)**,** 846-849.

Swerdlow, H., and Gesteland, R. (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research* 18(6)**,** 1415-1419.

Swerdlow, H., Wu, S.L., Harke, H., and Dovichi, N.J. (1990). Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of Chromatography* 516(1)**,** 61-67.

Tarr, C.L., Patel, J.S., Puhr, N.D., Sowers, E.G., Bopp, C.A., and Strockbine, N.A. (2007). Identification of *Vibrio* isolates by a multiplex PCR assay and *rpoB* sequence determination. *Journal of Clinical Microbiology* 45(1)**,** 134-140.

Thompson, F.L., Gevers, D., Thompson, C.C., Dawyndt, P., Naser, S., Hoste, B., et al. (2005). Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. *Applied and Environmental Microbiology* 71(9)**,** 5107-5115.

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

Turcatti, G., Romieu, A., Fedurco, M., and Tairi, A.P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research* 36(4)**,** e25.

Urdaci, M.C., Marchand, M., Ageron, E., Arcos, J.M., Sesma, B., and Grimont, P.A. (1991). *Vibrio navarrensis* sp. nov., a species from sewage. *International Journal of Systematic Bacteriology* 41(2)**,** 290-294.

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K., and Swings, J. (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiological Reviews* 60(2)**,** 407-438.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Vezzulli, L., Stauder, M., Grande, C., Pezzati, E., Verheye, H.M., Owens, N.J., et al. (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(4)**,** e0123983.

Wagner-Döbler, I., and Biebl, H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annual Review of Microbiology* 60(1)**,** 255-280.

Wang, A., and Ash, G.J. (2015). Whole genome phylogeny of *Bacillus* by feature frequency profiles (FFP). *Scientific Reports* 5**,** 13644.

Wang, L.T., Lee, F.L., Tai, C.J., and Kasai, H. (2007). Comparison of *gyrB* gene sequences, 16S rRNA gene sequences and DNA-DNA hybridization in the *Bacillus subtilis* group. *International Journal of Systematic and Evolutionary Microbiology* 57(8)**,** 1846-1850.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., et al. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic Bacteriology* 37(4)**,** 463-464.

Whitman, W.B. (2015). *Bergey's Manual of Systematics of Archaea and Bacteria.* Hoboken, NJ, USA: John Wiley & Sons, Inc.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.H., et al. (2008). The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology* 31(4)**,** 241-250.

Yoon, S.H., Ha, S.M., Lim, J., Kwon, S., and Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek* 110(10)**,** 1281-1286.

Zeigler, D.R. (2003). Gene sequences useful for predicting relatedness of whole genomes in bacteria. *International Journal of Systematic and Evolutionary Microbiology* 53(6)**,** 1893-1900.

Zhang, J., and Yang, J.R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* 16(7)**,** 409-420.

CHAPTER 2: *Vibrio cidicii*, a Novel Species of *Vibrio* Closely Related to *Vibrio navarrensis*

A version of this chapter has been published as:

"Orata, F.D., Xu, Y., Gladney, L.M., Rishishwar, L., Case, R.J., Boucher, Y., Jordan, I.K., Tarr, C.L. (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(10), 4148-4155."

F.D.O. and Y.B. designed the study; L.M.G. and C.L.T. provided the *V. cidicii* and *V. navarrensis* isolates and genome sequences; F.D.O. and L.R. performed bioinformatic analyses; Y.X., L.M.G., and R.J.C. performed biochemical tests; F.D.O. and Y.B. wrote the manuscript; and Y.B., I.K.J., and C.L.T. supervised the project.

CHAPTER 2

2.1. Abstract

Four *Vibrio* spp. isolates from the historical culture collection at the Centers for Disease Control and Prevention, obtained from human blood specimens (n = 3) and river water (n = 1), show characteristics distinct from those of isolates of the most closely related species, *Vibrio navarrensis* and *Vibrio vulnificus*, based on phenotypic and genotypic tests. They are specifically adapted to survival in both freshwater and seawater, being able to grow in rich media without added salts as well as salinities above that of seawater. Phenotypically, these isolates resemble *V. navarrensis*, their closest known relative with a validly published name, but the group of isolates is distinguished from *V. navarrensis* by the ability to utilize L-rhamnose. Average nucleotide identity and percent DNA–DNA hybridization values obtained from the pairwise comparisons of whole-genome sequences of these isolates to *V. navarrensis* range from 95.4–95.8% and 61.9–64.3%, respectively, suggesting that the group represents a different species. Phylogenetic analysis of the core genome, including four protein-coding housekeeping genes (*pyrH*, *recA*, *rpoA*, and *rpoB*), places these four isolates into their own monophyletic clade, distinct from *V. navarrensis* and *V. vulnificus*. Based on these differences, we propose these isolates represent a novel species of the genus *Vibrio*, for which the name *Vibrio cidicii* sp. nov. is proposed; strain LMG 29267^T (= CIP 111013^T = 2756-81^T), isolated from river water, is the type strain.

2.2. Introduction

The genus *Vibrio* consists of over 100 species of bacteria autochthonous to the aquatic environment (Gomez-Gil et al., 2014). Species of clinical significance, such as *Vibrio cholerae*,

Vibrio parahaemolyticus, and *Vibrio vulnificus*, have been studied in depth, as clear identification of pathogens is recognized as essential for the treatment of the disease and epidemiologic surveillance. *Vibrio navarrensis*, a species not previously associated with human illness, has received little attention since the original description of isolates from sewage in 1991 (Urdaci et al., 1991). However, this species has recently been the focus of investigation after the identification of *V. navarrensis* by the Centers for Disease Control and Prevention (CDC, Atlanta, GA, USA) among clinical isolates submitted for routine characterization (Gladney et al., 2014; Gladney and Tarr, 2014). Constructing an evolutionary framework that included contemporary *V. navarrensis* and phenotypically similar historical isolates led to the discovery of four isolates that could represent a novel species of *Vibrio* closely related to *V. navarrensis*. These isolates, three of which were recovered from human clinical specimens, are genetically distinct from *V. navarrensis* and *V. vulnificus* based on phylogenetic analysis of housekeeping gene sequences (Gladney and Tarr, 2014). The closest known relative of *V. navarrensis* with a validly published name is currently *V. vulnificus* (Thompson et al., 2005; Gomez-Gil et al., 2014). The latter is an opportunistic pathogen mainly associated with deaths related to seafood consumption, and it causes a fatality rate of over 50% in patients with septicemia (Jones and Oliver, 2009).

In this study, we employed a polyphasic approach to describe the novel species, such as extensive metabolic profiling of the four isolates, comparative genomic analysis to determine DNA–DNA relatedness, and multilocus sequence analysis (MLSA) of core genes. The name *Vibrio cidicii* sp. nov. is proposed for the new species.

2.3. Materials and Methods

Isolates of *V. cidicii* sp. nov. and *V. navarrensis* used in this study were obtained from the CDC (Table 2.1). Phenotypic characterization was performed on the four isolates of *V. cidicii* sp. nov. and three isolates of *V. navarrensis* (Table 2.2 and Table B.1). The isolates were streaked on tryptic soy broth (TSB; Becton Dickinson) with an added 1% NaCl (BDH), yielding a final concentration of 1.5% NaCl, and 1.5% agar (Becton Dickinson). Alternatively, the isolates were streaked on thiosulfate citrate bile salts sucrose (TCBS) agar (Becton Dickinson). The cultures were then incubated overnight (TSB agar) or for two days (TCBS agar) at 30°C. Single colonies from the TSB agar cultures were tested using the Analytical Profile Index (API) 20 NE (bioMérieux) and the Phenotype MicroArray 1 (PM1) MicroPlate (Biolog) according to the instructions of the manufacturers. A minor modification of the PM1 test was the addition of 1.0% NaCl to the inoculating fluid (Biolog) to obtain a final concentration of 1.5% NaCl. The API 20 NE strips and PM1 plates were incubated for 42 hours or 18 hours, respectively, at 30°C. Additional standard phenotypic tests for the routine identification of *Vibrio* not covered by API 20 NE and PM1 were also performed, including: fermentation using various substrates, citrate (in Simmons agar), DNase (at 25°C), H_2S production (in peptone iron agar and triple sugar iron agar), malonate utilization, methyl red, lysine and ornithine decarboxylase (in Moeller medium), motility (swimming and swarming), ONPG (o-nitrophenyl β-D-galactopyranoside), oxidase, phenylalanine deaminase, tyrosine clearing, and Voges-Proskauer (Farmer et al., 2005; Tarr et al., 2015). Permissive growth temperatures were determined in TSB with a final concentration of 1.5% NaCl and incubated at a range of 4–45°C, whereas permissive salinity concentrations were determined in TSB at 30°C in a range of 0–10% NaCl. TSB without NaCl was prepared with 17.0 g/L pancreatic digest of casein (Becton Dickinson), 3.0 g/L papaic digest of soybean

(Becton Dickinson), 2.5 g/L dextrose (Fisher Scientific), and 2.5 g/L dipotassium phosphate (BDH). Gram staining was conducted on the isolates of *V. cidicii* sp. nov. following the protocol of Claus (1992) and viewed under the light microscope (Carl Zeiss) at ×1,000 magnification.

Table 2.1. Source and year of isolation of the strains of *V. cidicii* sp. nov. and *V. navarrensis* used in this study.

Species and strain	Source	Country	Year of isolation	Accession number*	Reference/s
Vibrio cidicii sp. nov.					
LMG 29267^T (= CIP 111013^T = 2756-81^T)	River water	Not known	1981	LOMK00000000	This study
1048-83	Human blood	USA	1983	LOBP00000000	This study
2423-01	Human blood	USA	2001	LOBQ00000000	This study
2538-88	Human blood	USA	1988	LOBR00000000	This study
Vibrio navarrensis					
LMG 15976^T (= ATCC 51183^T = BCRC 15896^T = CAIM 609^T = CCRC 15896^T = CCUG 28805^T = CIP 103381^T = DSM 21557^T = NCIMB 13120^T = 1397-6^T = 2540-90^T)	Sewage	Spain	1982	JMCG00000000	Urdaci et al. (1991); Gladney et al. (2014)
2232 (= 2541-90)	Sewage	Spain	1983	JMCH00000000	This study; Urdaci et al. (1991)
0053-83	Human wound	USA	1983	JMCF00000000	Gladney et al. (2014)
08-2462	Human blood	USA	2008	JMCI00000000	Gladney et al. (2014)

*GenBank BioProject accession numbers: PRJNA304180 (*V. cidicii* sp. nov.) and PRJNA242769 (*V. navarrensis*).

For genotypic characterization, genomic DNA was extracted from overnight TSB cultures of the isolates of *V. cidicii* sp. nov. with the ArchivePure DNA Cell/Tissue Kit (5 PRIME). Whole-genome sequencing and assembly of the environmental isolate LMG 29267^T (= CIP 111013^T = 2756-81^T) were performed with the PacBio RS and the SMRT (Single-Molecule, Real-Time) Analysis software 2.0 (Pacific Biosciences), respectively, as previously described (Gladney et al., 2014). For the clinical isolates 1048-83 and 2538-88, 150-bp paired-end reads were generated on the MiSeq platform (Illumina), as previously described (Gladney et al., 2014).

The genome for clinical isolate 2423-01 was sequenced using 454 Sequencing on the Genome Sequencer FLX System (454 Life Sciences) and also on the Genome Analyzer IIx platform (Illumina), generating 70-bp single-end reads. *De novo* assemblies of the sequences from the clinical isolates were performed using the CG-Pipeline 0.4.1 (Kislyuk et al., 2010). The GenBank/EMBL/DDBJ accession numbers for the whole-genome sequences of *V. cidicii* sp. nov. LMG 29267^T, 1048-83, 2423-01, and 2538-88 are LOMK00000000, LOBP00000000, LOBQ00000000, and LOBR00000000, respectively (Table 2.1), under BioProject accession number PRJNA304180. Accession numbers for the *rpoB* sequences of the same strains are KU593643, KU593646, KU593645, and KU593644, respectively, and for the *rpoB* sequences of *V. navarrensis* LMG 15976^T, 2232, 0053-83, and 08-2462 are KU593635, KU593636, KU593629 and KU593637, respectively (Table B.2). The accession number for the 16S rRNA gene sequence of *V. cidicii* sp. nov. LMG 29267^T is KJ807108.

From the *V. cidcii* sp. nov. and *V. navarrensis* whole-genome sequences (Table 2.1), the G+C content was determined using Geneious 8.1.2 (Kearse et al., 2012). Pairwise average nucleotide identity (ANI) was calculated using the dnadiff program in MUMmer 3.0 (ANIm; Kurtz et al., 2004). Pairwise percent DNA–DNA hybridization (DDH) was also calculated *in silico* using the Genome-to-Genome Distance Calculator 2.0 (GGDC; Meier-Kolthoff et al., 2013). The genome sequences were annotated with RAST 2.0 (Rapid Annotation Using Subsystem Technology; Aziz et al., 2008) and Prodigal 1.2 (Prokaryotic Dynamic Programming Genefinding Algorithm; Hyatt et al., 2010). Orthologous protein-coding gene families were determined from the annotated genomes by pairwise bidirectional BLASTP (Altschul et al., 1990) using the OrthoMCL pipeline 2.0 (Li et al., 2003) with 30% identity cutoff (Rost, 1999). The gene families unique to *V. cidicii* sp. nov. or *V. navarrensis* were subsequently determined

using Intella 1.7.0 (https://www.vound-software.com). The predicted functions of these gene families were determined based on the Clusters of Orthologous Groups of proteins (COG) database (Tatusov et al., 2000) and by sequence similarity search in the GenBank database (National Center for Biotechnology Information) using BLASTP (Altschul et al., 1990).

With additional genome sequences from closely related species of the genus *Vibrio* obtained from the GenBank database (Table B.2), single-copy, protein-coding core gene families were determined using OrthoMCL (Li et al., 2003). The sequences were aligned using Clustal W 2.1 (Larkin et al., 2007), and the alignments were concatenated, stripping columns with at least one gap, using Geneious 8.1.2 (Kearse et al., 2012). This resulted in a single alignment with a total length of 446,032 bp, which was used to construct a maximum-likelihood tree with RAxML 8.2.8 (Stamatakis, 2014) using the GTR (general time reversible) nucleotide substitution model and gamma distribution pattern. Robustness of branching was estimated with 100 bootstrap replicates. Moreover, a subset of four housekeeping genes was selected for MLSA – *pyrH*, *recA*, *rpoA*, and *rpoB* (Thompson et al., 2005; Tarr et al., 2007; Gladney and Tarr, 2014). From the partial DNA sequences, a concatenated alignment of 2,313 bp was obtained and used to construct a maximum-likelihood tree, as described above. Patristic distances between species, the sum of the lengths of the branches that link two terminal nodes in a tree, were calculated from the latter tree using Geneious 8.1.2 (Kearse et al., 2012). In addition, whole-genome phylogeny was also reconstructed based on genomic similarity (ANI) between each pair of genome sequences. First, pairwise ANIm was computed using MUMmer 3.0 (Kurtz et al., 2004). The similarity obtained was then converted into average nucleotide distances (= 100 – ANI). The resulting distance matrix was utilized to quantify all against all pairwise species distances as well as in the

construction of a neighbor-joining tree (Saitou and Nei, 1987) using the software MEGA 7.0 (Molecular Evolutionary Genetics Analysis; Kumar et al., 2016).

2.4. Results and Discussion

All the isolates of *V. cidicii* sp. nov. and *V. navarrensis* studied exhibited growth in TSB without NaCl (Table 2.2 and Table B.1). This is contrary to a previous report of these isolates not exhibiting growth in nutrient broth without NaCl (Gladney and Tarr, 2014). This is possibly due to differences in the media used in both studies, as test conditions will dictate salt requirement (Farmer et al., 2005). Urdaci et al. (1991) reported seven out of ten isolates of *V. navarrensis* grew weakly in peptone water without NaCl. Growth in medium without NaCl was previously reported for a few species of the genus *Vibrio*, including *V. cholerae* and *Vibrio mimicus* (Farmer et al., 2005; Gomez-Gil et al., 2014), two species of great clinical significance. The ability of bacteria to survive in freshwater makes it more likely to come in contact with humans through ingestion (Boucher et al., 2015). The isolation of strains of *V. cidicii* sp. nov. and *V. navarrensis* from river water and sewage, respectively, suggests these species are also able to survive in low salt environments. Furthermore, both species are also able to survive at 40°C, a trait observed mostly in pathogenic vibrios that can survive inside the human body (Farmer et al., 2005; Gomez-Gil et al., 2014).

Vibrio cidicii sp. nov. resembles *V. navarrensis* in the majority of phenotypic characteristics tested (123 of 158 tests or 78%; Table 2.2 and Table B.1). However, a single phenotypic feature distinguished *V. cidicii* sp. nov. from *V. navarrensis*: it tested positive for the utilization of L-rhamnose as the sole carbon and energy source in both fermentation and assimilation tests. Although previous reports support our result of an L-rhamnose-negative *V.*

navarrensis (Farmer et al., 2005; Gomez-Gil et al., 2014), a recent study reported one isolate of *V. navarrensis* also capable of utilizing L-rhamnose (Gladney and Tarr, 2014). To our knowledge, this is the only reported isolate of *V. navarrensis* that is L-rhamnose-positive. *Vibrio vulnificus* is also not able to utilize this substrate (Farmer et al., 2005; Gomez-Gil et al., 2014). Other species of the genus *Vibrio* that are capable of utilizing this substrate are *Vibrio hispanicus*, *Vibrio natriegens*, and *Vibrio pectenicida*, all of which are very distantly related to *V. cidicii* sp. nov. and *V. navarrensis* (Thompson et al., 2005; Gomez-Gil et al., 2014). Comparison of the annotated genomes of isolates of *V. cidicii* sp. nov. and *V. navarrensis* revealed four genes found only in the former encoding proteins involved in L-rhamnose transport and metabolism: L-rhamnose isomerase, L-rhamnose mutarotase, L-rhamnose-proton symporter, and rhamnulose-1-phosphate aldolase (Table B.3; Wilson and Ajl, 1957; Sawada and Takagi, 1964; Ryu et al., 2004). L-rhamnose is produced in high levels by diatoms (Brown, 1991), a group of phytoplankton found in both marine and freshwater environments, where *V. cidicii* sp. nov. was also found. This suggests a physiological differentiation with *V. navarrensis*, in which *V. cidicii* sp. nov. is adapted to living on or near algae and exploiting their carbon exudates. An additional 24 genes with predicted functions were found in *V. cidicii* sp. nov. but were absent from *V. navarrensis* (Table B.3). These could encode other distinguishing characteristics for the species. However, no physiological tests were available to test the phenotypes they are predicted to encode.

Table 2.2. Summary of phenotypic test results for *V. cidicii* sp. nov., *V. navarrensis*, and *V. vulnificus*.

Strains: 1, *V. cidicii* sp. nov. LMG 29267^T; 2, *V. cidicii* sp. nov. 1048-83; 3, *V. cidicii* sp. nov. 2423-01; 4, *V. cidicii* sp. nov. 2538-88; 5, *V. navarrensis* LMG 15976^T; 6, *V. navarrensis* 0053-83; 7, *V. navarrensis* 08-2462; 8, *V. vulnificus*. +, Growth/positive test result; −, no growth/negative test result; v, variable results between tests; ND, not determined. All strains were positive for: indole production, methyl red, phenylalanine deaminase, swimming motility (37°C), gelatin hydrolysis, esculin hydrolysis, reduction of nitrate to nitrite, oxidase, DNase (25°C), acid production from D-glucose, growth in TSB with 6.5% NaCl (6% for *V. vulnificus*), and growth in TSB at 40°C. All strains were negative for Voges-Proskauer, H_2S production, urea hydrolysis, arginine dihydrolase, gas production from D-glucose, acid production from L-arabinose, glycerol, D-xylose, and growth in TSB at 4°C.

Phenotypic test/substrate tested	1	2	3	4	5	6	7	8†
Citrate (Simmons agar)	−	+	−	−	+	+	+	+
Lysine decarboxylase (Moeller medium)	−	−	−	−	−	−	−	+
Ornithine decarboxylase (Moeller medium)	−	−	−	−	−	−	−	+
Swarming (marine agar, 25°C)	−	+	−	+	−	−	−	−
Acid production from:								
L-Rhamnose	+	+	+	+	−	−	−	−
Salicin	−	−	+	+	−	−	−	+
Assimilation of:								
L-Arabinose*	v	v	v	v	v	v	v	−
Glycerol	+	+	+	+	+	+	+	ND
L-Rhamnose	+	+	+	+	−	−	−	−
D-Xylose	+	+	+	+	+	+	+	ND
Growth in TSB (at 30°C) with:								
0% NaCl	+	+	+	+	+	+	+	−
1.5% NaCl	+	+	+	+	+	+	+	− (1%)
Growth in TSB (with 1.5% NaCl) at:								
30°C	+	+	+	+	+	+	+	−

*Assimilation of L-arabinose (for *V. cidicii* sp. nov. and *V. navarrensis* isolates): positive with PM1, negative with API 20 NE.
†Results for *V. vulnificus* were obtained from Farmer et al. (2005) and Gomez-Gil et al. (2014).

Three additional characteristics distinguish *V. cidicii* sp. nov. from its closest relatives, although not universal among isolates (Table 2.2 and Table B.1). First, two of the four isolates of *V. cidicii* sp. nov. exhibited swarming on marine agar. Many vibrios have been reported to exhibit swarming motility, including *Vibrio cincinnatiensis* and *Vibrio proteolyticus*. However, *V. navarrensis* and *V. vulnificus* are negative for the phenotype (Farmer et al., 2005; Gomez-Gil et al., 2014). Second, two of the four isolates of *V. cidicii* sp. nov. tested positive for salicin fermentation, while none of the isolates of *V. navarrensis* did. This trait is also a key differentiating characteristic between *V. navarrensis* and *V. vulnificus*, with the latter also able to

utilize the substrate (Farmer et al., 2005). Lastly, three of the four isolates of *V. cidicii* sp. nov. are negative for the utilization of sodium citrate (citrate test), whereas both *V. navarrensis* and *V. vulnificus* are able to utilize the substrate (Farmer et al., 2005). Both *V. cidicii* sp. nov. and *V. navarrensis* can be differentiated from *V. vulnificus* by two characteristics; they test negative for lysine and ornithine decarboxylase (Farmer et al., 2005; Gomez-Gil et al., 2014).

We observed contradicting results between the PM1 and API 20 NE tests we conducted for the assimilation of L-arabinose, where all isolates of *V. cidicii* sp. nov. and *V. navarrensis* tested positive with the PM1 system and negative with the API 20 NE system. This difference can be attributed to the differences in methods between the tests (e.g., incubation period, NADH production and redox dye chemistry detection versus turbidity detection). We conclude that this test is not reliable for the identification of the species *V. cidicii* sp. nov. and *V. navarrensis*. Additionally, we observed a difference between our assimilation and fermentation tests (positive for all isolates with the former and negative with the latter) with two other substrates, glycerol and D-xylose. This difference is due to the different attributes being measured (i.e., NADH production versus acid production), suggesting that both species are capable of utilizing the substrates aerobically but not via fermentation.

Based on their whole-genome sequences, the G+C content of the four isolates of *V. cidicii* sp. nov. range from 47.9–48.2 mol%, which is within the known range for the genus *Vibrio* (38.0–51.0 mol%; Farmer et al., 2005). This eliminates the assignment of the isolates to other genera in the family *Vibrionaceae* such as *Aliivibrio* (38.0–42.0 mol%; Urbanczyk et al., 2007), *Photobacterium* (39.0–44.0 mol%; Thyssen and Ollevier, 2005), and *Salinivibrio* (49.4–50.5 mol%; Ventosa, 2005), and most genera in the family *Enterobacteriaceae* (50.0–67.0 mol%; Brenner and Farmer, 2005).

Various tools to measure DNA–DNA relatedness *in silico* are available to replace the traditional method of DDH (Konstantinidis and Tiedje, 2005a; Goris et al., 2007; Richter and Rosselló-Móra, 2009; Meier-Kolthoff et al., 2013). Here, we determined relatedness of organisms using ANI and DDH by pairwise comparisons of whole-genome sequences. The ANI between isolates within the species *V. cidicii* sp. nov. or *V. navarrensis* range from 97.4–100.0% (Figure 2.1 and Table B.4). In contrast, the ANI between *V. cidicii* sp. nov. and *V. navarrensis* range from 95.4–95.8%. Since the results are close to the cutoff of 96% ANI for two genomes to belong to the same species (Richter and Rosselló-Móra, 2009), we complemented our ANI results with percent DDH data. The GGDC package was used to calculate percent DDH *in silico* to mimic wet lab-based DDH (Meier-Kolthoff et al., 2013). Percent DDH within the species *V. cidicii* sp. nov. or *V. navarrensis* range from 75.8–88.7%, whereas they range from 61.9–64.3% between the two groups (Figure 2.1 and Table B.5). The determined same-species cutoff for DDH is 70% (Goris et al., 2007), supporting our ANI results and suggesting the two groups to be distinct from each other.

MLSA further supports our proposal of a novel species. Single-copy, protein-coding core genes are used as alternatives to 16S rRNA gene sequences for the identification and phylogenetic analysis of various species of the genus *Vibrio*, since there is a lack of species-level resolution using 16S rRNA gene sequences (Thompson et al., 2005; Gladney and Tarr, 2014). A core genome tree was constructed from 586 single-copy core genes that are shared by all strains used in this study (Vernikos et al., 2015). The four isolates of *V. cidicii* sp. nov. form a monophyletic clade that is distinct from the *V. navarrensis* and *V. vulnificus* clades, with 100% bootstrap support (Figure B.1). Since recombination is also apparent within the core genome and can occur at a high rate for very closely related species (Orata et al., 2015), we examined a

subset of four housekeeping genes (*pyrH*, *recA*, *rpoA*, and *rpoB*) that do not exhibit recombination among the isolates of *V. cidicii* sp. nov., *V. navarrensis*, and *V. vulnificus*. These genes have been shown to be reliable for the taxonomic characterization of vibrios (Thompson et al., 2005; Tarr et al., 2007; Gladney and Tarr, 2014). Phylogenetic analysis using the four housekeeping genes also distinguishes *V. cidicii* sp. nov. from *V. navarrensis* and *V. vulnificus* (Figure 2.2). The average patristic distance calculated from this tree between isolates of *V. cidicii* sp. nov. and *V. navarrensis* is 0.066, while lower average distances of 0.005 and 0.007 are obtained when comparing isolates within the species *V. cidicii* sp. nov. or *V. navarrensis*, respectively (Table B.6). To further demonstrate this distinction, a phylogeny was constructed based on whole-genome comparisons to account for whole-genome variation between isolates that would otherwise be excluded from the core genome (Figure B.2). This phylogeny also shows the distinct clustering of the isolates of *V. cidicii* sp. nov. from *V. navarrensis*. The average nucleotide distances calculated for this tree (Table B.7) show that the diversity within *V. cidicii* sp. nov (1.780) or *V. navarrensis* (2.160) is much lower than the diversity between the two species (4.414). Our phylogenetic analyses placed the *V. cidicii* sp. nov. lineage into the context of a larger *Vibrio* phylogeny, showing that the novel species is distinct from all *Vibrionaceae* that have been characterized to date (Figure 2.2, Figure B.1, and Figure B.2; Gladney and Tarr, 2014). On the other hand, the use of the 16S rRNA gene did not clearly distinguish *V. cidicii* sp. nov. and *V. navarrensis* (Gladney and Tarr, 2014).

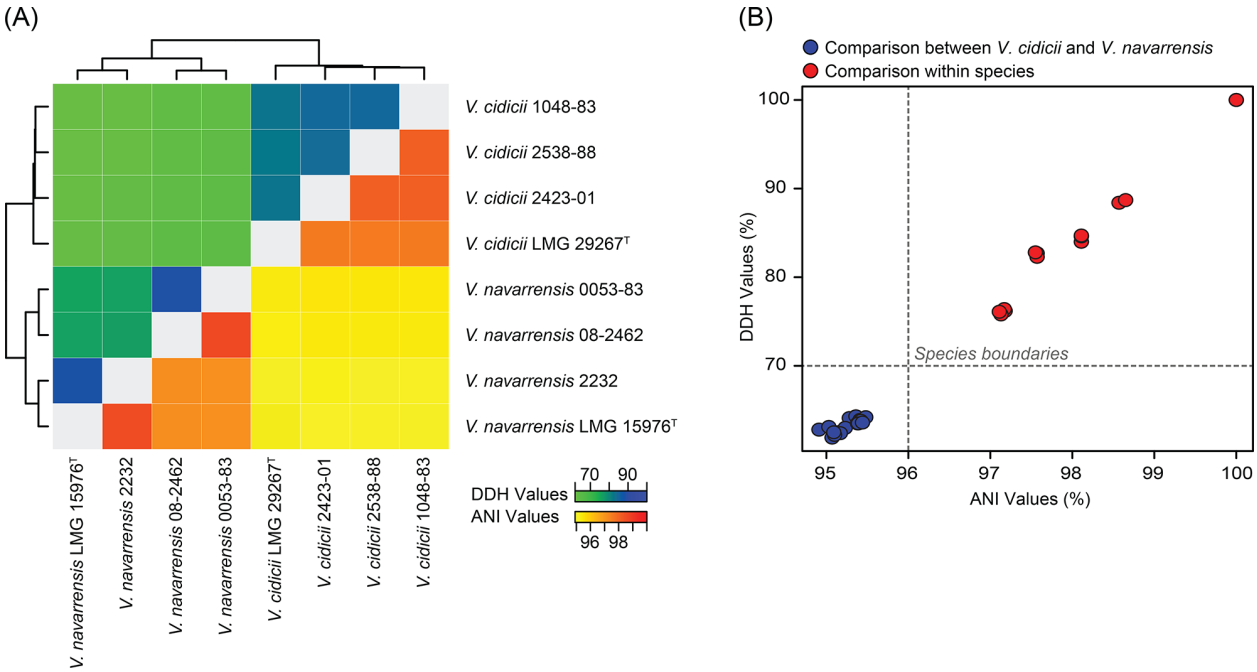

Figure 2.1. Whole-genome comparisons among isolates of *V. cidicii* sp. nov. and *V. navarrensis*.

(A) Heat map showing percent DDH and ANI along with the clustering-based dendrogram showing the relationship between *V. cidicii* sp. nov. and *V. navarrensis* genomes. (B) Pairwise percent DDH and ANI comparisons between isolates of *V. cidicii* sp. nov. and *V. navarrensis* fall below the species boundary cutoffs indicating that they are different species.

Overall, phylogenetic analyses confirm the position of *V. cidicii* sp. nov. in the genus *Vibrio*, which forms a monophyletic clade distinct from *V. navarrensis* and *V. vulnificus*, supporting its identification as a novel species of the genus *Vibrio*. This distinction is further confirmed by ANI and percent DDH below 96% and 70%, respectively, between species. The ability of *V. cidicii* sp. nov. to utilize L-rhamnose could be a feature that drove its speciation from a common ancestor shared with *V. navarrensis*. Further studies are needed to determine the prevalence of *V. cidicii* sp. nov. in various environments. The isolation of strains from human blood suggests it is capable of infecting humans and can be pathogenic. Two additional isolates were recovered since 2014 after the commencement of this study, underscoring the need to

identify and characterize isolates of this pathogen. It will be important to study the pathogenicity and epidemiology of this novel species for control, treatment, and prevention of disease.

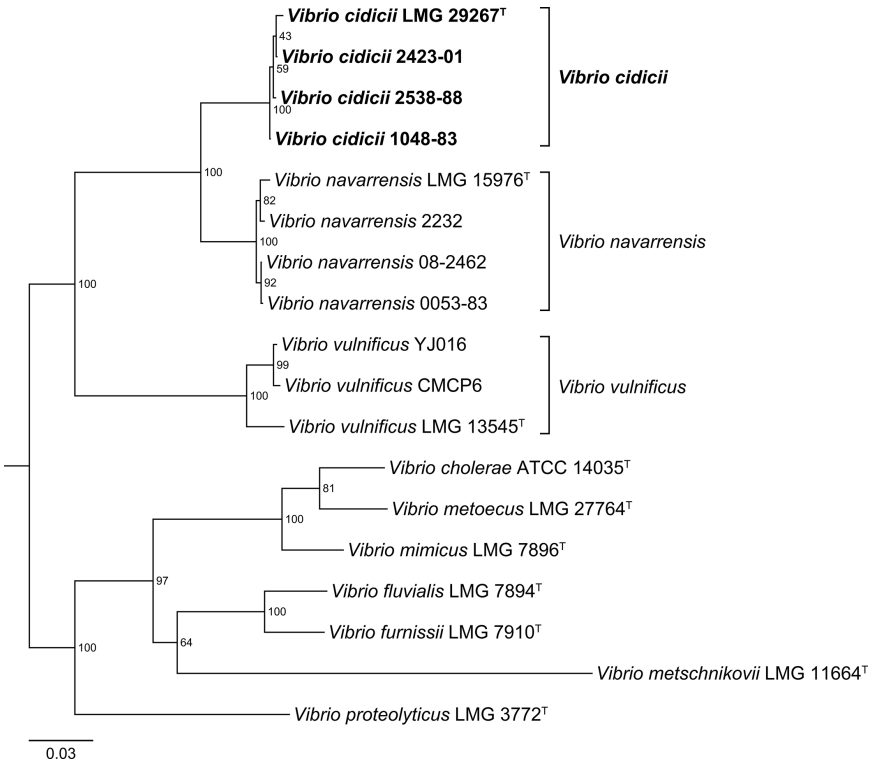

Figure 2.2. The phylogenetic relationship of *V. cidicii* sp. nov. and its closest relatives.

The tree was constructed from the concatenated alignment of partial DNA sequences of four protein-coding housekeeping genes (*pyrH*, *recA*, *rpoA*, and *rpoB*) with a total length of 2,313 bp. Bootstrap support is indicated on the nodes. Bar, 0.03 nucleotide substitutions per site.

2.5. Description of *Vibrio cidicii* sp. nov.

Vibrio cidicii (ci.di'ci.i. N.L. gen. n. *cidicii*, from the phonetics of the CDC, the acronym for the Centers for Disease Control and Prevention, where the species was initially identified).

Cells are Gram-negative, curved, motile rods, 0.64–0.78 × 1.48–1.68 μm in size, which produce convex, smooth, circular, entire, cream-coloured colonies on TSB agar and yellow colonies (sucrose-fermenting) on TCBS agar. Growth is observed in TSB at 30°C with salt

concentrations in the range of 0–6.5% NaCl, and up to 8% for some isolates (three out of four tested); no growth occurs in the presence of 10% NaCl. Growth is also observed in TSB with 1.5% total NaCl concentration at a temperature range of 30–40°C, and no growth occurs at 4°C and 45°C. The ability to utilize L-rhamnose as the sole carbon and energy source distinguishes *V. cidicii* sp. nov. from *V. navarrensis*, its closest relative. In addition, the following traits are variable across isolates: swarming on marine agar (two positive out of four tested; negative for *V. navarrensis*), salicin fermentation (two positive out of four tested; negative for *V. navarrensis*), and sodium citrate utilization (three negative out of four tested; positive for *V. navarrensis*). Positive results in tests for: indole production; methyl red test; phenylalanine deaminase; gelatin and esculin hydrolysis; reduction of nitrate to nitrite; oxidase and DNase; and acid production from D-glucose, cellobiose, maltose, mannitol, mannose, L-rhamnose, sucrose, and trehalose. Negative results in tests for: Voges-Proskauer test; H_2S production; urea hydrolysis; arginine dihydrolase, lysine decarboxylase, and ornithine decarboxylase; malonate utilization; gas production from D-glucose; and acid production from D-adonitol, L-arabinose, D-arabitol, dulcitol, erythritol, glycerol, myo-inositol, lactose, melibiose, mucic acid, raffinose, and D-xylose.

Utilizes the following substrates as sole carbon and energy sources: acetic acid, acetoacetic acid, N-acetyl-D-glucosamine, adenosine, D-alanine, L-alanine, L-alanyl-glycine, L-asparagine, L-aspartic acid, bromosuccinic acid, cellobiose, citric acid, 2'-deoxyadenosine, deoxyribonucelic acid, esculin, formic acid, D-fructose, fructose-6-phosphate, fumaric acid, gelatin, D-gluconic acid, D-glucose, glucose-1-phosphate, glucose-6-phosphate, L-glutamic acid, L-glutamine, glycerol, D,L-α-glycerol phosphate, glycyl-L-aspartic acid, glycyl-L-glutamic acid, glycyl-L-proline, α-hydroxy butyric acid, α-hydroxy glutaric acid-γ-lactone, inosine, α-

ketobutyric acid, α-ketoglutaric acid, L-lactic acid, L-lyxose, D-malic acid, L-malic acid, D,L-malic acid, D-maltose, maltotriose, D-mannitol, D-mannose, methyl pyruvate, β-methyl-D-glucoside, monomethyl succinate, phenylalanine, potassium gluconate, L-proline, propionic acid, D-psicose, pyruvic acid, L-rhamnose, D-ribose, L-serine, succinic acid, sucrose, L-threonine, thymidine, trehalose, L-tryptophan, Tween 40, Tween 80, uridine, and D-xylose. The following substrates that are not utilized (in four tested isolates): adipic acid, 2-aminoethanol, D-arabitol, L-arginine, D-aspartic acid, capric acid, erythritol, L-fucose, L-galactonic acid-γ-lactone, D-galacturonic acid, glucuronamide, D-glucosaminic acid, glyoxylic acid, m-hydroxyphenylacetic acid, p-hydroxyphenylacetic acid, myo-inositol, lysine, malonate, α-methyl-D-galactoside, ornithine, phenylacetic acid, phenylethylamine, 1,2-propanediol, raffinose, D-serine, m-tartaric acid, D-threonine, tricarballylic acid, trisodium citrate, tyramine, and urea.

The type strain is LMG 29267^{T} (= CIP 111013^{T} = 2756-81^{T}) isolated from river water in 1981 (country of origin unknown). The type strain displays all of the properties given above for the species. The G+C content of the type strain is 47.9 mol% based on whole-genome sequencing.

2.6. References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215(3)**,** 403-410.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9**,** 75.

Boucher, Y., Orata, F.D., and 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. *Frontiers in Microbiology* 6**,** 1120.

Brenner, D.J., and Farmer, J.J., III (2005). "Family I. *Enterobacteriaceae*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part B: The Gammaproteobacteria,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, M. Goodfellow, F.A. Rainey & K.H. Schleifer. 2nd ed. (New York, NY, USA: Springer), 587-850.

Brown, M.R. (1991). The amino-acid and sugar composition of 16 species of microalgae used in mariculture. *Journal of Experimental Marine Biology and Ecology* 145(1)**,** 79-99.

Claus, D. (1992). A standardized Gram staining procedure. *World Journal of Microbiology and Biotechnology* 8(4)**,** 451-452.

Farmer, J.J., III, Janda, M., Brenner, F.W., Cameron, D.N., and Birkhead, K.M. (2005). "Genus I. *Vibrio*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part B: The Gammaproteobacteria,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, M. Goodfellow, F.A. Rainey & K.H. Schleifer. 2nd ed. (New York, NY, USA: Springer), 494-546.

Gladney, L.M., Katz, L.S., Knipe, K.M., Rowe, L.A., Conley, A.B., Rishishwar, L., et al. (2014). Genome sequences of *Vibrio navarrensis*, a potential human pathogen. *Genome Announcements* 2(6)**,** e01188-14.

Gladney, L.M., and Tarr, C.L. (2014). Molecular and phenotypic characterization of *Vibrio navarrensis* isolates associated with human illness. *Journal of Clinical Microbiology* 52(11)**,** 4070-4074.

Gomez-Gil, B., Thompson, C.C., Matsumura, Y., Sawabe, T., Iida, T., Christen, R., et al. (2014). "The family *Vibrionaceae*," in *The Prokaryotes – Gammaproteobacteria,* eds. E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt & F. Thompson. 4th ed. (Berlin, Germany: Springer), 659-747.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57(1)**,** 81-91.

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

Jones, M.K., and Oliver, J.D. (2009). *Vibrio vulnificus*: disease and pathogenesis. *Infection and Immunity* 77(5)**,** 1723-1733.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12)**,** 1647-1649.

Kislyuk, A.O., Katz, L.S., Agrawal, S., Hagen, M.S., Conley, A.B., Jayaraman, P., et al. (2010). A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* 26(15)**,** 1819-1826.

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

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33(7)**,** 1870-1874.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biology* 5(2)**,** R12.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21)**,** 2947-2948.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* 13(9)**,** 2178-2189.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14**,** 60.

Orata, F.D., Kirchberger, P.C., Méheust, R., Barlow, E.J., Tarr, C.L., and 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(10)**,** 2941-2954.

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

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering* 12(2)**,** 85-94.

Ryu, K.S., Kim, C., Kim, I., Yoo, S., Choi, B.S., and Park, C. (2004). NMR application probes a novel and ubiquitous family of enzymes that alter monosaccharide configuration. *Journal of Biological Chemistry* 279(24)**,** 25544-25548.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4)**,** 406-425.

Sawada, H., and Takagi, Y. (1964). The metabolism of L-rhamnose in *Escherichia coli*: III. L-rhamnulose-phosphate aldolase. *Biochimica et Biophysica Acta* 92(1)**,** 26-32.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9)**,** 1312-1313.

Tarr, C.L., Bopp, C.A., and Farmer, J.J., III (2015). "*Vibrio* and related organisms," in *Manual of Clinical Microbiology,* eds. J.H. Jorgensen, M.A. Pfaller, K.C. Carroll, G. Funke, M.L. Landry, S.S. Richter & D.W. Warnock. 11th ed. (Washington, DC, USA: ASM Press), 762-772.

Tarr, C.L., Patel, J.S., Puhr, N.D., Sowers, E.G., Bopp, C.A., and Strockbine, N.A. (2007). Identification of *Vibrio* isolates by a multiplex PCR assay and *rpoB* sequence determination. *Journal of Clinical Microbiology* 45(1)**,** 134-140.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28(1)**,** 33-36.

Thompson, F.L., Gevers, D., Thompson, C.C., Dawyndt, P., Naser, S., Hoste, B., et al. (2005). Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. *Applied and Environmental Microbiology* 71(9)**,** 5107-5115.

Thyssen, A., and Ollevier, F. (2005). "Genus II. *Photobacterium*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part B: The Gammaproteobacteria,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, M. Goodfellow, F.A. Rainey & K.H. Schleifer. 2nd ed. (New York, NY, USA: Springer), 546-552.

Urbanczyk, H., Ast, J.C., Higgins, M.J., Carson, J., and Dunlap, P.V. (2007). Reclassification of *Vibrio fischeri*, *Vibrio logei*, *Vibrio salmonicida* and *Vibrio wodanis* as *Aliivibrio fischeri* gen. nov., comb. nov., *Aliivibrio logei* comb. nov., *Aliivibrio salmonicida* comb. nov. and *Aliivibrio wodanis* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* 57(12)**,** 2823-2829.

Urdaci, M.C., Marchand, M., Ageron, E., Arcos, J.M., Sesma, B., and Grimont, P.A. (1991). *Vibrio navarrensis* sp. nov., a species from sewage. *International Journal of Systematic Bacteriology* 41(2)**,** 290-294.

Ventosa, A. (2005). "Genus III. *Salinivibrio*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part B: The Gammaproteobacteria,* eds. D.J. Brenner,

N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, M. Goodfellow, F.A. Rainey & K.H. Schleifer. 2nd ed. (New York, NY, USA: Springer), 552-555.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Wilson, D.M., and Ajl, S. (1957). Metabolism of L-rhamnose by *Escherichia coli*. I. L-rhamnose isomerase. *Journal of Bacteriology* 73(3)**,** 410-414.

CHAPTER 3: Genomic taxonomy of *Rhodobacteraceae* and *Stappiaceae* fam. nov. in the order *Rhodobacterales*

A version of this chapter will be submitted for publication as:

"Orata, F.D., Liang, K., Boucher, Y., and Case, R.J. Whole genome-based taxonomy of *Rhodobacteraceae* and *Stappiaceae* fam. nov. in the order *Rhodobacterales*."

F.D.O., Y.B., and R.J.C. designed the study and wrote the manuscript; F.D.O. and K.L. performed bioinformatic analyses; and Y.B. and R.J.C. supervised the project.

CHAPTER 3

3.1. Abstract

Bacterial and archaeal taxonomy has changed substantially in the past few decades. It has embraced a polyphasic approach, which combines phenotypic, genotypic, and phylogenetic information, to make correct taxonomic attributions. With the advent of next-generation sequencing, genotypic characterizations have evolved from the use of a single to few genes to whole-genome sequences. As a consequence, new genome-based similarity indexes have been established, more so for species delineations [*in silico* (digital) DNA-DNA hybridization (dDDH) and average nucleotide identity (ANI)] than higher ranks. Average amino acid identity (AAI) is the proposed metric for the delineation of ranks beyond species. In this study, we performed AAI, as well as dDDH, ANI, and phylogenomic analysis using 287 genomes of members of the family *Rhodobacteraceae* (70 genera and 127 species). Within this family is the rosoebacter clade, a group of marine α-proteobacteria that makes up almost 20% of marine bacterioplankton and plays many important roles in the global carbon and sulfur cycles. Taxonomy within this family is inconsistent, exhibiting several non-monophyletic genera and species. We propose 60% AAI as the new family boundary and the reclassification of some strains into a newly proposed family, *Stappiaceae* fam. nov. We also propose a new genus boundary of 70% AAI and the reclassification of several members at the genus level. Changes at the species level are also proposed based on the systematic application of already recognized species cutoffs (70% dDDH and 95% ANI). By applying these boundaries, reclassifications resulted to 53 genera and 124 species. Identification and assignment of scientific names need to

be resolved within this group first and foremost to avoid confusion with ecological and evolutionary interpretations in subsequent studies.

3.2. Introduction

Taxonomy is the science of characterizing, naming, and classifying organisms. For microorganisms, this has changed substantially in the past few decades, embracing a polyphasic approach – phenotypic, genotypic, and phylogenetic – that considers a range of different traits for a systematic identification and description (Vandamme et al., 1996). A defined set of criteria has been put in place (Rosselló-Móra and Amann, 2001; Kämpfer et al., 2003; Tindall et al., 2010), although there is no universally accepted concept of species for microorganisms (Ereshefsky, 2010), and prokaryotic taxonomy is still largely a matter of scientific judgment (Sneath and Brenner, 1992; Garrity, 2016). Tests for identification include, among others, cell and colony morphology, biochemical tests, G+C content deviation, and DNA-DNA hybridization (DDH) efficiency, as well as phylogenetic analysis of the small subunit rRNA genes (16S rRNA) or multilocus sequence analysis (MLSA) using core housekeeping protein-coding genes (Tindall et al., 2010).

The availability of whole-genome sequences has provided tangible standards for the delineation of species. Experimental DDH is now being replaced by *in silico* (digital) DDH (dDDH), yet still maintaining the cutoff of 70% hybridization for two genomes to belong to the same species (Goris et al., 2007; Meier-Kolthoff et al., 2013). Average nucleotide identity (ANI) can also be determined by pairwise comparison of genomes to determine if they belong to the same species, if they display at least 95% ANI (Goris et al., 2007; Richter and Rosselló-Móra, 2009). On the other hand, there is still a need to define and delineate taxonomic ranks higher

than species (genera, families, etc.). Clustering by 16S rRNA gene sequences is the widely used approach for circumscribing higher taxonomic ranks (Krieg and Garrity, 2005), but it is not without limitations, such as poor discriminatory power for some genera (Janda and Abbott, 2007) and the existence of intragenomic heterogeneity between multiple copies found in a single cell (Klappenbach et al., 2001; Acinas et al., 2004). New metrics have been proposed to delineate organisms beyond the species level, such as average amino acid identity (AAI; Konstantinidis and Tiedje, 2005b) and percentage of conserved proteins (POCP; Qin et al., 2014), also by utilizing whole-genome sequence data.

In this study, we used available whole-genome sequences and the dDDH, ANI, AAI, and POCP metrics to systematically delineate members of the current family *Rhodobacteraceae* in the family, genus, and species ranks. The family *Rhodobacteraceae* belongs to the order *Rhodobacterales*, class *Alphaproteobacteria*, and phylum *Proteobacteria*, circumscribed on the basis of phylogenetic analysis of 16S rRNA gene sequences (Garrity et al., 2005). The *Bergey's Manual* lists 25 genera under this family as of its publication in 2005 (Garrity et al., 2005), whereas NCBI Taxonomy currently lists a total of 163 identified genera as of October 5, 2017, 110 of which have sequenced genomes available. From these genera, there are currently 254 identified species with available genome sequences.

Within the family *Rhodobacteraceae* is the roseobacter clade (not to be confused with the genus *Roseobacter*), a phylogenetically coherent group of predominantly marine bacteria that shares at least 89% identity of 16S rRNA gene sequences (Giovanni and Rappé, 2000; Buchan et al., 2005). This is one of the most abundant groups of bacteria, which can represent up to 20% of bacterial cells in coastal ecosystems and 3 to 5% in open ocean surface waters and inhabits a wide range of marine niches (Moran et al., 2007). Many representative roseobacters can be easily

cultivated in the laboratory, thus they have become one of the most extensively studied marine microorganisms (Buchan et al., 2005; Wagner-Döbler and Biebl, 2006). This group is valued in the field of oceanography because it was the only one for which cultivated strains are available that were closely related to cloned environmental sequences (Buchan et al., 2005). Additionally, members of this group play many important roles for the global carbon and sulfur cycles (Wagner-Döbler and Biebl, 2006).

The first species descriptions of roseobacters, those of *Roseobacter denitrificans* and *Roseobacter litoralis*, appeared in 1991 (Shiba, 1991), around the same time when the significance of 16S rRNA-based approaches as culture-independent methods for studying prokaryotic diversity was realized (Giovannoni et al., 1990; Ward et al., 1990). Thus, diversity studies of this group of marine microorganisms have been significantly linked to the use of 16S rRNA sequences (Buchan et al., 2005; Wagner-Döbler and Biebl, 2006). Despite the phylogenetic coherence of the roseobacter clade within the family *Rhodobacteraceae*, phylogeny within the clade is inconsistent and is still mainly based on the 16S rRNA gene. The inconsistency is evident with the genera *Leisingera*, *Phaeobacter*, *Rosoebacter*, and *Ruegeria*, which are polyphyletic or paraphyletic (Buchan et al., 2005; Wagner-Döbler and Biebl, 2006; Luo and Moran, 2014; Simon et al., 2017). Thus, the incoherent assignment of scientific names needs to be resolved to avoid confusion with ecological and evolutionary interpretations. Whereas previous studies have focused on resolving taxonomy only within the rosoebacter clade (Buchan et al., 2005; Wagner-Döbler and Biebl, 2006; Moran et al., 2007; Luo and Moran, 2014), an analysis of the phylogenetic affiliation of this group as part of the family *Rhodobacteraceae* has only been carried out recently (Simon et al., 2017). In this study, genotypic and phylogenetic analyses were systematically performed to all available high-quality

genomes within this family. This allowed the determination of a family boundary (60% AAI), which resulted in the proposed reclassification of some members of the family *Rhodobacteraceae* into a newly proposed family *Stappiaceae* fam. nov. A new genus boundary (70% AAI) is also proposed to delineate genera within the families. We also reclassified strains at the species level based on already recognized species boundaries (70% dDDH and 95% ANI; Goris et al., 2007).

3.3. Materials and Methods

3.3.1. Whole-genome sequences used in this study

Whole genome sequences submitted to the GenBank database (National Center for Biotechnology Information) prior to September 1, 2016 were collected for members of the family *Rhodobacteraceae* (Table C.1). In total, 313 *Rhodobacteraceae* genomes were obtained comprising 70 genera and 127 species, excluding those with only "sp." designations for unknown species (Table C.1). The genomes of three *Agrobacterium tumefaciens* strains were also used to serve as outgroup for phylogenetic analyses (Ettema and Andersson, 2009). Plasmid sequences were excluded from analyses.

3.3.2. Genome annotation and completeness

To ensure all open-reading frames were predicted with the same methodology for all organisms, all 316 genomes were re-annotated using RAST 2.0 (Rapid Annotations Using Subsystems Technology; Aziz et al., 2008). Completeness of the genomes was then determined by detecting the presence or absence of a subset of housekeeping genes using BLAST 2.5.0

(Altschul et al., 1990). One hundred eight single-copy core housekeeping genes were used to determine completeness (Table C.2; Luo and Moran, 2014), which was reported as the percentage of 108 housekeeping genes present in each genome (Table C.1). Protein sequences of the 108 housekeeping genes from *Dinoroseobacter shibae* DFL 12 were used as reference (Wagner-Döbler et al., 2010). Genomes that were at least 95% complete, a total of 290 genomes, were retained and used for further analyses.

3.3.3. 16S rRNA and core genome phylogenetic analyses

A single copy of the full-length 16S rRNA gene sequence was randomly selected from each of the re-annotated genomes. The sequences were then aligned with Clustal W 2.1 (Larkin et al., 2007). Poorly aligned positions were eliminated using GBlocks 0.91b (Castresana, 2000), and the final alignment (with 1,453 nucleotide positions) was used to reconstruct a 16S rRNA gene maximum-likelihood (ML) phylogenetic tree with RAxML 8.2.9 (Stamatakis, 2014). The GTR (general time reversible) nucleotide substitution model and gamma model of rate heterogeneity were used. Robustness of branching was estimated with 100 bootstrap replicates.

Orthologous protein-coding gene families were determined from the annotated genomes using the Bacterial Pan Genome Analysis (BPGA) tool 1.3.0 (Chaudhari et al., 2016), which employs USEARCH 9.2.64 (Edgar, 2010), and using 30% amino acid identity cutoff (Rost, 1999). The determined core genome, the set of genes present in all strains (Vernikos et al., 2015), was composed of 115 core gene families in total. For every core gene family, the amino acid sequences were aligned with Clustal W 2.1 (Larkin et al., 2007). The 115 core gene alignments were concatenated using Geneious 8.1.8 (Kearse et al., 2012). The final alignment

(with 55,414 amino acid positions) was used to reconstruct a core genome ML tree with RAxML 8.2.9 (Stamatakis, 2014).

3.3.4. Family, genus, and species delineations

Two metrics were used for family- and genus-level identification, AAI (Konstantinidis and Tiedje, 2005b) and POCP (Qin et al., 2014). Using amino acid sequences of all protein-coding genes, pairwise AAI was calculated with CompareM 0.0.21 (https://github.com/dparks1134/CompareM). The number of orthologous genes shared between two genomes, provided by CompareM, was then used to determine POCP. This was calculated as $[(2\times S)/(T1+T2)]\times 100\%$, where S represents the number of shared genes between genomes and T1 and T2 represent the total number of proteins in the two genomes being compared.

Genomes were also identified up to the species rank, based on their high average nucleotide identity (Konstantinidis and Tiedje, 2005a) and dDDH values (Meier-Kolthoff et al., 2013) against known type strains. Using whole-genome sequences, pairwise ANI was calculated with the ANI Calculator (Rodriguez-R and Konstantinidis, 2005) or JSpecies 1.2.1 (Richter and Rosselló-Móra, 2009), using default parameters. Meanwhile, dDDH was calculated using the Genome-to-Genome Distance Calculator 2.1 (GGDC; Meier-Kolthoff et al., 2013). Two genomes belonging to the same species would have an ANI of at least 95%, which corresponds to at least 70% DDH (Goris et al., 2007).

3.4. Results and Discussion

It was in the late 1980s that Woese (1987) first described the use of the 16S rRNA gene for phylogenetic studies. Since then, the gene has become a standard in bacterial taxonomic

classification (Tindall et al., 2010; Garrity, 2016). Discovery of marine roseobacters started around the same time when the use of the 16S rRNA gene in prokaryotic diversity studies was realized (Giovannoni et al., 1990; Ward et al., 1990; Shiba, 1991). Therefore, the majority of the species were classified based on 16S rRNA sequences, and this has resulted to several inconsistencies in nomenclature (Buchan et al., 2005; Wagner-Döbler and Biebl, 2006; Luo and Moran, 2014; Simon et al., 2017). For a systematic classification of strains, there is a need to re-examine and provide tangible standards for the delineation of genera and species within this group.

With the advent of next-generation sequencing technologies, taxonomy is now transitioning from being based on a single to a few genes to whole-genome sequences for identification (Konstantinidis and Tiedje, 2005a; b; Oren and Garrity, 2014; Varghese et al., 2015; Garrity, 2016). The use of whole-genome sequences in establishing species relationships is well established and used for modern species descriptions (Konstantinidis and Tiedje, 2005a; Goris et al., 2007; Tindall et al., 2010; Meier-Kolthoff et al., 2013; Varghese et al., 2015). However, genome-based classification beyond the species level is not fully utilized (Konstantinidis and Tiedje, 2005b). Here, we use various genome-based metrics, dDDH, ANI, AAI, and POCP, in identifying and reclassifying strains and establishing new genome-based genera and family boundaries for *Rhodobacteraceae*.

3.4.1. The 16S rRNA phylogeny results in a poorly resolved tree

To determine the impact of using the 16S rRNA gene as the main molecular marker for naming new species and genera, a phylogenetic analysis of the family *Rhodobacteraceae* was performed using this gene. We used full-length 16S rRNA sequences from the annotated

genomes, which are highly recommended for use in phylogenetic and taxonomic studies (Tindall et al., 2010). As expected, the 16S rRNA gene-based tree has poor resolution and low bootstrap support overall. This is even more evident when nodes with bootstrap support below 50% were collapsed, resulting in a poorly resolved tree backbone (Figure 3.1).

Multiple strains from the same genera do not cluster together in the tree (Figure 3.1). For example, some species from the genera *Leisingera*, *Phaeobacter*, *Roseovarius*, and *Sulfitobacter* are polyphyletic or paraphyletic. Multiple strains identified only up to the genus level ("sp." designations only) also do not cluster with the majority of the strains of their respective genera (e.g., *Ruegeria*, *Thalassobius*, etc.), indicating possible misclassification of these strains. The polyphyly and paraphyly exhibited in Figure 3.1 may be the consequence of a poorly resolved tree, suggesting that the 16S rRNA gene is not a robust gene marker for identification. It is important to note that the use of 16S rRNA gene sequences alone should not describe a taxon, but it may provide the first indication that the isolate could belong to a novel taxon (Tindall et al., 2010). It is also necessary to use full-length sequences during analysis (Tindall et al., 2010), especially if sequence identities are used as an indication of strains belonging to the same species or genus. The roseobacter clade is a good example, for its currently accepted basic definition is based on 16S rRNA sequence identity (at least 89%; Buchan et al., 2005). Since the majority of universal primers only target and amplify smaller regions of the 16S rRNA gene during PCR (Klindworth et al., 2013), additional steps such as amplifying and aligning sequences of different overlapping regions are needed to obtain the full gene. On the other hand, availability of whole genomes can readily provide full-length 16S rRNA gene sequences.

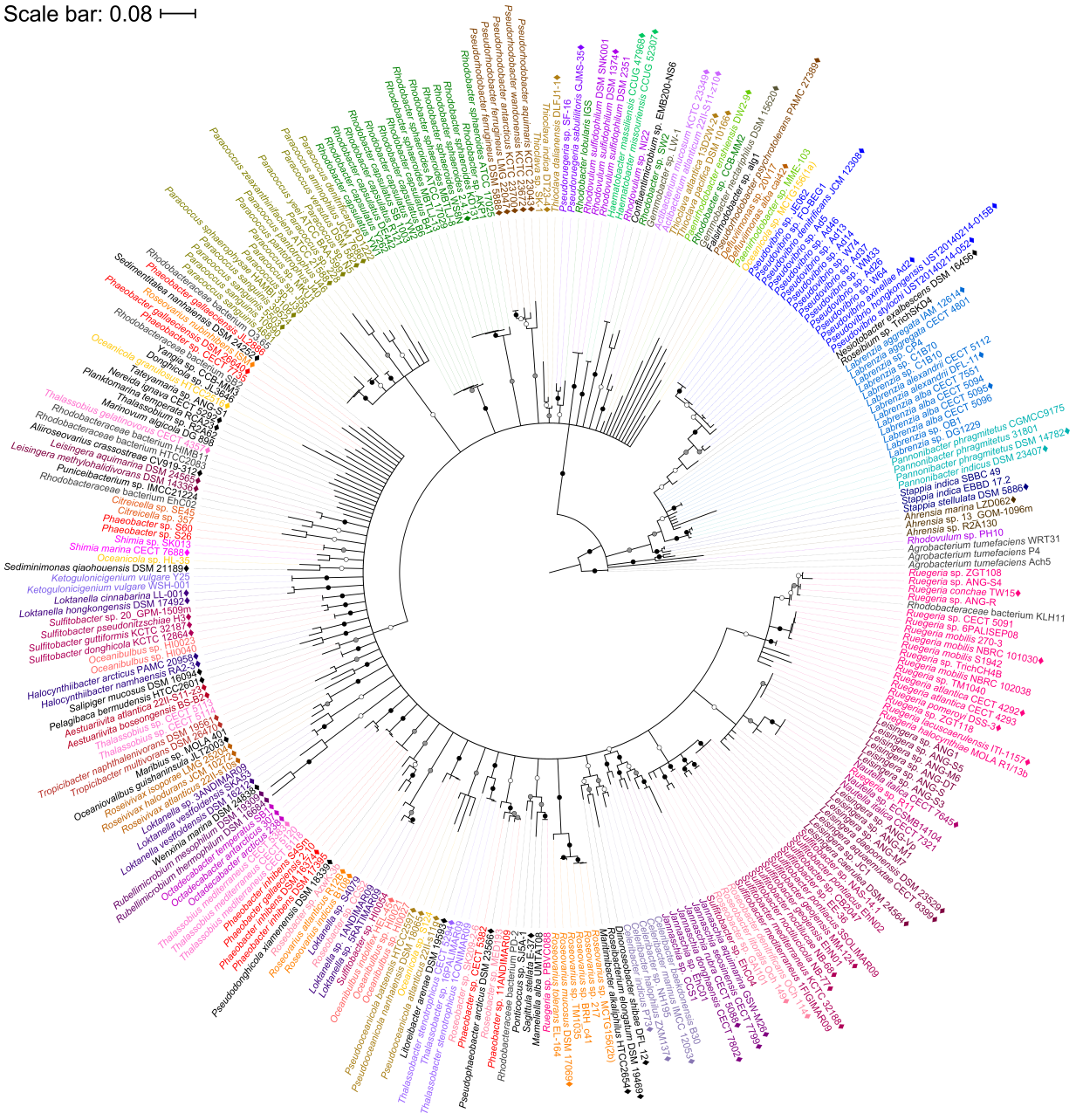

Figure 3.1. The phylogenetic relationship of *Rhodobacteraceae* based on 16S rRNA gene sequences.

The tree was constructed from the alignment of 16S rRNA sequences (1,493 nucleotide positions). Nodes with 50% or less bootstrap support are collapsed. Bootstrap support is indicated on the nodes as black circles (≥ 95%), gray circles (≥ 70%), or white circles (> 50%). Bar, 0.08 nucleotide substitutions per site. Excluding *A. tumefaciens* (outgroup), genera with at least two members are colored, genera with single members are black, and strains with no genus designation (e.g. *Rhodobacteraceae* bacterium) are gray. Diamonds after strain names indicate type strains.

3.4.2. A more robust phylogenetic tree is obtained from the core genome

The use of conserved protein-coding genes such as housekeeping genes can provide higher resolution than 16S rRNA gene sequences. Initially using only seven protein-coding genes, the concept of MLSA has been expanded to genome sequence data (Maiden et al., 2013), such as the core genome that is present in all strains of a given taxonomic group (Vernikos et al., 2015). In a previous study, a core genome phylogeny of the roseobacter clade was reconstructed using 108 core housekeeping genes (Luo and Moran, 2014). To determine the phylogenetic affiliation of this clade within the family *Rhodobacteraceae*, Simon et al. (2017) reconstructed the phylogeny of the entire family using 208 core genes from 106 strains. In this study, the core genome phylogeny of the family was reconstructed using 115 core genes from a dataset almost three times as large (290 strains including the *A. tumefaciens* outgroup) as the previous work (Simon et al., 2017), providing a more complete picture of the phylogenetic framework of *Rhodobacteraceae*. In contrast to the 16S rRNA gene tree, the core genome tree showed higher resolution with a more robust bootstrap support overall (≥ 95%) and a well-resolved backbone even within genera (Figure 3.2). A major limitation of this approach is the quality of assembled genomes used. Although we initially obtained 316 genomes from GenBank, we subsequently excluded 26 genomes after assessing their completeness (< 95% complete; Table C.1).

A significant amount of polyphyly and paraphyly from the 16S rRNA gene phylogeny (Figure 3.1) was immediately resolved by the core genome phylogeny (Figure 3.2), such as with the genus *Leisingera*, although several other genera are still polyphyletic or paraphyletic (*Phaeobacter*, *Sulifotbacter*, *Roseovarius*, *Ruegeria*). Surprisingly, members of the *Ruegeria* and *Rhodobacter* genera were split into two separate clades each. It is possible that several strains of *Ruegeria* and *Rhodobacter* have been misclassified based on 16S rRNA gene sequences, since

members of the same genus should be monophyletic (Rosselló-Móra and Amann, 2001). Thus, there is a need to review their classification. For one, *Ruegeria* sp. R11 (Case et al., 2011) has recently been reclassified as *Nautella* sp. R11 based on 16S rRNA sequence identity alone (Fernandes et al., 2011), and our core genome analysis verifies its phylogenetic placement with the rest of the strains of the genus *Nautella* (Figure 3.2). However, its name in the GenBank database has not been updated. The genus *Sulfitobacter* also shows paraphyly, with multiple strains of the genera *Oceanibulbus* and *Roseobacter* clustering together with *Sulfitobacter* (Figure 3.2). This includes *R. denitrificans* Och 114^T and *R. litoralis* Och 149^T, the first roseobacters to be officially described (Shiba, 1991).

Polyphyly and paraphyly are also found at the species level. For instance, our analysis includes three strains each of *Phaeobacter gallaeciensis* and *Phaeobacter inihibens*. One of the strains, submitted to GenBank as *P. gallaeciensis* JL2886 (Fu et al., 2016), clusters separately from type strain *P. gallaeciensis* DSM 26440^T (Martens et al., 2006) and the rest of the phaeobacters, but is instead monophyletic with members of the genus *Leisingera* (Figure 3.2). The genus *Phaeobacter*, an intensively studied group mainly due to their ability to produce secondary metabolites (Martens et al., 2007; Seyedsayamdost et al., 2011), poses some controversy. Some researchers have pointed out misnaming of the frequently used *P. inhibens* DSM 17395 as *P. gallaeciensis* initially, which could have been cause by a possible error in the deposition of this strain through various culture collections (Buddruhs et al., 2013; Breider et al., 2014; Frank et al., 2014). The core genome tree shows affinity of both *P. inhibens* DSM 17395 as well as *P. gallaeciensis* 2.10 with the *P. inhibens* cluster (Figure 3.2).

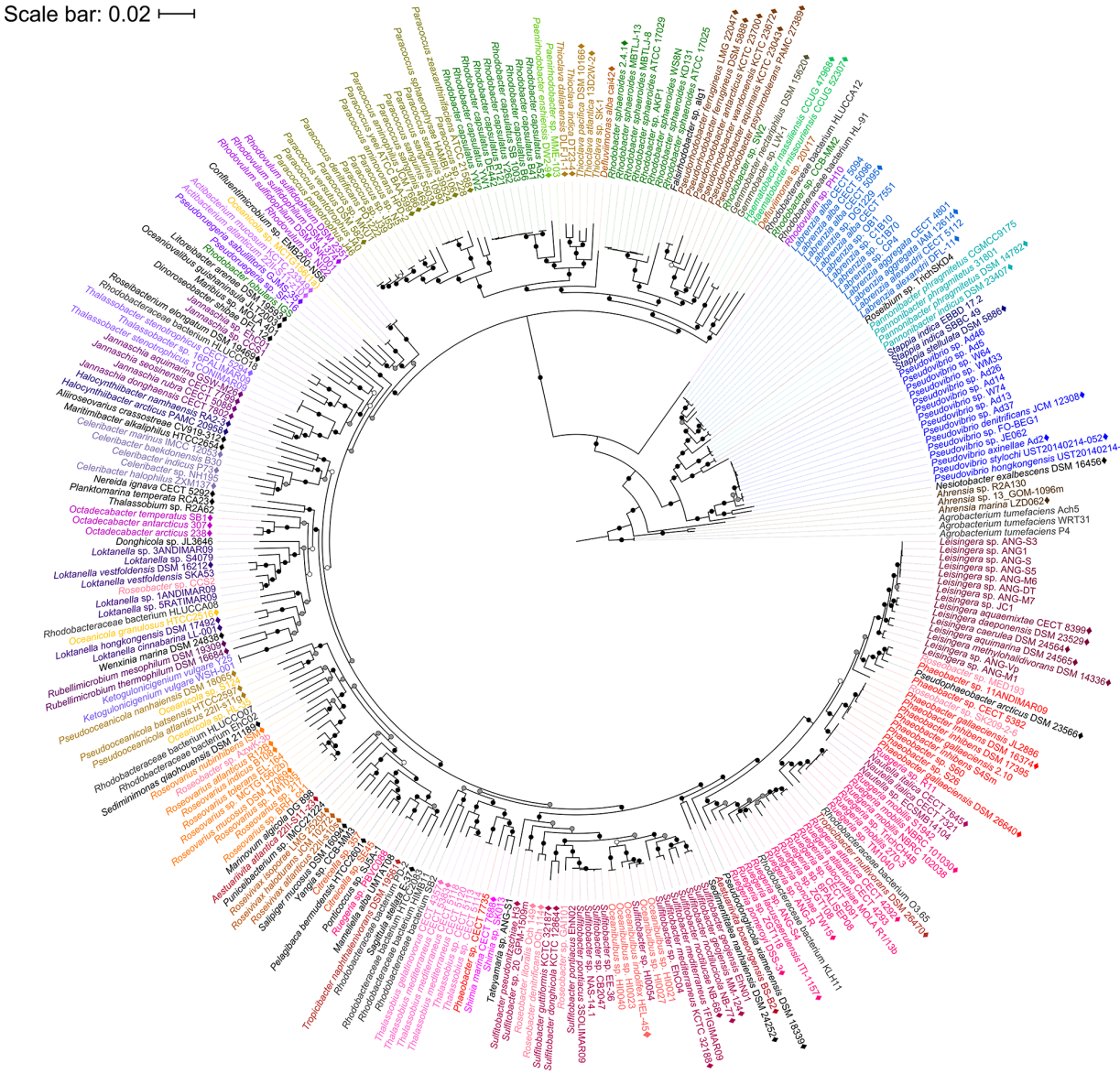

Figure 3.2. The phylogenetic relationship of *Rhodobacteraceae* based on the core genome.

The tree was constructed from the concatenated alignment of 115 single-copy, protein-coding core genes (55,414 amino acid positions). Nodes with 50% or less bootstrap support are collapsed. Bootstrap support is indicated on the nodes as black circles (≥ 95%), gray circles (≥ 70%), or white circles (> 50%). Bar, 0.02 amino acid substitutions per site. Excluding *A. tumefaciens* (outgroup), genera with at least two members are colored, genera with single members are black, and strains with no genus designation (e.g. *Rhodobacteraceae* bacterium) are gray. Diamonds after strain names indicate type strains.

Overall, our core genome phylogenetic analysis significantly improves on the 16S rRNA gene phylogeny by providing a tree with a robust bootstrap support and well-resolved backbone.

This shows placements of strains in the context of a phylogeny based on a large number of protein-coding genes. However, using their current names, polyphyly and paraphyly for multiple genera and species remains.

3.4.3. A new family boundary for the order *Rhodobacterales*

Although ANI is widely used to compare closely related strains using nucleotide sequences, comparisons using amino acid sequences is recommended for distantly related organisms because resolution is progressively lost at the nucleotide level (Rodriguez-R and Konstantinidis, 2014). Therefore, to determine family and genus boundaries, we employed metrics that utilize amino acid sequences (protein-coding genes) for comparisons, AAI (Konstantinidis and Tiedje, 2005b) and POCP (Qin et al., 2014).

The core genome phylogenetic tree (Figure 3.2) presents three well-supported major clades (clade 1, n = 187; clade 2, n = 59; clade 3, n = 37; Figure C.1), with clade 1 constituting the roseobacter clade (Buchan et al., 2005; Luo and Moran, 2014). AAI and POCP were determined for inter- and intra-clade pairwise comparisons (41,905 comparisons; Figure 3.3). Pairwise comparisons within the three major clades always show AAI above 60% (Figure 3.3). Comparisons between organisms found in clades 1 and 2 (Figure C.1) were also always above 60% AAI (Figure 3.3B). Comparisons between organisms from clades 1 and 3 or clades 2 and 3 (Figure C.1) were always below 60% AAI (Figure 3.3B). From a previous analysis of 410 genomes of various taxa, the determined boundaries would fall within the range 45–60% AAI (for order) and 60–80% AAI (for genus), with the majority of comparisons falling at 53% and 68–70% AAI, respectively (Luo et al., 2014). Surprisingly, Luo et al. (2014) did not perform comparisons within families, but should be expected to fall between 53–70% AAI based on their

comparisons within the order and genus levels. Our results indicate a distinct 60% AAI cutoff, which we propose as the new family boundary, where clades 1 and 2 would belong to the same family and clade 3 is part of a different family. On the other hand, our POCP analysis did not show clear boundaries between clades (Figure 3.3 and Figure C.2).

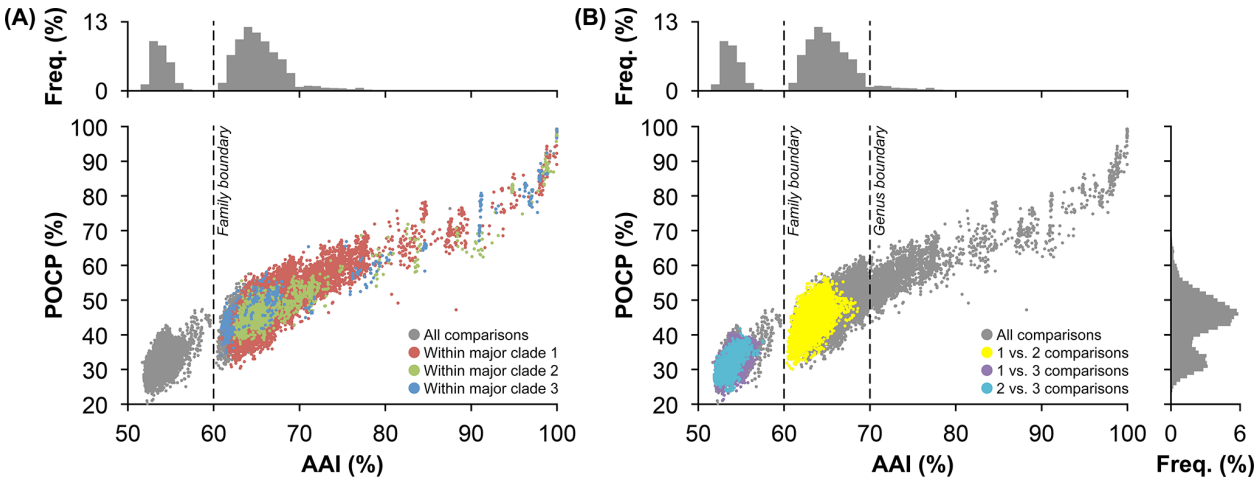

Figure 3.3. Plot of POCP vs. AAI comparisons within and between major clades of the core genome phylogenetic tree.

POCP vs. AAI was plotted for 41,905 pairwise comparisons (gray dots) and their frequencies (freq.) are indicated as bar graphs above (AAI) or on the right (POCP) of each plot. (A) Intra-clade pairwise comparisons are plotted after the core genome phylogenetic tree was split into three major clades: clade 1, 187 members (red); clade 2, 59 members (green); clade 3, 37 members (blue). (B) Plot for inter-clade comparisons: 1 vs. 2 (yellow), 1 vs. 3 (purple), 2 vs. 3 (blue). Broken lines indicate the proposed AAI boundaries for family (60%) and genus (70%) delineations.

To determine if the 60% AAI boundary is applicable to other families outside of *Rhodobacterales* but still within the *Alphaproteobacteria*, we reconstructed the core genome phylogeny and calculated the AAI between pairs for three strains each from five families (*Beijerinckiaceae*, *Bradyrhizobiaceae*, *Hyphomicrobiaceae*, *Phyllobacteriaceae*, and *Rhizobiaceae*) in the order *Rhizobiales* (Table C.3 and Figure C.3; Kuykendall, 2005). Our core genome phylogeny clusters members of three out of five families into monophyletic clades, whereas members of *Hyphomicrobiaceae* and *Phyllobacteriaceae* are polyphyletic (Figure

C.3A). Pairwise comparisons within members of the same family exhibit at least 60% AAI in four out of five families (except *Hyphomicrobiaceae*; Figure C.3B), and the majority of inter-family comparisons (*Hyphomicrobiaceae* excluded) shows AAI below 60% (Figure C.3C). This indicates that 60% AAI can serve as a potential family boundary even outside of our group of interest. Also, comparisons between members of *Phyllobacteriaceae* and *Rhizobiaceae* are all above 60% AAI, suggesting that they could be reclassified into one family based on this cutoff. A more detailed analysis of all available genomes should be performed for members of these families to determine if such reclassification is warranted.

Based on the new family boundary of 60% AAI, we propose clades 1 and 2 to be retained under family *Rhodobacteraceae*, with *Rhodobacter* as the type genus (Imhoff et al., 1984; Figure 3.4). On the other hand, clade 3 should be reclassified to a new family, with the newly proposed name *Stappiaceae* fam. nov., after the type genus *Stappia*, the first genus described in this new family (Rüger and Höfle, 1992; Uchino et al., 1998). In addition, a few strains were excluded from both families. The strains originally identified as *Rhodovulum* sp. PH10 and *Ahrensia* sp. R2A130 did not exhibit at least 60% AAI when compared with any of the other strains, and are therefore reclassified as *Rhodobacterales* bacterium PH10 and R2A130, respectively (Figure 3.4 and Table C.4). *Ahrensia marina* LZD062^{T} and *Ahrensia* sp. 13_GOM-1096m potentially fall into their own distinct family as well. These four strains could potentially be part of yet to be determined families, but there are currently not enough available genomes within these groups to perform further analyses. Interestingly, all our inter-clade comparisons exhibited at least 52% AAI, which could be a potential order boundary, although this needs to be verified in the future by conducting comparisons between orders within class *Alphaproteobacteria*.

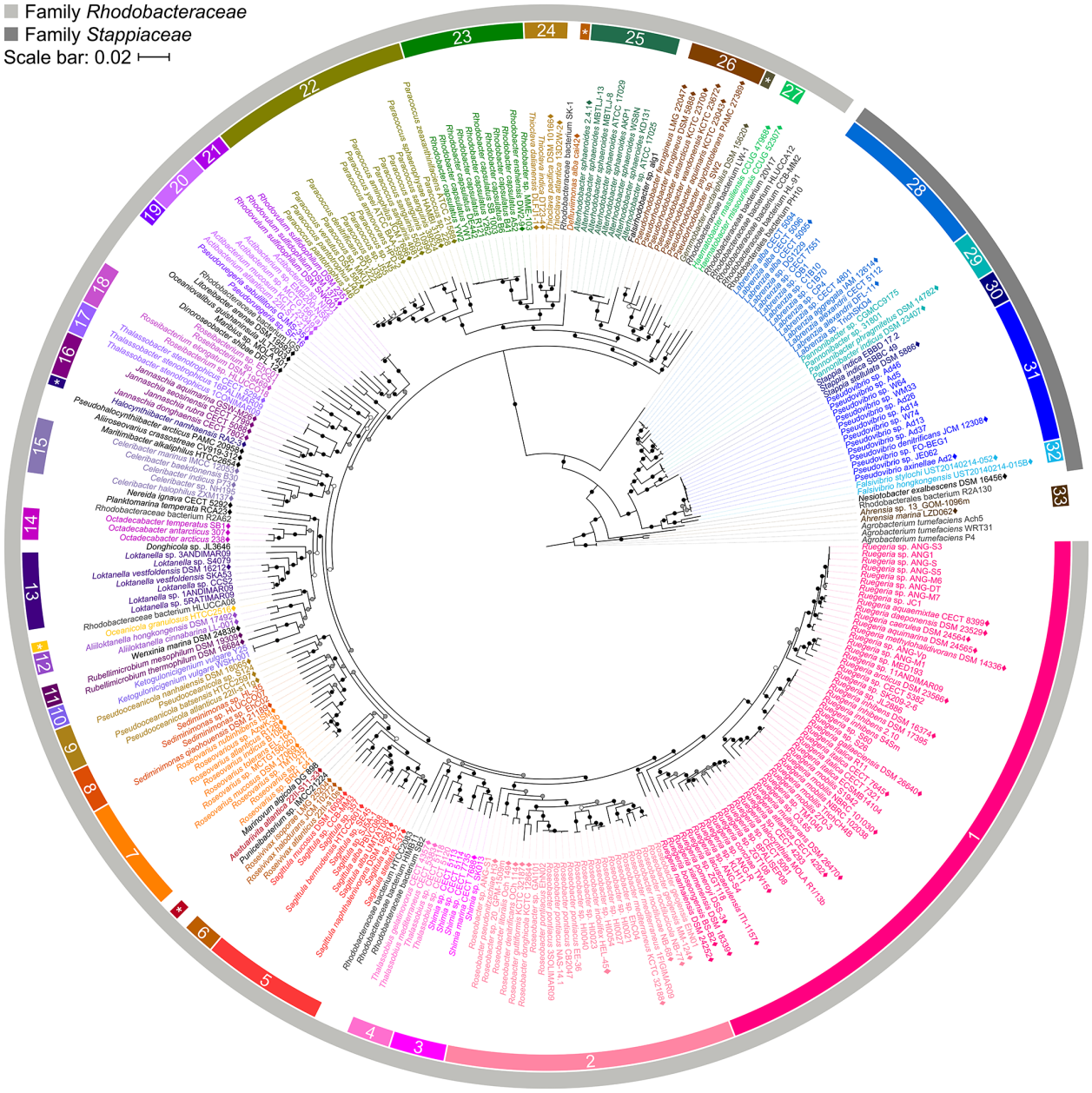

Figure 3.4. The phylogenetic relationship of *Rhodobacteraceae* and *Stappiaceae* fam. nov. based on the core genome.

(Figure legend continued on the next page)

Figure 3.4. The phylogenetic relationship of *Rhodobacteraceae* and *Stappiaceae* fam. nov. based on the core genome. The core genome phylogenetic tree (Figure 3.2) was revised to indicate reclassified strains. Members of *Rhodobacteraceae* or *Stappiaceae* fam. nov. are indicated by the outermost gray circles (light gray and dark gray, respectively). Excluding *A. tumefaciens* (outgroup), genera with at least two members are colored and indicated by the outer coloured bars also of the same color; genera with single members are black, and strains with no genus designation (e.g. *Rhodobacteraceae* or *Rhodobacterales* bacterium) are gray. Colored strains that are marked by asterisks indicate genera with at least two members previously but have been reduced to only one member. Genera designations are as follows: 1, *Ruegeria*; 2, *Roseobacter*; 3, *Shimia*; 4, *Thalassobius*; 5, *Sagittula*; 6, *Roseivivax*; 7, *Roseovarius*; 8, *Sediminimonas*; 9, *Pseudooceanicola*; 10, *Ketogulonicigenium*; 11, *Rubellimicrobium*; 12, *Aliiloktanella* gen. nov.; 13, *Loktanella*; 14, *Octadecabacter*; 15, *Celeribacter*; 16, *Jannaschia*; 17, *Thalassobacter*; 18, *Roseibacterium*; 19, *Pseudoruegeria*; 20, *Actibacterium*; 21, *Rhodovulum*; 22, *Paracoccus*; 23, *Rhodobacter*; 24, *Thioclava*; 25, *Alterhodobacter* gen. nov.; 26, *Pseudorhodobacter*; 27, *Haematobacter*; 28, *Labrenzia*; 29, *Pannonibacter*; 30, *Stappia*; 31, *Pseudovibrio*; 32, *Falsivibrio* gen. nov.; 33, *Ahrensia*. Diamonds after strain names indicate type strains.

3.4.4. A new genus boundary within the families *Rhodobacteraceae* and *Stappiaceae* fam. nov.

Thirteen well-supported monophyletic clades representing different genera were found in the core genome tree (Figure C.1), and pairwise comparisons of organisms within and between each of these clades were performed to calculate AAI and POCP. The majority of pairwise comparisons between organisms belonging to the same clade resulted in at least 70% AAI, and those between clades fell below 70% (Figure C.4). For example, all comparisons within and between the clades *Leisingera*, *Phaeobacter*, *Ruegeria* 1 and *Ruegeria* 2 (not all *Ruegeria* strains are monophyletic and they were therefore split into two clades; Figure C.1), were above 70% AAI (Figure C.5). However, any comparison of these clades against their sister taxon, *Sulfitobacter*, fell below 70% AAI (but above 60%). Only a very small number of comparisons of strains belonging to the same clade yielded values lower than 70% AAI (Figure C.4). This includes comparisons against *Pseudovibrio hongkongensis* UST20140214-015B^T, *Pseudovibrio stylochi* UST20140214-052^T, *Thioclava* sp. SK-1, *Paracoccus zeaxanthinifaciens* ATCC 21588^T, and *Paracoccus* sp. 228. As mentioned previously, the genus boundary proposed by Luo et al. (2014) would fall in the 68–70% AAI range. We propose the upper limit of 70% AAI as the new

genus boundary for this group. For the majority of our 41,905 comparisons, the 70% AAI cutoff reassigns several misclassified strains and removes any paraphyly within genera (Figure 3.4). On the other hand, we could not obtain a clear cutoff value for the POCP within and between clades (Figure 3.3 and Figure C.2). For a clade containing multiple genera that needs to be combined into a single genus, the name of the first genus described in that clade was chosen.

As a consequence, *Leisingera*, *Phaeobacter*, *Nautella*, and *Ruegeria* should be combined into genus *Ruegeria* (Figure 3.4 and Table C.4), chosen because it is the oldest described genus in this group (Rüger and Höfle, 1992; Uchino et al., 1998). This also includes the reclassification of *Tropicibacter multivorans* DSM 26470^T (Lucena et al., 2012), *Pseudodonghicola xiamenensis* DSM 18339^T (Hameed et al., 2014), *Aestuariivita boseongensis* BS-B2^T (Park et al., 2014), and *Sedimentitalea nanhaiensis* DSM 24252^T (Breider et al., 2014) into *Ruegeria multivorans* sp. nov., *Ruegeria xiamenensis* sp. nov., *Ruegria boseongensis* sp. nov., and *Ruegeria nanhaiensis* sp. nov., respectively (Figure 3.4 and Table C.4). Meanwhile, all members of the *Sulfitobacter-Oceanibulbus-Rosoebacter* clade and *Tateyamaria* sp. ANG-S1 will be reassigned to the genus *Rosoebacter* (Shiba, 1991; Figure 3.4 and Table C.4). Selecting the lower limit determined by Luo et al. (2014) as the genus boundary (68% AAI) would reassign members of this clade into *Ruegeria* (Figure C.5), which would violate the monophyly rule (Rosselló-Móra and Amann, 2001).

There are polyphyletic genera needing to be reclassified into different genera, which involve the proposal of new names (Figure 3.4 and Table C.4). For example, *Loktanella cinnabarina* LL-001^T (Tsubouchi et al., 2013) and *Loktanella hongkongensis* DSM 17492^T (Lau et al., 2004) cluster together (AAI = 87%) but separately from *Loktanella vestfoldensis* DSM 16212^T (Van Trappen et al., 2004) and other *Loktanella* strains (AAI = 67–68%), and only one of

these clusters could keep the genus name. As such, we propose the former two strains to be renamed *Aliiloktanella cinnabarina* gen. nov., sp. nov. and *Aliiloktanella hongkongensis* gen. nov., sp. nov., since *L. vestfoldensis* was described before the other two (Van Trappen et al., 2004). In addition, *Halocynthiibacter arcticus* PAMC 20958^T (Baek et al., 2015) only shows 65.8% AAI against *Halocynthiibacter namhaensis* RA2-3^T (Kim et al., 2014b). Even though both are monophyletic in the core genome tree, we propose the former to be reclassified into *Pseudohalocynthiibacter arcticus* gen. nov., sp. nov. based on AAI and that it was described later than *H. namhaensis* (Kim et al., 2014b; Baek et al., 2015).

Strikingly, one of the most studied genera of *Rhodobacteraceae*, *Rhodobacter*, is split into two clades, *R. capsulatus* and *R. sphaeroides*, which do not form a monophyletic group (Figure 3.3). Comparisons between the clades only show 67–68% AAI. Since *R. capsulatus* was described before *R. sphaeroides* and is the type species for the genus (Imhoff et al., 1984), the former should retain the genus name (Figure 3.4 and Table C.4). The *Rhodobacter* clade should also include the *Paenirhodobacter* strains, *P. enshiensis* DW2-9^T (Wang et al., 2014) and *Paenirhodobacter* sp. MME-103 (AAI = 71–75%), renaming them into *Rhodobacter enshiensis* sp. nov. and *Rhodobacter* sp., respectively. We therefore propose the reclassification of the *R. sphaeroides* clade into *Alterhodobacter sphaeroides* gen. nov., sp. nov. Lastly, *Pseudovibrio stylochi* UST20140214-052^T (Zhang et al., 2016) and *Pseudovibrio hongkongensis* UST20140214-015B^T (Xu et al., 2015), which share 71% AAI, were reassigned into *Falsivibrio sylochi* gen. nov., sp. nov. and *Falsivibrio hongkongensis* gen. nov., sp. nov. because of only 66–67% AAI against other pseudovibrios (Figure 3.4, Figure C.4, and Table C.4).

For borderline cases, such as comparisons of *Paracoccus zeaxanthinifaciens* ATCC 21588^T (Berry et al., 2003) or *Paracoccus* sp. 228 against other strains of *Paracoccus* (AAI =

69–71%; Figure C.4), we retained the names of both strains (Figure 3.4), since our results based on AAI were inconclusive. Such cases can be further verified by other tests (e.g., phenotypic) or whole genome-based analyses upon the availability of more genomes from these species.

Setting standards for cutoffs for designating ranks will always be somewhat arbitrary (Konstantinidis and Tiedje, 2005b), and the family and genus boundaries proposed in this study may or may not apply to other families and genera. For example, no clear boundaries were called for both family and genus from the POCP analysis (Figure 3.3 and Figure C.2), but Pannekoek et al. (2016) were able to apply the 50% POCP genus boundary to their study of the order *Chlamydiales* and proposed the unification of *Chlamydia* and *Chlamydophila* into *Chlamydia*. Thus, to set organism-specific standards and to determine if the same boundaries apply or not, a similar whole genome-based analysis should be performed for other groups of organisms if or when sufficient genome sequences are available.

3.4.5. Species delineations within the families *Rhodobacteraceae* and *Stappiaceae* fam. nov.

After establishing family and genus boundaries for these groups and resolving most of the paraphyly within genera in the core genome phylogenetic tree (Figure 3.4), it was also imperative to identify members up to the species level. Nucleotide-level comparisons of genome sequences were performed between all strains belonging to the same genus. Using established species boundaries, two organisms were considered to belong to the same species if they had at least 70% dDDH and 95% ANI (Konstantinidis and Tiedje, 2005a; Goris et al., 2007; Richter and Rosselló-Móra, 2009; Meier-Kolthoff et al., 2013).

In the genus *Ruegeria*, forming the largest clade in our dataset after genera reclassifications, several strains (ANG-S3, ANG1, ANG-S, ANG-S5, ANG-M6, ANG-DT;

formerly *Leiseingera*), should be part of a yet to be determined species based on the high dDDH (81–100%) and ANI (98–100%) values among them. This species will have to be officially described in the future. On the other hand, this group of strains only have 38–39% dDDH and 89–90% ANI against *Ruegeria* sp. ANG-M7 and JC1 (also formerly *Leiseingera*), so the latter two strains should not be part of this species. Meanwhile, strains CECT 7645^T and CECT 7321 are joined by strains R11 (formerly *Ruegeria* sp. R11) and ECSMB14104 (formerly *Nautella* sp. ECSMB14104) in the species *Ruegeria italica* sp. nov. (formerly *Nautella italica*) based on 84–98% dDDH and 98–100% ANI.

In some cases, ANI and dDDH values are close to the cutoff for species. The core genome phylogeny, as well as other characteristics, is necessary to determine species-level classification. For example, *Ruegeria* sp. CECT 4293 was initially identified as *Ruegeria atlantica*. Although ANI between this strain and the type strain CECT 4292^T is 95%, dDDH is only 60–66% (with 95% confidence; Meier-Kolthoff et al., 2013), which is well below the dDDH cutoff for species. Therefore, we have currently designated CECT 4293 as an unidentified *Ruegeria* species. On the other hand, comparisons between *Ruegeria inhibens* sp. nov. strains [DSM 16374^T, DSM 17395, S4Sm (formerly *P. inhibens*), and 2.10 (formerly *P. gallaeciensis*)], show 73–80% dDDH and 96–98% ANI, except between DSM 17395 and S4Sm which is only 68.5% dDDH. The dDDH range by the GGDC for this comparison is 66–71%, and S4Sm was therefore classified under the same species. Throughout the core genome phylogenetic tree, our analysis revealed several unidentified strains that should be part of already known species, identified strains that should be reclassified into their own species, and unidentified strains that should belong to species that are yet to be officially described (Figure 3.4 and Table C.4).

3.4.6. Rules on classification and reclassification of strains

It seems necessary for a genome of a strain to be sequenced before it is attributed an official taxonomic classification (Whitman, 2011). Based on this assumption, several criteria need to be applied when performing taxonomy based on genome sequence comparison (Figure 3.5). The AAI boundaries proposed in this study (for family and genus) may or may not apply to other families and genera. An exhaustive analysis of the distribution of AAI (and the POCP) values between organisms of interest should be performed to determine which cutoff can be used to assign family and genera to those organisms. For the family *Rhodobacteraceae* and the newly proposed *Stappiaceae* fam. nov. the following criteria were applied to obtain a robust and systematic taxonomic classification:

1. No strains belonging to the same family, genus, or species should exist as polyphyletic or paraphyletic (i.e., they should be monophyletic) in the core genome tree (Rosselló-Móra and Amann, 2001), and monophyly should be supported by robust bootstrap values (at least 70%; Hillis and Bull, 1993).
2. Strains for which pairwise comparison shows at least 60% AAI are assigned to the same family.
3. Strains for which pairwise comparison shows at least 70% AAI are assigned to the same genus.
4. Strains for which pairwise comparison shows at least 70% dDDH and 95% ANI are assigned to the same species.
5. Monophyly is necessary and takes precedence over other criteria (Rosselló-Móra and Amann, 2001).

6. For two species that are assigned to different genera but should belong to the same genus, the name for the genus that was first described officially (based on publication record) takes precedence over newer names. The same is true for two strains that are assigned to different species but should be the same species.

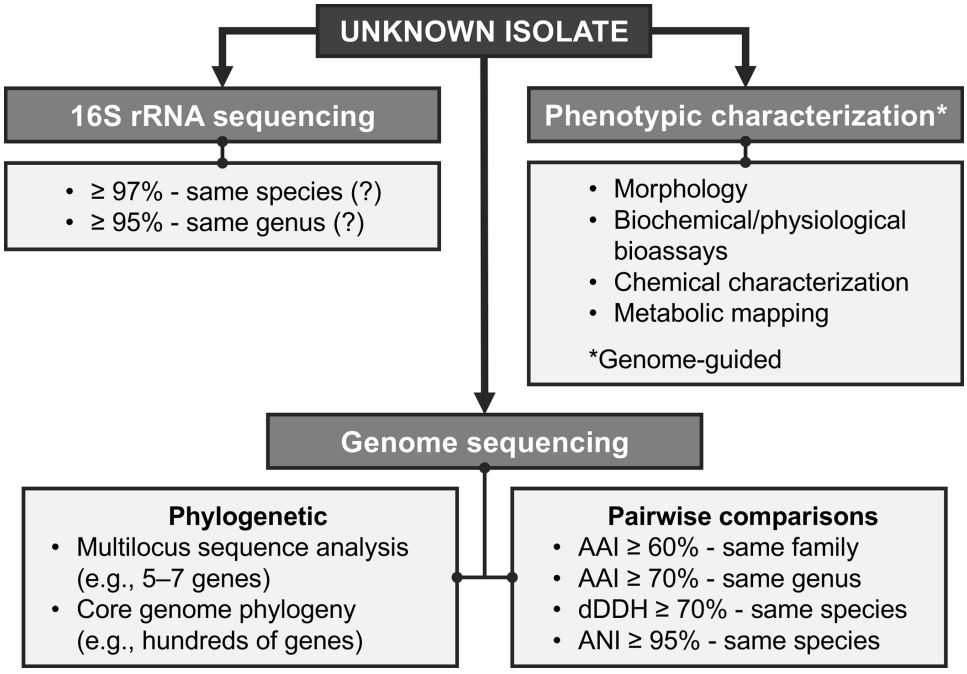

Figure 3.5. General flow chart of the polyphasic identification of microorganisms.

Given an unknown isolate, 16S rRNA sequencing can serve as a preliminary test to infer if the isolate is part of a known species/genus or can be classified into a novel species/genus. Genome sequencing will provide a set of protein-coding housekeeping genes for multilocus sequence analysis and the establishment of phylogeny based on a few or hundreds of genes (e.g., core genome). Phenotypic characteristics will also allow for the inclusion of the isolate with an established taxon based on common characteristics or the establishment of a novel taxon based on distinguishing characteristics from other taxa. Phenotypic characterization should be genome-guided. That is, regardless of phenotype, inclusion of an isolate to a known species should be supported by phylogenetic analysis (i.e., monophyly) and whole genome-based comparison metrics such as average amino acid identity (AAI), *in silico* (digital) DNA-DNA hybridization (dDDH), and average nucleotide identity (ANI).

3.5. Conclusion

Our study highlights several issues with the taxonomic classification of bacterial strains, specifically within the families *Rhodobacteraceae* and the newly proposed *Stappiaceae* fam.

nov. of the order *Rhodobacterales*. Overall, our whole genome-based analysis removed polyphyly and paraphyly within the original 70 genera by consolidating or splitting them based on the 70% AAI genus boundary, which reduced the number of genera to 53. On the other hand, the number of species was not significantly reduced. From 127 previously described species (excluding *Rhodobacteraceae* bacterium and "sp." only designations), reclassifications resulted in 124 species (Table C.4).

The role of whole-genome sequencing (WGS) and analysis is becoming more and more prominent and significant in taxonomy and should be included in every official species description (Whitman, 2011). Of the 114 type strains included in this study, 79 were described in official species descriptions from 2005, when next-generation sequencing technologies revolutionized DNA sequencing (Mardis, 2013). All 79 publications relied at least on 16S rRNA sequences to justify the creation of new taxa, whereas only 17 of these publications incorporated whole-genome analysis. These 17 publications were from 2012 onward, suggesting that WGS is only being incorporated recently into official species descriptions. A polyphasic approach to species identification is still necessary and should be the standard. The genotypic boundaries we propose here for *Rhodobacteraceae* and *Stappiaceae* fam. nov. cannot be taken as absolute unless they agree with the whole phenotypic and phylogenetic information.

3.6. References

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *Journal of Bacteriology* 186(9)**,** 2629-2635.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215(3)**,** 403-410.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9**,** 75.

Baek, K., Lee, Y.M., Shin, S.C., Hwang, K., Hwang, C.Y., Hong, S.G., et al. (2015). *Halocynthiibacter arcticus* sp. nov., isolated from Arctic marine sediment. *International Journal of Systematic and Evolutionary Microbiology* 65(11)**,** 3861-3865.

Berry, A., Janssens, D., Hümbelin, M., Jore, J.P., Hoste, B., Cleenwerck, I., et al. (2003). *Paracoccus zeaxanthinifaciens* sp. nov., a zeaxanthin-producing bacterium. *International Journal of Systematic and Evolutionary Microbiology* 53(1)**,** 231-238.

Breider, S., Scheuner, C., Schumann, P., Fiebig, A., Petersen, J., Pradella, S., et al. (2014). Genome-scale data suggest reclassifications in the *Leisingera-Phaeobacter* cluster including proposals for *Sedimentitalea* gen. nov. and *Pseudophaeobacter* gen. nov. *Frontiers in Microbiology* 5**,** 416.

Buchan, A., González, J.M., and Moran, M.A. (2005). Overview of the marine *Roseobacter* lineage. *Applied and Environmental Microbiology* 71(10)**,** 5665-5677.

Buddruhs, N., Pradella, S., Göker, M., Päuker, O., Pukall, R., Spröer, C., et al. (2013). Molecular and phenotypic analyses reveal the non-identity of the *Phaeobacter gallaeciensis* type strain deposits CIP 105210T and DSM 17395. *International Journal of Systematic and Evolutionary Microbiology* 63(11)**,** 4340-4349.

Case, R.J., Longford, S.R., Campbell, A.H., Low, A., Tujula, N., Steinberg, P.D., et al. (2011). Temperature induced bacterial virulence and bleaching disease in a chemically defended marine macroalga. *Environmental Microbiology* 13(2)**,** 529-537.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17(4)**,** 540-552.

Chaudhari, N.M., Gupta, V.K., and Dutta, C. (2016). BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports* 6**,** 24373.

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

Ereshefsky, M. (2010). Microbiology and the species problem. *Biology & Philosophy* 25(4)**,** 553-568.

Ettema, T.J.G., and Andersson, S.G.E. (2009). The α-proteobacteria: the Darwin finches of the bacterial world. *Biology Letters* 5(3)**,** 429-432.

Fernandes, N., Case, R.J., Longford, S.R., Seyedsayamdost, M.R., Steinberg, P.D., Kjelleberg, S., et al. (2011). Genomes and virulence factors of novel bacterial pathogens causing bleaching disease in the marine red alga *Delisea pulchra*. *PLoS One* 6(12)**,** e27387.

Frank, O., Pradella, S., Rohde, M., Scheuner, C., Klenk, H.P., Göker, M., et al. (2014). Complete genome sequence of the *Phaeobacter gallaeciensis* type strain CIP 105210T (= DSM 26640T = BS107T). *Standards in Genomic Sciences* 9(3)**,** 914-932.

Fu, Y., Wang, R., Zhang, Z., and Jiao, N. (2016). Complete genome sequence of the D-amino acid catabolism bacterium *Phaeobacter* sp. strain JL2886, isolated from deep seawater of the South China Sea. *Genome Announcements* 4(5)**,** e00913-16.

Garrity, G.M. (2016). A new genomics-driven taxonomy of *Bacteria* and *Archaea*: are we there yet? *Journal of Clinical Microbiology* 54(8)**,** 1956-1963.

Garrity, G.M., Bell, J.A., and Lilburn, T. (2005). "Order III. *Rhodobacterales* ord. nov.," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part C: The Alpha-, Beta-, Delta-, and Epsilonproteobacteria,* eds. D.J. Brenner, N.R. Krieg & J.T. Staley. 2nd ed. (New York, NY, USA: Springer), 161-229.

Giovanni, S.J., and Rappé, M.S. (2000). "Evolution, diversity, and molecular ecology of marine prokaryotes," in *Microbial Ecology of the Oceans,* ed. D.L. Kichman. (New York, NY, USA: John Wiley & Sons, Inc.), 47-84.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and Field, K.G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345(6270)**,** 60-63.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57(1)**,** 81-91.

Hameed, A., Shahina, M., Lin, S.Y., Nakayan, P., Liu, Y.C., Lai, W.A., et al. (2014). *Youngimonas vesicularis* gen. nov., sp. nov., of the family *Rhodobacteraceae*, isolated from surface seawater, reclassification of *Donghicola xiamenensis* Tan et al. 2009 as *Pseudodonghicola xiamenensis* gen. nov., comb. nov. and emended description of the genus *Donghicola* Yoon et al. 2007. *International Journal of Systematic and Evolutionary Microbiology* 64(8)**,** 2729-2737.

Hillis, D.M., and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42(2)**,** 182-192.

Imhoff, J.F., Trüper, H.G., and Pfennig, N. (1984). Rearrangement of the species and genera of the phototrophic "purple nonsulfur bacteria". *International Journal of Systematic and Evolutionary Microbiology* 34(3)**,** 340-343.

Janda, J.M., and Abbott, S.L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology* 45(9)**,** 2761-2764.

Kämpfer, P., Buczolits, S., Albrecht, A., Busse, H.J., and Stackebrandt, E. (2003). Towards a standardized format for the description of a novel species (of an established genus): *Ochrobactrum gallinifaecis* sp. nov. *International Journal of Systematic and Evolutionary Microbiology* 53(3)**,** 893-896.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12)**,** 1647-1649.

Kim, Y.O., Park, S., Kim, H., Park, D.S., Nam, B.H., Kim, D.G., et al. (2014b). *Halocynthiibacter namhaensis* gen. nov., sp. nov., a novel alphaproteobacterium isolated from sea squirt *Halocynthia roretzi*. *Antonie van Leeuwenhoek* 105(5)**,** 881-889.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Research* 29(1)**,** 181-184.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research* 41(1)**,** e1.

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

Konstantinidis, K.T., and Tiedje, J.M. (2005b). Towards a genome-based taxonomy for prokaryotes. *Journal of Bacteriology* 187(18)**,** 6258-6264.

Krieg, N.R., and Garrity, G.M. (2005). "On using the *Manual*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part A: Introductory Essays,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley & G.M. Garrity. 2nd ed. (New York, NY, USA: Springer), 15-19.

Kuykendall, L.D. (2005). "Order VI. *Rhizobiales* ord. nov.," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part C: The Alpha-, Beta-, Delta-, and Epsilonproteobacteria,* eds. D.J. Brenner, N.R. Krieg & J.T. Staley. 2nd ed. (New York, NY, USA: Springer), 324-574.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21)**,** 2947-2948.

Lau, S.C., Tsoi, M.M., Li, X., Plakhotnikova, I., Wu, M., Wong, P.K., et al. (2004). *Loktanella hongkongensis* sp. nov., a novel member of the *α-Proteobacteria* originating from marine biofilms in Hong Kong waters. *International Journal of Systematic and Evolutionary Microbiology* 54(6)**,** 2281-2284.

Lucena, T., Pujalte, M.J., Ruvira, M.A., Garay, E., Macián, M.C., and Arahal, D.R. (2012). *Tropicibacter multivorans* sp. nov., an aerobic alphaproteobacterium isolated from surface seawater. *International Journal of Systematic and Evolutionary Microbiology* 62(4)**,** 844-848.

Luo, C., Rodriguez-R, L.M., and Konstantinidis, K.T. (2014). MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research* 42(8)**,** e73.

Luo, H., and Moran, M.A. (2014). Evolutionary ecology of the marine *Roseobacter* clade. *Microbiology and Molecular Biology Reviews* 78(4)**,** 573-587.

Maiden, M.C., Jansen van Rensburg, M.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A., et al. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology* 11(10)**,** 728-736.

Mardis, E.R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* 6(1)**,** 287-303.

Martens, T., Gram, L., Grossart, H.P., Kessler, D., Müller, R., Simon, M., et al. (2007). Bacteria of the *Roseobacter* clade show potential for secondary metabolite production. *Microbial Ecology* 54(1)**,** 31-42.

Martens, T., Heidorn, T., Pukall, R., Simon, M., Tindall, B.J., and Brinkhoff, T. (2006). Reclassification of *Roseobacter gallaeciensis* Ruiz-Ponte et al. 1998 as *Phaeobacter gallaeciensis* gen. nov., comb. nov., description of *Phaeobacter inhibens* sp. nov., reclassification of *Ruegeria algicola* (Lafay et al. 1995) Uchino et al. 1999 as *Marinovum algicola* gen. nov., comb. nov., and emended descriptions of the genera *Roseobacter*, *Ruegeria* and *Leisingera*. *International Journal of Systematic and Evolutionary Microbiology* 56(6)**,** 1293-1304.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14**,** 60.

Moran, M.A., Belas, R., Schell, M.A., Gonzalez, J.M., Sun, F., Sun, S., et al. (2007). Ecological genomics of marine roseobacters. *Applied and Environmental Microbiology* 73(14)**,** 4559-4569.

Oren, A., and Garrity, G.M. (2014). Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. *Antonie van Leeuwenhoek* 106(1)**,** 43-56.

Pannekoek, Y., Qi-Long, Q., Zhang, Y.Z., and Van Der Ende, A. (2016). Genus delineation of *Chlamydiales* by analysis of the percentage of conserved proteins justifies the reunifying of the genera *Chlamydia* and *Chlamydophila* into one single genus *Chlamydia*. *Pathogens and Disease* 74(6)**,** ftw071.

Park, S., Won, S.M., Kim, H., Park, D.S., and Yoon, J.H. (2014). *Aestuariivita boseongensis* gen. nov., sp. nov., isolated from a tidal flat sediment. *International Journal of Systematic and Evolutionary Microbiology* 64(9)**,** 2969-2974.

Qin, Q.L., Xie, B.B., Zhang, X.Y., Chen, X.L., Zhou, B.C., Zhou, J., et al. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of Bacteriology* 196(12)**,** 2210-2215.

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

Rodriguez-R, L.M., and Konstantinidis, K.T. (2005). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints* 4(1)**,** e1900v1.

Rodriguez-R, L.M., and Konstantinidis, K.T. (2014). Bypassing cultivation to identify bacterial species. *Microbe* 9(3)**,** 111-118.

Rosselló-Móra, R., and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiology Reviews* 25(1)**,** 39-67.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering* 12(2)**,** 85-94.

Rüger, H.J., and Höfle, M.G. (1992). Marine star-shaped-aggregate-forming bacteria: *Agrobacterium atlanticum* sp. nov.; *Agrobacterium meteori* sp. nov.; *Agrobacterium ferrugineum* sp. nov., nom. rev.; *Agrobacterium gelatinovorum* sp. nov., nom. rev.; and *Agrobacterium stellulatum* sp. nov., nom. rev. *International Journal of Systematic Bacteriology* 42(1)**,** 133-143.

Seyedsayamdost, M.R., Case, R.J., Kolter, R., and Clardy, J. (2011). The Jekyll-and-Hyde chemistry of *Phaeobacter gallaeciensis*. *Nature Chemistry* 3(4)**,** 331-335.

Shiba, T. (1991). *Roseobacter litoralis* gen. nov., sp. nov., and *Roseobacter denitrificans* sp. nov., aerobic pink-pigmented bacteria which contain bacteriochlorophyll a. *Systematic and Applied Microbiology* 14(2)**,** 140-145.

Simon, M., Scheuner, C., Meier-Kolthoff, J.P., Brinkhoff, T., Wagner-Döbler, I., Ulbrich, M., et al. (2017). Phylogenomics of *Rhodobacteraceae* reveals evolutionary adaptation to marine and non-marine habitats. *ISME Journal* 11(6)**,** 1483-1499.

Sneath, P.H.A., and Brenner, D.J. (1992). "Official" nomenclature lists. *ASM News* 58(1)**,** 175.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9)**,** 1312-1313.

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

Tsubouchi, T., Shimane, Y., Mori, K., Miyazaki, M., Tame, A., Uematsu, K., et al. (2013). *Loktanella cinnabarina* sp. nov., isolated from a deep subseafloor sediment, and emended description of the genus *Loktanella*. *International Journal of Systematic and Evolutionary Microbiology* 63(4)**,** 1390-1395.

Uchino, Y., Hirata, A., Yokota, A., and Sugiyama, J. (1998). Reclassification of marine *Agrobacterium* species: proposals of *Stappia stellulata* gen. nov., comb. nov., *Stappia aggregata* sp. nov., nom. rev., *Ruegeria atlantica* gen. nov., comb. nov., *Ruegeria gelatinovora* comb. nov., *Ruegeria algicola* comb. nov., and *Ahrensia kieliense* gen. nov., sp. nov., nom. rev. *Journal of General and Applied Microbiology* 44(3)**,** 201-210.

Van Trappen, S., Mergaert, J., and Swings, J. (2004). *Loktanella salsilacus* gen. nov., sp. nov., *Loktanella fryxellensis* sp. nov. and *Loktanella vestfoldensis* sp. nov., new members of the *Rhodobacter* group, isolated from microbial mats in Antarctic lakes. *International Journal of Systematic and Evolutionary Microbiology* 54(4)**,** 1263-1269.

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K., and Swings, J. (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiological Reviews* 60(2)**,** 407-438.

Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C., et al. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research* 43(14)**,** 6761-6771.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Wagner-Döbler, I., Ballhausen, B., Berger, M., Brinkhoff, T., Buchholz, I., Bunk, B., et al. (2010). The complete genome sequence of the algal symbiont *Dinoroseobacter shibae*: a hitchhiker's guide to life in the sea. *ISME Journal* 4(1)**,** 61-77.

Wagner-Döbler, I., and Biebl, H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annual Review of Microbiology* 60(1)**,** 255-280.

Wang, D., Liu, H., Zheng, S., and Wang, G. (2014). *Paenirhodobacter enshiensis* gen. nov., sp. nov., a non-photosynthetic bacterium isolated from soil, and emended descriptions of the genera *Rhodobacter* and *Haematobacter*. *International Journal of Systematic and Evolutionary Microbiology* 64(2)**,** 551-558.

Ward, D.M., Weller, R., and Bateson, M.M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* 345(6270)**,** 63-65.

Whitman, W.B. (2011). Intent of the nomenclatural Code and recommendations about naming new species based on genomic sequences. *Bulletin of BISMiS* 2(2)**,** 135-139.

Woese, C.R. (1987). Bacterial evolution. *Microbiological Reviews* 51(2)**,** 221-271.

Xu, Y., Li, Q., Tian, R., Lai, Q., and Zhang, Y. (2015). *Pseudovibrio hongkongensis* sp. nov., isolated from a marine flatworm. *Antonie van Leeuwenhoek* 108(1)**,** 127-132.

Zhang, Y., Li, Q., Tian, R., Lai, Q., and Xu, Y. (2016). *Pseudovibrio stylochi* sp. nov., isolated from a marine flatworm. *International Journal of Systematic and Evolutionary Microbiology* 66(5)**,** 2025-2029.

CHAPTER 4: Genetic Promiscuity Between Close Relatives

Vibrio cholerae and *Vibrio metoecus* – Part 1

A version of this chapter has been published as:

"Orata, F.D., Kirchberger, P.C., Méheust, R., Barlow, E.J., Tarr, C.L., and 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(10), 2941-2954."

F.D.O. and Y.B. designed the study and wrote the manuscript; Y.B. performed sampling; C.L.T. provided the clinical *V. metoecus* genome sequences; F.D.O. performed whole-genome sequencing and analysis; P.C.K. performed phylogenetic analysis; R.M. quantified interspecies gene transfers; E.J.B. provided multiple scripts used in data analysis; and Y.B. supervised the project.

CHAPTER 4

4.1. 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 has 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.

4.2. 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, USA). 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.

4.3. Materials and Methods

4.3.1. 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 Table D.1. Environmental strains of *V. metoecus* and *V. cholerae* were isolated from Oyster Pond (Falmouth, MA, USA) on August and September 2009 using previously described methods (Boucher et al., 2011). Isolates were grown overnight at 30°C in tryptic soy broth (Becton Dickinson) with 1% NaCl (BDH). The sequences of the clinical *V. metoecus* strains were determined by the CDC. Additional sequences were obtained from the GenBank database (National Center for Biotechnology Information).

4.3.2. Genomic DNA extraction and quantitation

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

4.3.3. 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). The contiguous sequences were assembled *de novo* with the CLC Genomics Workbench 7.5.2 (CLC Bio). Functional annotations of the draft genomes were done in RAST 2.0 (Rapid Annotation Using Subsystem Technology; Aziz et al., 2008).

The whole-genome sequences generated in this study were deposited in the GenBank database under the BioProject accession number PRJNA281423. The individual genome accession numbers are listed in Table D.1.

4.3.4. 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).

4.3.5. 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 2.0 (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 1.0.1 (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.

4.3.6. 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 if 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).

4.3.7. 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 (Improved Structural Annotation of *attC*) software (Szamosi,

2012). The *intI4* and gene cassette sequences were used to calculate the average nucleotide identity (ANI) (Konstantinidis and Tiedje, 2005a; Goris et al., 2007) between strains (intra- and interspecies) in JSpecies 1.2.1 (Richter and Rosselló-Móra, 2009), using the bidirectional best BLAST hits between nucleotides. The ANI of the integron region was compared to the ANI of 1,560 single-copy core ORFs (≈ 1.42 Mb).

4.3.8. Phylogenetic analyses

Using the PhyloPhlAn pipeline 0.99 (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 Clustal W 2.1 (Larkin et al., 2007), and an ML tree was constructed using RAxML 8.1.17 (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 1.99-11 (Schliep, 2011) in R 3.1.2 (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 8.1.2 (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 8.1.17 (Stamatakis, 2014), as described above.

4.4. Results and Discussion

Vibrio cholerae is widely studied, and the genomes of globally diverse clinical and environmental isolates are available (Table D.1). On the other hand, there are currently only two *V. metoecus* genomes available. Strain RC341 was isolated from Chesapeake Bay (MD, USA) in 1998. It was presumptively identified as a variant *V. cholerae* based on 16S 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, USA. 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 USA 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 (Figure 4.1).

4.4.1. *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 (Figure 4.2).

On average, *V. metoecus* shares 84% of its ORFs with *V. cholerae*, whereas 89–91% ORFs are shared between strains of the same species. 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*.

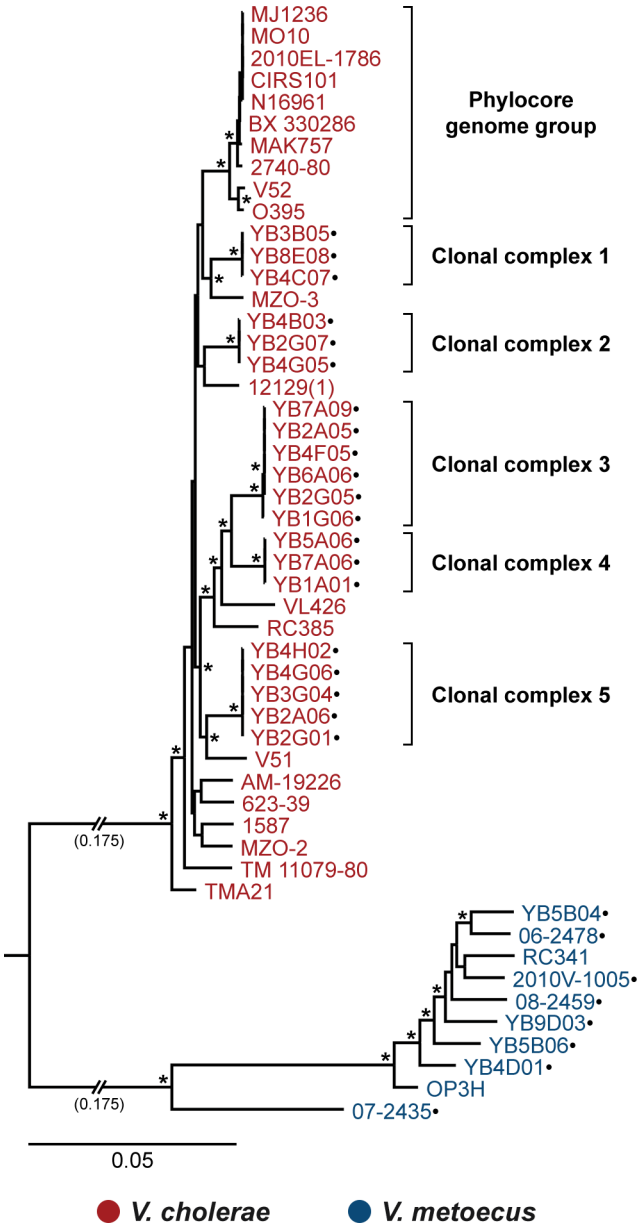

Figure 4.1. The phylogenetic relationship of *V. metoecus* and *V. cholerae*.

The phylogenetic tree was constructed from the concatenated sequence alignment of single-copy core gene families (771,455 bp). All reliable bootstrap support are indicated with * and are at least 97% for this tree. Bar, 0.05 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 belong to the same clonal complex.

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, 2005a; Goris et al., 2007). The ANI of the core genome is 86–87% between species and 98–100% within species (Figure 4.3A), 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, 2005a). 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* (Figure 4.3A).

4.4.2. 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 identity of genes found in the integron region does not seem to differ within and between species (Figure 4.2).

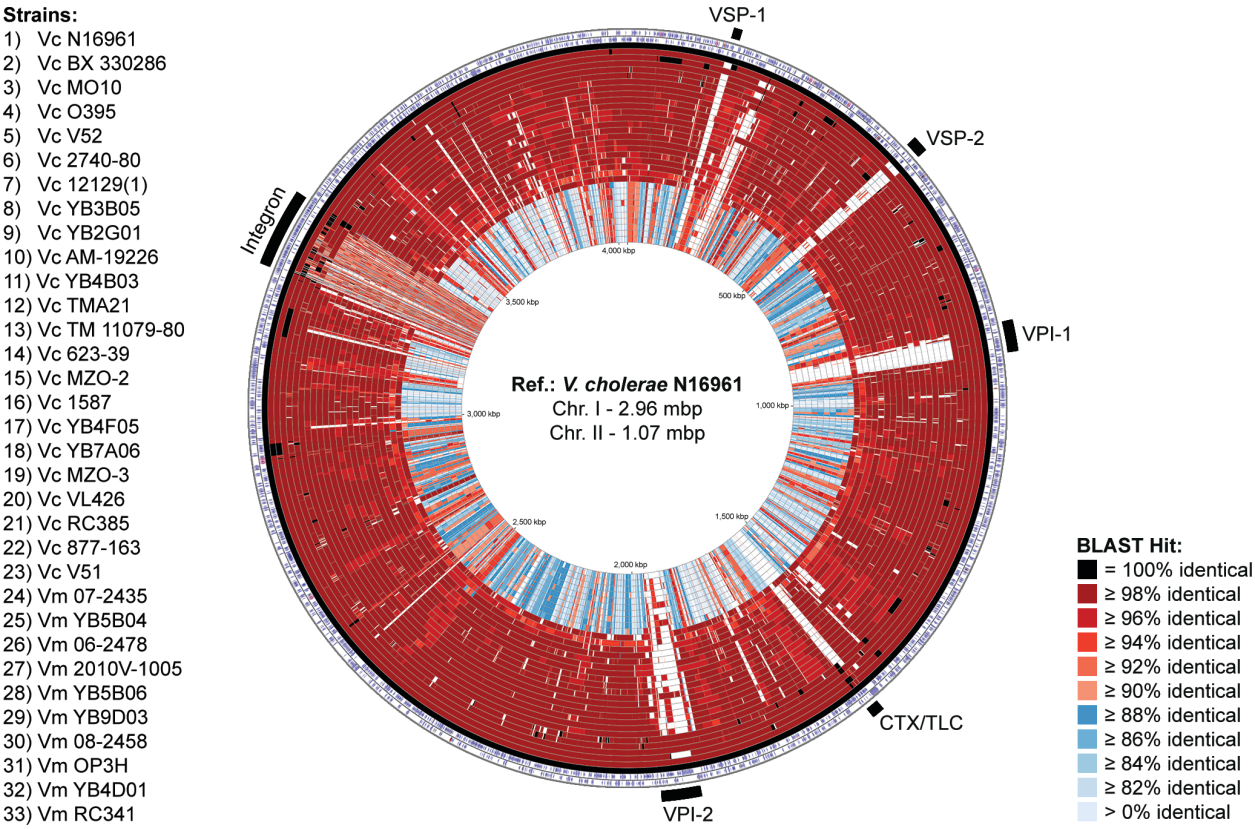

Figure 4.2. The *V. metoecus* (Vm) and *V. cholerae* (Vc) BLAST atlas.

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.

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 (Table D.2). The ANI of the integron region was determined between pairs of strains and compared to the ANI of the core genome (Figure 4.3). Although ANI is 86–87% between species and 98–100% within species for the core genome (Figure 4.3A), the

integron region displays an average pairwise ANI of 93–94%, both within and between species (Figure 4.3B). 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 (Figure 4.2 and Figure 4.3B). 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.

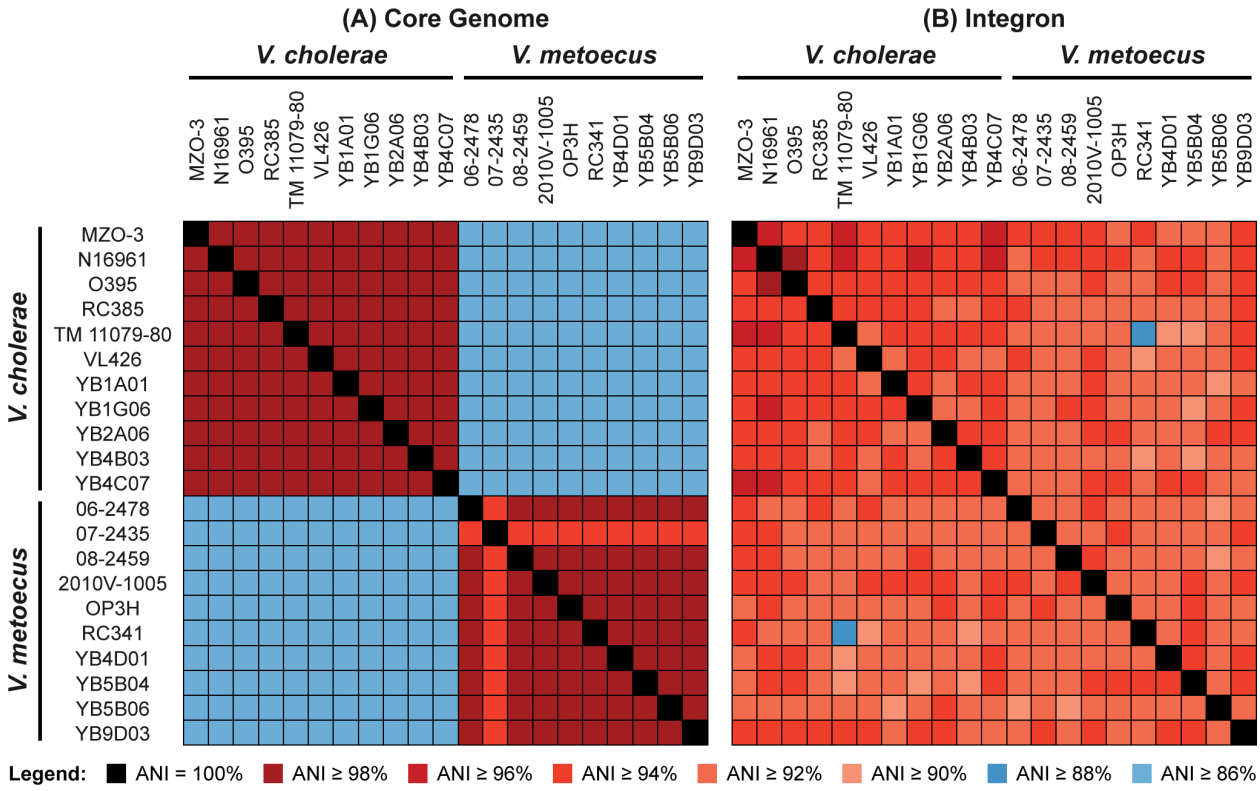

Figure 4.3. ANI of the core genome versus chromosomal integron region of *V. metoecus* and *V. cholerae.*

(A) Intra- and interspecies pairwise comparisons of the 1,560 single-copy core genes (≈ 1.42 Mb). (B) Intra- and interspecies pairwise comparisons of the integron gene cassettes.

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 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 (Figure D.1), 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).

4.4.3. 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 (Figure 4.4). 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%). 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 (Table D.3).

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 (Figure 4.1). They also exhibit 100% ANI only with each other but not with isolates from other clonal complexes (Table D.4). Indeed, members of the same clonal complex always cluster together in all the individual gene trees examined (Figure 4.4). We therefore randomly chose one isolate from each *V. cholerae* clonal complex from Oyster Pond, yielding a final dataset of five genomes from each species. A total of 224 interspecies gene transfer events were detected in this environment-specific dataset, where 192 (86%; $P = 0.0012$) involved *V. metoecus* as the recipient and only 32 (14%) with *V. cholerae* as the recipient (Table 4.1). 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.

Table 4.1. HGT count for *V. metoecus* and representative *V. cholerae* strains from Oyster Pond.

Values are based on 376 single-copy core genes with inferred HGT events. 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, $P = 0.0012$.

Species and strain	HGT count	Percent of total
Vibrio metoecus		
OP3H	55	25
YB4D01	43	19
YB5B06	37	17
YB5B04	30	13
YB9D03	27	12
Vibrio metoecus total	192	86
Vibrio cholerae		
YB2G01 (CC 5)	16	7
YB4F05 (CC 3)	9	4
YB4B03 (CC 2)	4	2
YB7A06 (CC 4)	2	1
YB3B05 (CC 1)	1	0
Vibrio cholerae total	32	14
Total	224	100

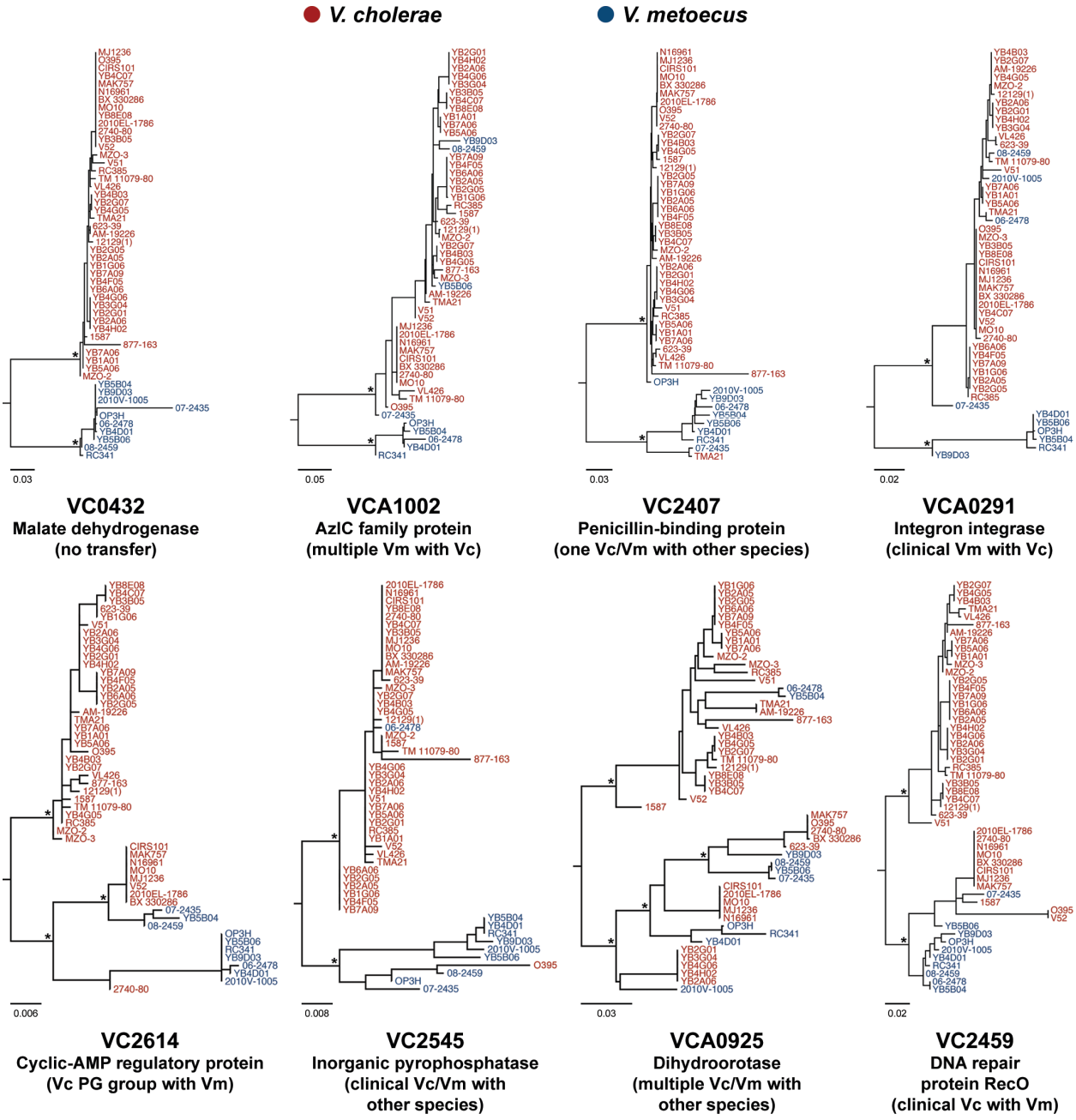

Figure 4.4. Representative HGT events between *V. metoecus* (Vm) and *V. cholerae* (Vc).

The phylogenetic trees are representative 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 *. Bar, nucleotide substitutions per site.

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* (Figure 4.4). 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 (Figure 4.1; 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 (*ctxB1*). *Vibrio cholerae* with the *ctxB* genotype 7 (*ctxB7*) outcompeted *ctxB1* 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., 2016).

4.4.4. 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 (Figure 4.5). 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* (Figure 4.5A). The absence of CTX and TCP in *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 non-pandemic *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 sub-clusters we call "islets," as indicated in Figure 4.5A. These four islets match the previous description of Jermyn and Boyd (2002) for VPI-2: (A) a type-1 restriction-modification system for protection against viral infection, (B) a *nan-nag* cluster for sialic acid metabolism, (C) a Mu phage-like region, and (D) 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, 2013). 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* (Figure 4.5B and Figure D.2). 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 (Figure 4.5A), 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*. Lastly, 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 (Figure 4.5B). 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 2010V-1005 and environmental RC341 (Figure 4.5A), 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 (Figure 4.5B and Figure D.3). 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 (Figure 4.5A); 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.

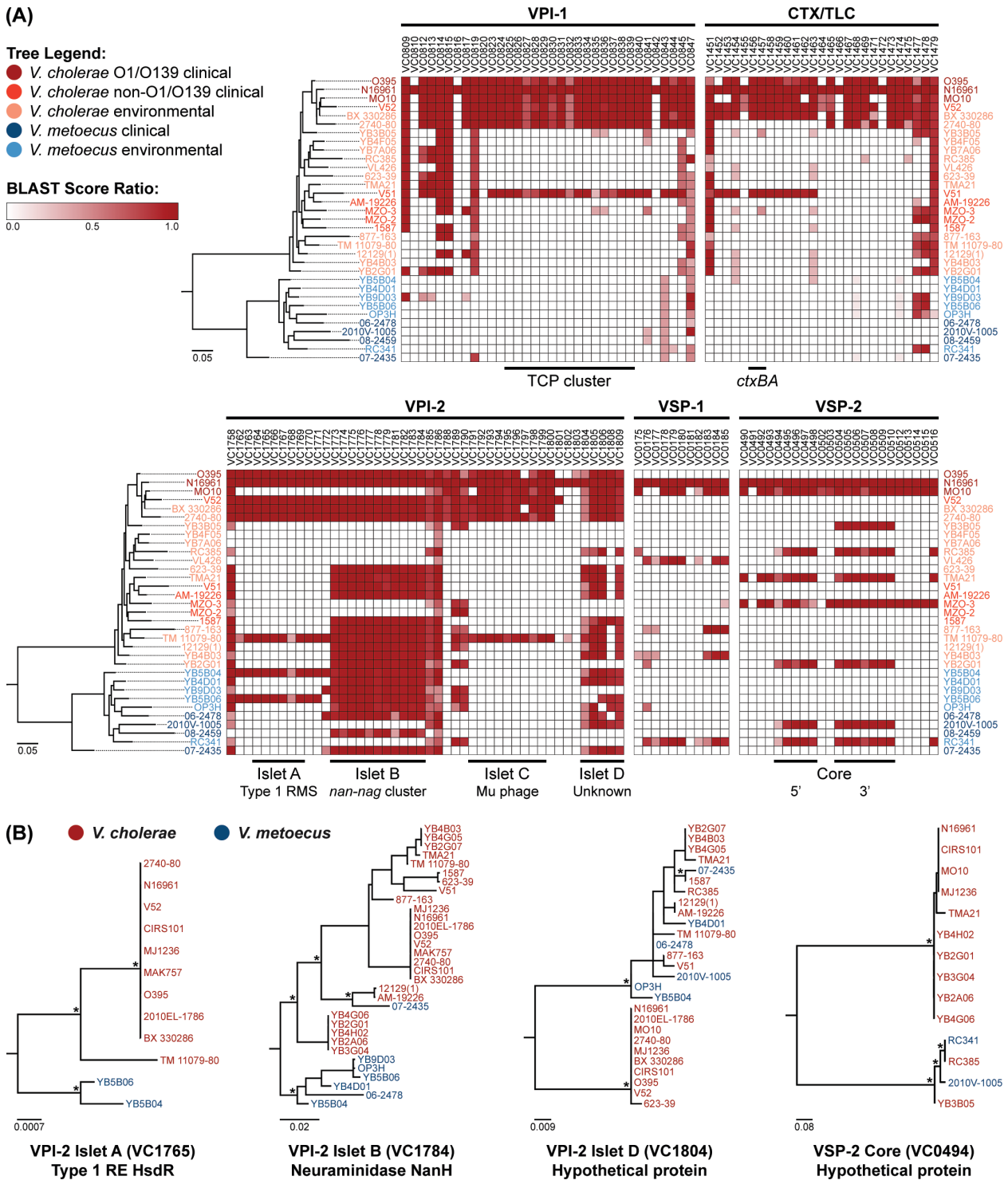

Figure 4.5. Virulence factors present in *V. metoecus* and *V. cholerae*.

(Figure legend continued on the next page)

Figure 4.5. 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 phylogenetic tree was constructed using 3,978 amino acid positions based on 400 universally conserved bacterial and archaeal proteins. Bar, 0.05 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 colours, 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 phylogenetic trees of orthologous gene families of the VPI-2 islets and the VSP-2 core. Relevant bootstrap support (≥ 70%) is indicated with *. Bar, nucleotide substitutions per site. RE, restriction endonuclease.

4.4.5. 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, we first compiled their pan- and core genomes (Figure D.4). 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 (Figure D.4A). 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 dataset. The reduced core genome size is expected since 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 (Figure D.4B). 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 (Figure D.5). Function was predicted for each gene family based on the COG database (Figure D.6). *Vibrio metoecus* contains more unique gene families than *V. cholerae* that are involved in carbohydrate transport and metabolism (Figure D.6A). 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 dataset, 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 (Figure D.6B). 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 dataset because they are not unique to clinical strains, but are also present in some environmental ones (Figure 4.5A). 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.

4.5. 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.

4.6. References

Ali, M., Lopez, A.L., You, Y.A., Kim, Y.E., Sah, B., Maskery, B., et al. (2012). The global burden of cholera. *Bulletin of the World Health Organization* 90(3)**,** 209-218A.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215(3)**,** 403-410.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9**,** 75.

Beyhan, S., and Yildiz, F.H. (2007). Smooth to rugose phase variation in *Vibrio cholerae* can be mediated by a single nucleotide change that targets c-di-GMP signalling pathway. *Molecular Microbiology* 63(4)**,** 995-1007.

Blokesch, M., and Schoolnik, G.K. (2008). The extracellular nuclease Dns and its role in natural transformation of *Vibrio cholerae*. *Journal of Bacteriology* 190(21)**,** 7232-7240.

Boucher, Y., Cordero, O.X., Takemura, A., Hunt, D.E., Schliep, K., Bapteste, E., et al. (2011). Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *MBio* 2(2)**,** e00335-10.

Boucher, Y., Labbate, M., Koenig, J.E., and Stokes, H.W. (2007). Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends in Microbiology* 15(7)**,** 301-309.

Choopun, N. (2004). *The population structure of Vibrio cholerae in Chesapeake Bay*. Doctor of Philosophy Thesis, University of Maryland.

Chun, J., Grim, C.J., Hasan, N.A., Lee, J.H., Choi, S.Y., Haley, B.J., et al. (2009). Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 106(36)**,** 15442-15447.

Collis, C.M., and Hall, R.M. (1995). Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrobial Agents and Chemotherapy* 39(1)**,** 155-162.

Davis, B.R., Fanning, G.R., Madden, J.M., Steigerwalt, A.G., Bradford, H.B., Jr., Smith, H.L., Jr., et al. (1981). Characterization of biochemically atypical *Vibrio cholerae* strains and designation of a new pathogenic species, *Vibrio mimicus*. *Journal of Clinical Microbiology* 14(6)**,** 631-639.

De la Cruz, F., and Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in Microbiology* 8(3)**,** 128-133.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology* 2(5)**,** 414-424.

Dziejman, M., Balon, E., Boyd, D., Fraser, C.M., Heidelberg, J.F., and Mekalanos, J.J. (2002). Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proceedings of the National Academy of Sciences of the United States of America* 99(3)**,** 1556-1561.

Freeman, S., and Herron, J.C. (2013). *Evolutionary Analysis.* San Francisco, CA, USA: Benjamin Cummings.

Gaasbeek, E.J., Wagenaar, J.A., Guilhabert, M.R., Wösten, M.M.S.M., Van Putten, J.P., Van Der Graaf-Van Bloois, L., et al. (2009). A DNase encoded by integrated element CJIE1 inhibits natural transformation of *Campylobacter jejuni*. *Journal of Bacteriology* 191(7)**,** 2296-2306.

García, B., Latasa, C., Solano, C., García-del Portillo, F., Gamazo, C., and Lasa, I. (2004). Role of the GGDEF protein family in *Salmonella* cellulose biosynthesis and biofilm formation. *Molecular Microbiology* 54(1)**,** 264-277.

Ge, F., Wang, L.S., and Kim, J. (2005). The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biology* 3(10)**,** e316.

Gomez-Gil, B., Thompson, C.C., Matsumura, Y., Sawabe, T., Iida, T., Christen, R., et al. (2014). "The family *Vibrionaceae*," in *The Prokaryotes – Gammaproteobacteria,* eds. E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt & F. Thompson. 4th ed. (Berlin, Germany: Springer), 659-747.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57(1)**,** 81-91.

Grant, J.R., Arantes, A.S., and Stothard, P. (2012). Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics* 13**,** 202.

Haley, B.J., Grim, C.J., Hasan, N.A., Choi, S.Y., Chun, J., Brettin, T.S., et al. (2010). Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. *BMC Microbiology* 10**,** 154.

Hasan, N.A., Grim, C.J., Haley, B.J., Chun, J., Alam, M., Taviani, E., et al. (2010). Comparative genomics of clinical and environmental *Vibrio mimicus*. *Proceedings of the National Academy of Sciences of the United States of America* 107(49)**,** 21134-21139.

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., et al. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406(6795)**,** 477-483.

Hillis, D.M., and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42(2)**,** 182-192.

Iqbal, N., Guérout, A.M., Krin, E., Le Roux, F., and Mazel, D. (2015). Comprehensive functional analysis of the 18 *Vibrio cholerae* N16961 toxin-antitoxin systems substantiates their role in stabilizing the superintegron. *Journal of Bacteriology* 197(13)**,** 2150-2159.

Jermyn, W.S., and Boyd, E.F. (2002). Characterization of a novel *Vibrio* pathogenicity island (VPI-2) encoding neuraminidase (*nanH*) among toxigenic *Vibrio cholerae* isolates. *Microbiology* 148(11)**,** 3681-3693.

Karaolis, D.K., Somara, S., Maneval, D.R., Jr., Johnson, J.A., and Kaper, J.B. (1999). A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. *Nature* 399(6734)**,** 375-379.

Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., et al. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4(4)**,** e00398-13.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12)**,** 1647-1649.

Kim, S., Kurokawa, D., Watanabe, K., Makino, S., Shirahata, T., and Watarai, M. (2004). *Brucella abortus* nicotinamidase (PncA) contributes to its intracellular replication and infectivity in mice. *FEMS Microbiology Letters* 234(2)**,** 289-295.

Kirchberger, P.C., Turnsek, M., Hunt, D.E., Haley, B.J., Colwell, R.R., Polz, M.F., et al. (2014). *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. *International Journal of Systematic and Evolutionary Microbiology* 64(9)**,** 3208-3214.

Koenig, J.E., Boucher, Y., Charlebois, R.L., Nesbo, C., Zhaxybayeva, O., Bapteste, E., et al. (2008). Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. *Environmental Microbiology* 10(4)**,** 1024-1038.

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

Langille, M.G., and Brinkman, F.S. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 25(5)**,** 664-665.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21)**,** 2947-2948.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* 13(9)**,** 2178-2189.

Lombardo, M.J., Michalski, J., Martinez-Wilson, H., Morin, C., Hilton, T., Osorio, C.G., et al. (2007). An *in vivo* expression technology screen for *Vibrio cholerae* genes expressed in human volunteers. *Proceedings of the National Academy of Sciences of the United States of America* 104(46)**,** 18229-18234.

Louis, P., and Doré, J. (2014). "Functional metagenomics of human intestinal microbiome β-glucuronidase activity," in *Encyclopedia of Metagenomics,* ed. K.E. Nelson. (New York, NY, USA: Springer), 1-8.

Magnelli, P., Bielik, A., and Guthrie, E. (2012). "Identification and characterization of protein glycosylation using specific endo- and exoglycosidases," in *Protein Expression in Mammalian Cells: Methods and Protocols (Methods in Molecular Biology),* ed. J.L. Hartley. (New York, NY, USA: Humana Press), 189-211.

Mathur, J., and Waldor, M.K. (2004). The *Vibrio cholerae* ToxR-regulated porin OmpU confers resistance to antimicrobial peptides. *Infection and Immunity* 72(6)**,** 3577-3583.

Mazel, D. (2006). Integrons: agents of bacterial evolution. *Nature Reviews Microbiology* 4(8)**,** 608-620.

Mazel, D., Dychinco, B., Webb, V.A., and Davies, J. (1998). A distinctive class of integron in the *Vibrio cholerae* genome. *Science* 280(5363)**,** 605-608.

Nesterchuk, M.V., Sergiev, P.V., and Dontsova, O.A. (2011). Posttranslational modifications of ribosomal proteins in *Escherichia coli*. *Acta Naturae* 3(2)**,** 22-33.

O'Shea, Y.A., Finnan, S., Reen, F.J., Morrissey, J.P., O'Gara, F., and Boyd, E.F. (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(12)**,** 4053-4063.

Petrack, B., Greengard, P., Craston, A., and Sheppy, F. (1965). Nicotinamide deamidase from mammalian liver. *Journal of Biological Chemistry* 240(4)**,** 1725-1730.

Provenzano, D., Schuhmacher, D.A., Barker, J.L., and Klose, K.E. (2000). The virulence regulatory protein ToxR mediates enhanced bile resistance in *Vibrio cholerae* and other pathogenic *Vibrio* species. *Infection and Immunity* 68(3)**,** 1491-1497.

Purser, J.E., Lawrenz, M.B., Caimano, M.J., Howell, J.K., Radolf, J.D., and Norris, S.J. (2003). A plasmid-encoded nicotinamidase (PncA) is essential for infectivity of *Borrelia burgdorferi* in a mammalian host. *Molecular Microbiology* 48(3)**,** 753-764.

R Development Core Team (2014). "R: a language and environment for statistical computing". 3.1.2 ed. (Vienna, Austria: R Foundation for Statistical Computing).

Rashed, S.M., Mannan, S.B., Johura, F.T., Islam, M.T., Sadique, A., Watanabe, H., et al. (2012). Genetic characteristics of drug-resistant *Vibrio cholerae* O1 causing endemic cholera in Dhaka, 2006-2011. *Journal of Medical Microbiology* 61(12)**,** 1736-1745.

Rashid, M.U., Rashed, S.M., Islam, T., Johura, F.T., Watanabe, H., Ohnishi, M., et al. (2016). *CtxB1* outcompetes *CtxB7* in *Vibrio cholerae* O1, Bangladesh. *Journal of Medical Microbiology* 65(1)**,** 101-103.

Rasko, D.A., Myers, G.S., and Ravel, J. (2005). Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics* 6**,** 2.

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

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering* 12(2)**,** 85-94.

Rowe-Magnus, D.A., and Mazel, D. (2002). The role of integrons in antibiotic resistance gene capture. *International Journal of Medical Microbiology* 292(2)**,** 115-125.

Samant, S., Lee, H., Ghassemi, M., Chen, J., Cook, J.L., Mankin, A.S., et al. (2008). Nucleotide biosynthesis is critical for growth of bacteria in human blood. *PLoS Pathogens* 4(2)**,** e37.

Schliep, K., Lopez, P., Lapointe, F.J., and Bapteste, É. (2011). Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Molecular Biology and Evolution* 28(4)**,** 1393-1405.

Schliep, K.P. (2011). Phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4)**,** 592-593.

Schmidt, H., and Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clinical Microbiology Reviews* 17(1)**,** 14-56.

Seed, K.D., Yen, M., Shapiro, B.J., Hilaire, I.J., Charles, R.C., Teng, J.E., et al. (2014). Evolutionary consequences of intra-patient phage predation on microbial populations. *Elife* 3**,** e03497.

Segata, N., Börnigen, D., Morgan, X.C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications* 4**,** 2304.

Skorupski, K., and Taylor, R.K. (1997). Cyclic AMP and its receptor protein negatively regulate the coordinate expression of cholera toxin and toxin-coregulated pilus in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 94(1)**,** 265-270.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9)**,** 1312-1313.

Stokes, H.W., and Hall, R.M. (1989). A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Molecular Microbiology* 3(12)**,** 1669-1683.

Szamosi, J.C. (2012). *ISAAC: an improved structural annotation of attC and an initial application thereof.* Master of Science Thesis, McMaster University.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28(1)**,** 33-36.

Taviani, E., Grim, C.J., Choi, J., Chun, J., Haley, B., Hasan, N.A., et al. (2010). Discovery of novel *Vibrio cholerae* VSP-II genomic islands using comparative genomic analysis. *FEMS Microbiology Letters* 308(2)**,** 130-137.

Taylor, R.K., Miller, V.L., Furlong, D.B., and Mekalanos, J.J. (1987). Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proceedings of the National Academy of Sciences of the United States of America* 84(9)**,** 2833-2837.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* 102(39)**,** 13950-13955.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Vesth, T., Wassenaar, T.M., Hallin, P.F., Snipen, L., Lagesen, K., and Ussery, D.W. (2010). On the origins of a *Vibrio* species. *Microbial Ecology* 59(1)**,** 1-13.

Waldor, M.K., and Mekalanos, J.J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272(5270)**,** 1910-1914.

Wang, G., and Maier, R.J. (2004). An NADPH quinone reductase of *Helicobacter pylori* plays an important role in oxidative stress resistance and host colonization. *Infection and Immunity* 72(3)**,** 1391-1396.

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., et al. (2014). PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* 30(9)**,** 1297-1299.

CHAPTER 5: Genetic Promiscuity Between Close Relatives

Vibrio cholerae and *Vibrio metoecus* – Part 2

A version of this chapter will be submitted for publication as:

"Orata, F.D., Liang, K., Nasreen, T., Hussain, N.A.S., and Boucher, Y. Differences in abundance and seasonal patterns of *Vibrio cholerae* and *Vibrio metoecus* lead to a directional bias in interspecies horizontal gene transfer."

F.D.O. and Y.B. designed the study and wrote the manuscript; Y.B. performed sampling; F.D.O. performed whole-genome sequencing and assembly; F.D.O. and K.L. performed bioinformatic analyses; T.N. performed qPCR; F.D.O. and N.A.S.H. performed the natural transformation assay; and Y.B. supervised the project.

CHAPTER 5

5.1. Abstract

Although some genotypes of *Vibrio cholerae* are responsible for the potent diarrheal disease cholera, most members of this species are harmless inhabitants of coastal environments worldwide. Most of the virulence determinants of *V. cholerae* pathogenic genotypes have been acquired by horizontal gene transfer (HGT). Despite the role HGT plays in the evolution of virulence in this species, little is known about the mode and tempo of this process in its aquatic habitat. To get insight on HGT in a natural environment, *V. cholerae* was co-isolated with its closest known relative, *Vibrio metoecus*, from a cholera-free brackish coastal pond. The genomes of 17 strains from each species were sequenced, and the extent of their genetic interactions through HGT was determined within their core and accessory genomes. Interspecies HGT events could be inferred in 21% of the core genes and 45% of the accessory genes. A directional bias in gene transfer events was found in the core genome, where *V. metoecus* was a recipient of three times (75%) more core genes from *V. cholerae* than it was a donor (25%). Using quantitative PCR on DNA extracted from water sampled in the pond over the summers of two consecutive years, *V. cholerae* was found to be three times more abundant than *V. metoecus*. Whereas *V. cholerae* was present throughout the summer, *V. metoecus* was only ever detectable at the end of the season. This difference in abundance and marked seasonality could be major contributors to the bias in HGT between the two species.

5.2. Introduction

Horizontal gene transfer (HGT), the acquisition of genetic material from individuals that are not the direct parent of the recipient, plays a significant role in the evolution of bacteria, archaea, and eukaryotes (Soucy et al., 2015). HGT is responsible for the evolution of pathogenic variants of *Vibrio cholerae*, including the lineage that is the causative agent of the potent diarrheal disease cholera (Faruque and Mekalanos, 2003). The two major virulence factors of toxigenic *V. cholerae*, the cholera toxin and toxin-coregulated pilus, reside in genomic islands that have been acquired from phages or through undefined HGT events (Waldor and Mekalanos, 1996; Faruque et al., 2003). The integron, another genomic island which can occupy up to 3% of the *V. cholerae* genome (Heidelberg et al., 2000), is also a major HGT hotspot (Orata et al., 2015) and provides *V. cholerae* a wide range of biochemical functions (Mazel et al., 1998).

Genetic exchange in bacteria is widespread and can even occur between distantly related groups of bacteria. The main barrier is the cell membrane that can be crossed through transduction, transformation, or conjugation (Thomas and Nielsen, 2005; Soucy et al., 2015). Unless it is carried by an extra-chromosomal element such as a plasmid, foreign DNA has to be integrated in the genome, either through heterologous or homologous recombination, to result in a successful event of HGT. While heterologous recombination does not require similarity between the donor and recipient DNA, homologous recombination is dependent on short regions of high sequence conservation at the ends of donor-recipient DNA to initiate exchange (Majewski and Cohan, 1999). It is more likely to occur between closely related organisms than evolutionarily distant ones, its frequency having being shown to be directly proportional to relatedness between the donor and recipient (Majewski and Cohan, 1999). Despite numerous studies of vectors allowing DNA to cross the membrane and the process leading to its integration

in the chromosome (Thomas and Nielsen, 2005), there have been relatively few investigations of what is occurring in natural habitats. The main aspect of HGT which has been investigated in nature is the frequency of recombination relative to mutation. Using culture-dependent approaches, several studies have determined recombination to mutation ratios in single species populations, such as in coastal *V. cholerae*. Recombination rates have been found for populations of this species, ranging from 1.9 to 6.5× the mutation rate (Keymer and Boehm, 2011; Kirchberger et al., 2016), indicating the significance of recombination over mutation in maintaining genetic diversity and shaping a population. Interspecies HGT is even more rarely studied, given the logistical aspects of adequately sampling two organisms from a single location. *Vibrio cholerae* and its closest known relative, *Vibrio metoecus*, provide a unique opportunity to do so. They are found to co-occur in the environment and are phenotypically similar enough to be co-isolated on selective media (Haley et al., 2010; Kirchberger et al., 2014). Indeed, multiple strains of both species were recently co-isolated from a cholera-free, brackish coastal pond in the United States East Coast (Boucher et al., 2011; Kirchberger et al., 2016). Boucher et al. (2011) demonstrated that the integron gene cassettes of geographically co-occurring *V. cholerae* and *V. metoecus* are more similar than geographically distinct *V. cholerae* (i.e., USA and Bangladesh). Thus, their co-isolation has led us to hypothesize that both species are likely in constant interaction with each other, providing opportunities for gene transfer (Boucher et al., 2011; Orata et al., 2015).

There is evidence that HGT between *V. cholerae* and *V. metoecus* could be biased. Using a limited dataset of genomes and only looking at the core genes, *V. metoecus* was found to be a recipient, as opposed to a donor, of six times more genes than *V. cholerae* (Orata et al., 2015). Based on the recovery of more *V. cholerae* than *V. metoecus* in culture, it was hypothesized that

this possible directional bias in HGT could be the result of a greater abundance of *V. cholerae* than *V. metoecus* in the environment (Orata et al., 2015). However, that study did not have a meaningful ecological context, as geographically disparate environmental and clinical samples were considered. Also, culture-based quantification of vibrios is unreliable, as cells are frequently in the viable but nonculturable (VBNC) state (Xu et al., 1982; Huq et al., 1990; Alam et al., 2007). To reliably estimate directional bias of interspecies HGT, the genomes of 17 strains from each species, originating from the same water sample, were sequenced and screened for HGT events. Up to three times as many genes were found to have moved from *V. cholerae* to *V. metoecus* than vice versa. To determine if this bias correlated with differences in relative abundance, qPCR of *V. cholerae*- or *V. metoecus*-specific genes from environmental DNA samples was used to accurately quantify these species. *Vibrio cholerae* was systematically found to be more abundant than *V. metoecus*, with an average of three times as many cells in a water sample. Furthermore, *V. metoecus* was found to have strong seasonal abundance patterns, only appearing for two months at the end of summer during the two consecutive years of sampling. Higher and sustained abundance could be important factors in *V. cholerae* becoming a successful human pathogen after its split from a common ancestor with *V. metoecus*.

5.3. Materials and Methods

5.3.1. Water sample collection, strain isolation, growth, and genomic DNA extraction

Water samples were collected from Oyster Pond and the adjacent lagoon (Falmouth, MA, USA) on June to September 2008 and 2009. The water samples were filtered through 4.5-cm Durapore filters (Millipore), as previously described (Kirchberger et al., 2016). Total DNA

extraction from the biomass on the filters using the DNeasy Blood and Tissue Kit (QIAGEN) was performed as follows: 0.25 g of sterile zirconia/silica beads (BioSpec) were added to cut-up filter pieces in a 1.5 mL screw-capped tube, followed by the addition of 360 µL Cell Lysis Buffer ATL (QIAGEN) and bead-beating for 30 s at maximum speed using the FastPrep-24 homogenizer (MP Biomedicals). Forty microliters of Proteinase K (QIAGEN) were added, and the tubes were vortexed for several seconds. Further steps followed the instructions of the manufacturer.

Environmental strains of *V. cholerae* and *V. metoecus* were isolated from the August and September 2009 filters, as previously described (Kirchberger et al., 2016). The isolates were grown in tryptic soy broth (Becton Dickinson) with 1.0% NaCl (BDH), incubated at 30°C and 200 rpm. Genomic DNA was extracted from overnight cultures using the DNeasy Blood and Tissue Kit (QIAGEN) and quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Molecular Probes) and the Synergy H1 microplate reader (BioTek).

5.3.2. Whole-genome sequencing, assembly, and annotation

Whole-genome sequences were obtained from 17 strains each of *V. cholerae* and *V. metoecus* (Table 5.1). The genomes of seven strains (two *V. cholerae* and five *V. metoecus*) had been sequenced previously (Kirchberger et al., 2014; Orata et al., 2015). For the other 27 strains, genomic DNA extracts were prepared for sequencing using the Nextera XT DNA Library Preparation Kit (Illumina). Whole-genome sequencing was performed using the NextSeq 500/550 High Output Kit v2 (for 300 cycles) and the NextSeq sequencing technology (Illumina), generating 150-bp paired-end reads.

Table 5.1. The *V. cholerae* and *V. metoecus* Oyster Pond isolates used in this study.

Clonal complex and sequence type (ST) information are based on multilocus sequence typing with seven housekeeping genes (Kirchberger et al., 2016). Singletons that do not belong to a clonal complex are indicated by ST numbers instead. Genome information was generated from the *de novo* assembly of sequencing reads. Completeness was calculated by determining the presence of 104 core housekeeping genes. ND – not determined.

Species and strain	Isolation date	Clonal complex	Genome size (bp)	GC (%)	Completeness (%)	Accession number*
Vibrio cholerae						
OYP1G01	August, 2009	10	3,969,671	47.4	100	NMTO00000000
OYP2A12	August, 2009	4	4,068,380	47.4	100	NMTN00000000
OYP2E01	August, 2009	2	3,966,741	47.7	100	NMTK00000000
OYP3B05†	August, 2009	13	4,014,368	47.5	100	LBGB00000000
OYP3F10	August, 2009	17	3,937,113	47.6	100	NMTJ00000000
OYP4B01	August, 2009	5	3,929,714	47.6	100	NMTI00000000
OYP4C07†	August, 2009	13	4,015,430	47.5	100	LBGE00000000
OYP4G08	August, 2009	11	3,927,912	47.5	100	NMTH00000000
OYP4H06	August, 2009	5	3,907,548	47.6	99	NMTG00000000
OYP4H11	August, 2009	2	3,934,959	47.7	100	NMTE00000000
OYP6D06	August, 2009	12	4,036,442	47.6	100	NMTC00000000
OYP6E07	August, 2009	16	3,957,612	47.5	100	NMTB00000000
OYP6F08	August, 2009	2	3,912,172	47.7	100	NMTA00000000
OYP6F10	August, 2009	1	3,860,908	47.6	100	NMSZ00000000
OYP7C09	August, 2009	1	3,869,397	47.6	99	NMSX00000000
OYP8C06	September, 2009	3	4,033,034	47.4	100	NMSV00000000
OYP8F12	September, 2009	3	4,038,901	47.4	100	NMSU00000000
Vibrio metoecus						
OP3H†	2006	ND	3,963,175	46.9	100	JJMN00000000
OYP4D01†	August, 2009	ST 25	3,982,476	46.9	100	LBGO00000000
OYP4E03	August, 2009	1	4,098,312	46.8	100	NMST00000000
OYP5B04†	August, 2009	ST 1	4,045,661	46.9	100	LBGP00000000
OYP5B06†	August, 2009	ST 9	3,938,456	46.9	100	LBGQ00000000
OYP5H08	August, 2009	7	3,986,064	46.9	100	NMSR00000000
OYP8G05	September, 2009	3	4,040,107	46.9	100	NMSQ00000000
OYP8G09	September, 2009	ST 10	3,939,012	46.9	100	NMSP00000000
OYP8G12	September, 2009	ST 27	3,890,048	46.9	100	NMSO00000000
OYP8H05	September, 2009	6	4,006,704	46.8	100	NMSN00000000
OYP9B03	September, 2009	6	4,016,639	46.8	100	NMSM00000000
OYP9B09	September, 2009	5	4,027,326	46.9	100	NMSL00000000
OYP9C12	September, 2009	ST 2	3,924,188	46.9	100	NMSK00000000
OYP9D03†	September, 2009	2	3,963,180	46.8	100	LBGR00000000
OYP9D09	September, 2009	1	4,100,806	46.8	100	NMSJ00000000
OYP9E03	September, 2009	2	3,950,427	46.8	100	NMSI00000000
OYP9E10	September, 2009	2	3,914,005	46.8	100	NMSH00000000

*****GenBank BioProject accession number: PRJNA281423.
†Sequenced previously by Kirchberger et al. (2014) and Orata et al. (2015).

Using the CLC Genomics Workbench 7.5.2 (CLC bio), reads were first filtered for quality with the following parameters: quality score limit = 0.05, maximum number of ambiguous nucleotides = 0, discard reads below length = 15. *De novo* assembly of filtered reads into contiguous sequences (contigs) was then performed using the following parameters: word size = 45, bubble size = 98, minimum contig length = 1,000, mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.9, similarity fraction = 0.96. The draft genomes were annotated with RAST 2.0 (Rapid Annotations Using Subsystems Technology; Aziz et al., 2008).

The whole-genome sequences generated in this study have been deposited in the GenBank database under the BioProject accession number PRJNA281423. The individual genome accession numbers are listed in Table 5.1.

5.3.3. Genome completeness and identity

Completeness of the sequenced genomes was assessed by determining the presence or absence of a subset of housekeeping genes using BLAST 2.5.0 (Basic Local Alignment Search Tool; Altschul et al., 1990). Completeness, in this context, was reported as the percentage of 104 single-copy core housekeeping genes present in each genome. The set of genes used was modified from Luo and Moran (2014) with the addition of genes used for multilocus sequence analysis of vibrios (Table E.1; Thompson et al., 2005; Tarr et al., 2007; Gladney and Tarr, 2014; Kirchberger et al., 2014; Orata et al., 2016). Protein sequences encoded by genes from *V. cholerae* N16961 were used as reference Heidelberg et al., 2000.

The identity of the sequenced genomes were confirmed by calculating pairwise average nucleotide identity (ANI) against reference genomes, *V. cholerae* N16961 (Heidelberg et al., 2000) and *V. metoecus* OP3H (Kirchberger et al., 2014), and against each other using JSpecies

1.2.1 (Richter and Rosselló-Móra, 2009). Two genomes belonging to the same species would have an ANI of at least 95% (Goris et al., 2007).

5.3.4. Phylogenetic analysis

The 34 *V. cholerae* and *V. metoecus* genomes were aligned using Mugsy 1.2.3 (Angiuoli and Salzberg, 2011). Locally collinear blocks less than 500 bp were removed using Galaxy 16.04 (Goecks et al., 2010), and alignment positions with at least one gap were then stripped with Geneious 8.1.2 (Kearse et al., 2012). The resulting core-genome alignment, 2,801,207 bp in total length, was used to reconstruct a maximum-likelihood (ML) phylogenetic tree with RAxML 8.0.19 (Stamatakis, 2014) using the GTR (general time reversible) nucleotide substitution model and gamma model of rate heterogeneity. Robustness of branching was estimated with 100 bootstrap replicates.

5.3.5. Quantification of interspecies gene transfers

Orthologous protein-coding gene families were determined from the annotated genomes using the Bacterial Pan Genome Analysis (BPGA) tool 1.3.0 (Chaudhari et al., 2016), which employs USEARCH 9.2.64 (Edgar, 2010), using 30% amino acid identity cutoff (Rost, 1999). The BPGA tool (Chaudhari et al., 2016) was also used to determine the core, accessory, and unique gene families.

For every gene family, nucleotide sequences were aligned with Clustal W 2.1 (Larkin et al., 2007). An ML tree was then reconstructed for each gene alignment with RAxML 8.0.19 (Stamatakis, 2014), as described above. Interspecies core gene transfer events were determined and quantified by tree topology comparisons (Schliep et al., 2011). The trees were partitioned

into clades and visually inspected to determine whether the clades were perfect or not. Following the definition by Schliep et al. (2011), a perfect clade is a partition that is both complete and homogeneous for a given taxonomic category (e.g., a clade with all *V. cholerae* members and only *V. cholerae*). At least one gene transfer event was hypothesized if a tree showed perfect clades for neither *V. cholerae* nor *V. metoecus* (i.e., in a rooted tree, *V. cholerae* and *V. metoecus* are both polyphyletic). The direction of transfer was then inferred if within a clade of one species (the donor) there was a strain from the other species (the recipient). For example, gene transfer from *V. cholerae* to *V. metoecus* is inferred if a strain of *V. metoecus* clusters within the *V. cholerae* clade.

5.3.6. Culture-independent quantification of *V. cholerae* and *V. metoecus* by qPCR

Real-time multiplex qPCR was performed targeting genes specific to *V. cholerae* or *V. metoecus* for the enumeration of each species in the water samples obtained from Oyster Pond and the adjacent lagoon. For *V. cholerae*, the *viuB* gene encoding vibriobactin utilization protein B was targeted using the probe, 5'-/56-FAM/TCA TTT GGC/ZEN/CAG AGC ATA AAC CGG T/3IABkFQ/-3', and forward and reverse primers, 5'-TCG GTA TTG TCT AAC GGT AT-3' and 5'-CGA TTC GTG AGG GTG ATA-3', respectively, for a 77-bp product. For *V. metoecus*, a gene generally annotated as a methyl-accepting chemotaxis protein (MCP) was targeted using the probe, 5'-/5Cy5/TTG TCC GTT TCG ACA CTG AAA TCA/3IAbRQSp/-3', and forward and reverse primers, 5'-GCA GTC TCT TAC CGA AAC ACT A-3' and 5'-ATG AAC AGC TTA TCT TGC CAT TC-3', respectively, for an 81-bp product. The probe and primers were evaluated using the Dynamite qPCR Mastermix (Molecular Biology Service Unit, Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada), which contains Tris (pH

8.3), KCl, $MgCl_2$, glycerol, Tween 20, DMSO, dNTPs, ROX as a normalizing dye, and an antibody-inhibited *Taq* polymerase (Platinum, Thermo Fisher Scientific). A qPCR reaction with 10 µL total volume was prepared containing 5 µL of 2× Dynamite qPCR Mastermix, 1 µL each of 10 mM primer-probe mix, 1 µL of molecular biology-grade water, and 2 µL of DNA template. qPCR assays were performed using the Eco Real-Time PCR System (Illumina) under the following reaction conditions: initial denaturation at 95°C for 2 min, followed by 40 cycles of 95°C for 15 s and 60°C for 1 min. The assay included standards of known copy number and no-template control for each master mix to assure the reaction was contamination-free. The limit of detection was determined to be three copies per reaction.

5.3.7. Natural transformation of *V. cholerae* and *V. metoecus* on chitin flakes

Overnight cultures of *V. cholerae* and *V. metoecus* were diluted to 1:100 in Luria-Bertani broth (LB; Becton Dickinson) and incubated with shaking (200 rpm) at 30°C for 2.5 h to an OD_{600} of 0.5–0.8. Eight hundred microliters of each culture were transferred to a 1.5 mL tube and centrifuged at maximum speed (16,873 × g) in an Eppendorf 5418 centrifuge. The supernatant was discarded, and 1 mL of defined artificial seawater (DASW; Meibom et al., 2005), supplemented with HEPES and vitamins (Gibco MEM vitamin solution 100×; Thermo Fisher Scientific) in a 1:100 dilution, was added to resuspend the pellet. To a 1.5 mL tube containing 80 mg of sterile chitin flakes (Sigma-Aldrich), 500 µL each of resuspended culture and DASW were added. The tubes were incubated at 30°C for 24 h.

Genomic DNA was extracted from a *V. cholerae* C6706 transposon mutant (Cameron et al., 2008) using the GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich). The mutated gene, VC0290, encoding a factor-for-inversion stimulation protein, was selected based on high

nucleotide identity of end regions from the gene alignment of *V. cholerae* and *V. metoecus* sequences (Figure E.1). From this, 2.2 μg of genomic DNA were added to the tubes, and the tubes were re-incubated at 30°C for 24 h.

The tubes were vortexed for 10 s, and 200 μL of supernatant were transferred into the first row of a 96-well plate, avoiding any chitin flakes. The cultures were serially diluted (10-fold) in LB up to a dilution of 1×10^{-6}. Five microliters from all dilutions were spot-plated on LB plates with and without kanamycin (50 μg/mL; Sigma-Aldrich), and the plates were incubated at 30°C for 16 h. The colonies were then counted, and the transformation frequency was calculated by dividing the number of transformants (that grew on LB-kanamycin plates) by the total number of bacteria (that grew on LB-only plates).

Vibrio cholerae A1552 was included as a control in the experiment, with and without added genomic DNA. This strain was previously shown to be competent for natural transformation during growth on a chitin surface (Meibom et al., 2005). All samples were done in duplicate.

5.4. Results and Discussion

Vibrio cholerae is a well-studied organism with 747 whole-genome sequences available in public databases, as of October 1, 2017. Multiple comparative genomic studies have been performed with *V. cholerae* genomes to determine the population structure and genetic diversity of the species, with most studies focusing on clinical cases. In a more specific case, comparative genomics proved to be a very useful tool in determining the source of the cholera outbreak in Haiti (Chin et al., 2011; Hendriksen et al., 2011; Reimer et al., 2011; Katz et al., 2013; Eppinger et al., 2014; Orata et al., 2014). On the other hand, *V. metoecus*, the closest known relative to *V.*

cholerae, is a more recently described species (Kirchberger et al., 2014) and is poorly understood. The genomes of only 11 strains from disparate geographical sources had been sequenced prior to this study (Haley et al., 2010; Kirchberger et al., 2014; Orata et al., 2015; Carda Diéguez, 2016).

Here, we obtained genome sequences from *V. cholerae* and *V. metoecus* co-isolated from the same location, Oyster Pond (Falmouth, MA, USA), a cholera-free environment (Boucher et al., 2011; Kirchberger et al., 2016). Although using a single population does not allow for an accurate description of the diversity within the species, it enables evolutionarily and ecologically meaningful observations resulting from the interaction of these two species. The limiting factor for the size of the dataset was the isolation of *V. metoecus* strains, for which a total of 17 strains with genome sequences of high quality was obtained. *Vibrio cholerae* genomes of an equal number of strains isolated from the same samples were therefore selected, matching the genetic diversity of the *V. metoecus* genomes to avoid bias (Table 5.1). A phylogenetic tree reconstructed from the core-genome alignment of the 34 genomes clearly placed our *V. cholerae* and *V. metoecus* isolates in clades distinct from each other (Figure 5.1). Pairwise ANI was calculated to confirm the identity of the sequenced genomes. Any pair of genomes belonging to the same species should have at least have 95% ANI (Goris et al., 2007). Our ANI results indicate proper species assignment, with at least 97–98% ANI within the *V. cholerae* or *V. metoecus* species, and 86–87% between species (Table E.2 and Table E.3).

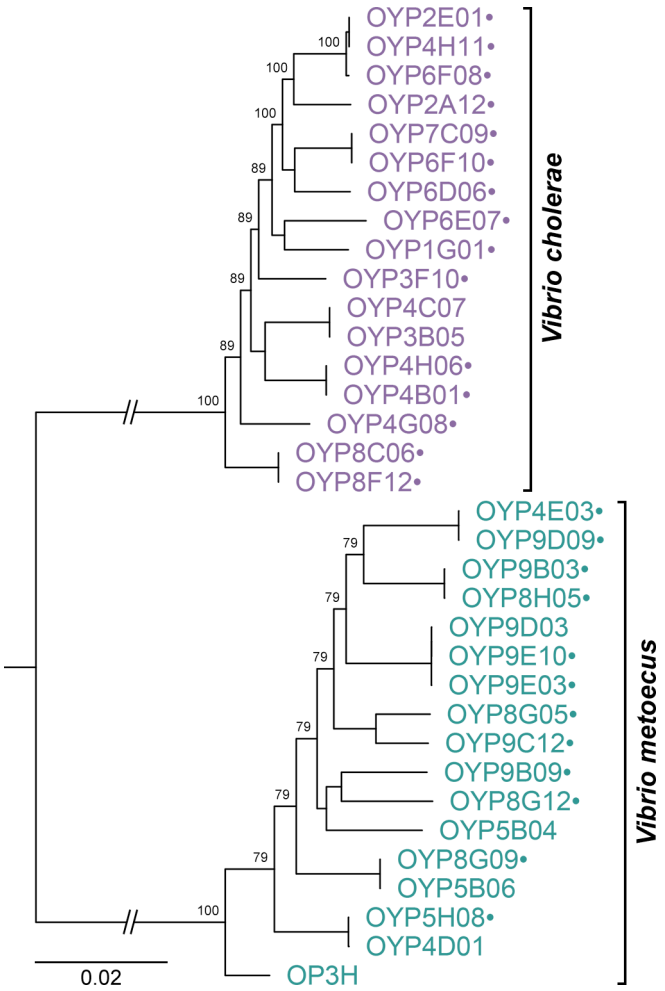

Figure 5.1. The phylogenetic relationship of *V. cholerae* and *V. metoecus*.

The phylogenetic tree was reconstructed from a core-genome alignment (≈ 2.8 Mb). Bootstrap support values are indicated on the nodes. Bar, 0.02 nucleotide substitutions per site. Parallel lines indicate shortened branch lengths, approximately 5× the scale bar. Dots after the strain names indicate isolates with their genomes sequenced in this study.

5.4.1. There is a bias in gene transfer from *V. cholerae* to *V. metoecus*

The core genome, which consists of genes that are shared by all strains (Vernikos et al., 2015), was determined for our dataset of 17 *V. cholerae* and 17 *V. metoecus* genomes (Figure 5.2). An ML tree was reconstructed for each of the 2,675 core gene families. Phylogenetic trees were also reconstructed for the accessory genes, defined as genes present in a subset of strains (Vernikos et al., 2015). We only examined accessory gene families with at least 17 members

(half of the number of genomes, 621 gene families in total; Figure 5.2). A gene transfer event and its directionality were inferred in a tree, whether from *V. cholerae* to *V. metoecus* or vice versa, if a member of one species, the recipient, clustered with the clade of the other species, the donor, with robust bootstrap support (≥ 70%; Hillis and Bull, 1993). If HGT had occurred between the two species, the tree could not be partitioned into two perfect clades, each containing members from one species and only of that species (Schliep et al., 2011). Out of the 2,675 core gene family trees, 554 (20.7%) could not be partitioned into perfect clades (i.e., has at least one HGT event; Figure 5.3A and Table E.4). From the 554 trees exhibiting HGT, 1,368 gene transfer events were inferred. *Vibrio metoecus* was a recipient of genes from *V. cholerae* in 1,027 (75.1%) of those transfer events, whereas *V. cholerae* was a recipient in only 341 (24.9%). Additionally, from the 621 accessory gene family trees, 228 were excluded from our analysis, as they consisted of members from one species only or did not have robust bootstrap support (< 70%) to infer HGT. From the remaining 393 trees, 178 (45.3%) exhibited HGT, more than twice the proportion found in the core genome (Figure 5.3B and Table E.5). Here, *V. metoecus* was a recipient in 429 out of 783 (54.8%) transfer events, whereas *V. cholerae* was a recipient in 354 (45.2%). Overall, there is a higher number of gene transfer events from *V. cholerae* to *V. metoecus*, indicating a bias in the direction of gene transfer in the Oyster Pond populations. This is more prominent within the core genome, where *V. metoecus* was three times (3.0×) more often the recipient of DNA from *V. cholerae* as it was the donor (only 1.2× in the accessory genome). The accessory genes might encode supplementary functions that are not necessarily essential for growth but may offer selective advantages, such as niche adaptation, antibiotic resistance, or host colonization (Vernikos et al., 2015). On the contrary, the core genome, which contains genes essential for growth, is under strong selective pressure, limiting the extent of sequence changes

and preventing gene loss (Lapierre and Gogarten, 2009). For this reason, transfer of core genes is mostly limited to homologous recombination while HGT of accessory genes can occur through various mechanisms (homologous and heterologous recombination) and is found to be more frequent for most species (Lapierre and Gogarten, 2009). The more pronounced HGT directional bias observed for core genes as opposed to accessory genes could be linked to the fact that they are more abundant, being present in all members of the population (Vernikos et al., 2015). Because of their ubiquity and a mechanistically more uniform mode of DNA integration (homologous recombination), core gene HGT is likely to be more intimately connected to the abundance of the donor DNA than accessory gene HGT.

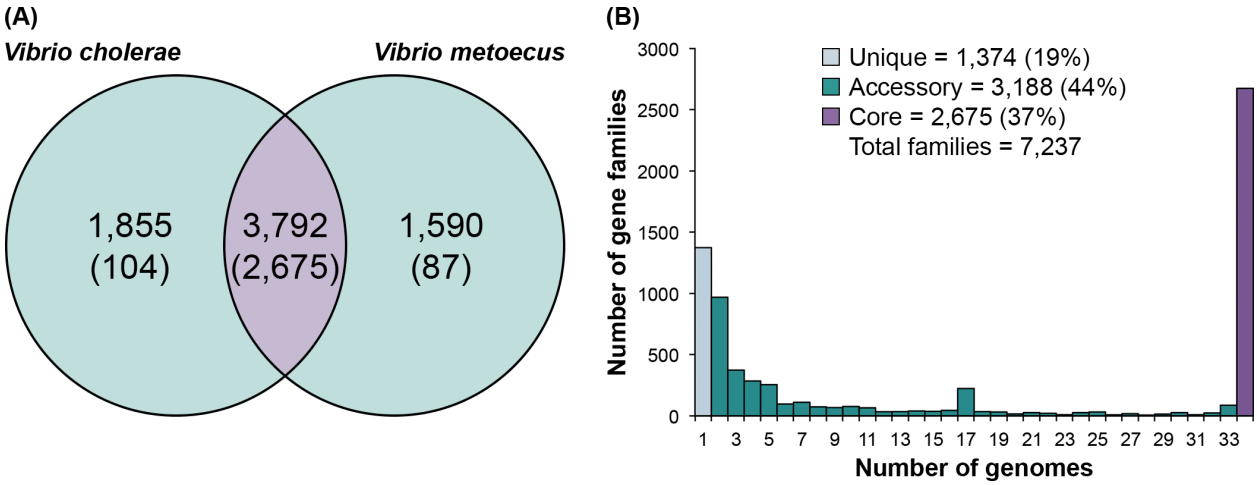

Figure 5.2. Distribution of unique, accessory, and core genes among the *V. cholerae* and *V. metoecus* isolates.

(A) Venn diagram showing the number of genes shared between *V. cholerae* and *V. metoecus* or uniquely present within each species. Top numbers indicate total gene families, including accessory genes and singletons, present in some or all strains of *V. cholerae*, *V. metoecus*, or both; bottom numbers in parentheses indicate gene families exclusively present in all strains of *V. cholerae*, *V. metoecus*, or both. (B) Histogram showing the number of unique, accessory, and core gene families against the number of genomes present in these gene families.

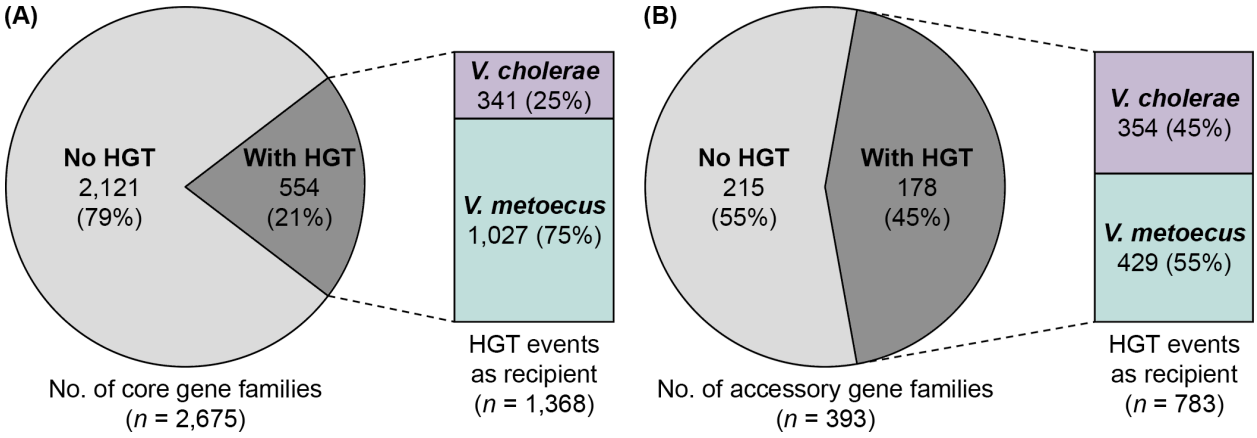

Figure 5.3. Horizontal gene transfer events in the core and accessory genomes.

The pie charts show the number of (A) core and (B) accessory gene families with and without HGT. The bars show the number of transfer events within the gene families exhibiting HGT, showing the frequency of *V. cholerae* or *V. metoecus* as recipients.

5.4.2. *Vibrio cholerae* is more abundant and present over longer periods of time than *V. metoecus*

The directional bias in gene transfer events could be due to a higher abundance of *V. cholerae* than *V. metoecus* in Oyster Pond, as isolation data has suggested that *V. cholerae* is ten times more abundant than *V. metoecus* at that location (Kirchberger et al., 2016). However, isolation rarely recovers all strains present in a given environmental sample (Amann et al., 1995) and vibrios have been shown to be susceptible to a VBNC state (Xu et al., 1982; Huq et al., 1990; Alam et al., 2007). To limit the bias in estimating the abundance of both species, the amount of *V. cholerae* or *V. meteocus* in Oyster Pond and the adjacent lagoon was quantified by multiplex qPCR, a culture-independent approach. Two sets of qPCR primers were designed to target only *V. cholerae* or *V. metoecus* based on our dataset of genes unique for each species (Figure 5.2). *Vibrio cholerae* and *V. metoecus* were quantified in DNA extracted from water sampled in the summers of 2008 and 2009 (June to September). The presence of 1.6×10^4 to 8.2×10^5 *V. cholerae*/L were detected in Oyster Pond water across all eight months of sampling (two

consecutive summers), we were only able to detect 2.8×10^4 to 1.0×10^5 *V. metoecus*/L of water from August and September in both years (Figure 5.4A and C). Consistently, *V. cholerae* was also detected in all lagoon water samples (2.7×10^4 to 7.2×10^5/L of water) and *V. metoecus* only in August and September in both years (4.4×10^4 to 2.5×10^5/L of water; Figure 5.4B and D). For the two months in which both species are detected, *V. cholerae* is, on average, three times more abundant than *V. metoecus* in both Oyster Pond (2.8×) and the nearby lagoon (3.0×; Figure 5.4). This suggests that the directionality bias in HGT could be caused by a higher abundance of *V. cholerae* in the environment. This higher abundance would lead to having more DNA from *V. cholerae* readily available for acquisition by *V. metoecus*, making it a decisive factor in biasing HGT directionality.

We cannot ignore the possibility that *V. cholerae* could be more refractory to HGT than *V. metoecus*, as the former may contain more barriers to gene uptake or the latter contains more DNA uptake systems making it more permissive for gene acquisition. From our annotated genomes, we surveyed the presence or absence of genes that are involved in the regulation of competence or providing barriers to DNA uptake (Thomas and Nielsen, 2005; Seitz and Blokesch, 2013; Borgeaud et al., 2015; Watve et al., 2015). Major regulators and genes of the DNA uptake system (Seitz and Blokesch, 2013; Borgeaud et al., 2015; Watve et al., 2015) are present in all of our isolates (Figure E.2). On the other hand, some isolates of *V. cholerae* and *V. metoecus* were missing some or all genes of the restriction-modification system, an immune system in bacteria that recognize self from non-self (foreign) DNA (Vasu and Nagaraja, 2013). Nuclease activity can also inhibit natural transformation (Blokesch and Schoolnik, 2008; Dalia et al., 2015), and the genes encoding the deoxyribonucleases Dns and Xds are present in all our isolates (Figure E.2). Finally, we tested the competence of our isolates when grown on chitin

flakes (Meibom et al., 2005). Only two of our isolates were competent using this assay (Table E.6), showing that, under the conditions tested, there is no significant difference in competence between *V. cholerae* and *V. metoecus*. Since all strains have a full complement of competence genes (Figure E.2), it is likely that the assay conditions simply do not allow for expression of the machinery. In fact, Bernardy et al. (2016) demonstrated previously that natural transformation in clinical and environmental isolates of *V. cholerae* is rare under tested experimental conditions.

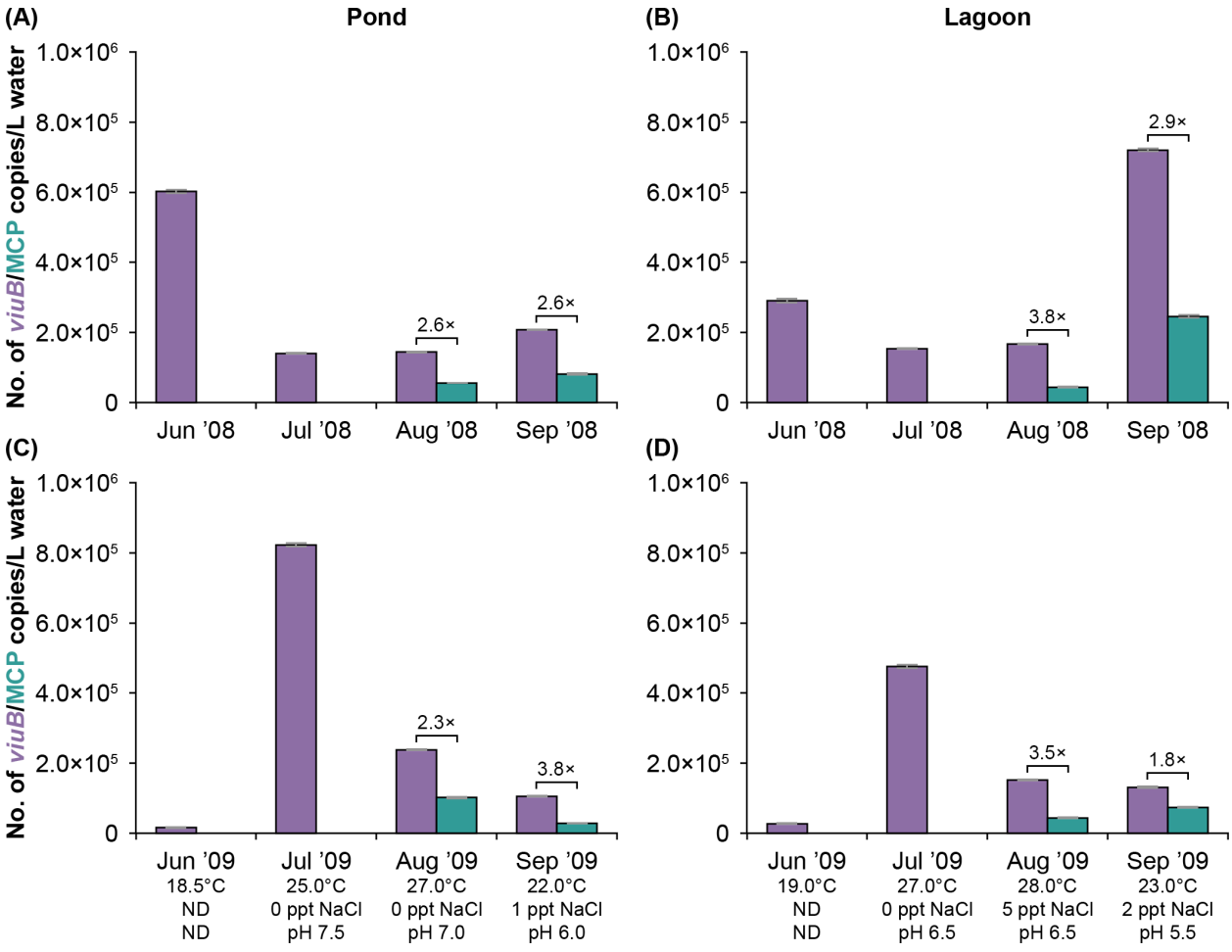

Figure 5.4. Quantification of *V. cholerae* and *V. metoecus* in Oyster Pond and the adjacent lagoon by qPCR.

(Figure legend continued on the next page)

Figure 5.4. Quantification of *V. cholerae* and *V. metoecus* in Oyster Pond and the adjacent lagoon by qPCR. Multiplex qPCR was conducted using DNA extracted from water samples collected from (A, C) Oyster Pond and (B, D) the lagoon from June to September (A, B) 2008 and (C, D) 2009. The genes encoding the vibriobactin utilization protein B (*viuB*) and a methyl-accepting chemotaxis protein (MCP), which are uniquely present in *V. cholerae* or *V. metoecus*, respectively, were targeted during qPCR. Mean values from triplicate qPCR runs are shown. Error bars indicate the standard deviation between mean values. The relative abundance of *V. cholerae* over *V. metoecus* is indicated for August and September 2008 and 2009 (average = 2.8×, pond; 3.0×, lagoon; 2.9×, pond and lagoon). ND, not determined.

5.5. Conclusion

Although seasonal abundance patterns have been found for *V. cholerae*, which is more abundant in the warmer months of the year and rarely detectable in winter (Jiang and Fu, 2001), those observed for *V. metoecus* are much more striking. The species is essentially limited to two months of the year (August and September). Water temperature alone cannot explain this limited temporal range, as July (before bloom) and August (start of bloom) do not differ significantly in temperature, at least in 2009 (Figure 5.4). There is also no appreciable change in salinity or pH between July and September in Oyster Pond (Figure 5.4). *Vibrio metoecus* could be tracking a specific eukaryotic host (phytoplankton or zooplankton) and/or a set of nutrients that becomes abundant in August/September. This bloom is remarkable, as *V. metoecus* would need to compete with established *V. cholerae* strains in the same environment. This competition may be facilitated by methods such as antagonistic interactions; both species have been observed to kill each other via their type VI secretion system (T6SS; N.A.S. Hussain, unpublished). The T6SS is used by bacteria to inject toxins (effectors) into neighbouring cells to lyse and eliminate competitors that do not express the cognate immunity proteins against these effectors (Joshi et al., 2017). Since the T6SS is co-regulated with competence in *V. cholerae* (Borgeaud et al., 2015), it is possible that strong competition from *V. metoecus* when it is blooming would release *V. cholerae* genetic material into the environment. The uptake of DNA by *V. metoecus* from its

victim during its rapid rise in August could also help to explain the directional bias of HGT observed.

To our knowledge, this is the first time that a relationship between the rate of HGT and the abundance of donor and recipient species has been shown. Although the direct link between abundance and rate of HGT has not been experimentally demonstrated, there is an exact correlation between these two factors in a natural environment. A careful genetic screen also shows that there is no reason to suspect a differential ability between the two species to uptake DNA, making differential abundance the most likely explanation for the bias. If a direct relationship between relative abundance of a species in a community and its intake of foreign DNA is discovered for other organisms, it would have major implication for the dynamics of speciation and the spread of various fitness characters between species.

5.6. References

Alam, M., Sultana, M., Nair, G.B., Siddique, A.K., Hasan, N.A., Sack, R.B., et al. (2007). Viable but nonculturable *Vibrio cholerae* O1 in biofilms in the aquatic environment and their role in cholera transmission. *Proceedings of the National Academy of Sciences of the United States of America* 104(45)**,** 17801-17806.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215(3)**,** 403-410.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiological Reviews* 59(1)**,** 143-169.

Angiuoli, S.V., and Salzberg, S.L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27(3)**,** 334-342.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9**,** 75.

Bernardy, E.E., Turnsek, M.A., Wilson, S.K., Tarr, C.L., and Hammer, B.K. (2016). Diversity of clinical and environmental isolates of *Vibrio cholerae* in natural transformation and contact-dependent bacterial killing indicative of type VI secretion system activity. *Applied and Environmental Microbiology* 82(9)**,** 2833-2842.

Blokesch, M., and Schoolnik, G.K. (2008). The extracellular nuclease Dns and its role in natural transformation of *Vibrio cholerae*. *Journal of Bacteriology* 190(21)**,** 7232-7240.

Borgeaud, S., Metzger, L.C., Scrignari, T., and Blokesch, M. (2015). The type VI secretion system of *Vibrio cholerae* fosters horizontal gene transfer. *Science* 347(6217)**,** 63-67.

Boucher, Y., Cordero, O.X., Takemura, A., Hunt, D.E., Schliep, K., Bapteste, E., et al. (2011). Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *MBio* 2(2)**,** e00335-10.

Cameron, D.E., Urbach, J.M., and Mekalanos, J.J. (2008). A defined transposon mutant library and its use in identifying motility genes in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 105(25)**,** 8736-8741.

Carda Diéguez, M. (2016). *Vibrio vulnificus: from water to host.* Doctor of Philosophy Thesis, University of Valencia.

Chaudhari, N.M., Gupta, V.K., and Dutta, C. (2016). BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports* 6**,** 24373.

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., et al. (2011). The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 364(1)**,** 33-42.

Dalia, A.B., Seed, K.D., Calderwood, S.B., and Camilli, A. (2015). A globally distributed mobile genetic element inhibits natural transformation of *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 112(33)**,** 10485-10490.

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

Eppinger, M., Pearson, T., Koenig, S.S., Pearson, O., Hicks, N., Agrawal, S., et al. (2014). Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *MBio* 5(6)**,** e01721-14.

Faruque, S.M., and Mekalanos, J.J. (2003). Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends in Microbiology* 11(11)**,** 505-510.

Faruque, S.M., Zhu, J., Asadulghani, Kamruzzaman, M., and Mekalanos, J.J. (2003). Examination of diverse toxin-coregulated pilus-positive *Vibrio cholerae* strains fails to demonstrate evidence for *Vibrio* pathogenicity island phage. *Infection and Immunity* 71(6)**,** 2993-2999.

Gladney, L.M., and Tarr, C.L. (2014). Molecular and phenotypic characterization of *Vibrio navarrensis* isolates associated with human illness. *Journal of Clinical Microbiology* 52(11)**,** 4070-4074.

Goecks, J., Nekrutenko, A., Taylor, J., and The Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11(8)**,** R86.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57(1)**,** 81-91.

Haley, B.J., Grim, C.J., Hasan, N.A., Choi, S.Y., Chun, J., Brettin, T.S., et al. (2010). Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. *BMC Microbiology* 10**,** 154.

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., et al. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406(6795)**,** 477-483.

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., et al. (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* 2(4)**,** e00157-11.

Hillis, D.M., and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42(2)**,** 182-192.

Huq, A., Colwell, R.R., Rahman, R., Ali, A., Chowdhury, M.A., Parveen, S., et al. (1990). Detection of *Vibrio cholerae* O1 in the aquatic environment by fluorescent-monoclonal antibody and culture methods. *Applied and Environmental Microbiology* 56(8)**,** 2370-2373.

Jiang, S.C., and Fu, W. (2001). Seasonal abundance and distribution of *Vibrio cholerae* in coastal waters quantified by a 16S-23S intergenic spacer probe. *Microbial Ecology* 42(4)**,** 540-548.

Joshi, A., Kostiuk, B., Rogers, A., Teschler, J., Pukatzki, S., and Yildiz, F.H. (2017). Rules of engagement: the type VI secretion system in *Vibrio cholerae*. *Trends in Microbiology* 25(4)**,** 267-279.

Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., et al. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4(4)**,** e00398-13.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12)**,** 1647-1649.

Keymer, D.P., and Boehm, A.B. (2011). Recombination shapes the structure of an environmental *Vibrio cholerae* population. *Applied and Environmental Microbiology* 77(2)**,** 537-544.

Kirchberger, P.C., Orata, F.D., Barlow, E.J., Kauffman, K.M., Case, R.J., Polz, M.F., et al. (2016). A small number of phylogenetically distinct clonal complexes dominate a coastal *Vibrio cholerae* population. *Applied and Environmental Microbiology* 82(18)**,** 5576-5586.

Kirchberger, P.C., Turnsek, M., Hunt, D.E., Haley, B.J., Colwell, R.R., Polz, M.F., et al. (2014). *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. *International Journal of Systematic and Evolutionary Microbiology* 64(9)**,** 3208-3214.

Lapierre, P., and Gogarten, J.P. (2009). Estimating the size of the bacterial pan-genome. *Trends in Genetics* 25(3)**,** 107-110.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21)**,** 2947-2948.

Luo, H., and Moran, M.A. (2014). Evolutionary ecology of the marine *Roseobacter* clade. *Microbiology and Molecular Biology Reviews* 78(4)**,** 573-587.

Majewski, J., and Cohan, F.M. (1999). DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* 153(4)**,** 1525-1533.

Mazel, D., Dychinco, B., Webb, V.A., and Davies, J. (1998). A distinctive class of integron in the *Vibrio cholerae* genome. *Science* 280(5363)**,** 605-608.

Meibom, K.L., Blokesch, M., Dolganov, N.A., Wu, C.Y., and Schoolnik, G.K. (2005). Chitin induces natural competence in *Vibrio cholerae*. *Science* 310(5755)**,** 1824-1827.

Orata, F.D., Keim, P.S., and Boucher, Y. (2014). The 2010 cholera outbreak in Haiti: how science solved a controversy. *PLoS Pathogens* 10(4)**,** e1003967.

Orata, F.D., Kirchberger, P.C., Méheust, R., Barlow, E.J., Tarr, C.L., and 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(10)**,** 2941-2954.

Orata, F.D., Xu, Y., Gladney, L.M., Rishishwar, L., Case, R.J., Boucher, Y., et al. (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(10)**,** 4148-4155.

Reimer, A.R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., et al. (2011). Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerging Infectious Diseases* 17(11)**,** 2113-2121.

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

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering* 12(2)**,** 85-94.

Schliep, K., Lopez, P., Lapointe, F.J., and Bapteste, É. (2011). Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Molecular Biology and Evolution* 28(4)**,** 1393-1405.

Seitz, P., and Blokesch, M. (2013). DNA-uptake machinery of naturally competent *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 110(44)**,** 17987-17992.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* 16(8)**,** 472-482.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9)**,** 1312-1313.

Tarr, C.L., Patel, J.S., Puhr, N.D., Sowers, E.G., Bopp, C.A., and Strockbine, N.A. (2007). Identification of *Vibrio* isolates by a multiplex PCR assay and *rpoB* sequence determination. *Journal of Clinical Microbiology* 45(1)**,** 134-140.

Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology* 3(9)**,** 711-721.

Thompson, F.L., Gevers, D., Thompson, C.C., Dawyndt, P., Naser, S., Hoste, B., et al. (2005). Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. *Applied and Environmental Microbiology* 71(9)**,** 5107-5115.

Vasu, K., and Nagaraja, V. (2013). Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiology and Molecular Biology Reviews* 77(1)**,** 53-72.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Waldor, M.K., and Mekalanos, J.J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272(5270)**,** 1910-1914.

Watve, S.S., Thomas, J., and Hammer, B.K. (2015). CytR is a global positive regulator of competence, type VI secretion, and chitinases in *Vibrio cholerae*. *PLoS One* 10(9)**,** e0138834.

Xu, H.S., Roberts, N., Singleton, F.L., Attwell, R.W., Grimes, D.J., and Colwell, R.R. (1982). Survival and viability of nonculturable *Escherichia coli* and *Vibrio cholerae* in the estuarine and marine environment. *Microbial Ecology* 8(4)**,** 313-323.

CHAPTER 6: Determining the Origin of the 2010 Cholera Outbreak in Haiti

---

A version of this chapter has been published as:

Orata, F.D., Keim, P.S., and Boucher, Y. (2014). The 2010 cholera outbreak in Haiti: how science solved a controversy. *PLoS Pathogens* 10(4), e1003967.

F.D.O. and Y.B. wrote the manuscript, with revisions from P.S.K.; and F.D.O. created the figures.

CHAPTER 6

6.1. The 2010 earthquake and cholera outbreak in Haiti

On January 12, 2010, a catastrophic 7.0 magnitude earthquake struck Haiti, affecting 3,500,000 people (United States Geological Survey, 2010; Farmer, 2013). This severely damaged an already marginal public sanitation system, creating ideal conditions for outbreaks of major infectious diseases. In October 2010, nine months after the earthquake, an outbreak of cholera started, which quickly spread all across the country (Delva, 2010). As of January 7, 2014, 8,534 deaths and 697,256 cholera cases have been reported by the Haitian Ministry of Public Health and Population (Ministère de la Santé Publique et de la Population, 2014). Prior to 2010, there was no reported history of cholera in Haiti, despite devastating outbreaks in the Caribbean region in the 19th century (Jenson et al., 2011). Many wondered where the cholera in Haiti came from. Two hypotheses as to its origin were presented. The climatic hypothesis argued that nonpathogenic *Vibrio cholerae*, indigenous in the coastal waters of Haiti, was given the right environmental circumstances and evolved into a pathogenic strain (Parker, 2010). On the other hand, the human transmission hypothesis suggested that cholera was introduced to Haiti by individuals infected in a foreign country.

6.2. Cholera and *V. cholerae*

Cholera, caused by *V. cholerae*, is a disease characterized by very severe diarrhea and dehydration, which can lead to death in less than 48 hours if left untreated. Cholera is treatable through oral rehydration salt solutions, intravenous fluids, or antibiotics, depending on severity (Harris et al., 2012). Ingestion of contaminated water is the main vehicle for human infection.

The principal virulence determinant is the potent cholera toxin, encoded by the *ctxAB* genes on the bacteriophage CTXΦ (Waldor and Mekalanos, 1996) found in toxigenic *V. cholerae* genomes. The toxin, together with other virulence factors, leads to the harmful effects of the *V. cholerae* infection (Figure 6.1). These auxiliary virulence factors are encoded in clusters of genes called genomic islands, which are acquired by environmental *V. cholerae* through horizontal gene transfer (Figure 6.1; De la Cruz and Davies, 2000). It is also important to note that infection can be asymptomatic, and these cases play a major role in the transmission of the disease (Benčić and Sinha, 1972). *Vibrio cholerae* is of major public health concern because of its potential to cause pandemics. Seven such pandemics have been recorded since 1817, when cholera first spread beyond the Indian subcontinent, all presumably caused by *V. cholerae* belonging to the O1 serogroup. *Vibrio cholerae* of the classical biotype dominated the previous six pandemics and was replaced by the El Tor biotype in the currently ongoing seventh pandemic, which originated in Southeast Asia in 1961 (Harris et al., 2012). In 1992, a new serogroup of *V. cholerae*, O139, was first identified after causing cholera epidemics in India and Bangladesh (Shimada et al., 1993). Cholera has been eliminated from industrialized countries by efficient water and sewage treatments but not in less-developed countries with poor water sanitation.

6.3. Initial studies support the human transmission hypothesis

Rumors spread on October 27, 2010, pointing blame for the outbreak at the United Nations Stabilization Mission in Haiti (MINUSTAH) troops from Nepal who had recently set up camp in Meille, a small village 2 km south of Mirebalais (Figure 6.2A). This followed revelations by news reporters showing improper sewage waste disposal in the camp (Al Jazeera English, 2010; Katz, 2010).

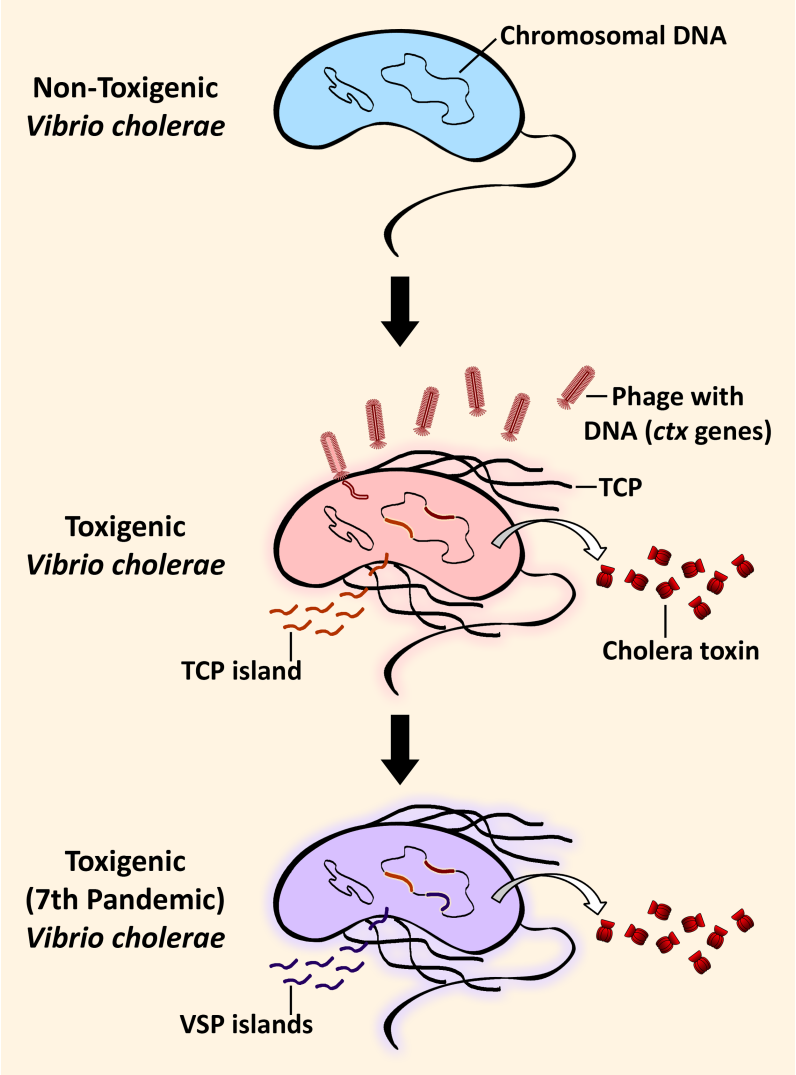

Figure 6.1. Steps in the evolution of the seventh pandemic *V. cholerae*.

Environmental V. cholerae indigenous in coastal waters can harbor genomic islands (GIs) by horizontal gene transfer, rendering it pathogenic. Pathogenesis of toxigenic (toxin-producing) V. cholerae critically depends on the production of the cholera toxin, which is responsible for the cholera symptoms, and the toxin-coregulated pilus (TCP). The genes for the cholera toxin (ctx) are from the filamentous bacteriophage, CTXΦ, that has been incorporated into the genome. The genes in the TCP island encode factors necessary for the colonization of the small intestine in the human host after ingestion of contaminated water. Additionally, seventh pandemic strains are distinguishable from pre-seventh pandemic strains due to the acquisition of additional GIs, the Vibrio seventh pandemic (VSP) islands.

The stool samples collected by the Haiti National Public Health Laboratory from cholera patients at the start of the outbreak were sent to the Centers for Disease Control and Prevention (CDC) for analysis. On November 13, the CDC reported that *V. cholerae* El Tor O1 was isolated

from the samples and independent isolates were indistinguishable by multiple rapid phenotypic and molecular characterization methods, suggesting that a single strain caused the outbreak and was likely introduced into Haiti in one event (Centers for Disease Control and Prevention, 2010).

A study by Piarroux and colleagues made use of all available epidemiological data, checking hospital records, doing field surveys, and applying statistics for spatiotemporal analysis, to trace the source and spread of the outbreak (Piarroux et al., 2011). The findings of their investigation confirmed the claims of news reporters. Based on all gathered evidence, they put together a likely scenario – the MINUSTAH camp contaminated the Meille tributary with fecal matter by their unsanitary practice of sewage drainage, and the Meille tributary connects downstream to the Latem River that goes through the town of Mirebalais, the site of the first reported cholera case (Ivers and Walton, 2012). The Latem River eventually connects to the Artibonite River, the longest as well as the most important river that spans Haiti (Figure 6.2A). The movement and spread of cholera in the early onset of the epidemic was closely linked to proximity with the Artibonite River.

It had been reported that Kathmandu, the capital of Nepal, where the troops trained shortly before being stationed to Haiti, experienced a cholera outbreak on September 23 (Maharjan, 2010). The first batch of troops arrived in Haiti on October 8 (Lantagne et al., 2014), and the first cholera case was reported on October 12 (Figure 6.2B; Ivers and Walton, 2012). Because none of the troops apparently exhibited symptoms of cholera during the pre-deployment medical examination, the MINUSTAH chief medical officer later revealed that no follow-up tests were done (BBC News, 2010a). However, the absence of symptoms did not prove that the troops were *V. cholerae*-free, as they could have been infected in the days following the medical examination and prior to deployment, or they could have been asymptomatic carriers (Benčić

and Sinha, 1972; Piarroux et al., 2011). Unfortunately, other than that done by the MINUSTAH, no independent testing was done of the troops to confirm the presence or absence of *V. cholerae*.

6.4. Comparative genomics traced a single source for the epidemic

The first molecular study on the origin of *V. cholerae* in Haiti was published on December 9, 2010 (Chin et al., 2011). Chin and colleagues sequenced the complete genomes of two Haitian strains obtained from the outbreak, as well as epidemic strains from South America and Bangladesh, and compared them to those of epidemic-associated strains available in public databases. Comparison of single-nucleotide variations and hypervariable chromosomal elements in the genomes showed both of these Haitian strains to be genetically identical. While this is a small sample size, it was consistent with a clonal source for the outbreak. In addition, the study was able to genotype the two strains at polymorphic loci previously used for population genetic studies of *V. cholerae* (Lam et al., 2010), and this subtype had been previously observed in a broad region that included South Asia, Thailand, and Africa, but not the Americas. The study by Chin and colleagues suggested that cholera was introduced into Haiti through human transmission from a distant geographic source, most probably from South Asia (i.e., Bangladesh), although their conclusions were based upon a very limited strain analysis from both Haitian and global populations.

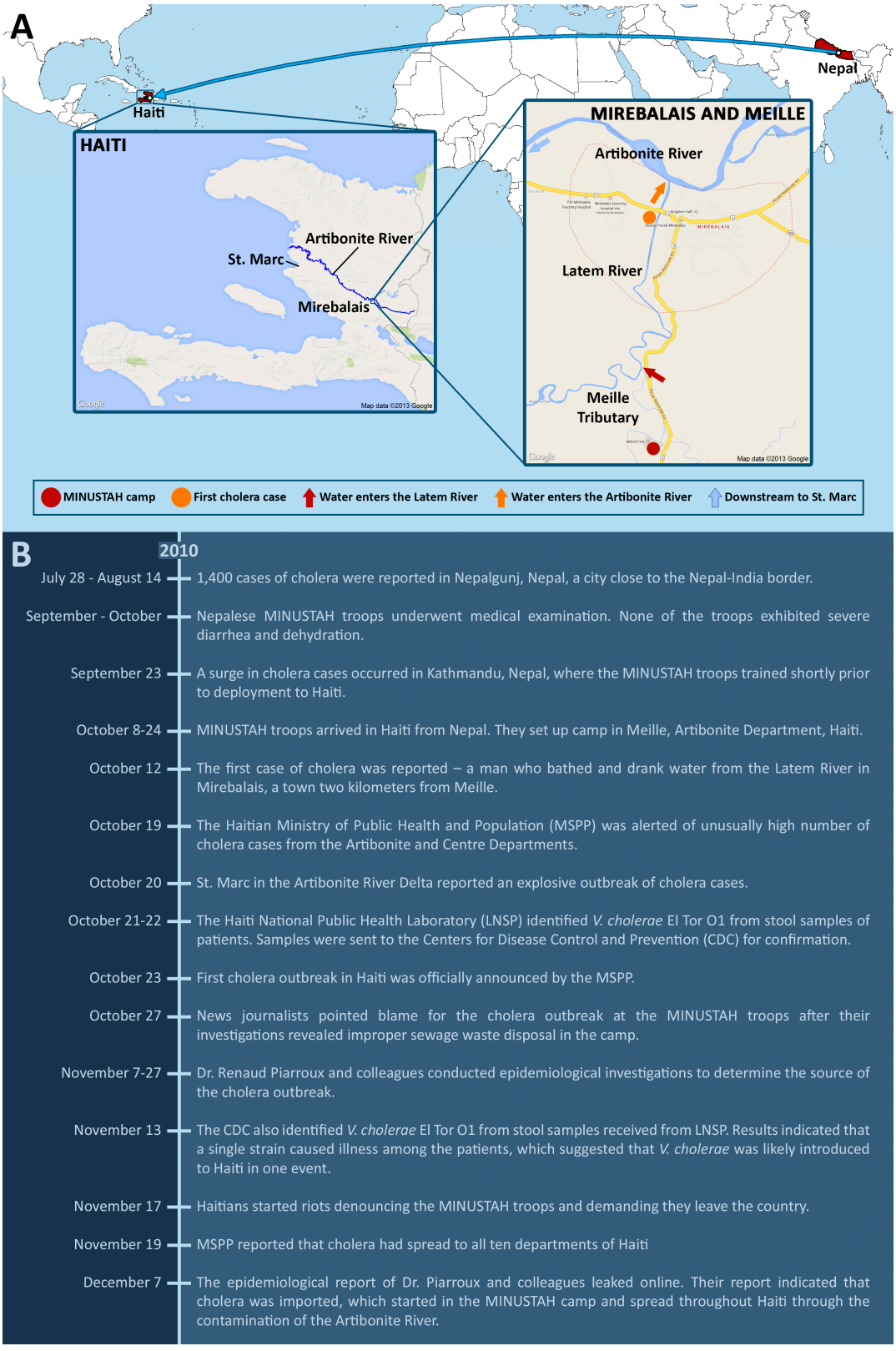

Figure 6.2. How the Haiti cholera outbreak started.

(Figure legend continued on the next page)

Figure 6.2. How the Haiti cholera outbreak started. (A) MINUSTAH troops from Nepal were stationed in Haiti starting on October 8, 2010, and set up camp in Meille (red circle). Improper disposal of sewage led to the contamination of the Meille tributary, which connects downstream to the Latem River (red arrow). The first case of cholera occurred on October 12 along the Latem River in Mirebalais (orange circle), 2 km north of Meille. Water from the Latem River enters the Artibonite River (orange arrow), the major river that spans across Haiti, which flows downstream to St. Marc (blue arrow). The Artibonite River played a significant role in the rapid spread of cholera. During the early onset of the epidemic, reported cases were linked to proximity with the river. (B) A chronological timeline of events involving the Haiti cholera outbreak from July to December 2010. A more extensive timeline can be found in Table F.1.

Two subsequent and larger genomic studies used 23 (Reimer et al., 2011) and 154 (Mutreja et al., 2011) whole genome sequences to document the repeated historical spread of *V. cholerae* O1 from South Asia. These studies used up to nine more Haitian isolates and placed them into the context of the expanded strain genome collection. They found phylogenetic affinity between the 2010 Haitian strains and those seen in previous years from Cameroon, Bangladesh, India, and Pakistan. The Haitian isolates were nearly identical and again consistent with a single clonal outbreak. Contemporary (i.e., 2010) *V. cholerae* strains from Nepal were not included in these studies and the genomic association between Haitian and Nepalese *V. cholerae* was not differentiated from other South Asian or even African locations.

The first study to include strains from Nepal was published by Hendriksen and colleagues on August 23, 2011 (Hendriksen et al., 2011). It compared the genomes of 24 strains isolated from five geographic regions in Nepal (between July 30 and November 1, 2010) with ten genomes of previously sequenced *V. cholerae*, including three from Haiti. All strains from Nepal, Haiti, and Bangladesh clustered together in a single monophyletic group (i.e., they shared a common ancestor). More importantly, the three Haitian and three Nepalese strains formed a very tight subgroup within the cluster, and these were almost identical, with only one or two nucleotide difference(s) in their core genome. This study, coupled with classical epidemiology

(Piarroux et al., 2011; Lantagne et al., 2014), showed convincing evidence that cholera was introduced into Haiti from an external source, with Nepal being the most likely origin.

Despite such strong evidence supporting the human transmission hypothesis, some scientists still stood by the climatic hypothesis. A study published on June 18, 2012 by Hasan and colleagues suggested that indigenous non-O1/O139 Haitian strains were involved in the outbreak (Hasan et al., 2012). The study entailed the identification and comparison of *V. cholerae* from 81 stool samples taken from the beginning of the outbreak by traditional methods and comparative genomics. *Vibrio cholerae* O1 was found in 48% of the samples, but more surprisingly, non-O1/O139 strains were identified in 21% of the samples. In addition, both O1 and non-O1/O139 strains were co-cultured in 7% of the samples, suggesting that non-O1/O139 strains may have played a role in the epidemic, whether alone or in concert with O1 strains. The authors stated that the assignment of attribution for cholera in Haiti remains controversial. However, scientists and Haitian public health officials supporting the human transmission hypothesis criticized the work, pointing out the unreliability of sampling methods (Mekalanos et al., 2012) and that the study did not offer evidence that non-O1/O139 played a notable role in the epidemic (Frerichs et al., 2012).

A recent study led by scientists from the CDC and published on July 2, 2013 provided additional strong evidence to refute the climatic hypothesis (Katz et al., 2013). Katz and colleagues sequenced the genomes of *V. cholerae* strains isolated from different time points within a two-year period since the start of the outbreak. The genomic affinity of the Haitian and Nepalese strains was reaffirmed; they were clearly distinct from isolates circulating elsewhere in the world, and there was no evidence of novel gene acquisition by horizontal gene transfer. A molecular clock was calculated, and the date of the most recent common ancestor (MRCA) was

estimated to be between July 23 to October 17, 2010 (with a 95% range credibility). The human transmission hypothesis suggesting Nepalese origins would stipulate that the MRCA was in Nepal prior to deployment of the MINUSTAH troops. Additionally, this time interval encompasses the reported cholera outbreak in Nepal (September 23; Maharjan, 2010), the arrival of the Nepalese soldiers in Haiti (October 8; Lantagne et al., 2014), and the first reported cholera cases (October 12 and 17; Ivers and Walton, 2012; Lantagne et al., 2014), supporting the time frame of the outbreak proposed by previous studies (Figure 6.2B; Piarroux et al., 2011; Lantagne et al., 2014).

6.5. Whole-genome sequencing as a tool for molecular epidemiology

The investigations of the cholera outbreak in Haiti illustrated how traditional epidemiological investigations can be greatly enhanced by genomic sequencing and phylogenetic analysis. In this case, the analysis of hospital case records established a spatiotemporal pattern to the outbreak but failed to differentiate between the two competing hypotheses. The subsequent genomic analyses provided very strong evidence to support the human transmission hypothesis; thus, the climatic hypothesis could be rejected. These analyses, because of their ability to detect minute differences between different strains, also allowed the determination of the exact source of the outbreak. This can help enormously with prevention of future outbreaks and can also have legal implications. In a recent development, a lawsuit has been filed against the United Nations in the United States Federal Court for damages caused by the cholera outbreak (Gladstone, 2013). With the potential for legal action, genomic analysis methods have to be very rigorous, as the judicial system will require high standards to accept them as evidence.

In retrospect, the identification of the origin of cholera in Haiti was limited by two essential factors. First was that genome sequencing was used only in the later steps of the investigation, not as a "first responder" screening method for identification of infectious agents. We are clearly entering an era where genomic or even metagenomic screening will become a part of medical diagnostics. The second limitation was the absence of a public database containing sufficient genome sequences of recurring pathogens from various geographical locations. These two elements are not out of reach, and major efforts are underway to remove both limitations (Aarestrup et al., 2012). For example, pipelines using current technologies to fully sequence a genome and perform key analyses for typing and identification within 24 hours are now becoming available (Heger, 2013). With the price of sequencing declining, comprehensive, geographically-informed genome sequence databases of pathogens could soon be a reality.

6.6. References

Aarestrup, F.M., Brown, E.W., Detter, C., Gerner-Smidt, P., Gilmour, M.W., Harmsen, D., et al. (2012). Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. *Emerging Infectious Diseases* 18(11)**,** e1.

Al Jazeera English (2010). *UN investigates Haiti outbreak* [Online]. Doha, Qatar: Al Jazeera Satellite Network. Available: http://www.aljazeera.com/news/africa/2010/10/201010284141214196 7.html [Accessed 25 March 2014].

BBC News (2010a). *Haiti cholera outbreak: Nepal troops not tested* [Online]. London, UK: British Broadcasting Corporation. Available: http://www.bbc.co.uk/news/world-south-asia-11949181 [Accessed 25 March 2014].

Benčić, Z., and Sinha, R. (1972). Cholera carriers and circulation of cholera vibrios in the community. *International Journal of Epidemiology* 1(1)**,** 13-14.

Centers for Disease Control and Prevention (2010). Update: cholera outbreak — Haiti, 2010. *MMWR: Morbidity and Mortality Weekly Report* 59(45)**,** 1473-1479.

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., et al. (2011). The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 364(1)**,** 33-42.

De la Cruz, F., and Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in Microbiology* 8(3)**,** 128-133.

Delva, J.G. (2010). *Quake-hit Haiti battles cholera epidemic, 150 dead* [Online]. New York, NY, USA: Thomson Reuters. Available: http://www.reuters.com/article/2010/10/22/us-haiti-cholera-idUSTRE69L21520101022 [Accessed 25 March 2014].

Farmer, P. (2013). *Key statistics - facts and figures about the earthquake, cholera, and development challenges in Haiti* [Online]. New York, NY, USA: United Nations Office of the Secretary-General's Special Adviser for Community-Based Medicine and Lessons from Haiti. Available: http://www.lessonsfromhaiti.org/relief-and-recovery/key-statistics. [Accessed 25 March 2014].

Frerichs, R.R., Boncy, J., Barrais, R., Keim, P.S., and Piarroux, R. (2012). Source attribution of 2010 cholera epidemic in Haiti. *Proceedings of the National Academy of Sciences of the United States of America* 109(47)**,** E3208.

Gladstone, R. (2013). *Rights advocates suing U.N. over the spread of cholera in Haiti* [Online]. New York, NY, USA: The New York Times Company. Available: http://www.nytimes.com/2013/10/09/world/americas/rights-advocates-suing-un-over-the-spread-of-cholera-in-haiti.html [Accessed 25 March 2014].

Harris, J.B., LaRocque, R.C., Qadri, F., Ryan, E.T., and Calderwood, S.B. (2012). Cholera. *Lancet* 379(9835)**,** 2466-2476.

Hasan, N.A., Choi, S.Y., Eppinger, M., Clark, P.W., Chen, A., Alam, M., et al. (2012). Genomic diversity of 2010 Haitian cholera outbreak strains. *Proceedings of the National Academy of Sciences of the United States of America* 109(29)**,** E2010-E2017.

Heger, M. (2013). *McGill University team develops rapid genome sequencing technique for outbreak monitoring* [Online]. New York, NY, USA: GenomeWeb LLC. Available: http://www.genomeweb.com/sequencing/mcgill-university-team-develops-rapid-genome-sequencing-technique-outbreak-monit [Accessed January 28 2014].

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., et al. (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* 2(4)**,** e00157-11.

Ivers, L.C., and Walton, D.A. (2012). The "first" case of cholera in Haiti: lessons for global health. *American Journal of Tropical Medicine and Hygiene* 86(1)**,** 36-38.

Jenson, D., Szabo, V., and Duke FHI Haiti Humanities Laboratory Student Research Team (2011). Cholera in Haiti and other Caribbean regions, 19th century. *Emerging Infectious Diseases* 17(11)**,** 2130-2135.

Katz, J.M. (2010). *UN probes base as source of Haiti cholera outbreak* [Online]. Washington, DC, USA: The Washington Times, LLC. Available: http://www.washingtontimes.com/news/2010/oct/27/un-probes-base-as-source-of-haiti-cholera-outbreak [Accessed 25 March 2014].

Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., et al. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4(4)**,** e00398-13.

Lam, C., Octavia, S., Reeves, P., Wang, L., and Lan, R. (2010). Evolution of seventh cholera pandemic and origin of 1991 epidemic, Latin America. *Emerging Infectious Diseases* 16(7)**,** 1130-1132.

Lantagne, D., Nair, G.B., Lanata, C.F., and Cravioto, A. (2014). "The cholera outbreak in Haiti: where and how did it begin?," in *Cholera Outbreaks: Current Topics in Microbiology and Immunology,* eds. B.G. Nair & Y. Takeda. (Berlin, Germany: Springer-Verlag), 145-164.

Maharjan, L. (2010). *Cholera outbreak looms over capital* [Online]. Kathmandu, Nepal: International Media Network Nepal Pvt. Ltd. Available: http://www.thehimalayantimes.com/fullNews.php?headline=Cholera+outbreak+looms+over+capital&NewsID=258974 [Accessed 25 March 2014].

Mekalanos, J.J., Robins, W., Ussery, D.W., Davis, B.M., Schadt, E., and Waldor, M.K. (2012). Non-O1 *Vibrio cholerae* unlinked to cholera in Haiti. *Proceedings of the National Academy of Sciences of the United States of America* 109(47)**,** E3206.

Ministère de la Santé Publique et de la Population (2014). *Rapport journalier MSPP du 07 Janvier 2014* [Online]. Port-au-Prince, Haiti: Ministère de la Santé Publique et de la Population. Available: http://mspp.gouv.ht/site/downloads/Rapport Web_07.01_Avec_Courbes_Departementales.pdf [Accessed 28 January 2014].

Mutreja, A., Kim, D.W., Thomson, N.R., Connor, T.R., Lee, J.H., Kariuki, S., et al. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365)**,** 462-465.

Parker, A.A. (2010). *Cholera in Haiti — the climate connection* [Online]. Traverse City, MI, USA: Circle of Blue. Available: http://www.circleofblue.org/waternews/2010/world/hold-cholera-in-haiti-the-climate-connection [Accessed 22 March 2014].

Piarroux, R., Barrais, R., Faucher, B., Haus, R., Piarroux, M., Gaudart, J., et al. (2011). Understanding the cholera epidemic, Haiti. *Emerging Infectious Diseases* 17(7)**,** 1161-1168.

Reimer, A.R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., et al. (2011). Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerging Infectious Diseases* 17(11)**,** 2113-2121.

Shimada, T., Nair, G.B., Deb, B.C., Albert, M.J., Sack, R.B., and Takeda, Y. (1993). Outbreak of *Vibrio cholerae* non-O1 in India and Bangladesh. *Lancet* 341(8856)**,** 1347.

United States Geological Survey (2010). *Magnitude 7.0 - Haiti region* [Online]. Reston, VA, USA: United States Geological Survey. Available: http://earthquake.usgs.gov/earthquakes/eqinthenews/2010/us2010rja6/us2010rja6.php [Accessed 25 March 2014].

Waldor, M.K., and Mekalanos, J.J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272(5270)**,** 1910-1914.

CHAPTER 7: General Discussion and Concluding Remarks

CHAPTER 7

7.1. Advantages and disadvantages of whole-genome sequencing in taxonomy

The work conducted in this thesis demonstrates the power and usefulness of whole-genome sequencing (WGS) in taxonomy. In Chapter 2, *Vibrio cidicii* was only differentiated from *Vibrio navarrensis* in one out of 158 phenotypic tests, and 35 tests showed variable results across isolates or between the kits used. On the other hand, *in silico* DNA–DNA hybridization (dDDH) and multilocus sequence analysis (MLSA) of core genes provided conclusive results, justifying the proposal of a new species. In another case, several members of *Rhodobacteraceae* were identified based on phenotypic and 16S rRNA sequence identity (Buchan et al., 2005; Wagner-Döbler and Biebl, 2006; Luo and Moran, 2014). Here, a whole genome-based system in classifying these organisms was proposed to remove paraphyly in several genera and prevent further taxonomic inconsistencies (Chapter 3).

Despite advancements and improvements in WGS technologies that have reached a point where it may now be used fully in prokaryotic taxonomic classification, the current practice does not yet integrate genomic data. In fact, incorporation of data from genomes [e.g., core genome phylogeny, average nucleotide identity (ANI), dDDH, etc.] is not mandatory in officially describing a novel taxon (Tindall et al., 2010). Five novel species of *Rhodobacteraceae* (e.g., *Actibacterium pelagium*, *Paraphaeobacter pallidus*, *Phaeobacter piscinae*, *Phaeobacter porticola*, *Ruegeria marisrubri*, *Ruegeria profundi*) and a novel genus (e.g., *Paraphaeobacter*) were recently described (Breider et al., 2017; Cai et al., 2017; Guo et al., 2017; Sonnenschein et al., 2017; Zhang et al., 2017). While five of the six species descriptions incorporated genome data in their analysis, one description (for *P. pallidus*) relied solely on 16S rRNA sequence

analysis and phenotypic tests to validate the novel species and genus. However, the *P. pallidus* type strain shares a high 16S rRNA sequence identity (97%) with the type strain of *Phaeobacter gallaeciensis* (Cai et al., 2017). This high sequence identity should have warranted further genotypic tests such as genomic comparisons with know type strains to confirm identity. The use of genome sequences is highly recommended (Whitman, 2011), and metrics such as ANI and dDDH may eventually replace traditional DDH (Goris et al., 2007; Tindall et al., 2010; Meier-Kolthoff et al., 2013; Figueras et al., 2014; Beaz-Hidalgo et al., 2015).

There are, however, disadvantages to using genomics for taxonomy. Completely sequenced genomes for major lineages are lacking (Klenk and Göker, 2010), and the increasing number of genomes remains highly biased towards strains of biotechnological or medical importance (Zhi et al., 2012). Genome sequences have been obtained mostly from three phyla, *Proteobacteria*, *Firmicutes*, and *Actinobacteria* (Sentausa and Fournier, 2013). Even if genome sequences are available, they are not type strains in many cases and must be used with caution, since the current taxonomic practice is highly reliant on comparisons with type strains (Tindall et al., 2010). A type strain is a representative unit of a microbial species, chosen when the species is established based on extensive taxonomic and phenotypic data, isolation source metadata, and other criteria. As of 2015, there were 12,981 bacterial and archaeal species with validly published names, with 650 new type strains added (on average) every year (Kyrpides et al., 2014; Garrity, 2016). Efforts in sequencing the genomes of type strains and making them publicly available are underway, such as the massive Genomic Encyclopedia of *Bacteria* and *Archaea* sequencing project (Wu et al., 2009; Kyrpides et al., 2014; Whitman et al., 2015; Mukherjee et al., 2017). Despite the importance of type strains, only 1,003 genomes from 974

bacterial and 29 archaeal type strains have been sequenced so far through this project (Mukherjee et al., 2017).

Genome sequences vary greatly in quality, and most are only available as unfinished or draft genomes and may be less informative than finished ones (Klassen and Currie, 2012; Ricker et al., 2012). Genomes need to be screened for quality before use, and standard guidelines are being developed (Gargis et al., 2012). My dataset of *Rhodobacteraceae* genomes (Chapter 3) initially started with 316 genomes, which was reduced to 290 to only keep high-quality ones. I assessed for quality by determining genome completeness, reported as the percentage of 100 or so core housekeeping genes present in a genome. This method was used previously to assess for genome quality (Luo and Moran, 2014; Yeoh et al., 2016), and has proven to be effective in my study. Core genome determination using all 316 genomes resulted in only two core genes, but removing 26 genomes (< 95% complete) significantly increased the core genes to 115.

The current procedure used to describe novel taxa should be revised to introduce new publication formats (Sutcliffe et al., 2012). There is a recommendation on the routine description of species on the basis of their genome sequences (Whitman, 2011; Thompson et al., 2015), which would allow for type strains to be uniquely and unambiguously identified and avoid redundancies and inconsistencies in nomenclature. However, it is still recommended that a genomic approach to descriptions of novel taxa must not abandon the fundamental principles of taxonomy, including the incorporation of phenotypic data (Tindall et al., 2010; Kämpfer and Glaeser, 2012) and the deposition of type strains in culture collections (Tindall, 2008; Tindall and Garrity, 2008).

7.2. Reconciling horizontal gene transfer with the biological species concept

Interspecies horizontal gene transfer (HGT) has been demonstrated between two closely related species, *V. cholerae* and *V. metoecus*, co-occurring in the same environment. A significant amount of recombination events was shown to occur between the two species, specifically in their integron region (Chapter 4). Interestingly directional HGT does not happen in equilibrium, and *V. cholerae* contributes three times more genes to the latter (Chapter 5). This bias is more likely due to *V. cholerae* being significantly more abundant (three times more) than *V. metoecus*, as well as the former being abundant throughout the summer season but the latter only appearing towards the end of the season. These works have provided insights on the natural ecological interactions between close relatives and the role HGT plays in their evolution and diversification.

Evolution, in its simplest form, is a change in allele frequencies in a set of organisms over time (Madigan et al., 2014). How these changes spread through populations and trigger speciation remains controversial. There is evidence of gene-specific sweeps, where genes or regions of the genome spread through populations independently through homologous recombination (Guttman and Dykhuizen, 1994; Papke et al., 2007; Coleman and Chisholm, 2010; Denef et al., 2010; Shapiro et al., 2012). However, other studies have also shown that homologous recombination is generally low enough that it does not interfere with the formation of ecological clusters of species (Melendrez et al., 2015). In this case, the force that drives speciation is periodic selection, where natural selection favours beneficial mutations but purges deleterious mutations and therefore genetic diversity throughout the genome (i.e., genome-wide sweep) resulting in an ecologically homogeneous species or ecotype (Bendall et al., 2016; Cohan, 2016).

Shapiro et al. (2012) investigated possible mechanisms that could lead to differentiation (and possibly speciation) of closely related organisms by looking into two recently diverged populations of *Vibrio cyclitrophicus* with clearly delineated habitat associations, zoo- and phytoplankton (large size fractions of filtered seawater) or suspended organic particles (small size fractions). Comparison of these two populations shows that differentiation between these populations is restricted to a few small patches of the core genome (i.e., gene-specific sweeps), therefore rejecting ecological partition. Also, genomic fragments can sweep through populations without purging genome-wide variation. For example, chromosome II can sweep through a subset of the small size fraction population without affecting the diversity of chromosome I. Lastly, recent recombination events were more preferential for members within, rather than between, habitats, but older events are not, reinforcing the notion that these populations are on independent evolutionary trajectories. This study therefore reconciled the two conflicting theories by suggesting that gene-centered sweeps may eventually lead to a pattern characteristic of genome-wide sweeps.

Although HGT has been reported between distantly related organisms, it is actually more likely to occur between closely related organisms than evolutionary distant ones, due to homologous recombination relying on regions of sequence conservation between incoming DNA and the genome of the recipient bacteria (Majewski and Cohan, 1999). This need for genetic relatedness is reflective of what is happening in higher organisms, where there is genetic isolation within species than distantly related organisms. Thus, to some extent, the biological species concept may be applicable to some bacteria after all (Doolittle and Zhaxybayeva, 2009).

7.3. The uncultivated majority and their own species problem

A key limitation in the current taxonomic practice is that results are usually derived from geographically disparate and ecologically irrelevant strains. This does not allow for significant ecological inferences in terms of population structure and dynamics and how the species functions as a group. Current technologies in high-throughput metagenomic sequencing has now made it possible to study ecologically driven species boundaries without the need for cultivation, which has helped advance the definition of bacterial species (Handelsman et al., 2007; Konstantinidis et al., 2017). Such approach has led to the recovery of 7,903 bacterial and archaeal genomes, as well as the discovery of 17 bacterial and 3 archaeal candidate phyla, that has filled the gaps and expanded our knowledge and understanding of the diverse microbial world (Parks et al., 2017).

Metagenomic approaches have also provided information on how microbial communities organize into genetically and ecologically cohesive units, which possess attributes expected for a species (Oh et al., 2011; Caro-Quintero and Konstantinidis, 2012). From the metagenomic study of the microbial community of Lake Lanier (GA, USA), a striking pattern emerged such that communities tend to cluster into discrete populations with high genomic relatedness (Caro-Quintero and Konstantinidis, 2012). These "sequence-discrete" populations were not clonal but contained significant intra-population diversity, ranging from < 1% to ~ 5% genome-aggregate average nucleotide sequence divergence (i.e., 95–100% ANI), whereas comparisons between clusters fell below 90% ANI. The sequence divergence within sequence-discrete populations corresponds to thousands of generations from the last common ancestor (Lawrence and Ochman, 1998), suggestive of populations that likely represent long-lived entities shaped by long-term ecological and genetic selection pressures (Caro-Quintero and Konstantinidis, 2012).

Additionally, the genotype of members of a sequence-discrete population were more uniform in terms of evolutionary relatedness, gene content, and gene expression patterns than members of species that have been identified using traditional methods (Caro-Quintero and Konstantinidis, 2012). This preliminary finding is interesting since it is the first time ecologically relevant "species" have been described in a natural microbial community. It will be interesting to study more metagenomic datasets representing other communities to determine if similar patterns emerge.

Half of all known bacterial diversity represents organisms that might not be cultivated in the laboratory and are therefore not eligible for formal species descriptions (Hug et al., 2016). Therefore, metagenomics challenges the current taxonomic standards due to the large amount of bacterial diversity that cannot be formally recognized. This brings up the necessity to review current description standards to shift the emphasis to molecular data and the understanding of underlying evolutionary concepts that drive diversity into more naturally cohesive units (Fraser et al., 2009; Thompson et al., 2015).

7.4. Whole-genome sequencing in epidemiology

The ability to discriminate between organisms is essential in the medical, pharmacological, industrial, or evolutionary contexts. Genome sequences are important in epidemiological surveillance and diagnostics of infectious diseases. Rapid identification of pathogens by WGS is now seen as an important component of an effective response during an epidemic, as demonstrated in the case of the cholera outbreak in Haiti (Chapter 6), which was determined to be an imported case (Hendriksen et al., 2011; Katz et al., 2013). The high resolution of data provided by genomes allowed for the distinction of the Haitian *V. cholerae*

strains against the Nepalese strains, which only differed in 1–2 nucleotides within their core genome (Hendriksen et al., 2011). As such, genome sequences have provided a genome-level typing tool that serves the same purpose of earlier typing tools but with much higher resolution.

Multilocus sequence typing (MLST) was used to type pathogens using a small segment of their genomes (i.e., five to seven genes; Maiden et al., 1998; Maiden, 2006), but the concept has been applied to the core genome (cg), termed as cgMLST. This whole genome-based typing method was used during the *Escherichia coli* outbreak in Germany in 2011, which affected 3,842 and killed 53 people (Burger, 2012). Mellmann et al. (2011) used 1,144 core genes to genetically characterize and differentiate the 2011 outbreak strains from historical ones. The cgMLST scheme in *Listeria monocytogenes*, a major human foodborne pathogen that causes the fatal disease listeriosis, is a relatively new typing scheme for the species (Pightling et al., 2015; Ruppitsch et al., 2015; Chen et al., 2016; Moura et al., 2016). Using 1,748 core genes, the Insititut Pasteur (Paris, France), Centers for Disease Control and Prevention (CDC, Atlanta, GA, USA), Public Health England (London, UK), and Public Health Agency of Canada (Winnipeg, MB, Canada) were able to type 1,696 isolates of *L. monocytogenes* isolated from diverse geographical locations globally, providing a universal genome-based typing scheme necessary for monitoring *L. monocytogenes* outbreaks worldwide (Moura et al., 2016). The cgMLST typing scheme using 2,254 core genes for *Vibrio parahaemolyticus*, also a significant foodborne pathogen, has been initiated (Gonzalez-Escalona et al., 2017). Unfortunately, a cgMLST scheme is currently not available for *V. cholerae*, which could be of major importance due to ongoing cholera epidemics in Haiti, Asia, and Africa (World Health Organization, 2016). Most notably, the ongoing cholera outbreak in Yemen has reported 862,858 cases and 2,177 deaths, as of October 22, 2017 (World Health Organization, 2017).

7.5. Future research

Prokaryotic taxonomy is constantly evolving with new technologies, and the availability of genome sequences has increased the need for considerable reshuffling of taxa. The taxonomic scheme based on 16S rRNA gene sequences and phenotypic tests sometimes result in classifications that are not supported by genomic studies (ANI, dDDH, and MLSA). This has resulted in an ongoing process of complete shifting and reordering of species, genera, and complete families. Taxa above the rank of species have received very little attention in terms of determining tangible delineation standards, and this aspect of taxonomy will greatly benefit from genome sequence data. The whole genome-based taxonomic system employed for *Rhodobacteraceae* and the newly proposed *Stappiaceae* fam. nov. (Chapter 3) can be used in the future for classification of more species belonging to these families. As of December 1, 2017, there are currently 832 genome sequences available in the GenBank database classified under *Rhodobacteraceae*, including those of several novel species described after the commencement my work. Also, most of the methods described in this thesis can be implemented in other taxa of interest if sufficient genome sequences are available. The family *Vibrionaceae* (Gomez-Gil et al., 2014) of the order *Vibrionales* and class *Gammaproteobacteria* is one of the most well-studied families of bacteria, which could benefit from this classification system. The family consists of 2,361 genomes in the GenBank database (as of December 1, 2017), the majority of which are from species of clinical significance such as *V. cholerae* (n = 748), *V. parahaemolyticus* (n = 800), and *V. vulnificus* (n = 77).

The newly described species, *V. cidicii* (Chapter 2) and *V. metoecus* (Chapter 4 and Chapter 5), are also of clinical interest due to the isolation of contemporary strains from human specimens in the US by the CDC (Gladney and Tarr, 2014; Kirchberger et al., 2014). The

availability of whole-genome sequences from these species will make identification of future isolates easier and can be used to further explore their biology.

The demonstrated bias in HGT events between *V. cholerae* and *V. metoecus* is interesting in an ecological perspective (Chapter 4 and Chapter 5). Specifically, it will be interesting to determine any specific mechanisms of interaction between these two species in the environment. An antagonistic mode of interaction is possible by their contact-dependent type VI secretion systems (T6SS; Kirchberger et al., 2017). Both species have active T6SS, which they can use to compete for resources in the environment by killing (lysing) neighboring cells. The genetic material from the lysed cells is then released into the environment. Interestingly, the T6SS is co-regulated with genes of the DNA uptake machinery to facilitate the acquisition of genetic material (Borgeaud et al., 2015). The major challenge for now is to optimize the natural transformation protocol performed in this study, to be able to perform T6SS and natural transformation assays in the laboratory similar to the assays performed by Borgeaud et al. (2015) with *V. cholerae*.

The population pan-genome, the entire genomic repertoire of all members of a taxon (Tettelin et al., 2005; Vernikos et al., 2015), can be studied for *V. cholerae* with the increasing number of genome sequences from specific locations becoming available for the species. From the 439 *V. cholerae* isolates from Oyster Pond, MA, USA (Kirchberger et al., 2016), I have sequenced the genomes of 91 strains. These sequences can be compared to those of isolates obtained from other sites, such as the Bay of Bengal. This region in South Asia appears to be the single global source of cholera (Mutreja et al., 2011). The datasets obtained from these regions could represent cholera-free (Oyster Pond) and cholera-infected (Bay of Bengal) environments. The pan-genome profiles can be compared to determine full genetic content and diversity of a

species in a given environment, as well as provide insights on the pathogenic potential of each environment (i.e., the presence or absence of virulence genes from the pan-genomes).

Lastly, a cgMLST scheme for *V. cholerae* can be established due to the availability of hundreds of genomes for the species. My initial analysis of 609 high-quality *V. cholerae* genomes (≥ 98% complete) out of 767 genomes, available at the time from the GenBank database and private collections, revealed 1,830 core genes that could potentially be utilized to develop a cgMLST scheme. However, since the commencement of this work, there is now an additional 157 genomes that were deposited in GenBank (as of December 1, 2017) for a total of 924 *V. cholerae* genomes. My initial analysis can be updated to incorporate these newly deposited genomes for a more universal core genome dataset.

7.6. Conclusion

The various studies presented herein demonstrate applications of WGS in taxonomy, population genetics, and microbial epidemiology, made possible through advances in next-generation sequencing technologies and sequence analysis methods. In taxonomy, isolates were attributed to a novel species, *V. cidicii*, based on their genomic differences from *V. navarrensis*. Also, genome-based similarity metrics for genus and family were proposed for *Rhodobacteraceae* and *Stappiaceae* fam. nov. In population genetics, the role of HGT was shown through interspecies genetic interactions between *V. cholerae* and *V. metoecus*, where the directional HGT bias from the former to latter was attributed to differences in abundance between the two (i.e., more abundant *V. cholerae* than *V. metoecus*). In epidemiology, the power of genomics was highlighted; the source of the cholera outbreak in Haiti was deciphered through genome sequence analysis. We are only beginning to appreciate the significance of WGS, and its

full potential in several other fields will soon be realized as the price of sequencing continually drops.

7.7. References

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

Bendall, M.L., Stevens, S.L., Chan, L.K., Malfatti, S., Schwientek, P., Tremblay, J., et al. (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME Journal* 10.1038/ismej.2015.241**,** 1-13.

Borgeaud, S., Metzger, L.C., Scrignari, T., and Blokesch, M. (2015). The type VI secretion system of *Vibrio cholerae* fosters horizontal gene transfer. *Science* 347(6217)**,** 63-67.

Breider, S., Freese, H.M., Sproer, C., Simon, M., Overmann, J., and Brinkhoff, T. (2017). *Phaeobacter porticola* sp. nov., an antibiotic-producing bacterium isolated from a sea harbour. *International Journal of Systematic and Evolutionary Microbiology* 67(7)**,** 2153-2159.

Buchan, A., González, J.M., and Moran, M.A. (2005). Overview of the marine *Roseobacter* lineage. *Applied and Environmental Microbiology* 71(10)**,** 5665-5677.

Burger, R. (2012). "EHEC O104:H4 in Germany 2011: large outbreak of bloody diarrhea and haemolytic uraemic syndrome by shiga toxin-producing *E. coli* via contaminated food," in *Improving Food Safety Through a One Health Approach: Workshop Summary,* eds. E.R. Choffnes, D.A. Relman, L. Olsen, R. Hutton & A. Mack. (Washington, DC, USA: National Academies Press), 115-130.

Cai, X., Wang, Y., Yang, X., Liu, J., Wu, Y., and Zhang, X.H. (2017). *Paraphaeobacter pallidus* gen. nov., sp. nov., isolated from seawater. *International Journal of Systematic and Evolutionary Microbiology* 67(8)**,** 2520-2526.

Caro-Quintero, A., and Konstantinidis, K.T. (2012). Bacterial species may exist, metagenomics reveal. *Environmental Microbiology* 14(2)**,** 347-355.

Chen, Y., Gonzalez-Escalona, N., Hammack, T.S., Allard, M.W., Strain, E.A., and Brown, E.W. (2016). Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of *Listeria monocytogenes*. *Applied and Environmental Microbiology* 82(20)**,** 6258-6272.

Cohan, F.M. (2016). Bacterial speciation: genetic sweeps in bacterial species. *Current Biology* 26(3)**,** R112-R115.

Coleman, M.L., and Chisholm, S.W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences of the United States of America* 107(43)**,** 18634-18639.

Denef, V.J., Kalnejais, L.H., Mueller, R.S., Wilmes, P., Baker, B.J., Thomas, B.C., et al. (2010). Proteogenomic basis for ecological divergence of closely related bacteria in natural

acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America* 107(6)**,** 2383-2390.

Doolittle, W.F., and Zhaxybayeva, O. (2009). On the origin of prokaryotic species. *Genome Research* 19(5)**,** 744-756.

Figueras, M.J., Beaz-Hidalgo, R., Hossain, M.J., and Liles, M.R. (2014). Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. *Genome Announcements* 2(6)**,** e00927-14.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323(5915)**,** 741-746.

Gargis, A.S., Kalman, L., Berry, M.W., Bick, D.P., Dimmock, D.P., Hambuch, T., et al. (2012). Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature Biotechnology* 30(11)**,** 1033-1036.

Garrity, G.M. (2016). A new genomics-driven taxonomy of *Bacteria* and *Archaea*: are we there yet? *Journal of Clinical Microbiology* 54(8)**,** 1956-1963.

Gladney, L.M., and Tarr, C.L. (2014). Molecular and phenotypic characterization of *Vibrio navarrensis* isolates associated with human illness. *Journal of Clinical Microbiology* 52(11)**,** 4070-4074.

Gomez-Gil, B., Thompson, C.C., Matsumura, Y., Sawabe, T., Iida, T., Christen, R., et al. (2014). "The family *Vibrionaceae*," in *The Prokaryotes – Gammaproteobacteria,* eds. E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt & F. Thompson. 4th ed. (Berlin, Germany: Springer), 659-747.

Gonzalez-Escalona, N., Jolley, K.A., Reed, E., and Martinez-Urtaza, J. (2017). Defining a core genome multilocus sequence typing scheme for the global epidemiology of *Vibrio parahaemolyticus*. *Journal of Clinical Microbiology* 55(6)**,** 1682-1697.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57(1)**,** 81-91.

Guo, L.L., Wu, Y.H., Xu, X.W., Huang, C.J., Xu, Y.Y., Cheng, H., et al. (2017). *Actibacterium pelagium* sp. nov., a novel alphaproteobacterium, and emended description of the genus *Actibacterium*. *International Journal of Systematic and Evolutionary Microbiology* 10.1099/ijsem.0.002417.

Guttman, D.S., and Dykhuizen, D.E. (1994). Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* 138(4)**,** 993-1003.

Handelsman, J., Tiedje, J., Alvarez-Cohen, L., Ashburner, M., Cann, I.K.O., DeLong, E.F., et al. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Washington, DC, USA: National Academies Press.

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., et al. (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* 2(4)**,** e00157-11.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., et al. (2016). A new view of the tree of life. *Nature Microbiology* 1(5)**,** 16048.

Kämpfer, P., and Glaeser, S.P. (2012). Prokaryotic taxonomy in the sequencing era—the polyphasic approach revisited. *Environmental Microbiology* 14(2)**,** 291-317.

Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., et al. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4(4)**,** e00398-13.

Kirchberger, P.C., Orata, F.D., Barlow, E.J., Kauffman, K.M., Case, R.J., Polz, M.F., et al. (2016). A small number of phylogenetically distinct clonal complexes dominate a coastal *Vibrio cholerae* population. *Applied and Environmental Microbiology* 82(18)**,** 5576-5586.

Kirchberger, P.C., Turnsek, M., Hunt, D.E., Haley, B.J., Colwell, R.R., Polz, M.F., et al. (2014). *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. *International Journal of Systematic and Evolutionary Microbiology* 64(9)**,** 3208-3214.

Kirchberger, P.C., Unterweger, D., Provenzano, D., Pukatzki, S., and 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.

Klassen, J.L., and Currie, C.R. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* 13**,** 14.

Klenk, H.P., and Göker, M. (2010). En route to a genome-based classification of *Archaea* and *Bacteria*? *Systematic and Applied Microbiology* 33(4)**,** 175-182.

Konstantinidis, K.T., Rosselló-Móra, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. *ISME Journal* 11(11)**,** 2399-2406.

Kyrpides, N.C., Hugenholtz, P., Eisen, J.A., Woyke, T., Göker, M., Parker, C.T., et al. (2014). Genomic encyclopedia of *Bacteria* and *Archaea*: sequencing a myriad of type strains. *PLoS Biology* 12(8)**,** e1001920.

Lawrence, J.G., and Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences of the United States of America* 95(16)**,** 9413-9417.

Luo, H., and Moran, M.A. (2014). Evolutionary ecology of the marine *Roseobacter* clade. *Microbiology and Molecular Biology Reviews* 78(4)**,** 573-587.

Madigan, M.T., Martinko, J.M., Bender, K.S., Buckley, D.H., and Stahl, D.A. (2014). *Brock Biology of Microorganisms.* San Francisco, CA, USA: Benjamin Cummings.

Maiden, M.C. (2006). Multilocus sequence typing of bacteria. *Annual Review of Microbiology* 60(1)**,** 561-588.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* 95(6)**,** 3140-3145.

Majewski, J., and Cohan, F.M. (1999). DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* 153(4)**,** 1525-1533.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14**,** 60.

Melendrez, M.C., Becraft, E.D., Wood, J.M., Olsen, M.T., Bryant, D.A., Heidelberg, J.F., et al. (2015). Recombination does not hinder formation or detection of ecological species of *Synechococcus* inhabiting a hot spring cyanobacterial mat. *Frontiers in Microbiology* 6**,** 1540.

Mellmann, A., Harmsen, D., Cummings, C.A., Zentz, E.B., Leopold, S.R., Rico, A., et al. (2011). Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6(7)**,** e22751.

Moura, A., Criscuolo, A., Pouseele, H., Maury, M.M., Leclercq, A., Tarr, C., et al. (2016). Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nature Microbiology* 2(2)**,** 16185.

Mukherjee, S., Seshadri, R., Varghese, N.J., Eloe-Fadrosh, E.A., Meier-Kolthoff, J.P., Göker, M., et al. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology* 35(7)**,** 676-683.

Mutreja, A., Kim, D.W., Thomson, N.R., Connor, T.R., Lee, J.H., Kariuki, S., et al. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365)**,** 462-465.

Oh, S., Caro-Quintero, A., Tsementzi, D., DeLeon-Rodriguez, N., Luo, C., Poretsky, R., et al. (2011). Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Applied and Environmental Microbiology* 77(17)**,** 6000-6011.

Papke, R.T., Zhaxybayeva, O., Feil, E.J., Sommerfeld, K., Muise, D., and Doolittle, W.F. (2007). Searching for species in haloarchaea. *Proceedings of the National Academy of Sciences of the United States of America* 104(35)**,** 14092-14097.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* 2(11)**,** 1533-1542.

Pightling, A.W., Petronella, N., and Pagotto, F. (2015). The *Listeria monocytogenes* Core-Genome Sequence Typer (LmCGST): a bioinformatic pipeline for molecular characterization with next-generation sequence data. *BMC Microbiology* 15**,** 224.

Ricker, N., Qian, H., and Fulthorpe, R.R. (2012). The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. *Genomics* 100(3)**,** 167-175.

Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H.L., Allerberger, F., et al. (2015). Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Listeria monocytogenes*. *Journal of Clinical Microbiology* 53(9)**,** 2869-2876.

Sentausa, E., and Fournier, P.E. (2013). Advantages and limitations of genomics in prokaryotic taxonomy. *Clinical Microbiology and Infection* 19(9)**,** 790-795.

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabo, G., et al. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science* 336(6077)**,** 48-51.

Sonnenschein, E.C., Phippen, C.B.W., Nielsen, K.F., Mateiu, R.V., Melchiorsen, J., Gram, L., et al. (2017). *Phaeobacter piscinae* sp. nov., a species of the *Roseobacter* group and potential aquaculture probiont. *International Journal of Systematic and Evolutionary Microbiology* 10.1099/ijsem.0.002331.

Sutcliffe, I.C., Trujillo, M.E., and Goodfellow, M. (2012). A call to arms for systematists: revitalising the purpose and practises underpinning the description of novel microbial taxa. *Antonie van Leeuwenhoek* 101(1)**,** 13-20.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* 102(39)**,** 13950-13955.

Thompson, C.C., Amaral, G.R., Campeao, M., Edwards, R.A., Polz, M.F., Dutilh, B.E., et al. (2015). Microbial taxonomy in the post-genomic era: rebuilding from scratch? *Archives of Microbiology* 197(3)**,** 359-370.

Tindall, B.J. (2008). Confirmation of deposit, but confirmation of what? *International Journal of Systematic and Evolutionary Microbiology* 58(8)**,** 1785-1787.

Tindall, B.J., and Garrity, G.M. (2008). Proposals to clarify how type strains are deposited and made available to the scientific community for the purpose of systematic research. *International Journal of Systematic and Evolutionary Microbiology* 58(8)**,** 1987-1990.

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

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Wagner-Döbler, I., and Biebl, H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annual Review of Microbiology* 60(1)**,** 255-280.

Whitman, W.B. (2011). Intent of the nomenclatural Code and recommendations about naming new species based on genomic sequences. *Bulletin of BISMiS* 2(2)**,** 135-139.

Whitman, W.B., Woyke, T., Klenk, H.P., Zhou, Y., Lilburn, T.G., Beck, B.J., et al. (2015). Genomic Encyclopedia of Bacterial and Archaeal Type Strains, Phase III: the genomes of soil and plant-associated and newly described type strains. *Standards in Genomic Sciences* 10(5)**,** 26.

World Health Organization (2016). Cholera, 2015. *Weekly Epidemiological Record* 91(38)**,** 433-440.

World Health Organization (2017). *Daily epidemiology bulletin, 22 October 2017* [Online]. Sana'a, Yemen: World Health Organization. Available: http://www.emro.who.int/images/stories/yemen/Yemen_Cholera_Response_-_Daily_Epidemiological_Update_-_2017-10-22.pdf?ua=1 [Accessed October 24 2017].

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., et al. (2009). A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea*. *Nature* 462(7276)**,** 1056-1060.

Yeoh, Y.K., Sekiguchi, Y., Parks, D.H., and Hugenholtz, P. (2016). Comparative genomics of candidate phylum TM6 suggests that parasitism is widespread and ancestral in this lineage. *Molecular Biology and Evolution* 33(4)**,** 915-927.

Zhang, G., Haroon, M.F., Zhang, R., Dong, X., Wang, D., Liu, Y., et al. (2017). *Ruegeria profundi* sp. nov. and *Ruegeria marisrubri* sp. nov., isolated from the brine-seawater interface at Erba Deep in the Red Sea. *International Journal of Systematic and Evolutionary Microbiology* 10.1099/ijsem.0.002344.

Zhi, X.Y., Zhao, W., Li, W.J., and Zhao, G.P. (2012). Prokaryotic systematics in the genomics era. *Antonie van Leeuwenhoek* 101(1)**,** 21-34.

REFERENCES

REFERENCES

Aarestrup, F.M., Brown, E.W., Detter, C., Gerner-Smidt, P., Gilmour, M.W., Harmsen, D., et al. (2012). Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. *Emerging Infectious Diseases* 18(11)**,** e1.

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *Journal of Bacteriology* 186(9)**,** 2629-2635.

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P., et al. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research* 28(20)**,** e87.

Al Jazeera English (2010). *UN investigates Haiti outbreak* [Online]. Doha, Qatar: Al Jazeera Satellite Network. Available: http://www.aljazeera.com/news/africa/2010/10/201010284141214196 7.html [Accessed 25 March 2014].

Alam, M., Sultana, M., Nair, G.B., Siddique, A.K., Hasan, N.A., Sack, R.B., et al. (2007). Viable but nonculturable *Vibrio cholerae* O1 in biofilms in the aquatic environment and their role in cholera transmission. *Proceedings of the National Academy of Sciences of the United States of America* 104(45)**,** 17801-17806.

Ali, M., Lopez, A.L., You, Y.A., Kim, Y.E., Sah, B., Maskery, B., et al. (2012). The global burden of cholera. *Bulletin of the World Health Organization* 90(3)**,** 209-218A.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215(3)**,** 403-410.

Amann, R.I., Lin, C., Key, R., Montgomery, L., and Stahl, D.A. (1992). Diversity among *Fibrobacter* isolates: towards a phylogenetic classification. *Systematic and Applied Microbiology* 15(1)**,** 23-31.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiological Reviews* 59(1)**,** 143-169.

Andersson, J.O. (2005). Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences* 62(11)**,** 1182-1197.

Angiuoli, S.V., and Salzberg, S.L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27(3)**,** 334-342.

Auch, A.F., Klenk, H.P., and Göker, M. (2010). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Standards in Genomic Sciences* 2(1)**,** 142-148.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9**,** 75.

Baek, K., Lee, Y.M., Shin, S.C., Hwang, K., Hwang, C.Y., Hong, S.G., et al. (2015). *Halocynthiibacter arcticus* sp. nov., isolated from Arctic marine sediment. *International Journal of Systematic and Evolutionary Microbiology* 65(11)**,** 3861-3865.

Baum, D. (2008). Reading a phylogenetic tree: the meaning of monophyletic groups. *Nature Education* 1(1)**,** 190.

Bavykin, S.G., Lysov, Y.P., Zakhariev, V., Kelly, J.J., Jackman, J., Stahl, D.A., et al. (2004). Use of 16S rRNA, 23S rRNA, and *gyrB* gene sequence analysis to determine phylogenetic relationships of *Bacillus cereus* group microorganisms. *Journal of Clinical Microbiology* 42(8)**,** 3711-3730.

BBC News (2010a). *Haiti cholera outbreak: Nepal troops not tested* [Online]. London, UK: British Broadcasting Corporation. Available: http://www.bbc.co.uk/news/world-south-asia-11949181 [Accessed 25 March 2014].

BBC News (2010b). *UN appeals for calm after cholera riots in Haiti* [Online]. London, UK: British Broadcasting Corporation. Available: http://www.bbc.co.uk/news/world-latin-america-11772283 [Accessed 25 March 2014].

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

Benčić, Z., and Sinha, R. (1972). Cholera carriers and circulation of cholera vibrios in the community. *International Journal of Epidemiology* 1(1)**,** 13-14.

Bendall, M.L., Stevens, S.L., Chan, L.K., Malfatti, S., Schwientek, P., Tremblay, J., et al. (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME Journal* 10.1038/ismej.2015.241**,** 1-13.

Bernardy, E.E., Turnsek, M.A., Wilson, S.K., Tarr, C.L., and Hammer, B.K. (2016). Diversity of clinical and environmental isolates of *Vibrio cholerae* in natural transformation and contact-dependent bacterial killing indicative of type VI secretion system activity. *Applied and Environmental Microbiology* 82(9)**,** 2833-2842.

Bernstein, H., Byerly, H.C., Hopf, F.A., and Michod, R.E. (1985). Sex and the emergence of species. *Journal of Theoretical Biology* 117(4)**,** 665-690.

Berry, A., Janssens, D., Hümbelin, M., Jore, J.P., Hoste, B., Cleenwerck, I., et al. (2003). *Paracoccus zeaxanthinifaciens* sp. nov., a zeaxanthin-producing bacterium. *International Journal of Systematic and Evolutionary Microbiology* 53(1)**,** 231-238.

Beyhan, S., and Yildiz, F.H. (2007). Smooth to rugose phase variation in *Vibrio cholerae* can be mediated by a single nucleotide change that targets c-di-GMP signalling pathway. *Molecular Microbiology* 63(4)**,** 995-1007.

Blokesch, M., and Schoolnik, G.K. (2008). The extracellular nuclease Dns and its role in natural transformation of *Vibrio cholerae*. *Journal of Bacteriology* 190(21)**,** 7232-7240.

Bochner, B.R. (2003). New technologies to assess genotype-phenotype relationships. *Nature Reviews Genetics* 4(4)**,** 309-314.

Bochner, B.R. (2009). Global phenotypic characterization of bacteria. *FEMS Microbiology Reviews* 33(1)**,** 191-205.

Bochner, B.R., Gadzinski, P., and Panomitros, E. (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Research* 11(7)**,** 1246-1255.

Borgeaud, S., Metzger, L.C., Scrignari, T., and Blokesch, M. (2015). The type VI secretion system of *Vibrio cholerae* fosters horizontal gene transfer. *Science* 347(6217)**,** 63-67.

Bosshard, P.P., Zbinden, R., Abels, S., Boddinghaus, B., Altwegg, M., and Bottger, E.C. (2006). 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. *Journal of Clinical Microbiology* 44(4)**,** 1359-1366.

Boucher, Y. (2016). Sustained local diversity of *Vibrio cholerae* O1 biotypes in a previously cholera-free country. *MBio* 7(3)**,** e00570-16.

Boucher, Y., Cordero, O.X., Takemura, A., Hunt, D.E., Schliep, K., Bapteste, E., et al. (2011). Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *MBio* 2(2)**,** e00335-10.

Boucher, Y., Labbate, M., Koenig, J.E., and Stokes, H.W. (2007). Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends in Microbiology* 15(7)**,** 301-309.

Boucher, Y., Orata, F.D., and 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. *Frontiers in Microbiology* 6**,** 1120.

Breider, S., Freese, H.M., Sproer, C., Simon, M., Overmann, J., and Brinkhoff, T. (2017). *Phaeobacter porticola* sp. nov., an antibiotic-producing bacterium isolated from a sea harbour. *International Journal of Systematic and Evolutionary Microbiology* 67(7)**,** 2153-2159.

Breider, S., Scheuner, C., Schumann, P., Fiebig, A., Petersen, J., Pradella, S., et al. (2014). Genome-scale data suggest reclassifications in the *Leisingera-Phaeobacter* cluster including proposals for *Sedimentitalea* gen. nov. and *Pseudophaeobacter* gen. nov. *Frontiers in Microbiology* 5**,** 416.

Brenner, D.J. (1973). Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria. *International Journal of Systematic Bacteriology* 23(4)**,** 298-307.

Brenner, D.J., and Farmer, J.J., III (2005). "Family I. *Enterobacteriaceae*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part B: The Gammaproteobacteria,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, M. Goodfellow, F.A. Rainey & K.H. Schleifer. 2nd ed. (New York, NY, USA: Springer), 587-850.

Brenner, D.J., Martin, M.A., and Hoyer, B.H. (1967). Deoxyribonucleic acid homologies among some bacteria. *Journal of Bacteriology* 94(2)**,** 486-487.

Brown, M.R. (1991). The amino-acid and sugar composition of 16 species of microalgae used in mariculture. *Journal of Experimental Marine Biology and Ecology* 145(1)**,** 79-99.

Buchan, A., González, J.M., and Moran, M.A. (2005). Overview of the marine *Roseobacter* lineage. *Applied and Environmental Microbiology* 71(10)**,** 5665-5677.

Buddruhs, N., Pradella, S., Göker, M., Päuker, O., Pukall, R., Spröer, C., et al. (2013). Molecular and phenotypic analyses reveal the non-identity of the *Phaeobacter gallaeciensis* type strain deposits CIP 105210T and DSM 17395. *International Journal of Systematic and Evolutionary Microbiology* 63(11)**,** 4340-4349.

Burger, R. (2012). "EHEC O104:H4 in Germany 2011: large outbreak of bloody diarrhea and haemolytic uraemic syndrome by shiga toxin-producing *E. coli* via contaminated food," in *Improving Food Safety Through a One Health Approach: Workshop Summary,* eds. E.R. Choffnes, D.A. Relman, L. Olsen, R. Hutton & A. Mack. (Washington, DC, USA: National Academies Press), 115-130.

Cai, X., Wang, Y., Yang, X., Liu, J., Wu, Y., and Zhang, X.H. (2017). *Paraphaeobacter pallidus* gen. nov., sp. nov., isolated from seawater. *International Journal of Systematic and Evolutionary Microbiology* 67(8)**,** 2520-2526.

Cameron, D.E., Urbach, J.M., and Mekalanos, J.J. (2008). A defined transposon mutant library and its use in identifying motility genes in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 105(25)**,** 8736-8741.

Carda Diéguez, M. (2016). *Vibrio vulnificus: from water to host.* Doctor of Philosophy Thesis, University of Valencia.

Caro-Quintero, A., and Konstantinidis, K.T. (2012). Bacterial species may exist, metagenomics reveal. *Environmental Microbiology* 14(2)**,** 347-355.

Case, R.J., Longford, S.R., Campbell, A.H., Low, A., Tujula, N., Steinberg, P.D., et al. (2011). Temperature induced bacterial virulence and bleaching disease in a chemically defended marine macroalga. *Environmental Microbiology* 13(2)**,** 529-537.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17(4)**,** 540-552.

Centers for Disease Control and Prevention (2010). Update: cholera outbreak — Haiti, 2010. *MMWR: Morbidity and Mortality Weekly Report* 59(45)**,** 1473-1479.

Chaudhari, N.M., Gupta, V.K., and Dutta, C. (2016). BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports* 6**,** 24373.

Chen, C.Y. (2014). DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Frontiers in Microbiology* 5**,** 305.

Chen, Y., Gonzalez-Escalona, N., Hammack, T.S., Allard, M.W., Strain, E.A., and Brown, E.W. (2016). Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of *Listeria monocytogenes*. *Applied and Environmental Microbiology* 82(20)**,** 6258-6272.

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., et al. (2011). The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 364(1)**,** 33-42.

Choopun, N. (2004). *The population structure of Vibrio cholerae in Chesapeake Bay*. Doctor of Philosophy Thesis, University of Maryland.

Chun, J., Grim, C.J., Hasan, N.A., Lee, J.H., Choi, S.Y., Haley, B.J., et al. (2009). Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 106(36)**,** 15442-15447.

Claus, D. (1992). A standardized Gram staining procedure. *World Journal of Microbiology and Biotechnology* 8(4)**,** 451-452.

Cohan, F.M. (2002). What are bacterial species? *Annual Review of Microbiology* 56(1)**,** 457-487.

Cohan, F.M. (2016). Bacterial speciation: genetic sweeps in bacterial species. *Current Biology* 26(3)**,** R112-R115.

Coleman, M.L., and Chisholm, S.W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences of the United States of America* 107(43)**,** 18634-18639.

Collins, M.D., Rodrigues, U., Ash, C., Aguirre, M., Farrow, J.A.E., Martinez-Murcia, A., et al. (1991). Phylogenetic analysis of the genus *Lactobacillus* and related lactic acid bacteria as

determined by reverse transcriptase sequencing of 16S rRNA. *FEMS Microbiology Letters* 77(1)**,** 5-12.

Collis, C.M., and Hall, R.M. (1995). Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrobial Agents and Chemotherapy* 39(1)**,** 155-162.

Colwell, R.R. (1996). Global climate and infectious disease: the cholera paradigm. *Science* 274(5295)**,** 2025-2031.

Dalia, A.B., Seed, K.D., Calderwood, S.B., and Camilli, A. (2015). A globally distributed mobile genetic element inhibits natural transformation of *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 112(33)**,** 10485-10490.

Davis, B.R., Fanning, G.R., Madden, J.M., Steigerwalt, A.G., Bradford, H.B., Jr., Smith, H.L., Jr., et al. (1981). Characterization of biochemically atypical *Vibrio cholerae* strains and designation of a new pathogenic species, *Vibrio mimicus*. *Journal of Clinical Microbiology* 14(6)**,** 631-639.

De la Cruz, F., and Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in Microbiology* 8(3)**,** 128-133.

De Ley, J. (1968). "Molecular biology and bacterial phylogeny," in *Evolutionary Biology,* eds. T. Dobzhansky, M.K. Hecht & W.C. Steere. (Boston, MA, USA: Springer), 103-156.

De Meeûs, T., Prugnolle, F., and Agnew, P. (2007). Asexual reproduction: genetics and evolutionary aspects. *Cellular and Molecular Life Sciences* 64(11)**,** 1355-1372.

De Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences of the United States of America* 102(Suppl. 1)**,** 6600-6607.

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology* 56(6)**,** 879-886.

Delva, J.G. (2010). *Quake-hit Haiti battles cholera epidemic, 150 dead* [Online]. New York, NY, USA: Thomson Reuters. Available: http://www.reuters.com/article/2010/10/22/us-haiti-cholera-idUSTRE69L21520101022 [Accessed 25 March 2014].

Denef, V.J., Kalnejais, L.H., Mueller, R.S., Wilmes, P., Baker, B.J., Thomas, B.C., et al. (2010). Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America* 107(6)**,** 2383-2390.

Desvarieux, J. (2010). *At the heart of Haiti's cholera riots, anger at the U.N.* [Online]. New York, NY, USA: Time, Inc. Available: http://content.time.com/time/world/article/0,8599,2032437,00.html [Accessed 25 March 2014].

Detter, J.C., Johnson, S.L., Bishop-Lilly, K.A., Chain, P.S., Gibbons, H.S., Minogue, T.D., et al. (2014). "Nucleic acid sequencing for characterizing infectious and/or novel agents in complex samples," in *Biological Identification: DNA Amplification and Sequencing, Optical Sensing, Lab-On-Chip and Portable Systems,* ed. R.P. Schaudies. (Sawston, UK: Woodhead Publishing), 3-53.

Dewhirst, F.E., Shen, Z., Scimeca, M.S., Stokes, L.N., Boumenna, T., Chen, T., et al. (2005). Discordant 16S and 23S rRNA gene phylogenies for the genus *Helicobacter*: implications for phylogenetic inference and systematics. *Journal of Bacteriology* 187(17)**,** 6106-6118.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology* 2(5)**,** 414-424.

Doolittle, W.F., and Zhaxybayeva, O. (2009). On the origin of prokaryotic species. *Genome Research* 19(5)**,** 744-756.

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., and Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* 100(15)**,** 8817-8822.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* 102(40)**,** 14338-14343.

Dunning Hotopp, J.C. (2011). Horizontal gene transfer between bacteria and animals. *Trends in Genetics* 27(4)**,** 157-163.

Dziejman, M., Balon, E., Boyd, D., Fraser, C.M., Heidelberg, J.F., and Mekalanos, J.J. (2002). Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proceedings of the National Academy of Sciences of the United States of America* 99(3)**,** 1556-1561.

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

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910)**,** 133-138.

Eisen, J.A. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *Journal of Molecular Evolution* 41(6)**,** 1105-1123.

Eppinger, M., Pearson, T., Koenig, S.S., Pearson, O., Hicks, N., Agrawal, S., et al. (2014). Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *MBio* 5(6)**,** e01721-14.

Ereshefsky, M. (2010). Microbiology and the species problem. *Biology & Philosophy* 25(4)**,** 553-568.

Ettema, T.J.G., and Andersson, S.G.E. (2009). The α-proteobacteria: the Darwin finches of the bacterial world. *Biology Letters* 5(3)**,** 429-432.

Farmer, J.J., III, Janda, M., Brenner, F.W., Cameron, D.N., and Birkhead, K.M. (2005). "Genus I. *Vibrio*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part B: The Gammaproteobacteria,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, M. Goodfellow, F.A. Rainey & K.H. Schleifer. 2nd ed. (New York, NY, USA: Springer), 494-546.

Farmer, P. (2013). *Key statistics - facts and figures about the earthquake, cholera, and development challenges in Haiti* [Online]. New York, NY, USA: United Nations Office of the Secretary-General's Special Adviser for Community-Based Medicine and Lessons from Haiti. Available: http://www.lessonsfromhaiti.org/relief-and-recovery/key-statistics. [Accessed 25 March 2014].

Faruque, S.M., and Mekalanos, J.J. (2003). Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends in Microbiology* 11(11)**,** 505-510.

Faruque, S.M., and Mekalanos, J.J. (2012). "Epidemiology and molecular biology of *Vibrio cholerae*," in *Foodborne and Waterborne Bacterial Pathogens: Epidemiology, Evolution and Molecular Biology,* ed. S.M. Faruque. (Norfolk, UK: Caister Academic Press), 211-230.

Faruque, S.M., Zhu, J., Asadulghani, Kamruzzaman, M., and Mekalanos, J.J. (2003). Examination of diverse toxin-coregulated pilus-positive *Vibrio cholerae* strains fails to demonstrate evidence for *Vibrio* pathogenicity island phage. *Infection and Immunity* 71(6)**,** 2993-2999.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* 34(3)**,** e22.

Fernandes, N., Case, R.J., Longford, S.R., Seyedsayamdost, M.R., Steinberg, P.D., Kjelleberg, S., et al. (2011). Genomes and virulence factors of novel bacterial pathogens causing bleaching disease in the marine red alga *Delisea pulchra*. *PLoS One* 6(12)**,** e27387.

Figueras, M.J., Beaz-Hidalgo, R., Hossain, M.J., and Liles, M.R. (2014). Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. *Genome Announcements* 2(6)**,** e00927-14.

Fox, G.E., Wisotzkey, J.D., and Jurtshuk, P., Jr. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic Bacteriology* 42(1)**,** 166-170.

Frank, O., Pradella, S., Rohde, M., Scheuner, C., Klenk, H.P., Göker, M., et al. (2014). Complete genome sequence of the *Phaeobacter gallaeciensis* type strain CIP 105210T (= DSM 26640T = BS107T). *Standards in Genomic Sciences* 9(3)**,** 914-932.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323(5915)**,** 741-746.

Freeman, S., and Herron, J.C. (2013). *Evolutionary Analysis.* San Francisco, CA, USA: Benjamin Cummings.

Frerichs, R.R., Boncy, J., Barrais, R., Keim, P.S., and Piarroux, R. (2012). Source attribution of 2010 cholera epidemic in Haiti. *Proceedings of the National Academy of Sciences of the United States of America* 109(47)**,** E3208.

Fu, Y., Wang, R., Zhang, Z., and Jiao, N. (2016). Complete genome sequence of the D-amino acid catabolism bacterium *Phaeobacter* sp. strain JL2886, isolated from deep seawater of the South China Sea. *Genome Announcements* 4(5)**,** e00913-16.

Gaasbeek, E.J., Wagenaar, J.A., Guilhabert, M.R., Wösten, M.M.S.M., Van Putten, J.P., Van Der Graaf-Van Bloois, L., et al. (2009). A DNase encoded by integrated element CJIE1 inhibits natural transformation of *Campylobacter jejuni*. *Journal of Bacteriology* 191(7)**,** 2296-2306.

García, B., Latasa, C., Solano, C., García-del Portillo, F., Gamazo, C., and Lasa, I. (2004). Role of the GGDEF protein family in *Salmonella* cellulose biosynthesis and biofilm formation. *Molecular Microbiology* 54(1)**,** 264-277.

Gargis, A.S., Kalman, L., Berry, M.W., Bick, D.P., Dimmock, D.P., Hambuch, T., et al. (2012). Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature Biotechnology* 30(11)**,** 1033-1036.

Garrity, G.M. (2016). A new genomics-driven taxonomy of *Bacteria* and *Archaea*: are we there yet? *Journal of Clinical Microbiology* 54(8)**,** 1956-1963.

Garrity, G.M., Bell, J.A., and Lilburn, T. (2005). "Order III. *Rhodobacterales* ord. nov.," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part C: The Alpha-, Beta-, Delta-, and Epsilonproteobacteria,* eds. D.J. Brenner, N.R. Krieg & J.T. Staley. 2nd ed. (New York, NY, USA: Springer), 161-229.

Ge, F., Wang, L.S., and Kim, J. (2005). The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biology* 3(10)**,** e316.

Gertz, E.M., Yu, Y.K., Agarwala, R., Schaffer, A.A., and Altschul, S.F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology* 4**,** 41.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., et al. (2005). Opinion: re-evaluating prokaryotic species. *Nature Reviews Microbiology* 3(9)**,** 733-739.

Giovanni, S.J., and Rappé, M.S. (2000). "Evolution, diversity, and molecular ecology of marine prokaryotes," in *Microbial Ecology of the Oceans,* ed. D.L. Kichman. (New York, NY, USA: John Wiley & Sons, Inc.), 47-84.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and Field, K.G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345(6270)**,** 60-63.

Gladney, L.M., Katz, L.S., Knipe, K.M., Rowe, L.A., Conley, A.B., Rishishwar, L., et al. (2014). Genome sequences of *Vibrio navarrensis*, a potential human pathogen. *Genome Announcements* 2(6)**,** e01188-14.

Gladney, L.M., and Tarr, C.L. (2014). Molecular and phenotypic characterization of *Vibrio navarrensis* isolates associated with human illness. *Journal of Clinical Microbiology* 52(11)**,** 4070-4074.

Gladstone, R. (2013). *Rights advocates suing U.N. over the spread of cholera in Haiti* [Online]. New York, NY, USA: The New York Times Company. Available: http://www.nytimes.com/2013/10/09/world/americas/rights-advocates-suing-un-over-the-spread-of-cholera-in-haiti.html [Accessed 25 March 2014].

Glaeser, S.P., and Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology* 38(4)**,** 237-245.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11(5)**,** 759-769.

Goecks, J., Nekrutenko, A., Taylor, J., and The Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11(8)**,** R86.

Gomez-Gil, B., Thompson, C.C., Matsumura, Y., Sawabe, T., Iida, T., Christen, R., et al. (2014). "The family *Vibrionaceae*," in *The Prokaryotes – Gammaproteobacteria,* eds. E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt & F. Thompson. 4th ed. (Berlin, Germany: Springer), 659-747.

Gonzalez-Escalona, N., Jolley, K.A., Reed, E., and Martinez-Urtaza, J. (2017). Defining a core genome multilocus sequence typing scheme for the global epidemiology of *Vibrio parahaemolyticus*. *Journal of Clinical Microbiology* 55(6)**,** 1682-1697.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17(6)**,** 333-351.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57(1)**,** 81-91.

Grant, J.R., Arantes, A.S., and Stothard, P. (2012). Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics* 13**,** 202.

Guo, L.L., Wu, Y.H., Xu, X.W., Huang, C.J., Xu, Y.Y., Cheng, H., et al. (2017). *Actibacterium pelagium* sp. nov., a novel alphaproteobacterium, and emended description of the genus *Actibacterium*. *International Journal of Systematic and Evolutionary Microbiology* 10.1099/ijsem.0.002417.

Guttman, D.S., and Dykhuizen, D.E. (1994). Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* 138(4)**,** 993-1003.

Haley, B.J., Grim, C.J., Hasan, N.A., Choi, S.Y., Chun, J., Brettin, T.S., et al. (2010). Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. *BMC Microbiology* 10**,** 154.

Hameed, A., Shahina, M., Lin, S.Y., Nakayan, P., Liu, Y.C., Lai, W.A., et al. (2014). *Youngimonas vesicularis* gen. nov., sp. nov., of the family *Rhodobacteraceae*, isolated from surface seawater, reclassification of *Donghicola xiamenensis* Tan et al. 2009 as *Pseudodonghicola xiamenensis* gen. nov., comb. nov. and emended description of the genus *Donghicola* Yoon et al. 2007. *International Journal of Systematic and Evolutionary Microbiology* 64(8)**,** 2729-2737.

Handelsman, J., Tiedje, J., Alvarez-Cohen, L., Ashburner, M., Cann, I.K.O., DeLong, E.F., et al. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Washington, DC, USA: National Academies Press.

Harris, J.B., LaRocque, R.C., Qadri, F., Ryan, E.T., and Calderwood, S.B. (2012). Cholera. *Lancet* 379(9835)**,** 2466-2476.

Hasan, N.A., Choi, S.Y., Eppinger, M., Clark, P.W., Chen, A., Alam, M., et al. (2012). Genomic diversity of 2010 Haitian cholera outbreak strains. *Proceedings of the National Academy of Sciences of the United States of America* 109(29)**,** E2010-E2017.

Hasan, N.A., Grim, C.J., Haley, B.J., Chun, J., Alam, M., Taviani, E., et al. (2010). Comparative genomics of clinical and environmental *Vibrio mimicus*. *Proceedings of the National Academy of Sciences of the United States of America* 107(49)**,** 21134-21139.

Heger, M. (2013). *McGill University team develops rapid genome sequencing technique for outbreak monitoring* [Online]. New York, NY, USA: GenomeWeb LLC. Available: http://www.genomeweb.com/sequencing/mcgill-university-team-develops-rapid-genome-sequencing-technique-outbreak-monit [Accessed January 28 2014].

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., et al. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406(6795)**,** 477-483.

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., et al. (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* 2(4)**,** e00157-11.

Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K., and Schuster, S.C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics* 21(10)**,** 2329-2335.

Hillis, D.M., and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42(2)**,** 182-192.

Holmes, B., Willcox, W.R., and Lapage, S.P. (1978). Identification of *Enterobacteriaceae* by the API 20E system. *Journal of Clinical Pathology* 31(1)**,** 22-30.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., et al. (2016). A new view of the tree of life. *Nature Microbiology* 1(5)**,** 16048.

Hunkapiller, T., Kaiser, R.J., Koop, B.F., and Hood, L. (1991). Large-scale and automated DNA sequence determination. *Science* 254(5028)**,** 59-67.

Hunt, D.E., Klepac-Ceraj, V., Acinas, S.G., Gautier, C., Bertilsson, S., and Polz, M.F. (2006). Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Applied and Environmental Microbiology* 72(3)**,** 2221-2225.

Huq, A., Colwell, R.R., Rahman, R., Ali, A., Chowdhury, M.A., Parveen, S., et al. (1990). Detection of *Vibrio cholerae* O1 in the aquatic environment by fluorescent-monoclonal antibody and culture methods. *Applied and Environmental Microbiology* 56(8)**,** 2370-2373.

Huynen, M.A., and Bork, P. (1998). Measuring genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* 95(11)**,** 5849-5856.

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

Imhoff, J.F., Trüper, H.G., and Pfennig, N. (1984). Rearrangement of the species and genera of the phototrophic "purple nonsulfur bacteria". *International Journal of Systematic and Evolutionary Microbiology* 34(3)**,** 340-343.

Iqbal, N., Guérout, A.M., Krin, E., Le Roux, F., and Mazel, D. (2015). Comprehensive functional analysis of the 18 *Vibrio cholerae* N16961 toxin-antitoxin systems substantiates their role in stabilizing the superintegron. *Journal of Bacteriology* 197(13)**,** 2150-2159.

Ivers, L.C., and Walton, D.A. (2012). The "first" case of cholera in Haiti: lessons for global health. *American Journal of Tropical Medicine and Hygiene* 86(1)**,** 36-38.

Janda, J.M., and Abbott, S.L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology* 45(9)**,** 2761-2764.

Jenson, D., Szabo, V., and Duke FHI Haiti Humanities Laboratory Student Research Team (2011). Cholera in Haiti and other Caribbean regions, 19th century. *Emerging Infectious Diseases* 17(11)**,** 2130-2135.

Jermyn, W.S., and Boyd, E.F. (2002). Characterization of a novel *Vibrio* pathogenicity island (VPI-2) encoding neuraminidase (*nanH*) among toxigenic *Vibrio cholerae* isolates. *Microbiology* 148(11)**,** 3681-3693.

Jiang, S.C., and Fu, W. (2001). Seasonal abundance and distribution of *Vibrio cholerae* in coastal waters quantified by a 16S-23S intergenic spacer probe. *Microbial Ecology* 42(4)**,** 540-548.

Johnson, J.L. (1973). Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. *International Journal of Systematic Bacteriology* 23(4)**,** 308-315.

Johnson, J.L., and Ordal, E.J. (1968). Deoxyribonucleic acid homology in bacterial taxonomy: effect of incubation temperature on reaction specificity. *Journal of Bacteriology* 95(3)**,** 893-900.

Jones, M.K., and Oliver, J.D. (2009). *Vibrio vulnificus*: disease and pathogenesis. *Infection and Immunity* 77(5)**,** 1723-1733.

Jores, J., Appel, B., and Lewin, A. (2007). *Vibrio navarrensis* biotype *pommerensis*: a new biotype of *V. navarrensis* isolated in the German Baltic Sea. *Systematic and Applied Microbiology* 30(1)**,** 27-30.

Joshi, A., Kostiuk, B., Rogers, A., Teschler, J., Pukatzki, S., and Yildiz, F.H. (2017). Rules of engagement: the type VI secretion system in *Vibrio cholerae*. *Trends in Microbiology* 25(4)**,** 267-279.

Kämpfer, P., Buczolits, S., Albrecht, A., Busse, H.J., and Stackebrandt, E. (2003). Towards a standardized format for the description of a novel species (of an established genus): *Ochrobactrum gallinifaecis* sp. nov. *International Journal of Systematic and Evolutionary Microbiology* 53(3)**,** 893-896.

Kämpfer, P., and Glaeser, S.P. (2012). Prokaryotic taxonomy in the sequencing era—the polyphasic approach revisited. *Environmental Microbiology* 14(2)**,** 291-317.

Karaolis, D.K., Somara, S., Maneval, D.R., Jr., Johnson, J.A., and Kaper, J.B. (1999). A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. *Nature* 399(6734)**,** 375-379.

Katz, J.M. (2010). *UN probes base as source of Haiti cholera outbreak* [Online]. Washington, DC, USA: The Washington Times, LLC. Available: http://www.washingtontimes.com/news/2010/oct/27/un-probes-base-as-source-of-haiti-cholera-outbreak [Accessed 25 March 2014].

Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., et al. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4(4)**,** e00398-13.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12)**,** 1647-1649.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9(8)**,** 605-618.

Keymer, D.P., and Boehm, A.B. (2011). Recombination shapes the structure of an environmental *Vibrio cholerae* population. *Applied and Environmental Microbiology* 77(2)**,** 537-544.

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

Kim, S., Kurokawa, D., Watanabe, K., Makino, S., Shirahata, T., and Watarai, M. (2004). *Brucella abortus* nicotinamidase (PncA) contributes to its intracellular replication and infectivity in mice. *FEMS Microbiology Letters* 234(2)**,** 289-295.

Kim, Y.O., Park, S., Kim, H., Park, D.S., Nam, B.H., Kim, D.G., et al. (2014b). *Halocynthiibacter namhaensis* gen. nov., sp. nov., a novel alphaproteobacterium isolated from sea squirt *Halocynthia roretzi*. *Antonie van Leeuwenhoek* 105(5)**,** 881-889.

Kirchberger, P.C., Orata, F.D., Barlow, E.J., Kauffman, K.M., Case, R.J., Polz, M.F., et al. (2016). A small number of phylogenetically distinct clonal complexes dominate a coastal *Vibrio cholerae* population. *Applied and Environmental Microbiology* 82(18)**,** 5576-5586.

Kirchberger, P.C., Turnsek, M., Hunt, D.E., Haley, B.J., Colwell, R.R., Polz, M.F., et al. (2014). *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. *International Journal of Systematic and Evolutionary Microbiology* 64(9)**,** 3208-3214.

Kirchberger, P.C., Unterweger, D., Provenzano, D., Pukatzki, S., and 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.

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing – concepts and limitations. *Bioessays* 32(6)**,** 524-536.

Kislyuk, A.O., Katz, L.S., Agrawal, S., Hagen, M.S., Conley, A.B., Jayaraman, P., et al. (2010). A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* 26(15)**,** 1819-1826.

Kitahara, K., and Miyazaki, K. (2013). Revisiting bacterial phylogeny: natural and experimental evidence for horizontal gene transfer of 16S rRNA. *Mobile Genetic Elements* 3(1)**,** e24210.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Research* 29(1)**,** 181-184.

Klassen, J.L., and Currie, C.R. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* 13**,** 14.

Klenk, H.P., and Göker, M. (2010). En route to a genome-based classification of *Archaea* and *Bacteria*? *Systematic and Applied Microbiology* 33(4)**,** 175-182.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research* 41(1)**,** e1.

Koenig, J.E., Boucher, Y., Charlebois, R.L., Nesbo, C., Zhaxybayeva, O., Bapteste, E., et al. (2008). Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. *Environmental Microbiology* 10(4)**,** 1024-1038.

Konstantinidis, K.T., Rosselló-Móra, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. *ISME Journal* 11(11)**,** 2399-2406.

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

Konstantinidis, K.T., and Tiedje, J.M. (2005b). Towards a genome-based taxonomy for prokaryotes. *Journal of Bacteriology* 187(18)**,** 6258-6264.

Koonin, E.V., and Novozhilov, A.S. (2017). Origin and evolution of the universal genetic code. *Annual Review of Genetics* 51(1)**,** 45-62.

Korlach, J., Bjornson, K.P., Chaudhuri, B.P., Cicero, R.L., Flusberg, B.A., Gray, J.J., et al. (2010). Real-time DNA sequencing from single polymerase molecules. *Methods in Enzymology* 472**,** 431-455.

Krieg, N.R. (1988). Bacterial classification: an overview. *Canadian Journal of Microbiology* 34(4)**,** 536-540.

Krieg, N.R., and Garrity, G.M. (2005). "On using the *Manual*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part A: Introductory Essays,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley & G.M. Garrity. 2nd ed. (New York, NY, USA: Springer), 15-19.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33(7)**,** 1870-1874.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biology* 5(2)**,** R12.

Kuykendall, L.D. (2005). "Order VI. *Rhizobiales* ord. nov.," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part C: The Alpha-, Beta-, Delta-, and Epsilonproteobacteria,* eds. D.J. Brenner, N.R. Krieg & J.T. Staley. 2nd ed. (New York, NY, USA: Springer), 324-574.

Kyrpides, N.C., Hugenholtz, P., Eisen, J.A., Woyke, T., Göker, M., Parker, C.T., et al. (2014). Genomic encyclopedia of *Bacteria* and *Archaea*: sequencing a myriad of type strains. *PLoS Biology* 12(8)**,** e1001920.

Lam, C., Octavia, S., Reeves, P., Wang, L., and Lan, R. (2010). Evolution of seventh cholera pandemic and origin of 1991 epidemic, Latin America. *Emerging Infectious Diseases* 16(7)**,** 1130-1132.

Langille, M.G., and Brinkman, F.S. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 25(5)**,** 664-665.

Lantagne, D., Nair, G.B., Lanata, C.F., and Cravioto, A. (2014). "The cholera outbreak in Haiti: where and how did it begin?," in *Cholera Outbreaks: Current Topics in Microbiology and Immunology,* eds. B.G. Nair & Y. Takeda. (Berlin, Germany: Springer-Verlag), 145-164.

Lapierre, P., and Gogarten, J.P. (2009). Estimating the size of the bacterial pan-genome. *Trends in Genetics* 25(3)**,** 107-110.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21)**,** 2947-2948.

Lau, S.C., Tsoi, M.M., Li, X., Plakhotnikova, I., Wu, M., Wong, P.K., et al. (2004). *Loktanella hongkongensis* sp. nov., a novel member of the *α-Proteobacteria* originating from marine biofilms in Hong Kong waters. *International Journal of Systematic and Evolutionary Microbiology* 54(6)**,** 2281-2284.

Lawrence, J.G., and Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences of the United States of America* 95(16)**,** 9413-9417.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* 13(9)**,** 2178-2189.

Lombardo, M.J., Michalski, J., Martinez-Wilson, H., Morin, C., Hilton, T., Osorio, C.G., et al. (2007). An *in vivo* expression technology screen for *Vibrio cholerae* genes expressed in human volunteers. *Proceedings of the National Academy of Sciences of the United States of America* 104(46)**,** 18229-18234.

Louis, P., and Doré, J. (2014). "Functional metagenomics of human intestinal microbiome β-glucuronidase activity," in *Encyclopedia of Metagenomics,* ed. K.E. Nelson. (New York, NY, USA: Springer), 1-8.

Lucena, T., Pujalte, M.J., Ruvira, M.A., Garay, E., Macián, M.C., and Arahal, D.R. (2012). *Tropicibacter multivorans* sp. nov., an aerobic alphaproteobacterium isolated from surface seawater. *International Journal of Systematic and Evolutionary Microbiology* 62(4)**,** 844-848.

Ludwig, W., and Schleifer, K.H. (1994). Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiology Reviews* 15(2-3)**,** 155-173.

Luo, C., Rodriguez-R, L.M., and Konstantinidis, K.T. (2014). MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research* 42(8)**,** e73.

Luo, H., and Moran, M.A. (2014). Evolutionary ecology of the marine *Roseobacter* clade. *Microbiology and Molecular Biology Reviews* 78(4)**,** 573-587.

Madigan, M.T., Martinko, J.M., Bender, K.S., Buckley, D.H., and Stahl, D.A. (2014). *Brock Biology of Microorganisms.* San Francisco, CA, USA: Benjamin Cummings.

Magnelli, P., Bielik, A., and Guthrie, E. (2012). "Identification and characterization of protein glycosylation using specific endo- and exoglycosidases," in *Protein Expression in Mammalian Cells: Methods and Protocols (Methods in Molecular Biology),* ed. J.L. Hartley. (New York, NY, USA: Humana Press), 189-211.

Maharjan, L. (2010). *Cholera outbreak looms over capital* [Online]. Kathmandu, Nepal: International Media Network Nepal Pvt. Ltd. Available: http://www.thehimalayantimes.com/fullNews.php?headline=Cholera+outbreak+looms+over+capital&NewsID=258974 [Accessed 25 March 2014].

Maiden, M.C. (2006). Multilocus sequence typing of bacteria. *Annual Review of Microbiology* 60(1)**,** 561-588.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* 95(6)**,** 3140-3145.

Maiden, M.C., Jansen van Rensburg, M.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A., et al. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology* 11(10)**,** 728-736.

Majewski, J., and Cohan, F.M. (1999). DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* 153(4)**,** 1525-1533.

Mallet, J. (1995). A species definition for the modern synthesis. *Trends in Ecology & Evolution* 10(7)**,** 294-299.

Mardis, E.R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* 6(1)**,** 287-303.

Mardis, E.R. (2017). DNA sequencing technologies: 2006-2016. *Nature Protocols* 12(2)**,** 213-218.

Martens, T., Gram, L., Grossart, H.P., Kessler, D., Müller, R., Simon, M., et al. (2007). Bacteria of the *Roseobacter* clade show potential for secondary metabolite production. *Microbial Ecology* 54(1)**,** 31-42.

Martens, T., Heidorn, T., Pukall, R., Simon, M., Tindall, B.J., and Brinkhoff, T. (2006). Reclassification of *Roseobacter gallaeciensis* Ruiz-Ponte et al. 1998 as *Phaeobacter gallaeciensis* gen. nov., comb. nov., description of *Phaeobacter inhibens* sp. nov., reclassification of *Ruegeria algicola* (Lafay et al. 1995) Uchino et al. 1999 as *Marinovum algicola* gen. nov., comb. nov., and emended descriptions of the genera *Roseobacter*, *Ruegeria* and *Leisingera*. *International Journal of Systematic and Evolutionary Microbiology* 56(6)**,** 1293-1304.

Martinez-Murcia, A.J., Benlloch, S., and Collins, M.D. (1992). Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. *International Journal of Systematic Bacteriology* 42(3)**,** 412-421.

Martinez-Murcia, A.J., and Collins, M.D. (1990). A phylogenetic analysis of the genus *Leuconostoc* based on reverse transcriptase sequencing of 16S rRNA. *FEMS Microbiology Letters* 58(1)**,** 73-83.

Mathur, J., and Waldor, M.K. (2004). The *Vibrio cholerae* ToxR-regulated porin OmpU confers resistance to antimicrobial peptides. *Infection and Immunity* 72(6)**,** 3577-3583.

Mayr, E. (1942). *Systematics and the Origin of Species*. New York, NY, USA: Columbia University Press.

Mazel, D. (2006). Integrons: agents of bacterial evolution. *Nature Reviews Microbiology* 4(8)**,** 608-620.

Mazel, D., Dychinco, B., Webb, V.A., and Davies, J. (1998). A distinctive class of integron in the *Vibrio cholerae* genome. *Science* 280(5363)**,** 605-608.

McCarthy, B.J., and Bolton, E.T. (1963). An approach to the measurement of genetic relatedness among organisms. *Proceedings of the National Academy of Sciences of the United States of America* 50**,** 156-164.

Meibom, K.L., Blokesch, M., Dolganov, N.A., Wu, C.Y., and Schoolnik, G.K. (2005). Chitin induces natural competence in *Vibrio cholerae*. *Science* 310(5755)**,** 1824-1827.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14**,** 60.

Meier-Kolthoff, J.P., Klenk, H.P., and Göker, M. (2014). Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. *International Journal of Systematic and Evolutionary Microbiology* 64(2)**,** 352-356.

Mekalanos, J.J., Robins, W., Ussery, D.W., Davis, B.M., Schadt, E., and Waldor, M.K. (2012). Non-O1 *Vibrio cholerae* unlinked to cholera in Haiti. *Proceedings of the National Academy of Sciences of the United States of America* 109(47)**,** E3206.

Melendrez, M.C., Becraft, E.D., Wood, J.M., Olsen, M.T., Bryant, D.A., Heidelberg, J.F., et al. (2015). Recombination does not hinder formation or detection of ecological species of *Synechococcus* inhabiting a hot spring cyanobacterial mat. *Frontiers in Microbiology* 6**,** 1540.

Mellmann, A., Harmsen, D., Cummings, C.A., Zentz, E.B., Leopold, S.R., Rico, A., et al. (2011). Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6(7)**,** e22751.

Metzker, M.L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* 11(1)**,** 31-46.

Mignard, S., and Flandrois, J.P. (2006). 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *Journal of Microbiological Methods* 67(3)**,** 574-581.

Ministère de la Santé Publique et de la Population (2014). *Rapport journalier MSPP du 07 Janvier 2014* [Online]. Port-au-Prince, Haiti: Ministère de la Santé Publique et de la Population. Available: http://mspp.gouv.ht/site/downloads/Rapport Web_07.01_Avec_Courbes_Departementales.pdf [Accessed 28 January 2014].

Mitra, R.D., and Church, G.M. (1999). *In situ* localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research* 27(24)**,** e34.

Moran, M.A., Belas, R., Schell, M.A., Gonzalez, J.M., Sun, F., Sun, S., et al. (2007). Ecological genomics of marine roseobacters. *Applied and Environmental Microbiology* 73(14)**,** 4559-4569.

Moura, A., Criscuolo, A., Pouseele, H., Maury, M.M., Leclercq, A., Tarr, C., et al. (2016). Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nature Microbiology* 2(2)**,** 16185.

Mukherjee, S., Seshadri, R., Varghese, N.J., Eloe-Fadrosh, E.A., Meier-Kolthoff, J.P., Göker, M., et al. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology* 35(7)**,** 676-683.

Mutreja, A., Kim, D.W., Thomson, N.R., Connor, T.R., Lee, J.H., Kariuki, S., et al. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365)**,** 462-465.

Nesterchuk, M.V., Sergiev, P.V., and Dontsova, O.A. (2011). Posttranslational modifications of ribosomal proteins in *Escherichia coli*. *Acta Naturae* 3(2)**,** 22-33.

O'Shea, Y.A., Finnan, S., Reen, F.J., Morrissey, J.P., O'Gara, F., and Boyd, E.F. (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(12)**,** 4053-4063.

Oh, S., Caro-Quintero, A., Tsementzi, D., DeLeon-Rodriguez, N., Luo, C., Poretsky, R., et al. (2011). Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Applied and Environmental Microbiology* 77(17)**,** 6000-6011.

Orata, F.D., Keim, P.S., and Boucher, Y. (2014). The 2010 cholera outbreak in Haiti: how science solved a controversy. *PLoS Pathogens* 10(4)**,** e1003967.

Orata, F.D., Kirchberger, P.C., Méheust, R., Barlow, E.J., Tarr, C.L., and 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(10)**,** 2941-2954.

Orata, F.D., Xu, Y., Gladney, L.M., Rishishwar, L., Case, R.J., Boucher, Y., et al. (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(10)**,** 4148-4155.

Oren, A., and Garrity, G.M. (2014). Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. *Antonie van Leeuwenhoek* 106(1)**,** 43-56.

Overman, T.L., Kessler, J.F., and Seabolt, J.P. (1985). Comparison of API 20E, API rapid E, and API rapid NFT for identification of members of the family *Vibrionaceae*. *Journal of Clinical Microbiology* 22(5)**,** 778-781.

Pannekoek, Y., Qi-Long, Q., Zhang, Y.Z., and Van Der Ende, A. (2016). Genus delineation of *Chlamydiales* by analysis of the percentage of conserved proteins justifies the reunifying of the genera *Chlamydia* and *Chlamydophila* into one single genus *Chlamydia*. *Pathogens and Disease* 74(6)**,** ftw071.

Papke, R.T., Zhaxybayeva, O., Feil, E.J., Sommerfeld, K., Muise, D., and Doolittle, W.F. (2007). Searching for species in haloarchaea. *Proceedings of the National Academy of Sciences of the United States of America* 104(35)**,** 14092-14097.

Park, S., Won, S.M., Kim, H., Park, D.S., and Yoon, J.H. (2014). *Aestuariivita boseongensis* gen. nov., sp. nov., isolated from a tidal flat sediment. *International Journal of Systematic and Evolutionary Microbiology* 64(9)**,** 2969-2974.

Parker, A.A. (2010). *Cholera in Haiti — the climate connection* [Online]. Traverse City, MI, USA: Circle of Blue. Available: http://www.circleofblue.org/waternews/2010/world/hold-cholera-in-haiti-the-climate-connection [Accessed 22 March 2014].

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* 2(11)**,** 1533-1542.

Petrack, B., Greengard, P., Craston, A., and Sheppy, F. (1965). Nicotinamide deamidase from mammalian liver. *Journal of Biological Chemistry* 240(4)**,** 1725-1730.

Piarroux, R., Barrais, R., Faucher, B., Haus, R., Piarroux, M., Gaudart, J., et al. (2011). Understanding the cholera epidemic, Haiti. *Emerging Infectious Diseases* 17(7)**,** 1161-1168.

Pightling, A.W., Petronella, N., and Pagotto, F. (2015). The *Listeria monocytogenes* Core-Genome Sequence Typer (LmCGST): a bioinformatic pipeline for molecular characterization with next-generation sequence data. *BMC Microbiology* 15**,** 224.

Provenzano, D., Schuhmacher, D.A., Barker, J.L., and Klose, K.E. (2000). The virulence regulatory protein ToxR mediates enhanced bile resistance in *Vibrio cholerae* and other pathogenic *Vibrio* species. *Infection and Immunity* 68(3)**,** 1491-1497.

Purser, J.E., Lawrenz, M.B., Caimano, M.J., Howell, J.K., Radolf, J.D., and Norris, S.J. (2003). A plasmid-encoded nicotinamidase (PncA) is essential for infectivity of *Borrelia burgdorferi* in a mammalian host. *Molecular Microbiology* 48(3)**,** 753-764.

Qin, Q.L., Xie, B.B., Zhang, X.Y., Chen, X.L., Zhou, B.C., Zhou, J., et al. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of Bacteriology* 196(12)**,** 2210-2215.

R Development Core Team (2014). "R: a language and environment for statistical computing". 3.1.2 ed. (Vienna, Austria: R Foundation for Statistical Computing).

Rashed, S.M., Mannan, S.B., Johura, F.T., Islam, M.T., Sadique, A., Watanabe, H., et al. (2012). Genetic characteristics of drug-resistant *Vibrio cholerae* O1 causing endemic cholera in Dhaka, 2006-2011. *Journal of Medical Microbiology* 61(12)**,** 1736-1745.

Rashid, M.U., Rashed, S.M., Islam, T., Johura, F.T., Watanabe, H., Ohnishi, M., et al. (2016). *CtxB1* outcompetes *CtxB7* in *Vibrio cholerae* O1, Bangladesh. *Journal of Medical Microbiology* 65(1)**,** 101-103.

Rasko, D.A., Myers, G.S., and Ravel, J. (2005). Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics* 6**,** 2.

Reimer, A.R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., et al. (2011). Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerging Infectious Diseases* 17(11)**,** 2113-2121.

Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015). High-throughput sequencing technologies. *Molecular Cell* 58(4)**,** 586-597.

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

Ricker, N., Qian, H., and Fulthorpe, R.R. (2012). The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. *Genomics* 100(3)**,** 167-175.

Roberts, R.J., Carneiro, M.O., and Schatz, M.C. (2013). The advantages of SMRT sequencing. *Genome Biology* 14(7)**,** 405.

Rodriguez-R, L.M., and Konstantinidis, K.T. (2005). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints* 4(1)**,** e1900v1.

Rodriguez-R, L.M., and Konstantinidis, K.T. (2014). Bypassing cultivation to identify bacterial species. *Microbe* 9(3)**,** 111-118.

Rosselló-Móra, R., and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiology Reviews* 25(1)**,** 39-67.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering* 12(2)**,** 85-94.

Rowe-Magnus, D.A., and Mazel, D. (2002). The role of integrons in antibiotic resistance gene capture. *International Journal of Medical Microbiology* 292(2)**,** 115-125.

Rüger, H.J., and Höfle, M.G. (1992). Marine star-shaped-aggregate-forming bacteria: *Agrobacterium atlanticum* sp. nov.; *Agrobacterium meteori* sp. nov.; *Agrobacterium ferrugineum* sp. nov., nom. rev.; *Agrobacterium gelatinovorum* sp. nov., nom. rev.; and *Agrobacterium stellulatum* sp. nov., nom. rev. *International Journal of Systematic Bacteriology* 42(1)**,** 133-143.

Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H.L., Allerberger, F., et al. (2015). Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Listeria monocytogenes*. *Journal of Clinical Microbiology* 53(9)**,** 2869-2876.

Ryu, K.S., Kim, C., Kim, I., Yoo, S., Choi, B.S., and Park, C. (2004). NMR application probes a novel and ubiquitous family of enzymes that alter monosaccharide configuration. *Journal of Biological Chemistry* 279(24)**,** 25544-25548.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4)**,** 406-425.

Samant, S., Lee, H., Ghassemi, M., Chen, J., Cook, J.L., Mankin, A.S., et al. (2008). Nucleotide biosynthesis is critical for growth of bacteria in human blood. *PLoS Pathogens* 4(2)**,** e37.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., et al. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265(5596)**,** 687-695.

Santos, S.R., and Ochman, H. (2004). Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environmental Microbiology* 6(7)**,** 754-759.

Sawada, H., and Takagi, Y. (1964). The metabolism of L-rhamnose in *Escherichia coli*: III. L-rhamnulose-phosphate aldolase. *Biochimica et Biophysica Acta* 92(1)**,** 26-32.

Schliep, K., Lopez, P., Lapointe, F.J., and Bapteste, É. (2011). Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Molecular Biology and Evolution* 28(4)**,** 1393-1405.

Schliep, K.P. (2011). Phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4)**,** 592-593.

Schmidt, H., and Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clinical Microbiology Reviews* 17(1)**,** 14-56.

Schouls, L.M., Schot, C.S., and Jacobs, J.A. (2003). Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. *Journal of Bacteriology* 185(24)**,** 7241-7246.

Seed, K.D., Yen, M., Shapiro, B.J., Hilaire, I.J., Charles, R.C., Teng, J.E., et al. (2014). Evolutionary consequences of intra-patient phage predation on microbial populations. *Elife* 3**,** e03497.

Segata, N., Börnigen, D., Morgan, X.C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications* 4**,** 2304.

Segata, N., and Huttenhower, C. (2011). Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS One* 6(9)**,** e24704.

Seitz, P., and Blokesch, M. (2013). DNA-uptake machinery of naturally competent *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 110(44)**,** 17987-17992.

Sentausa, E., and Fournier, P.E. (2013). Advantages and limitations of genomics in prokaryotic taxonomy. *Clinical Microbiology and Infection* 19(9)**,** 790-795.

Seyedsayamdost, M.R., Case, R.J., Kolter, R., and Clardy, J. (2011). The Jekyll-and-Hyde chemistry of *Phaeobacter gallaeciensis*. *Nature Chemistry* 3(4)**,** 331-335.

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabo, G., et al. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science* 336(6077)**,** 48-51.

Shayegani, M., Maupin, P.S., and McGlynn, D.M. (1978). Evaluation of the API 20E system for identification of nonfermentative Gram-negative bacteria. *Journal of Clinical Microbiology* 7(6)**,** 539-545.

Shea, A., Wolcott, M., Daefler, S., and Rozak, D.A. (2012). "Biolog phenotye microarrays," in *Microbial Systems Biology. Methods in Molecular Biology (Methods and Protocols),* ed. A. Navid. (Totowa, NJ, USA: Humana Press), 331-373.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology* 26(10)**,** 1135-1145.

Shendure, J.A., Porreca, G.J., Church, G.M., Gardner, A.F., Hendrickson, C.L., Kieleczawa, J., et al. (2011). Overview of DNA sequencing strategies. *Current Protocols in Molecular Biology* 96(7.1)**,** 7.1.1-7.1.23.

Shiba, T. (1991). *Roseobacter litoralis* gen. nov., sp. nov., and *Roseobacter denitrificans* sp. nov., aerobic pink-pigmented bacteria which contain bacteriochlorophyll a. *Systematic and Applied Microbiology* 14(2)**,** 140-145.

Shimada, T., Nair, G.B., Deb, B.C., Albert, M.J., Sack, R.B., and Takeda, Y. (1993). Outbreak of *Vibrio cholerae* non-O1 in India and Bangladesh. *Lancet* 341(8856)**,** 1347.

Simon, M., Scheuner, C., Meier-Kolthoff, J.P., Brinkhoff, T., Wagner-Döbler, I., Ulbrich, M., et al. (2017). Phylogenomics of *Rhodobacteraceae* reveals evolutionary adaptation to marine and non-marine habitats. *ISME Journal* 11(6)**,** 1483-1499.

Skorupski, K., and Taylor, R.K. (1997). Cyclic AMP and its receptor protein negatively regulate the coordinate expression of cholera toxin and toxin-coregulated pilus in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 94(1)**,** 265-270.

Sneath, P.H.A., and Brenner, D.J. (1992). “Official” nomenclature lists. *ASM News* 58(1)**,** 175.

Snel, B., Bork, P., and Huynen, M.A. (1999). Genome phylogeny based on gene content. *Nature Genetics* 21(1)**,** 108-110.

Sonnenschein, E.C., Phippen, C.B.W., Nielsen, K.F., Mateiu, R.V., Melchiorsen, J., Gram, L., et al. (2017). *Phaeobacter piscinae* sp. nov., a species of the *Roseobacter* group and potential aquaculture probiont. *International Journal of Systematic and Evolutionary Microbiology* 10.1099/ijsem.0.002331.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* 16(8)**,** 472-482.

Stackebrandt, E., and Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today* 6(4)**,** 152-155.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A., Kampfer, P., Maiden, M.C., et al. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* 52(3)**,** 1043-1047.

Stackebrandt, E., and Goebel, B.M. (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 and Evolutionary Microbiology* 44(4)**,** 846-849.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9)**,** 1312-1313.

Stokes, H.W., and Hall, R.M. (1989). A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Molecular Microbiology* 3(12)**,** 1669-1683.

Sutcliffe, I.C., Trujillo, M.E., and Goodfellow, M. (2012). A call to arms for systematists: revitalising the purpose and practises underpinning the description of novel microbial taxa. *Antonie van Leeuwenhoek* 101(1)**,** 13-20.

Swerdlow, H., and Gesteland, R. (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research* 18(6)**,** 1415-1419.

Swerdlow, H., Wu, S.L., Harke, H., and Dovichi, N.J. (1990). Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of Chromatography* 516(1)**,** 61-67.

Szamosi, J.C. (2012). *ISAAC: an improved structural annotation of attC and an initial application thereof.* Master of Science Thesis, McMaster University.

Tarr, C.L., Bopp, C.A., and Farmer, J.J., III (2015). "*Vibrio* and related organisms," in *Manual of Clinical Microbiology,* eds. J.H. Jorgensen, M.A. Pfaller, K.C. Carroll, G. Funke, M.L. Landry, S.S. Richter & D.W. Warnock. 11th ed. (Washington, DC, USA: ASM Press), 762-772.

Tarr, C.L., Patel, J.S., Puhr, N.D., Sowers, E.G., Bopp, C.A., and Strockbine, N.A. (2007). Identification of *Vibrio* isolates by a multiplex PCR assay and *rpoB* sequence determination. *Journal of Clinical Microbiology* 45(1)**,** 134-140.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28(1)**,** 33-36.

Taviani, E., Grim, C.J., Choi, J., Chun, J., Haley, B., Hasan, N.A., et al. (2010). Discovery of novel *Vibrio cholerae* VSP-II genomic islands using comparative genomic analysis. *FEMS Microbiology Letters* 308(2)**,** 130-137.

Taylor, R.K., Miller, V.L., Furlong, D.B., and Mekalanos, J.J. (1987). Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proceedings of the National Academy of Sciences of the United States of America* 84(9)**,** 2833-2837.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* 102(39)**,** 13950-13955.

Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology* 3(9)**,** 711-721.

Thompson, C.C., Amaral, G.R., Campeao, M., Edwards, R.A., Polz, M.F., Dutilh, B.E., et al. (2015). Microbial taxonomy in the post-genomic era: rebuilding from scratch? *Archives of Microbiology* 197(3)**,** 359-370.

Thompson, F.L., Gevers, D., Thompson, C.C., Dawyndt, P., Naser, S., Hoste, B., et al. (2005). Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. *Applied and Environmental Microbiology* 71(9)**,** 5107-5115.

Thyssen, A., and Ollevier, F. (2005). "Genus II. *Photobacterium*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part B: The Gammaproteobacteria,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, M. Goodfellow, F.A. Rainey & K.H. Schleifer. 2nd ed. (New York, NY, USA: Springer), 546-552.

Tindall, B.J. (2008). Confirmation of deposit, but confirmation of what? *International Journal of Systematic and Evolutionary Microbiology* 58(8)**,** 1785-1787.

Tindall, B.J., and Garrity, G.M. (2008). Proposals to clarify how type strains are deposited and made available to the scientific community for the purpose of systematic research. *International Journal of Systematic and Evolutionary Microbiology* 58(8)**,** 1987-1990.

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

Tsubouchi, T., Shimane, Y., Mori, K., Miyazaki, M., Tame, A., Uematsu, K., et al. (2013). *Loktanella cinnabarina* sp. nov., isolated from a deep subseafloor sediment, and emended description of the genus *Loktanella*. *International Journal of Systematic and Evolutionary Microbiology* 63(4)**,** 1390-1395.

Turcatti, G., Romieu, A., Fedurco, M., and Tairi, A.P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research* 36(4)**,** e25.

Uchino, Y., Hirata, A., Yokota, A., and Sugiyama, J. (1998). Reclassification of marine *Agrobacterium* species: proposals of *Stappia stellulata* gen. nov., comb. nov., *Stappia aggregata* sp. nov., nom. rev., *Ruegeria atlantica* gen. nov., comb. nov., *Ruegeria gelatinovora* comb. nov., *Ruegeria algicola* comb. nov., and *Ahrensia kieliense* gen. nov., sp. nov., nom. rev. *Journal of General and Applied Microbiology* 44(3)**,** 201-210.

United States Geological Survey (2010). *Magnitude 7.0 - Haiti region* [Online]. Reston, VA, USA: United States Geological Survey. Available: http://earthquake.usgs.gov/earthquakes/eqinthenews/2010/us2010rja6/us2010rja6.php [Accessed 25 March 2014].

Urbanczyk, H., Ast, J.C., Higgins, M.J., Carson, J., and Dunlap, P.V. (2007). Reclassification of *Vibrio fischeri*, *Vibrio logei*, *Vibrio salmonicida* and *Vibrio wodanis* as *Aliivibrio fischeri* gen. nov., comb. nov., *Aliivibrio logei* comb. nov., *Aliivibrio salmonicida* comb. nov. and *Aliivibrio wodanis* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* 57(12)**,** 2823-2829.

Urdaci, M.C., Marchand, M., Ageron, E., Arcos, J.M., Sesma, B., and Grimont, P.A. (1991). *Vibrio navarrensis* sp. nov., a species from sewage. *International Journal of Systematic Bacteriology* 41(2)**,** 290-294.

Van Trappen, S., Mergaert, J., and Swings, J. (2004). *Loktanella salsilacus* gen. nov., sp. nov., *Loktanella fryxellensis* sp. nov. and *Loktanella vestfoldensis* sp. nov., new members of the *Rhodobacter* group, isolated from microbial mats in Antarctic lakes. *International Journal of Systematic and Evolutionary Microbiology* 54(4)**,** 1263-1269.

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K., and Swings, J. (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiological Reviews* 60(2)**,** 407-438.

Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C., et al. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research* 43(14)**,** 6761-6771.

Vasu, K., and Nagaraja, V. (2013). Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiology and Molecular Biology Reviews* 77(1)**,** 53-72.

Ventosa, A. (2005). "Genus III. *Salinivibrio*," in *Bergey's Manual of Systematic Bacteriology – Volume Two: The Proteobacteria, Part B: The Gammaproteobacteria,* eds. D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, M. Goodfellow, F.A. Rainey & K.H. Schleifer. 2nd ed. (New York, NY, USA: Springer), 552-555.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Vesth, T., Wassenaar, T.M., Hallin, P.F., Snipen, L., Lagesen, K., and Ussery, D.W. (2010). On the origins of a *Vibrio* species. *Microbial Ecology* 59(1)**,** 1-13.

Vezzulli, L., Stauder, M., Grande, C., Pezzati, E., Verheye, H.M., Owens, N.J., et al. (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(4)**,** e0123983.

Wagner-Döbler, I., Ballhausen, B., Berger, M., Brinkhoff, T., Buchholz, I., Bunk, B., et al. (2010). The complete genome sequence of the algal symbiont *Dinoroseobacter shibae*: a hitchhiker's guide to life in the sea. *ISME Journal* 4(1)**,** 61-77.

Wagner-Döbler, I., and Biebl, H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annual Review of Microbiology* 60(1)**,** 255-280.

Waldor, M.K., and Mekalanos, J.J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272(5270)**,** 1910-1914.

Wang, A., and Ash, G.J. (2015). Whole genome phylogeny of *Bacillus* by feature frequency profiles (FFP). *Scientific Reports* 5**,** 13644.

Wang, D., Liu, H., Zheng, S., and Wang, G. (2014). *Paenirhodobacter enshiensis* gen. nov., sp. nov., a non-photosynthetic bacterium isolated from soil, and emended descriptions of the genera *Rhodobacter* and *Haematobacter*. *International Journal of Systematic and Evolutionary Microbiology* 64(2)**,** 551-558.

Wang, G., and Maier, R.J. (2004). An NADPH quinone reductase of *Helicobacter pylori* plays an important role in oxidative stress resistance and host colonization. *Infection and Immunity* 72(3)**,** 1391-1396.

Wang, L.T., Lee, F.L., Tai, C.J., and Kasai, H. (2007). Comparison of *gyrB* gene sequences, 16S rRNA gene sequences and DNA-DNA hybridization in the *Bacillus subtilis* group. *International Journal of Systematic and Evolutionary Microbiology* 57(8)**,** 1846-1850.

Ward, D.M., Weller, R., and Bateson, M.M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* 345(6270)**,** 63-65.

Watve, S.S., Thomas, J., and Hammer, B.K. (2015). CytR is a global positive regulator of competence, type VI secretion, and chitinases in *Vibrio cholerae*. *PLoS One* 10(9)**,** e0138834.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., et al. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic Bacteriology* 37(4)**,** 463-464.

Whitman, W.B. (2011). Intent of the nomenclatural Code and recommendations about naming new species based on genomic sequences. *Bulletin of BISMiS* 2(2)**,** 135-139.

Whitman, W.B. (2015). *Bergey's Manual of Systematics of Archaea and Bacteria.* Hoboken, NJ, USA: John Wiley & Sons, Inc.

Whitman, W.B., Woyke, T., Klenk, H.P., Zhou, Y., Lilburn, T.G., Beck, B.J., et al. (2015). Genomic Encyclopedia of Bacterial and Archaeal Type Strains, Phase III: the genomes of soil and plant-associated and newly described type strains. *Standards in Genomic Sciences* 10(5)**,** 26.

Wilson, D.M., and Ajl, S. (1957). Metabolism of L-rhamnose by *Escherichia coli*. I. L-rhamnose isomerase. *Journal of Bacteriology* 73(3)**,** 410-414.

Woese, C.R. (1987). Bacterial evolution. *Microbiological Reviews* 51(2)**,** 221-271.

World Health Organization (2016). Cholera, 2015. *Weekly Epidemiological Record* 91(38)**,** 433-440.

World Health Organization (2017). *Daily epidemiology bulletin, 22 October 2017* [Online]. Sana'a, Yemen: World Health Organization. Available: http://www.emro.who.int/images/stories/yemen/Yemen_Cholera_Response_-_Daily_Epidemiological_Update_-_2017-10-22.pdf?ua=1 [Accessed October 24 2017].

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., et al. (2009). A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea*. *Nature* 462(7276)**,** 1056-1060.

Xu, H.S., Roberts, N., Singleton, F.L., Attwell, R.W., Grimes, D.J., and Colwell, R.R. (1982). Survival and viability of nonculturable *Escherichia coli* and *Vibrio cholerae* in the estuarine and marine environment. *Microbial Ecology* 8(4)**,** 313-323.

Xu, Y., Li, Q., Tian, R., Lai, Q., and Zhang, Y. (2015). *Pseudovibrio hongkongensis* sp. nov., isolated from a marine flatworm. *Antonie van Leeuwenhoek* 108(1)**,** 127-132.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.H., et al. (2008). The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology* 31(4)**,** 241-250.

Yeoh, Y.K., Sekiguchi, Y., Parks, D.H., and Hugenholtz, P. (2016). Comparative genomics of candidate phylum TM6 suggests that parasitism is widespread and ancestral in this lineage. *Molecular Biology and Evolution* 33(4)**,** 915-927.

Yoon, S.H., Ha, S.M., Lim, J., Kwon, S., and Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek* 110(10)**,** 1281-1286.

Zeigler, D.R. (2003). Gene sequences useful for predicting relatedness of whole genomes in bacteria. *International Journal of Systematic and Evolutionary Microbiology* 53(6)**,** 1893-1900.

Zhang, G., Haroon, M.F., Zhang, R., Dong, X., Wang, D., Liu, Y., et al. (2017). *Ruegeria profundi* sp. nov. and *Ruegeria marisrubri* sp. nov., isolated from the brine-seawater interface at Erba Deep in the Red Sea. *International Journal of Systematic and Evolutionary Microbiology* 10.1099/ijsem.0.002344.

Zhang, J., and Yang, J.R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* 16(7)**,** 409-420.

Zhang, Y., Li, Q., Tian, R., Lai, Q., and Xu, Y. (2016). *Pseudovibrio stylochi* sp. nov., isolated from a marine flatworm. *International Journal of Systematic and Evolutionary Microbiology* 66(5)**,** 2025-2029.

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., et al. (2014). PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* 30(9)**,** 1297-1299.

Zhi, X.Y., Zhao, W., Li, W.J., and Zhao, G.P. (2012). Prokaryotic systematics in the genomics era. *Antonie van Leeuwenhoek* 101(1)**,** 21-34.

APPENDICES

Appendix A: List of Publications

This is the complete list of authored and co-authored publications in the course of my doctorate program (2012–2017). The list includes manuscripts that have been published, in press, or in preparation (i.e., full draft of manuscript available) as of the writing of this thesis. Individual author contributions are indicated after each publication.

A.1. Published

1. Liang, K., **Orata, F.D.**, Winkjer, N.S., Rowe, L.A., Tarr, C.L., and Boucher, Y. (2017). Complete genome sequence of *Vibrio* sp. strain 2521-89, a close relative of *Vibrio cholerae* isolated from lake water in New Mexico, USA. *Genome Announcements* 5(35)**,** e00905-17.
   - K.L. and F.D.O. performed bioinformatics analyses and wrote the manuscript; N.S.W., L.A.R., and C.L.T. sequenced and assembled the *Vibrio* sp. 2521-89 genome; K.L., F.D.O., and Y.B. wrote the manuscript; and Y.B. supervised the project.

2. Labbate, M., **Orata, F.D.**, Petty, N.K., Jayatilleke, N.D., King, W.L., Kirchberger, P.C., Allen, C., Mann, G., Mutreja, A., Thomson, N.R., Boucher, Y., and Charles, I.G. (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**,** 36891.
   - M.L. designed the study and wrote the manuscript; A.M. and N.R.T. sequenced and assembled 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; and M.L., G.M., and I.G.C. supervised the project.

3. **Orata, F.D.**, Xu, Y., Gladney, L.M., Rishishwar, L., Case, R.J., Boucher, Y., Jordan, I.K.*, Tarr, C.L.* (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(10)**,** 4148-4155.†
   - F.D.O. and Y.B. designed the study; L.M.G. and C.L.T. provided the *V. cidicii* and *V. navarrensis* isolates and genome sequences; F.D.O. and L.R. performed bioinformatic analyses; Y.X., L.M.G., and R.J.C. performed biochemical tests; F.D.O. and Y.B. wrote the manuscript; and Y.B., I.K.J., and C.L.T. supervised the project.

4. **Orata, F.D.***, Rosana, A.R.R.*, Xu, Y., Simkus, D.N., Bramucci, A.R., Boucher, Y., and Case, R.J. (2016). Draft genome sequences of four bacterial strains isolated from a polymicrobial culture of naked (N-type) *Emiliania huxleyi* CCMP1516. *Genome Announcements* 4(4)**,** e00674-16.

5. Rosana, A.R.R.*, **Orata, F.D.***, Xu, Y., Simkus, D.N., Bramucci, A.R., Boucher, Y., and Case, R.J. (2016). Draft genome sequences of seven bacterial strains isolated from a polymicrobial culture of coccolith-bearing (C-type) *Emiliania huxleyi* M217. *Genome Announcements* 4(4)**,** e00673-16.
   - For both 4 and 5 – F.D.O. performed whole-genome sequencing and assembly; F.D.O. and A.R.R.R. analyzed genome sequence data; A.R.R.R., Y.X., D.N.S., and A.R.B. isolated the strains and performed 16S rRNA sequence analysis; and Y.B. and R.J.C. supervised the project and wrote the manuscripts.

6. Kirchberger, P.C., **Orata, F.D.**, Barlow, E.J., Kauffman, K.M., Case, R.J., Polz, M.F., and Boucher, Y. (2016). A small number of phylogenetically distinct clonal complexes dominate a coastal *Vibrio cholerae* population. *Applied and Environmental Microbiology* 82(18)**,** 5576-5586.
   - M.F.P. and Y.B. designed the study; Y.B. and K.M.K. performed sampling; F.D.O. performed whole-genome sequencing and core genome analysis; E.J.B. provided multiple scripts used in data analysis; P.C.K performed all typing of strains, biochemical assays, and data analyses; P.C.K., R.J.C., and Y.B. wrote the manuscript; and Y.B. supervised the project.

7. Boucher, Y., **Orata, F.D.**, and 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. *Frontiers in Microbiology* 6**,** 1120.

- Y.B. wrote the manuscript; F.D.O. revised the manuscript and created the figure; and M.A. contributed the original hypothesis idea and revised the manuscript.

8. **Orata, F.D.**, Kirchberger, P.C., Méheust, R., Barlow, E.J., Tarr, C.L., and 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(10)**,** 2941-2954.†
   - F.D.O. and Y.B. designed the study and wrote the manuscript; Y.B. performed sampling; C.L.T. provided the clinical *V. metoecus* genome sequences; F.D.O. performed whole-genome sequencing and analyses; P.C.K. performed phylogenetic analysis; R.M. quantified interspecies gene transfers; E.J.B. provided multiple scripts used in data analysis; and Y.B. supervised the project.

9. **Orata, F.D.**, Keim, P.S., and Boucher, Y. (2014). The 2010 cholera outbreak in Haiti: how science solved a controversy. *PLoS Pathogens* 10(4)**,** e1003967.†
   - F.D.O. and Y.B. wrote the manuscript, with revisions from P.S.K.; and F.D.O. created the figures.

10. Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., Guo, Y., Wang, S., Paxinos, E.E., **Orata, F.**, Gladney, L.M., Stroika, S., Folster, J.P., Rowe, L., Freeman, M.M., Knox, N., Frace, M., Boncy, J., Graham, M., Hammer, B.K., Boucher, Y., Bashir, A., Hanage, W.P., Van Domselaar, G., and Tarr, C.L. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4(4)**,** e00398-13.
    - L.S.K. and C.L.T. designed the study and wrote the manuscript; J.Bo. provided the *V. cholerae* Haitian strains; L.S.K., A.P., S.T., M.A.T., Y.G., S.W., E.E.P., L.M.G., S.S., J.P.F., L.R., M.M.F., N.K., M.F., M.G., G.V.D. performed whole-genome sequencing and analyses; J.Be., A.B., and W.P.H. provided additional bioinformatic analyses; F.O. and Y.B. performed integron sequencing; E.S.A. and B.K.H. performed the natural transformation and quorum-sensing assays; and C.L.T. supervised the project.

A.2. In Press

11. Daas, M.S.*, Acedo, J.Z.*, Rosana, A.R.R.*, **Orata, F.D.**, Reiz, B., Zheng, J., Nateche, F., Case, R.J., Kebbouche-Gana, S., and Vederas, J.C. (2017). *Bacillus amyloliquefaciens* ssp. *plantarum* F11 isolated from an Algerian salty lake as a source of biosurfactants and bioactive lipopeptides. *FEMS Microbiology Letters*, fnx248.
    - M.S.D., J.Z.A., A.R.R.R., B.R., J.Z., and F.N. conducted all experiments; A.R.R.R. performed whole-genome sequencing; A.R.R.R. and F.D.O performed genome assembly; F.D.O. and R.J.C. performed bioinformatic analyses; M.S.D., J.Z.A., A.R.R.R., and J.C.V. wrote the manuscript; and S.K. and J.C.V. supervised the project.

A.3. In Preparation

12. **Orata, F.D.**, Liang, K., Boucher, Y., and Case, R.J. Whole genome-based taxonomy of *Rhodobacteraceae* and *Stappiaceae* fam. nov. in the order *Rhodobacterales*.†
    - F.D.O., Y.B., and R.J.C. designed the study and wrote the manuscript; F.D.O. and K.L. performed bioinformatic analyses; and Y.B. and R.J.C. supervised the project.

13. **Orata, F.D.**, Liang, K., Nasreen, T., Hussain, N.A.S., and Boucher, Y. Differences in abundance and seasonal patterns of *Vibrio cholerae* and *Vibrio metoecus* lead to a directional bias in horizontal gene transfer.†
    - F.D.O. and Y.B. designed the study and wrote the manuscript; Y.B. performed sampling; F.D.O. performed whole-genome sequencing and assembly; F.D.O. and K.L. performed bioinformatic analyses; T.N. performed qPCR; F.D.O. and N.A.S.H. performed the natural transformation assay; and Y.B. supervised the project.

14. Kirchberger, P.C., **Orata, F.D.**, Nasreen, T., Kauffman, K.M., Case, R.J., Polz, M.F., and Boucher, Y. 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.

- M.F.P. and Y.B. designed the study; K.M.K. and Y.B. performed sampling; P.C.K. developed all methods and performed *viuB* sequencing and analyses; F.D.O. performed whole-genome sequencing and analyses; T.N. performed qPCR; P.C.K., R.J.C., and Y.B. wrote the manuscript; and Y.B. supervised the project.

15. Bramucci, A.R.*, Labeeuw, L.*, **Orata, F.D.**, Case, R.J. *Phaeobacter inhibens* shapes the life history of its algal host *Emiliania huxleyi*.
    - A.R.B., L.L., and R.J.C. designed the study and wrote the manuscript; A.R.B. and L.L. conducted the experiments; F.D.O. performed whole-genome sequencing and isolate identification; and R.J.C. supervised the project.

---

*These authors contributed equally to this work.
†A version of this publication is included as a chapter in this thesis.

Appendix B: Supplementary Data for Chapter 2

Table B.1. Results of all phenotypic tests for the *V. cidicii* sp. nov. and *V. navarrensis* isolates conducted in this study.

+, Growth/positive test result; −, no growth/negative test result; v, variable results between tests.

Phenotypic test/substrate tested	*Vibrio cidicii* sp. nov.				*Vibrio navarrensis*		
	LMG 29267^T	1048-83	2423-01	2538-88	LMG 15976^T	0053-83	08-2462
Indole production	+	+	+	+	+	+	+
Methyl red	+	+	+	+	+	+	+
Voges-Proskauer	–	–	–	–	–	–	–
Citrate (Simmons agar)	–	+	–	–	+	+	+
H_2S production (peptone iron agar)	–	–	–	–	–	–	–
H_2S production (triple sugar iron agar)	–	–	–	–	–	–	–
Urea hydrolysis	–	–	–	–	–	–	–
Phenylalanine deaminase	+	+	+	+	+	+	+
Arginine dihydrolase	–	–	–	–	–	–	–
Lysine decarboxylase (Moeller medium)	–	–	–	–	–	–	–
Ornithine decarboxylase (Moeller medium)	–	–	–	–	–	–	–
Swimming motility (37°C)	+	+	+	+	+	+	+
Gelatin hydrolysis	+	+	+	+	+	+	+
Malonate utilization	–	–	–	–	–	+	–
Esculin hydrolysis	+	+	+	+	+	+	+
Reduction of nitrate to nitrite	+	+	+	+	+	+	+
Oxidase	+	+	+	+	+	+	+
DNase (25°C)	+	+	+	+	+	+	+
o-Nitrophenyl β-D-galactopyranoside (ONPG)	–	+	–	–	+	–	–
p-Nitrophenyl β-D-galactopyranoside (PNPG)	–	+	–	–	+	–	–
Tyrosine clearing	+	–	+	+	+	+	+
Swarming motility (marine agar, 25°C)	–	+	–	+	–	–	–
D-Glucose (acid production)	+	+	+	+	+	+	+
D-Glucose (gas production)	–	–	–	–	–	–	–
Acid production from:							
D-Adonitol	–	–	–	–	–	–	–
L-Arabinose	–	–	–	–	–	–	–
D-Arabitol	–	–	–	–	–	–	–
Cellobiose	+	+	+	+	+	+	+
Dulcitol	–	–	–	–	–	–	–
Erythritol	–	–	–	–	–	–	–
D-Galactose	–	+	–	–	–	–	–
Glycerol	–	–	–	–	–	–	–
myo-Inositol	–	–	–	–	–	–	–
Lactose	–	–	–	–	–	–	–
Maltose	+	+	+	+	+	+	+
Mannitol	+	+	+	+	+	+	+
Mannose	+	+	+	+	+	+	+
Melibiose	–	–	–	–	–	–	–
α-Methyl-D-glucoside	+	–	–	–	–	–	–
Mucic acid	–	–	–	–	–	–	–
Raffinose	–	–	–	–	–	–	–
L-Rhamnose	+	+	+	+	–	–	–

Phenotypic test/substrate tested	*Vibrio cidicii* sp. nov.				*Vibrio navarrensis*		
	LMG 29267^T	1048-83	2423-01	2538-88	LMG 15976^T	0053-83	08-2462
Acid production from:							
Salicin	–	–	+	+	–	–	–
D-Sorbitol	+	+	–	+	–	–	+
Sucrose	+	+	+	+	+	+	+
Trehalose	+	+	+	+	+	+	+
D-Xylose	–	–	–	–	–	–	–
Assimilation of:							
Acetic acid	+	+	+	+	+	+	+
Acetoacetic acid	+	+	+	+	–	+	+
N-Acetyl-D-glucosamine	+	+	+	+	+	+	+
N-Acetyl-β-D-mannosamine	–	–	+	–	–	+	–
Adenosine	+	+	+	+	+	+	+
Adipic acid	–	–	–	–	–	–	–
D-Adonitol	–	+	+	–	–	+	–
D-Alanine	+	+	+	+	+	+	+
L-Alanine	+	+	+	+	+	+	+
L-Alanyl-glycine	+	+	+	+	+	+	+
2-Aminoethanol	–	–	–	–	–	–	–
L-Arabinose*	v	v	v	v	v	v	v
L-Asparagine	+	+	+	+	+	+	+
D-Aspartic acid	–	–	–	–	–	–	–
L-Aspartic acid	+	+	+	+	+	+	+
Bromosuccinic acid	+	+	+	+	+	+	+
Capric acid	–	–	–	–	–	–	–
Cellobiose	+	+	+	+	+	+	+
Citric acid	+	+	+	+	+	+	+
2'-Deoxyadenosine	+	+	+	+	+	+	+
Dulcitol	–	+	–	–	–	–	–
Formic acid	+	+	+	+	+	+	+
D-Fructose	+	+	+	+	+	+	+
Fructose-6-phosphate	+	+	+	+	+	+	+
L-Fucose	–	–	–	–	–	–	–
Fumaric acid	+	+	+	+	+	+	+
L-Galactonic acid-γ-lactone	–	–	–	–	–	–	–
D-Galactonic acid-γ-lactone	–	+	–	–	–	–	–
D-Galactose	–	+	–	–	–	–	–
D-Galacturonic acid	–	–	–	–	–	+	–
D-Gluconic acid	+	+	+	+	+	+	+
Glucuronamide	–	–	–	–	–	–	–
D-Glucuronic acid	–	–	+	–	–	–	–
D-Glucosaminic acid	–	–	–	–	–	+	–
D-Glucose	+	+	+	+	+	+	+
Glucose-1-phosphate	+	+	+	+	+	+	+
Glucose-6-phosphate	+	+	+	+	+	+	+
L-Glutamic acid	+	+	+	+	+	+	+
L-Glutamine	+	+	+	+	+	+	+
Glycerol	+	+	+	+	+	+	+
D,L-α-Glycerol phosphate	+	+	+	+	+	+	+
Glycolic acid	–	+	–	–	–	+	–
Glycyl-L-aspartic acid	+	+	+	+	+	+	+
Glycyl-L-glutamic acid	+	+	+	+	+	+	+
Glycyl-L-proline	+	+	+	+	+	+	+

Phenotypic test/substrate tested	*Vibrio cidicii* sp. nov.				*Vibrio navarrensis*		
	LMG 29267^T	1048-83	2423-01	2538-88	LMG 15976^T	0053-83	08-2462
Assimilation of:							
Glyoxylic acid	–	–	–	–	–	–	–
α-Hydroxy butyric acid	+	+	+	+	+	+	+
α-Hydroxy glutaric acid-γ-lactone	+	+	+	+	+	+	+
m-Hydroxyphenylacetic acid	–	–	–	–	–	–	–
p-Hydroxyphenylacetic acid	–	–	–	–	–	–	–
Inosine	+	+	+	+	+	+	+
myo-Inositol	–	–	–	–	–	–	–
α-Ketobutyric acid	+	+	+	+	+	+	+
α-Ketoglutaric acid	+	+	+	+	+	+	+
L-Lactic acid	+	+	+	+	+	+	+
Lactose	–	+	–	–	–	+	–
Lactulose	–	+	–	–	–	+	–
L-Lyxose	+	+	+	+	+	+	+
D-Malic acid	+	+	+	+	+	+	+
L-Malic acid	+	+	+	+	+	+	+
D,L-Malic acid	+	+	+	+	+	+	+
D-Maltose	+	+	+	+	v[†]	+	+
Maltotriose	+	+	+	+	+	+	+
D-Mannitol	+	+	+	+	+	+	+
D-Mannose	+	+	+	+	v[†]	–	+
Melibiose	–	+	–	–	–	–	–
Methyl pyruvate	+	+	+	+	+	+	+
α-Methyl-D-galactoside	–	–	–	–	–	+	–
β-Methyl-D-glucoside	+	+	+	+	+	+	+
Monomethyl succinate	+	+	+	+	+	+	+
Mucic acid	–	+	–	–	–	+	–
Phenylacetic acid	–	–	–	–	–	–	–
Phenylethylamine	–	–	–	–	–	–	–
Potassium gluconate	+	+	+	+	+	+	+
L-Proline	+	+	+	+	+	+	+
1,2-Propanediol	–	–	–	–	–	–	–
Propionic acid	+	+	+	+	+	+	+
D-Psicose	+	+	+	+	+	+	+
Pyruvic acid	+	+	+	+	+	+	+
L-Rhamnose	+	+	+	+	–	–	–
D-Ribose	+	+	+	+	+	+	+
D-Saccharic acid	–	+	–	–	–	–	–
D-Serine	–	–	–	–	–	–	–
L-Serine	+	+	+	+	+	+	+
D-Sorbitol	+	+	–	+	–	–	+
Succinic acid	+	+	+	+	+	+	+
Sucrose	+	+	+	+	+	+	+
m-Tartaric acid	–	–	–	–	–	–	–
D-Threonine	–	–	–	–	–	–	–
L-Threonine	+	+	+	+	+	+	+
Thymidine	+	+	+	+	+	+	+
Trehalose	+	+	+	+	+	+	+
Tricarballylic acid	–	–	–	–	–	–	–
Trisodium citrate	–	–	–	–	–	–	–
Tween 20	+	–	–	+	+	+	+
Tween 40	+	+	+	+	+	+	+
Tween 80	+	+	+	+	+	+	+

Phenotypic test/substrate tested	*Vibrio cidicii* sp. nov.				*Vibrio navarrensis*		
	LMG 29267^T	1048-83	2423-01	2538-88	LMG 15976^T	0053-83	08-2462
Assimilation of:							
Tyramine	–	–	–	–	–	–	–
Uridine	+	+	+	+	+	+	+
D-Xylose	+	+	+	+	+	+	+
Growth in TSB (at 30°C) with:							
0% NaCl	+	+	+	+	+	+	+
1.5% NaCl	+	+	+	+	+	+	+
6.5% NaCl	+	+	+	+	+	+	+
8% NaCl	+	–	+	+	–	+	+
10% NaCl	–	–	–	–	–	–	–
Growth in TSB (with 1.5% NaCl) at:							
4°C	–	–	–	–	–	–	–
30°C	+	+	+	+	+	+	+
36°C	+	+	+	+	+	+	+
37°C	+	+	+	+	+	+	+
40°C	+	+	+	+	+	+	+
45°C	–	–	–	–	–	+	+

*For all isolates: assimilation of L-arabinose positive with PM1, negative with API 20 NE.
†For *V. navarrensis* LMG 15976^T: assimilation of D-maltose and D-mannose positive with PM1, negative with API 20 NE.

Table B.2. List of whole-genome and housekeeping gene sequences used in this study, including that of representative *Vibrio* species closely related to *V. cidicii* sp. nov. and *V. navarrensis*.

For *V. cidicii* sp. nov. and *V. navarrensis*, the *pyrH*, *recA*, and *rpoA* genes were sequenced previously (Gladney and Tarr, 2014), whereas the *rpoB* genes were sequenced in this study. For the other *Vibrio* species, the housekeeping gene sequences were obtained from their annotated whole-genome sequences.

Species and strain	Accession number/s	*pyrH* accession number	*recA* accession number	*rpoA* accession number	*rpoB* accession number
Vibrio cidicii sp. nov. LMG 29267^T	LOMK00000000	KJ807114	KJ807134	KJ807154	KU593643
Vibrio cidicii 1048-83	LOBP00000000	KJ807116	KJ807136	KJ807156	KU593646
Vibrio cidicii 2423-01	LOBQ00000000	KJ807125	KJ807145	KJ807165	KU593645
Vibrio cidicii 2538-88	LOBR00000000	KJ807122	KJ807142	KJ807162	KU593644
Vibrio navarrensis LMG 15976^T	JMCG00000000	KJ807123	KJ807143	KJ807163	KU593635
Vibrio navarrensis 2232	JMCH00000000	KJ807124	KJ807144	KJ807164	KU593636
Vibrio navarrensis 0053-83	JMCF00000000	KJ807115	KJ807135	KJ807155	KU593629
Vibrio navarrensis 08-2462	JMCI00000000	KJ807127	KJ807147	KJ807167	KU593637
Vibrio cholerae ATCC 14035^T	JHXR00000000				
Vibrio fluvialis LMG 7894^T	CP014034; CP014035				
Vibrio furnissii LMG 7910^T	ACZP00000000				
Vibrio metoecus LMG 27764^T	JJMN00000000				
Vibrio metschnikovii LMG 11664^T	ACZO00000000				
Vibrio mimicus LMG 7896^T	AOMO00000000				
Vibrio proteolyticus LMG 3772^T	BATJ00000000				
Vibrio vulnificus LMG 13545^T	AMQV00000000				
Vibrio vulnificus CMCP6	AE016795; AE016796				
Vibrio vulnificus YJ016	BA000037; BA000038				

Table B.3. Predicted functions of genes found in *V. cidicii* sp. nov. but not in *V. navarrensis*.

Functions are based on the COG database and/or BLASTP homology search. For hits with varying description between COG and BLASTP, the more specific description is used. Highlighted in gray are four genes with functions attributed to L-rhamnose transport and metabolism (Wilson and Ajl, 1957; Sawada and Takagi, 1964; Ryu et al., 2004). NA, not applicable – for a function that can not be attributed to a COG number; class and class description are still assigned based on function. A total of 67 putative unique genes were found in *V. cidicii* sp. nov. Not included in this table are genes that have general function predictions only (n = 8), encode hypothetical proteins (n = 15), or have no known hits (n = 16).

COG number	Description	Class/es	Class Description/s
COG0247	Fe-S oxidoreductase	C	Energy production and conversion
COG1150	Heterodisulfide reductase, subunit C	C	Energy production and conversion
NA	β-mannosidase	G	Carbohydrate transport and metabolism
COG4806	L-rhamnose isomerase	G	Carbohydrate transport and metabolism
COG3254	L-rhamnose mutarotase	G	Carbohydrate transport and metabolism
NA	L-rhamnose-proton symporter	G	Carbohydrate transport and metabolism
COG0235	Rhamnulose-1-phosphate aldolase	G	Carbohydrate transport and metabolism
COG0304	β-ketoacyl-ACP synthase II	IQ	Lipid transport and metabolism; Secondary metabolites biosynthesis, transport, and catabolism
COG0603	7-cyano-7-deazaguanine synthase QueC	J	Translation, ribosomal structure, and biogenesis
COG2207	AraC family transcriptional regulator	K	Transcription
COG0454	Histone acetyltransferase	K	Transcription
COG0583	LysR family transcriptional regulator	K	Transcription
COG1309	TetR family transcriptional regulator	K	Transcription
NA	DNA repair protein	L	Replication, recombination, and repair
COG4974	Site-specific recombinase XerD	L	Replication, recombination, and repair
NA	Transposase	L	Replication, recombination, and repair
COG0845	Hemolysin secretion protein D	M	Cell wall/membrane/envelope biogenesis
COG0668	Mechanosensitive ion channel protein MscS	M	Cell wall/membrane/envelope biogenesis
COG0840	Methyl-accepting chemotaxis protein	NT	Cell motility; Signal transduction mechanisms
COG1226	Potassium channel protein	P	Inorganic ion transport and metabolism
COG2897	Rhodanese-related sulfurtransferase	P	Inorganic ion transport and metabolism
COG2015	Alkyl sulfatase	Q	Secondary metabolites biosynthesis, transport, and catabolism
COG1335	Isochorismatase	Q	Secondary metabolites biosynthesis, transport, and catabolism
COG2199	FOG: GGDEF domain	T	Signal transduction mechanisms
COG2206	HD-GYP domain	T	Signal transduction mechanisms
COG4753	Response regulator containing CheY-like receiver domain and AraC-type DNA-binding domain	T	Signal transduction mechanisms
NA	Fluoroquinolone resistance protein	V	Defense mechanisms
COG1132	Multidrug ABC transporter	V	Defense mechanisms

Table B.4. ANI from the pairwise comparisons of whole-genome sequences among *V. cidicii* sp. nov. and *V. navarrensis* isolates.

Strains of *V. cidicii* sp. nov.: LMG 29267^T, 1048-83, 2423-01, and 2538-88. Strains of *V. navarrensis*: LMG 15976^T, 2232, 0053-83, and 08-2462.

	LMG 29267^T	**1048-83**	**2423-01**	**2538-88**	**LMG 15976^T**	**2232**	**0053-83**	**08-2462**
LMG 29267^T		98.18	98.13	98.08	95.55	95.56	95.76	95.75
1048-83	98.18		98.34	98.36	95.42	95.47	95.66	95.63
2423-01	98.13	98.34		98.23	95.50	95.52	95.74	95.69
2538-88	98.08	98.36	98.23		95.40	95.44	95.66	95.62
LMG 15976^T	95.55	95.42	95.50	95.40		98.74	97.36	97.40
2232	95.56	95.47	95.52	95.44	98.74		97.36	97.41
0053-83	95.76	95.66	95.74	95.66	97.36	97.36		98.77
08-2462	95.75	95.63	95.69	95.62	97.40	97.41	98.77	

Table B.5. Percent DDH from the pairwise comparisons of whole-genome sequences among *V. cidicii* sp. nov. and *V. navarrensis* isolates.

Strains of *V. cidicii* sp. nov.: LMG 29267^T, 1048-83, 2423-01, and 2538-88. Strains of *V. navarrensis*: LMG 15976^T, 2232, 0053-83, and 08-2462. Confidence interval (95%) is indicated for each DDH value.

	LMG 29267^T	**1048-83**	**2423-01**	**2538-88**	**LMG 15976^T**	**2232**	**0053-83**	**08-2462**
LMG 29267^T		82.8 ± 2.67	82.7 ± 2.67	82.3 ± 2.69	62.8 ± 2.86	63.1 ± 2.86	64.3 ± 2.88	64.1 ± 2.87
1048-83	82.8 ± 2.67		84.6 ± 2.56	84.7 ± 2.56	62.2 ± 2.85	62.5 ± 2.85	63.6 ± 2.87	63.5 ± 2.87
2423-01	82.7 ± 2.67	84.6 ± 2.56		84.0 ± 2.60	62.5 ± 2.85	63.0 ± 2.86	64.2 ± 2.88	63.9 ± 2.87
2538-88	82.3 ± 2.69	84.7 ± 2.56	84.0 ± 2.60		61.9 ± 2.84	62.4 ± 2.85	63.8 ± 2.87	63.5 ± 2.87
LMG 15976^T	62.8 ± 2.86	62.2 ± 2.85	62.5 ± 2.85	61.9 ± 2.84		88.4 ± 2.27	75.8 ± 2.89	76.2 ± 2.88
2232	63.1 ± 2.86	62.5 ± 2.85	63.0 ± 2.86	62.4 ± 2.85	88.4 ± 2.27		76.1 ± 2.88	76.4 ± 2.88
0053-83	64.3 ± 2.88	63.6 ± 2.87	64.2 ± 2.88	63.8 ± 2.87	75.8 ± 2.89	76.1 ± 2.88		88.7 ± 2.24
08-2462	64.1 ± 2.87	63.5 ± 2.87	63.9 ± 2.87	63.5 ± 2.87	76.2 ± 2.88	76.4 ± 2.88	88.7 ± 2.24	

Table B.6. Patristic distances calculated from the MLSA tree from the concatenated alignment of partial DNA sequences of four protein-coding housekeeping genes (*pyrH*, *recA*, *rpoA*, and *rpoB*).

Average distances are shown for the *V. cidicii* sp. nov., *V. navarrensis*, and *V. vulnificus* isolates. Species: 1, *V. cidicii* sp. nov.; 2, *V. navarrensis*; 3, *V. vulnificus*; 4, *V. cholerae*; 5, *V. fluvialis*; 6, *V. furnissii*; 7, *V. metoecus*; 8, *V. metschnikovii*; 9, *V. mimicus*; 10, *V. proteolyticus*.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Vibrio cidicii sp. nov.	0.005	0.066	0.192	0.282	0.255	0.254	0.284	0.380	0.263	0.238
Vibrio navarrensis	0.066	0.007	0.186	0.277	0.250	0.248	0.278	0.374	0.257	0.232
Vibrio vulnificus	0.192	0.186	0.038	0.284	0.257	0.256	0.286	0.382	0.265	0.240
Vibrio cholerae	0.282	0.277	0.284		0.191	0.189	0.063	0.315	0.077	0.247
Vibrio fluvialis	0.255	0.250	0.257	0.191		0.058	0.192	0.266	0.171	0.220
Vibrio furnissii	0.254	0.248	0.256	0.189	0.058		0.191	0.264	0.170	0.218
Vibrio metoecus	0.284	0.278	0.286	0.063	0.192	0.191		0.317	0.079	0.248
Vibrio metschnikovii	0.380	0.374	0.382	0.315	0.266	0.264	0.317		0.296	0.344
Vibrio mimicus	0.263	0.257	0.265	0.077	0.171	0.170	0.079	0.296		0.227
Vibrio proteolyticus	0.238	0.232	0.240	0.247	0.220	0.218	0.248	0.344	0.227	

Table B.7. Pairwise species distances from the comparisons of whole-genome sequences used in this study.

The values are percent nucleotide distances (= 100 − ANI). Average distances are shown for the *V. cidicii* sp. nov., *V. navarrensis*, and *V. vulnificus* isolates. Species: 1, *V. cidicii* sp. nov.; 2, *V. navarrensis*; 3, *V. vulnificus*; 4, *V. cholerae*; 5, *V. fluvialis*; 6, *V. furnissii*; 7, *V. metoecus*; 8, *V. metschnikovii*; 9, *V. mimicus*; 10, *V. proteolyticus*.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Vibrio cidicii sp. nov.	1.780	4.414	14.896	14.873	14.660	14.933	14.760	15.038	15.018	15.613
Vibrio navarrensis	4.414	2.160	15.476	15.153	14.558	15.060	15.120	14.718	15.418	15.755
Vibrio vulnificus	14.896	15.476	3.567	15.730	15.293	15.493	15.827	15.373	15.543	15.633
Vibrio cholerae	14.873	15.153	15.730		16.320	16.460	11.430	16.180	12.820	16.640
Vibrio fluvialis	14.660	14.558	15.293	16.320		12.200	16.140	16.260	16.000	15.650
Vibrio furnissii	14.933	15.060	15.493	16.460	12.200		16.330	16.370	16.180	15.830
Vibrio metoecus	14.760	15.120	15.827	11.430	16.140	16.330		16.280	11.840	16.320
Vibrio metschnikovii	15.038	14.718	15.373	16.180	16.260	16.370	16.280		16.140	15.850
Vibrio mimicus	15.018	15.418	15.543	12.820	16.000	16.180	11.840	16.140		16.020
Vibrio proteolyticus	15.613	15.755	15.633	16.640	15.650	15.830	16.320	15.850	16.020	

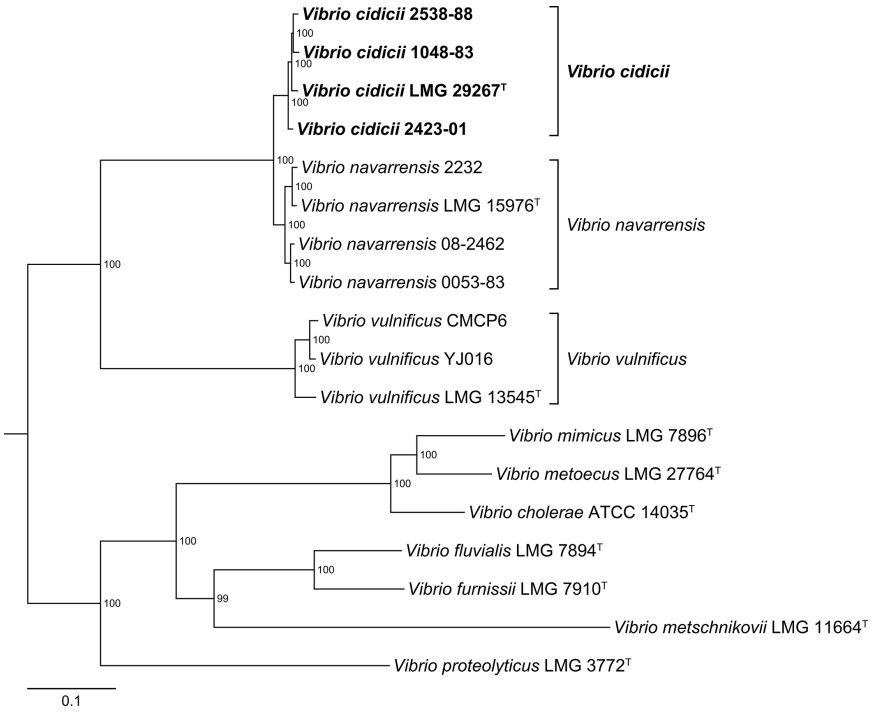

Figure B.1. The phylogenetic relationship of *V. cidicii* sp. nov. and its closest relatives based on their core genome.

The tree was constructed from the concatenated alignment of 586 single-copy, protein-coding core genes with a total length of 446,032 bp. Bootstrap support is indicated on the nodes. Bar, 0.1 nucleotide substitutions per site.

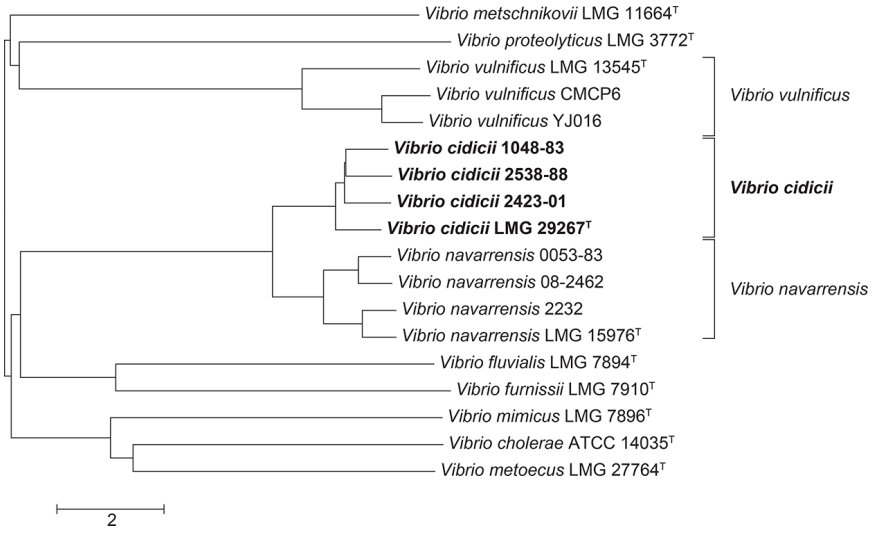

Figure B.2. The phylogenetic relationship of *V. cidicii* sp. nov. and its closest relatives based on average nucleotide distances.

The tree was constructed using the pairwise species distances (= 100 – ANI) from the comparisons of whole-genome sequences of these isolates. Bar, 2% nucleotide distance.

References

Gladney, L.M., and Tarr, C.L. (2014). Molecular and phenotypic characterization of *Vibrio navarrensis* isolates associated with human illness. *Journal of Clinical Microbiology* 52(11), 4070-4074.

Ryu, K.S., Kim, C., Kim, I., Yoo, S., Choi, B.S., and Park, C. (2004). NMR application probes a novel and ubiquitous family of enzymes that alter monosaccharide configuration. *Journal of Biological Chemistry* 279(24), 25544-25548.

Sawada, H., and Takagi, Y. (1964). The metabolism of L-rhamnose in *Escherichia coli*: III. L-rhamnulose-phosphate aldolase. *Biochimica et Biophysica Acta* 92(1), 26-32.

Wilson, D.M., and Ajl, S. (1957). Metabolism of L-rhamnose by *Escherichia coli*. I. L-rhamnose isomerase. *Journal of Bacteriology* 73(3), 410-414.

Appendix C: Supplementary Data for Chapter 3

Table C.1. Whole-genome sequences of *Rhodobacteraceae* strains used in this study.

The sequences were obtained from the GenBank database on September 1, 2016. Strains with genomes less than 95% complete were subsequently removed from downstream analyses.

Species and strain (as per GenBank annotation)	Accession number/s	No. of Contigs	Completeness
Actibacterium atlanticum 22II-S11-z10	AQQY00000000	24	98%
Actibacterium mucosum KCTC 23349	JFKE00000000	22	99%
Aestuariivita atlantica 22II-S11-z3	AQQZ00000000	56	100%
Aestuariivita boseongensis BS-B2	JXYH00000000	146	99%
Agrobacterium tumefaciens Ach5	CP011246; CP011247	2	99%
Agrobacterium tumefaciens P4	CM002258; CM002259	2	98%
Agrobacterium tumefaciens WRT31	CM002024; CM002025	2	99%
Ahrensia kielensis DSM 5890	ARFW00000000	16	76%
Ahrensia marina LZD062	JXMU00000000	65	99%
Ahrensia sp. 13_GOM-1096m	JIAX00000000	15	99%
Ahrensia sp. R2A130	AEEB00000000	59	98%
Aliiroseovarius crassostreae CV919-312	LKBA00000000	27	99%
Celeribacter baekdonensis B30	AMRK00000000	44	99%
Celeribacter halophilus ZXM137	LRUD00000000	71	100%
Celeribacter indicus P73	CP004393	1	99%
Celeribacter marinus IMCC 12053	CP012023	1	99%
Celeribacter sp. NH195	LRUC00000000	52	100%
Citreicella sp. 357	AJKJ00000000	180	96%
Citreicella sp. SE45	ACNW00000000	122	99%
Confluentimicrobium sp. EMB200-NS6	CP010869	1	99%
Defluviimonas alba cai42	CP012661	1	97%
Defluviimonas sp. 20V17	AYXI00000000	236	96%
Dinoroseobacter shibae DFL 12	CP000830	1	100%
Donghicola sp. JL3646	LZFQ00000000	44	99%
Falsirhodobacter sp. alg1	BBJC00000000	154	97%
Gemmobacter nectariphilus DSM 15620	AUCM00000000	59	97%
Gemmobacter sp. LW-1	LJSC00000000	71	98%
Haematobacter massiliensis CCUG 47968	JGYG00000000	53	98%
Haematobacter missouriensis CCUG 52307	JFGS00000000	136	98%
Halocynthiibacter arcticus PAMC 20958	CP014327	1	99%
Halocynthiibacter namhaensis RA2-3	JWIF00000000	74	100%
Jannaschia aquimarina GSW-M26	JYFE00000000	81	98%
Jannaschia donghaensis CECT 7802	CXSU00000000	13	99%
Jannaschia rubra CECT 5088	CXPG00000000	27	98%
Jannaschia seosinensis CECT 7799	CYPR00000000	253	95%
Jannaschia sp. CCS1	CP000264	1	100%
Jannaschia sp. EhC01	LXYJ00000000	96	99%
Ketogulonicigenium vulgare WSH-001	CP002018	1	96%
Ketogulonicigenium vulgare Y25	CP002224	1	95%
Labrenzia aggregata CECT 4801	CXST00000000	39	98%
Labrenzia aggregata IAM 12614	AAUW00000000	48	98%
Labrenzia alba CECT 5094	CXWA00000000	35	99%
Labrenzia alba CECT 5095	CXWE00000000	32	99%
Labrenzia alba CECT 5096	CXWC00000000	21	99%
Labrenzia alba CECT 7551	CXTY00000000	41	99%
Labrenzia alexandrii CECT 5112	CXWD00000000	47	98%

Species and strain (as per GenBank annotation)	Accession number/s	No. of Contigs	Completeness
Labrenzia alexandrii DFL-11	ACCU00000000	25	99%
Labrenzia sp. C1B10	AXBY00000000	70	98%
Labrenzia sp. C1B70	AXCE00000000	63	98%
Labrenzia sp. CP4	CP011927	1	98%
Labrenzia sp. DG1229	AYYG00000000	212	98%
Labrenzia sp. OB1	JSEP00000000	100	99%
Leisingera aquaemixtae CECT 8399	CYSR00000000	40	99%
Leisingera aquimarina DSM 24565	AXBE00000000	15	99%
Leisingera caerulea DSM 24564	AXBI00000000	21	99%
Leisingera daeponensis DSM 23529	AXBD00000000	12	98%
Leisingera methylohalidivorans DSM 14336	CP006773	1	99%
Leisingera sp. ANG-DT	JWLE00000000	91	98%
Leisingera sp. ANG-M1	JWLC00000000	92	99%
Leisingera sp. ANG-M6	JWLG00000000	54	99%
Leisingera sp. ANG-M7	JWLI00000000	58	99%
Leisingera sp. ANG-S	JWLM00000000	68	98%
Leisingera sp. ANG-S3	JWLF00000000	70	99%
Leisingera sp. ANG-S5	JWLH00000000	43	98%
Leisingera sp. ANG-Vp	JWLD00000000	143	99%
Leisingera sp. ANG1	AFCF00000000	20	99%
Leisingera sp. JC1	LYUZ00000000	168	100%
Litoreibacter arenae DSM 19593	AONI00000000	17	99%
Loktanella cinnabarina LL-001	BATB00000000	192	99%
Loktanella hongkongensis DSM 17492	APGJ00000000	9	99%
Loktanella sp. 1ANDIMAR09	LIGP00000000	18	97%
Loktanella sp. 3ANDIMAR09	LJAK00000000	34	98%
Loktanella sp. 5RATIMAR09	LJAL00000000	43	99%
Loktanella sp. S4079	JXYE00000000	43	100%
Loktanella vestfoldensis DSM 16212	ARNL00000000	49	96%
Loktanella vestfoldensis SKA53	AAMS00000000	14	98%
Mameliella alba UMTAT08	JSUQ00000000	62	100%
Maribius sp. MOLA 401	JQEY00000000	33	99%
Marinosulfonomonas sp. PRT-SC04	JPUR00000000	393	52%
Marinovum algicola DG 898	CP010855	1	99%
Maritimibacter alkaliphilus HTCC2654	AAMT00000000	46	98%
Maritimibacter sp. REDSEA-S28_B5	LUOO00000000	122	89%
Maritimibacter sp. REDSEA-S40_B3	LUOP00000000	28	73%
Nautella italica CECT 7321	CVRL00000000	46	98%
Nautella italica CECT 7645	CVRM00000000	37	98%
Nautella sp. ECSMB14104	LAUK00000000	40	99%
Nereida ignava CECT 5292	CVPC00000000	69	98%
Nesiotobacter exalbescens DSM 16456	AUGS00000000	40	98%
Oceanibulbus indolifex HEL-45	ABID00000000	105	100%
Oceanibulbus sp. HI0021	LWEP00000000	717	98%
Oceanibulbus sp. HI0023	LWEQ00000000	652	95%
Oceanibulbus sp. HI0027	LWER00000000	458	100%
Oceanibulbus sp. HI0040	LWEX00000000	611	96%
Oceanibulbus sp. HI0076	LWFQ00000000	875	94%
Oceanicola granulosus HTCC2516	AAOT00000000	85	98%
Oceanicola sp. HL-35	JAFT00000000	8	99%
Oceanicola sp. MCTG156(1a)	JQMY00000000	5	99%
Oceanicola sp. S124	AFPM00000000	339	96%
Oceaniovalibus guishaninsula JLT2003	AMGO00000000	68	99%
Octadecabacter antarcticus 307	CP003740	1	98%
Octadecabacter arcticus 238	CP003742	1	100%

Species and strain (as per GenBank annotation)	Accession number/s	No. of Contigs	Completeness
Octadecabacter temperatus SB1	CP012160	1	99%
Paenirhodobacter enshiensis DW2-9	JFZB00000000	112	99%
Paenirhodobacter sp. MME-103	BDDO00000000	202	98%
Pannonibacter indicus DSM 23407	CYHE00000000	36	98%
Pannonibacter phragmitetus 31801	CP013068	1	98%
Pannonibacter phragmitetus CGMCC9175	LGSQ00000000	40	98%
Pannonibacter phragmitetus DSM 14782	ARNQ00000000	46	98%
Paracoccus aminophilus JCM 7686	CP006650	1	96%
Paracoccus aminovorans HPD-2	LNBC00000000	55	97%
Paracoccus denitrificans PD1222	CP000489; CP000490	2	97%
Paracoccus halophilus JCM 14014	JRKN00000000	122	94%
Paracoccus pantotrophus J40	JAGK00000000	119	96%
Paracoccus pantotrophus J46	JAEM00000000	105	96%
Paracoccus sanguinis 10990	JRKR00000000	309	95%
Paracoccus sanguinis 39524	JRKP0000000	273	96%
Paracoccus sanguinis 4681	JRKT00000000	176	98%
Paracoccus sanguinis 5503	JRKQ00000000	265	98%
Paracoccus sp. 228	JYGY00000000	32	98%
Paracoccus sp. J39	JAEN00000000	50	97%
Paracoccus sp. J55	AZVA00000000	69	97%
Paracoccus sp. MKU1	LLWQ00000000	242	96%
Paracoccus sp. N5	AQUO00000000	3	98%
Paracoccus sp. PAMC 22219	BBPH00000000	265	94%
Paracoccus sp. S4493	JXYF00000000	586	89%
Paracoccus sp. TRP	AEPN00000000	119	94%
Paracoccus sphaerophysae HAMBI 3106	JRKS00000000	137	96%
Paracoccus versutus DSM 582	JRKO00000000	187	97%
Paracoccus yeei ATCC BAA-599	JHWH00000000	73	97%
Paracoccus zeaxanthinifaciens ATCC 21588	ATUJ00000000	35	97%
Pelagibaca bermudensis HTCC2601	AATQ00000000	103	97%
Phaeobacter gallaeciensis 2.10	CP002972	1	97%
Phaeobacter gallaeciensis DSM 26640	CP006966	1	98%
Phaeobacter gallaeciensis JL2886	CP015124	1	100%
Phaeobacter inhibens DSM 16374	AXBB00000000	8	97%
Phaeobacter inhibens DSM 17395	CP002976	1	97%
Phaeobacter inhibens S4Sm	LOHU00000000	80	97%
Phaeobacter sp. 11ANDIMAR09	LIKT00000000	101	99%
Phaeobacter sp. CECT 5382	CYSG00000000	56	99%
Phaeobacter sp. CECT 7735	CYTW00000000	12	99%
Phaeobacter sp. S26	JSWK00000000	59	98%
Phaeobacter sp. S60	JSWJ00000000	55	98%
Planktomarina temperata RCA23	CP003984	1	99%
Ponticoccus sp. SJ5A-1	LPXO00000000	116	100%
Pseudodonghicola xiamenensis DSM 18339	AUBS00000000	90	99%
Pseudooceanicola atlanticus 22II-s11g	AQQX00000000	47	98%
Pseudooceanicola batsensis HTCC2597	AAMO00000000	23	97%
Pseudooceanicola nanhaiensis DSM 18065	JHZF00000000	20	100%
Pseudophaeobacter arcticus DSM 23566	AXBF00000000	8	100%
P*seudorhodobacter antarcticus* KCTC 23700	LGHU00000000	129	96%
Pseudorhodobacter aquimaris KCTC 23043	LGHS00000000	77	96%
Pseudorhodobacter ferrugineus DSM 5888	ATVN00000000	42	98%
Pseudorhodobacter ferrugineus LMG 22047	LGHV00000000	46	98%
Pseudorhodobacter psychrotolerans PAMC 27389	LGIC00000000	93	99%
Pseudorhodobacter wandonensis KCTC 23672	LGHT00000000	45	95%
Pseudoruegeria sabulilitoris GJMS-35	LOAS00000000	148	99%

Species and strain (as per GenBank annotation)	Accession number/s	No. of Contigs	Completeness
Pseudoruegeria sp. SF-16	LNCI00000000	41	100%
Pseudovibrio axinellae Ad2	LMCB00000000	162	99%
Pseudovibrio denitrificans JCM 12308	BAZK00000000	94	99%
Pseudovibrio hongkongensis UST20140214-015B	LLWC00000000	39	98%
Pseudovibrio sp. Ad13	LMCC00000000	36	99%
Pseudovibrio sp. Ad14	LMCD00000000	57	99%
Pseudovibrio sp. Ad26	LMCE00000000	159	99%
Pseudovibrio sp. Ad37	LMCF00000000	224	99%
Pseudovibrio sp. Ad46	LMCG00000000	85	99%
Pseudovibrio sp. Ad5	LMCH00000000	66	99%
Pseudovibrio sp. FO-BEG1	CP003147	1	99%
Pseudovibrio sp. JCM 19062	BAXV00000000	286	82%
Pseudovibrio sp. JE062	ABXL00000000	53	99%
Pseudovibrio sp. POLY-S9	LCWX00000000	271	91%
Pseudovibrio sp. W64	LMCI00000000	49	99%
Pseudovibrio sp. W74	LMCJ00000000	64	99%
Pseudovibrio sp. WM33	LMCK00000000	159	99%
Pseudovibrio stylochi UST20140214-052	LLWE00000000	43	98%
Puniceibacterium sp. IMCC21224	LDPY00000000	9	99%
Rhodobacter capsulatus A52	LLVV00000000	70	97%
Rhodobacter capsulatus B41	LLVU00000000	64	97%
Rhodobacter capsulatus B6	AYQA00000000	113	97%
Rhodobacter capsulatus DE442	AYPR00000000	40	98%
Rhodobacter capsulatus R121	AYQC00000000	36	98%
Rhodobacter capsulatus SB 1003	CP001312	1	99%
Rhodobacter capsulatus Y262	AYQB00000000	51	99%
Rhodobacter capsulatus YW1	AYPY00000000	51	99%
Rhodobacter capsulatus YW2	AYPZ00000000	62	98%
Rhodobacter lobularis IGS	LFTY00000000	2	100%
Rhodobacter sp. AKP1	ANFS00000000	102	98%
Rhodobacter sp. BACL10 MAG-120419-bin15	LICP00000000	765	81%
Rhodobacter sp. BACL10 MAG-120910-bin24	LICH00000000	671	83%
Rhodobacter sp. BACL10 MAG-121220-bin24	LIBR00000000	796	64%
Rhodobacter sp. CACIA14H1	AYNO00000000	236	91%
Rhodobacter sp. CCB-MM2	LRRR00000000	136	98%
Rhodobacter sp. SW2	ACYY00000000	59	95%
Rhodobacter sphaeroides 2.4.1	CP000143; CP000144	2	99%
Rhodobacter sphaeroides ATCC 17025	CP000661	1	97%
Rhodobacter sphaeroides ATCC 17029	CP000577; CP000578	2	99%
Rhodobacter sphaeroides KD131	CP001150; CP001151	2	99%
Rhodobacter sphaeroides MBTLJ-13	CP015210; CP015211	2	99%
Rhodobacter sphaeroides MBTLJ-8	CP012960; CP012961	2	99%
Rhodobacter sphaeroides WS8N	CM001161; CM001162	2	99%
Rhodobacteraceae bacterium ASP10-04a	LAQD00000000	61	93%
Rhodobacteraceae bacterium EhC02	LXYH00000000	60	99%
Rhodobacteraceae bacterium HIMB11	AVDB00000000	34	99%
Rhodobacteraceae bacterium HL-91	LJSG00000000	24	97%
Rhodobacteraceae bacterium HLUCCA08	LJSF00000000	45	97%
Rhodobacteraceae bacterium HLUCCA09	LJNT00000000	226	89%
Rhodobacteraceae bacterium HLUCCA12	LJSV00000000	52	98%
Rhodobacteraceae bacterium HLUCCA24	LKOZ00000000	145	7%
Rhodobacteraceae bacterium HLUCCO07	LJSU00000000	35	100%
Rhodobacteraceae bacterium HLUCCO18	LJSY00000000	87	96%
Rhodobacteraceae bacterium HTCC2083	ABXE00000000	20	98%
Rhodobacteraceae bacterium HTCC2150	AAXZ00000000	25	94%

Species and strain (as per GenBank annotation)	Accession number/s	No. of Contigs	Completeness
Rhodobacteraceae bacterium KLH11	ACCW00000000	48	99%
Rhodobacteraceae bacterium O3.65	LPUY00000000	139	98%
Rhodobacteraceae bacterium PD-2	AWRV00000000	450	100%
Rhodobacteraceae bacterium REDSEA-S02_B3	LUPX00000000	113	71%
Rhodobacteraceae bacterium REDSEA-S03_B4	LUPY00000000	202	64%
Rhodobacteraceae bacterium REDSEA-S11_B6	LUQA00000000	192	65%
Rhodobacteraceae bacterium REDSEA-S29_B10	LUQB00000000	357	38%
Rhodobacteraceae bacterium REDSEA-S34_B6	LUQC00000000	111	86%
Rhodobacteraceae bacterium SB2	LGRT00000000	38	100%
Rhodovulum sp. NI22	JQFU00000000	120	100%
Rhodovulum sp. PH10	AKZI00000000	273	98%
Rhodovulum sulfidophilum DSM 1374	CP015418	1	98%
Rhodovulum sulfidophilum DSM 2351	AP014800	1	97%
Rhodovulum sulfidophilum DSM SNK001	CP015421	1	98%
Roseibacterium elongatum DSM 19469	CP004372	1	99%
Roseibium sp. TrichSKD4	AEFL00000000	108	99%
Roseivivax atlanticus 22II-s10s	AQQW00000000	64	98%
Roseivivax halodurans JCM 10272	JALZ00000000	150	99%
Roseivivax isoporae LMG 25204	JAME00000000	203	98%
Roseobacter denitrificans OCh 114	CP000362	1	99%
Roseobacter litoralis Och 149	CP002623	1	100%
Roseobacter sp. AzwK-3b	ABCR00000000	31	98%
Roseobacter sp. CCS2	AAYB00000000	11	95%
Roseobacter sp. GAI101	ABXS00000000	67	97%
Roseobacter sp. MED193	AANB00000000	19	100%
Roseobacter sp. SK209-2-6	AAYC00000000	29	99%
Roseovarius atlanticus R12B	LAXJ00000000	73	96%
Roseovarius indicus B108	LAXI00000000	122	97%
Roseovarius indicus EhC03	LXYQ00000000	168	94%
Roseovarius mucosus DSM 17069	AONH00000000	25	98%
Roseovarius nubinhibens ISM	AALY00000000	10	98%
Roseovarius sp. 217	AAMV00000000	37	99%
Roseovarius sp. BRH_c41	LADY00000000	71	100%
Roseovarius sp. MCTG156(2b)	JQLS00000000	8	99%
Roseovarius sp. TM1035	ABCL00000000	15	96%
Roseovarius tolerans EL-164	LGVV00000000	121	95%
Rubellimicrobium mesophilum DSM 19309	AOSK00000000	136	99%
Rubellimicrobium thermophilum DSM 16684	AOLV00000000	43	96%
Ruegeria atlantica CECT 4292	CYPU00000000	75	99%
Ruegeria atlantica CECT 4293	CYPS00000000	67	99%
Ruegeria conchae TW15	AEYW00000000	28	98%
Ruegeria halocynthiae MOLA R1/13b	JQEZ00000000	19	100%
Ruegeria lacuscaerulensis ITI-1157	ACNX00000000	47	98%
Ruegeria mobilis 270-3	LNVY00000000	81	100%
Ruegeria mobilis F1926	CP015230	1	90%
Ruegeria mobilis NBRC 101030	LNWY00000000	59	99%
Ruegeria mobilis NBRC 102038	LNWZ00000000	85	99%
Ruegeria mobilis S1942	JXYG00000000	93	99%
Ruegeria pomeroyi DSS-3	CP000031	1	97%
Ruegeria sp. 6PALISEP08	LGXZ00000000	42	99%
Ruegeria sp. ANG-R	JWLJ00000000	41	100%
Ruegeria sp. ANG-S4	JWLK00000000	20	98%
Ruegeria sp. CECT 5091	CYUD00000000	42	99%
Ruegeria sp. PBVC088	LZNT00000000	143	97%
Ruegeria sp. R11	ABXM00000000	17	96%

Species and strain (as per GenBank annotation)	Accession number/s	No. of Contigs	Completeness
Ruegeria sp. TM1040	CP000377	1	96%
Ruegeria sp. TrichCH4B	ACNZ00000000	129	99%
Ruegeria sp. ZGT108	LQBP00000000	33	99%
Ruegeria sp. ZGT118	LQBQ00000000	39	100%
Sagittula stellata E-37	AAYA00000000	39	99%
Salipiger mucosus DSM 16094	APVH00000000	84	100%
Sedimentitalea nanhaiensis DSM 24252	AXBG00000000	30	100%
Sediminimonas qiaohouensis DSM 21189	AUIJ00000000	68	100%
Shimia marina CECT 7688	CYPW00000000	45	99%
Shimia sp. SK013	LAJH00000000	28	100%
Stappia indica EBBD 17.2	MBQF00000000	45	99%
Stappia indica SBBC 49	MBQE00000000	17	99%
Stappia stellulata DSM 5886	AUIM00000000	15	98%
Sulfitobacter donghicola KCTC 12864	JASF00000000	9	100%
Sulfitobacter geojensis EhN01	LXYM00000000	46	100%
Sulfitobacter geojensis MM-124	JASE00000000	5	100%
Sulfitobacter guttiformis KCTC 32187	JASG00000000	4	100%
Sulfitobacter mediterraneus 1FIGIMAR09	JEMU00000000	55	100%
Sulfitobacter mediterraneus KCTC 32188	JASH00000000	31	98%
Sulfitobacter noctilucae NB-68	JASC00000000	14	100%
Sulfitobacter noctilucicola NB-77	JASD00000000	8	99%
Sulfitobacter pontiacus 3SOLIMAR09	AXZR00000000	25	99%
Sulfitobacter pontiacus EhN02	LXYK00000000	44	99%
Sulfitobacter pseudonitzschiae H3	JAMD00000000	80	100%
Sulfitobacter sp. 20_GPM-1509m	JIBC00000000	27	100%
Sulfitobacter sp. CB2047	JPOY00000000	12	99%
Sulfitobacter sp. EE-36	AALV00000000	15	100%
Sulfitobacter sp. EhC04	LXYI00000000	89	99%
Sulfitobacter sp. HI0054	LWFD00000000	555	95%
Sulfitobacter sp. NAS-14.1	AALZ00000000	27	100%
Tateyamaria sp. ANG-S1	JWLL00000000	32	100%
Thalassobacter sp. 16PALIMAR09	JHAK00000000	39	98%
Thalassobacter stenotrophicus 1CONIMAR09	JGVS00000000	28	98%
Thalassobacter stenotrophicus CECT 5294	CYRX00000000	33	97%
Thalassobium sp. R2A62	ACOA00000000	8	97%
Thalassobius gelatinovorus CECT 4357	CYSA00000000	30	97%
Thalassobius mediterraneus CECT 5118	CYSB00000000	48	100%
Thalassobius mediterraneus CECT 5120	CYSC00000000	47	100%
Thalassobius mediterraneus CECT 5383	CYSF00000000	19	99%
Thalassobius sp. CECT 5113	CYTO00000000	26	96%
Thalassobius sp. CECT 5114	CYUE00000000	26	97%
Thioclava atlantica 13D2W-2	AQRC00000000	47	98%
Thioclava dalianensis DLFJ1-1	JHEH00000000	98	95%
Thioclava indica DT23-4	AUNB00000000	105	96%
Thioclava pacifica DSM 10166	AUND00000000	42	98%
Thioclava sp. SK-1	MDHA00000000	106	97%
Tropicibacter multivorans DSM 26470	CYSD00000000	45	99%
Tropicibacter naphthalenivorans DSM 19561	CYSE00000000	33	99%
Wenxinia marina DSM 24838	ARAY00000000	42	99%
Yangia sp. CCB-MM3	CP014595; CP014596	2	97%

Table C.2. Core housekeeping genes used to determine genome completeness.

Completeness is the percentage of the 108 genes present in each genome. The locus tags and accession numbers of genes are from the reference genome *D. shibae* DFL 12^T.

Locus tag	Gene description	Accession no.
Dshi_1642	Alanyl-tRNA synthase	ABV93384
Dshi_1728	Arginyl-tRNA synthetase	ABV93470
Dshi_1022	3-Phosphoshikimate 1-carboxyvinyltransferase	ABV92764
Dshi_3275	Chorismate synthase	ABV95008
Dshi_2633	Aspartyl-tRNA synthetase	ABV94366
Dshi_1330	Biotin-[acetyl-CoA-carboxylase] ligase	ABV93072
Dshi_3449	Dephospho-CoA kinase	ABV95182
Dshi_1698	Cysteinyl-tRNA synthetase	ABV93440
Dshi_2976	Coenzyme A biosynthesis bifunctional protein	ABV94709
Dshi_0102	DNA polymerase III, alpha subunit	ABV91851
Dshi_3374	DNA polymerase III, beta subunit	ABV95107
Dshi_2145	Enolase	ABV93881
Dshi_0353	Signal recognition particle (SRP) protein	ABV92102
Dshi_2398	Dihydropteroate synthase	ABV94134
Dshi_1494	Ribosome recycling factor	ABV93236
Dshi_2725	Fused signal recognition particle (SRP) receptor	ABV94458
Dshi_0273	Translation elongation factor G	ABV92022
Dshi_1933	Isoleucyl-tRNA synthetase	ABV93674
Dshi_3563	Translation initiation factor IF-2	ABV95296
Dshi_1149	Translation initiation factor IF-3	ABV92891
Dshi_0809	GTP-binding protein LepA	ABV92554
Dshi_0057	Leucyl-tRNA synthetase	ABV91806
Dshi_2707	Methionyl-tRNA synthetase	ABV94440
Dshi_1350	Nucleoside-diphosphate kinase	ABV93092
Dshi_2249	N utilization substance protein B-like protein	ABV93985
Dshi_1682	Glucose-6-phosphate isomerase	ABV93424
Dshi_0235	Phenylalanyl-tRNA synthetase, alpha subunit	ABV91984
Dshi_0237	Phenylalanyl-tRNA synthetase, beta subunit	ABV91986
Dshi_3345	Nicotinate phosphoribosyltransferase	ABV95078
Dshi_2997	Polyribonucleotide nucleotidyltransferase	ABV94730
Dshi_1733	Peptide chain release factor RF-1	ABV93475
Dshi_1926	Primosomal protein N'	ABV93668
Dshi_0744	Prolyl-tRNA synthetase	ABV92489
Dshi_1535	Phosphate transport system permease protein	ABV93277
Dshi_1534	Phosphate transport system permease protein	ABV93276
Dshi_0947	Aminoacyl-tRNA hydrolase	ABV92692
Dshi_2742	Adenylosuccinate synthetase	ABV94475
Dshi_1076	Phosphoribosylaminoimidazole-succinocarboxamidesynthase	ABV92818
Dshi_3091	Aspartate carbamoyltransferase	ABV94824
Dshi_0615	Orotidine 5'-phosphate decarboxylase	ABV92361
Dshi_2744	CTP synthase	ABV94477
Dshi_1493	Uridylate kinase	ABV93235
Dshi_1643	Bacterial DNA recombination protein	ABV93385
Dshi_1510	ATP-dependent DNA helicase	ABV93252
Dshi_0200	Ribonuclease III	ABV91949
Dshi_0023	Ribonuclease HII	ABV91772
Dshi_0261	50S ribosomal subunit protein L1	ABV92010
Dshi_0285	50S ribosomal subunit protein L2	ABV92034
Dshi_0276	50S ribosomal subunit protein L3	ABV92025
Dshi_0277	50S ribosomal subunit protein L4/L1e	ABV92026

Locus tag	Gene description	Accession no.
Dshi_0297	50S ribosomal subunit protein L5	ABV92046
Dshi_0300	50S ribosomal subunit protein L6	ABV92049
Dshi_0260	50S ribosomal subunit protein L11	ABV92009
Dshi_0266	50S ribosomal subunit protein L7/12	ABV92015
Dshi_1602	50S ribosomal subunit protein L13	ABV93344
Dshi_0295	50S ribosomal subunit protein L14	ABV92044
Dshi_0304	50S ribosomal subunit protein L15	ABV92053
Dshi_0289	50S ribosomal subunit protein L16	ABV92038
Dshi_0310	50S ribosomal subunit protein L17	ABV92059
Dshi_0301	50S ribosomal subunit protein L18	ABV92050
Dshi_0341	50S ribosomal subunit protein L19	ABV92090
Dshi_0231	50S ribosomal subunit protein L20	ABV91980
Dshi_0287	50S ribosomal subunit protein L22	ABV92036
Dshi_0296	50S ribosomal subunit protein L24	ABV92045
Dshi_0309	DNA-directed RNA polymerase, alpha subunit	ABV92058
Dshi_0268	DNA-directed RNA polymerase, beta subunit	ABV92017
Dshi_1548	30S ribosomal subunit protein S2	ABV93290
Dshi_0288	30S ribosomal subunit protein S3	ABV92037
Dshi_2948	30S ribosomal subunit protein S4	ABV94681
Dshi_0302	30S ribosomal subunit protein S5	ABV92051
Dshi_0272	30S ribosomal subunit protein S7	ABV92021
Dshi_0299	30S ribosomal subunit protein S8	ABV92048
Dshi_1601	30S ribosomal subunit protein S9	ABV93343
Dshi_0275	30S ribosomal subunit protein S10	ABV92024
Dshi_0308	30S ribosomal subunit protein S11	ABV92057
Dshi_0271	30S ribosomal subunit protein S12	ABV92020
Dshi_0307	30S ribosomal subunit protein S13	ABV92056
Dshi_1104	Holliday junction DNA helicase RuvA	ABV92846
Dshi_1105	Holliday junction DNA helicase RuvB	ABV92847
Dshi_3567	Preprotein translocase, secA subunit	ABV95300
Dshi_0305	Preprotein translocase, SecY subunit	ABV92054
Dshi_1863	Seryl-tRNA synthetase	ABV93605
Dshi_0147	SsrA-binding protein	ABV91896
Dshi_2346	Queuine tRNA-ribosyltransferase	ABV94082
Dshi_0777	Threonyl-tRNA synthetase	ABV92522
Dshi_2190	Trigger factor	ABV93926
Dshi_2078	Triosephosphate isomerase	ABV93815
Dshi_0342	tRNA (guanine-N1)-methyltransferase	ABV92091
Dshi_0952	Tryptophan synthase, alpha subunit	ABV92697
Dshi_1800	Indole-3-glycerol phosphate synthase	ABV93542
Dshi_2836	Tryptophanyl-tRNA synthetase	ABV94569
Dshi_3042	tRNA pseudouridine synthase B	ABV94775
Dshi_1549	Translation elongation factor	ABV93291
Dshi_2151	Tyrosyl-tRNA synthetase	ABV93887
Dshi_1809	UvrABC system protein A	ABV93551
Dshi_2192	UvrABC system protein B	ABV93928
Dshi_2087	Valyl-tRNA synthetase	ABV93824
Dshi_2474	S-Adenosyl-L-methionine-dependent methyltransferase	ABV94208
Dshi_1495	Undecaprenyl diphosphate synthase	ABV93237
Dshi_1015	Conserved protein of unknown function UPF0054	ABV92757
Dshi_0954	GTP-binding protein YchF	ABV92699
Dshi_1362	FAD-dependent pyridine nucleotide-disulfide oxidoreductase	ABV93104
Dshi_2094	Translation elongation factor P	ABV93831
Dshi_1207	Glutamyl-tRNA(Gln) amidotransferase, B subunit	ABV92949
Dshi_3415	O-Sialoglycoprotein endopeptidase	ABV95148

Locus tag	Gene description	Accession no.
Dshi_1466	GTP-binding protein Obg/CgtA	ABV93208
Dshi_0328	Preprotein translocase subunit	ABV92077
Dshi_2831	Conserved hypothetical protein	ABV94564

Table C.3. Whole-genome sequences from strains of the order *Rhizobiales* used in this study.

Species and strain	Family	Accession number/s
Beijerinckia indica subsp. *indica* ATCC 9039	*Beijerinckiaceae*	CP001016
Chelatococcus daeguensis TAD1	*Beijerinckiaceae*	CP018095
Methylocella silvestris BL2	*Beijerinckiaceae*	CP001280
Bradyrhizobium japonicum USDA 6	*Bradyrhizobiaceae*	AP012206
Nitrobacter hamburgensis X14	*Bradyrhizobiaceae*	CP000319
Rhodopseudomonas palustris HaA2	*Bradyrhizobiaceae*	CP000250
Blastochloris viridis ATCC 19567	*Hyphomicrobiaceae*	CP012946
Hyphomicrobium denitrificans ATCC 51888	*Hyphomicrobiaceae*	CP002083
Rhodomicrobium vannielii ATCC 17100	*Hyphomicrobiaceae*	CP002292
Aminobacter aminovorans KCTC 2477	*Phyllobacteriaceae*	CP015005
Hoeflea phototrophica DFL-43	*Phyllobacteriaceae*	CM002917
Mesorhizobium loti NZP2037	*Phyllobacteriaceae*	CP016079
Agrobacterium tumefaciens Ach5	*Rhizobiaceae*	CP011246; CP011247
Rhizobium leguminosarum bv. *viciae* 3841	*Rhizobiaceae*	AM236080
Sinorhizobium meliloti 1021	*Rhizobiaceae*	AL591688

Table C.4. Proposed reclassification of strains based on dDDH, ANI, and AAI values.

Former species name and strain	New species name and strain
Family *Rhodobacteraceae*	
Actibacterium atlanticum 22II-S11-z10	*Actibacterium atlanticum* 22II-S11-z10
Actibacterium mucosum KCTC 23349	*Actibacterium mucosum* KCTC 23349
Confluentimicrobium sp. EMB200-NS6	*Actibacterium* sp. EMB200-NS6
Oceanicola sp. MCTG156(1a)	*Actibacterium* sp. MCTG156(1a)
Rhodovulum sp. NI22	*Actibacterium* sp. NI22
Aestuariivita atlantica 22II-S11-z3	*Aestuariivita atlantica* 22II-S11-z3
Loktanella cinnabarina LL-001	*Aliiloktanella cinnabarina* LL-001
Loktanella hongkongensis DSM 17492	*Aliiloktanella hongkongensis* DSM 17492
Aliiroseovarius crassostreae CV919-312	*Aliiroseovarius crassostreae* CV919-312
Rhodobacter sphaeroides ATCC 17025	*Alterhodobacter* sp. ATCC 17025
Rhodobacter sphaeroides 2.4.1	*Alterhodobacter sphaeroides* 2.4.1
Rhodobacter sp. AKP1	*Alterhodobacter sphaeroides* AKP1
Rhodobacter sphaeroides ATCC 17029	*Alterhodobacter sphaeroides* ATCC 17029
Rhodobacter sphaeroides KD131	*Alterhodobacter sphaeroides* KD131
Rhodobacter sphaeroides MBTLJ-13	*Alterhodobacter sphaeroides* MBTLJ-13
Rhodobacter sphaeroides MBTLJ-8	*Alterhodobacter sphaeroides* MBTLJ-8
Rhodobacter sphaeroides WS8N	*Alterhodobacter sphaeroides* WS8N
Celeribacter baekdonensis B30	*Celeribacter baekdonensis* B30
Celeribacter halophilus ZXM137	*Celeribacter halophilus* ZXM137
Celeribacter indicus P73	*Celeribacter indicus* P73
Celeribacter marinus IMCC 12053	*Celeribacter marinus* IMCC 12053
Celeribacter sp. NH195	*Celeribacter* sp. NH195
Defluviimonas alba cai42	*Defluviimonas alba* cai42
Dinoroseobacter shibae DFL 12	*Dinoroseobacter shibae* DFL 12
Donghicola sp. JL3646	*Donghicola* sp. JL3646
Falsirhodobacter sp. alg1	*Falsirhodobacter* sp. alg1
Gemmobacter nectariphilus DSM 15620	*Gemmobacter nectariphilus* DSM 15620
Haematobacter massiliensis CCUG 47968	*Haematobacter massiliensis* CCUG 47968
Haematobacter missouriensis CCUG 52307	*Haematobacter missouriensis* CCUG 52307
Halocynthiibacter namhaensis RA2-3	*Halocynthiibacter namhaensis* RA2-3
Jannaschia aquimarina GSW-M26	*Jannaschia aquimarina* GSW-M26
Jannaschia donghaensis CECT 7802	*Jannaschia donghaensis* CECT 7802
Jannaschia rubra CECT 5088	*Jannaschia rubra* CECT 5088
Jannaschia seosinensis CECT 7799	*Jannaschia seosinensis* CECT 7799
Ketogulonicigenium vulgare WSH-001	*Ketogulonicigenium vulgare* WSH-001
Ketogulonicigenium vulgare Y25	*Ketogulonicigenium vulgare* Y25
Litoreibacter arenae DSM 19593	*Litoreibacter arenae* DSM 19593
Loktanella sp. 1ANDIMAR09	*Loktanella* sp. 1ANDIMAR09
Loktanella sp. 3ANDIMAR09	*Loktanella* sp. 3ANDIMAR09
Loktanella sp. 5RATIMAR09	*Loktanella* sp. 5RATIMAR09
Roseobacter sp. CCS2	*Loktanella* sp. CCS2
Loktanella sp. S4079	*Loktanella* sp. S4079
Loktanella vestfoldensis DSM 16212	*Loktanella vestfoldensis* DSM 16212
*Loktanella vestfolden*sis SKA53	*Loktanella vestfoldensis* SKA53
Maribius sp. MOLA 401	*Maribius* sp. MOLA 401
Marinovum algicola DG 898	*Marinovum algicola* DG 898
Maritimibacter alkaliphilus HTCC2654	*Maritimibacter alkaliphilus* HTCC2654
Nereida ignava CECT 5292	*Nereida ignava* CECT 5292
Oceanicola granulosus HTCC2516	*Oceanicola granulosus* HTCC2516
*Oceaniovalibus guishanins*ula JLT2003	*Oceaniovalibus guishaninsula* JLT2003
Octadecabacter antarcticus 307	*Octadecabacter antarcticus* 307
Octadecabacter arcticus 238	*Octadecabacter arcticus* 238

Former species name and strain	New species name and strain
Family *Rhodobacteraceae*	
Octadecabacter temperatus SB1	*Octadecabacter temperatus* SB1
Paracoccus aminophilus JCM 7686	*Paracoccus aminophilus* JCM 7686
Paracoccus aminovorans HPD-2	*Paracoccus aminovorans* HPD-2
Paracoccus denitrificans PD1222	*Paracoccus denitrificans* PD1222
Paracoccus pantotrophus J40	*Paracoccus pantotrophus* J40
Paracoccus pantotrophus J46	*Paracoccus pantotrophus* J46
Paracoccus sanguinis 10990	*Paracoccus sanguinis* 10990
Paracoccus sanguinis 39524	*Paracoccus sanguinis* 39524
Paracoccus sanguinis 4681	*Paracoccus sanguinis* 4681
*Paracoccus sangu*inis 5503	*Paracoccus sanguinis* 5503
Paracoccus sp. 228	*Paracoccus* sp. 228
Paracoccus sp. J39	*Paracoccus* sp. J39
Paracoccus sp. J55	*Paracoccus* sp. J55
Paracoccus sp. MKU1	*Paracoccus* sp. MKU1
Paracoccus sp. N5	*Paracoccus* sp. N5
Paracoccus sphaerophysae HAMBI 3106	*Paracoccus sphaerophysae* HAMBI 3106
Paracoccus versutus DSM 582	*Paracoccus versutus* DSM 582
Paracoccus yeei ATCC BAA-599	*Paracoccus yeei* ATCC BAA-599
Paracoccus zeaxanthinifaciens ATCC 21588	*Paracoccus zeaxanthinifaciens* ATCC 21588
Planktomarina temperata RCA23	*Planktomarina temperata* RCA23
Halocynthiibacter arcticus PAMC 20958	*Pseudohalocynthiibacter arcticus* PAMC 20958
Pseudooceanicola atlanticus 22II-s11g	*Pseudooceanicola atlanticus* 22II-s11g
Pseudooceanicola batsensis HTCC2597	*Pseudooceanicola batsensis* HTCC2597
Pseudooceanicola nanhaiensis DSM 18065	*Pseudooceanicola nanhaiensis* DSM 18065
Oceanicola sp. S124	*Pseudooceanicola* sp. S124
Pseudorhodobacter antarcticus KCTC 23700	*Pseudorhodobacter antarcticus* KCTC 23700
Pseudorhodobacter aquimaris KCTC 23043	*Pseudorhodobacter aquimaris* KCTC 23043
Pseudorhodobacter ferrugineus DSM 5888	P*seudorhodobacter ferrugineus* DSM 5888
Pseudorhodobacter ferrugineus LMG 22047	*Pseudorhodobacter ferrugineus* LMG 22047
Pseudorhodobacter psychrotolerans PAMC 27389	*Pseudorhodobacter psychrotolerans* PAMC 27389
Rhodobacter sp. SW2	*Pseudorhodobacter* sp. SW2
Pseudorhodobacter wandonensis KCTC 23672	*Pseudorhodobacter wandonensis* KCTC 23672
Pseudoruegeria sabulilitoris GJMS-35	*Pseudoruegeria sabulilitoris* GJMS-35
Pseudoruegeria sp. SF-16	*Pseudoruegeria* sp. SF-16
Puniceibacterium sp. IMCC21224	*Puniceibacterium* sp. IMCC21224
Rhodobacter capsulatus A52	*Rhodobacter capsulatus* A52
Rhodobacter capsulatus B41	*Rhodobacter capsulatus* B41
Rhodobacter capsulatus B6	*Rhodobacter capsulatus* B6
Rhodobacter capsulatus DE442	*Rhodobacter capsulatus* DE442
Rhodobacter capsulatus R121	*Rhodobacter capsulatus* R121
Rhodobacter capsulatus SB 1003	*Rhodobacter capsulatus* SB 1003
Rhodobacter capsulatus Y262	*Rhodobacter capsulatus* Y262
Rhodobacter capsulatus YW1	*Rhodobacter capsulatus* YW1
Rhodobacter capsulatus YW2	*Rhodobacter capsulatus* YW2
Paenirhodobacter enshiensis DW2-9	*Rhodobacter enshiensis* DW2-9
Paenirhodobacter sp. MME-103	*Rhodobacter* sp. MME-103
Defluviimonas sp. 20V17	*Rhodobacteraceae* bacterium 20V17
Rhodobacter sp. CCB-MM2	*Rhodobacteraceae* bacterium CCB-MM2
Rhodobacteraceae bacterium HIMB11	*Rhodobacteraceae* bacterium HIMB11
Rhodobacteraceae bacterium HL-91	*Rhodobacteraceae* bacterium HL-91
Rhodobacteraceae bacterium HLUCCA08	*Rhodobacteraceae* bacterium HLUCCA08
Rhodobacteraceae bacterium HLUCCA12	*Rhodobacteraceae* bacterium HLUCCA12
Rhodobacteraceae bacterium HTCC2083	*Rhodobacteraceae* bacterium HTCC2083
Rhodobacter lobularis IGS	*Rhodobacteraceae* bacterium IGS

Former species name and strain	New species name and strain
Family *Rhodobacteraceae*	
Gemmobacter sp. LW-1	*Rhodobacteraceae* bacterium LW-1
Thalassobium sp. R2A62	*Rhodobacteraceae* bacterium R2A62
Rhodobacteraceae bacterium SB2	*Rhodobacteraceae* bacterium SB2
Thioclava sp. SK-1	*Rhodobacteraceae* bacterium SK-1
Rhodovulum sulfidophilum DSM 1374	*Rhodovulum sulfidophilum* DSM 1374
Rhodovulum sulfidophilum DSM 2351	*Rhodovulum sulfidophilum* DSM 2351
Rhodovulum sulfidophilum DSM SNK001	*Rhodovulum sulfidophilum* DSM SNK001
Roseibacterium elongatum DSM 19469	*Roseibacterium elongatum* DSM 19469
Jannaschia sp. CCS1	*Roseibacterium* sp. CCS1
Jannaschia sp. EhC01	*Roseibacterium* sp. EhC01
Rhodobacteraceae bacterium HLUCCO18	*Roseibacterium* sp. HLUCCO18
Roseivivax atlanticus 22II-s10s	*Roseivivax atlanticus* 22II-s10s
Roseivivax halodurans JCM 10272	*Roseivivax halodurans* JCM 10272
Roseivivax isoporae LMG 25204	*Roseivivax isoporae* LMG 25204
Roseobacter denitrificans OCh 114	*Roseobacter denitrificans* OCh 114
Sulfitobacter donghicola KCTC 12864	*Roseobacter donghicola* KCTC 12864
Sulfitobacter geojensis EhN01	*Roseobacter geojensis* EhN01
Sulfitobacter geojensis MM-124	*Roseobacter geojensis* MM-124
Sulfitobacter guttiformis KCTC 32187	*Roseobacter guttiformis* KCTC 32187
Oceanibulbus indolifex HEL-45	*Roseobacter indolifex* HEL-45
Roseobacter litoralis Och 149	*Roseobacter litoralis* Och 149
Sulfitobacter mediterraneus 1FIGIMAR09	*Roseobacter mediterraneus* 1FIGIMAR09
Sulfitobacter mediterraneus KCTC 32188	*Roseobacter mediterraneus* KCTC 32188
Sulfitobacter noctilucae NB-68	*Roseobacter noctilucae* NB-68
Sulfitobacter noctilucicola NB-77	*Roseobacter noctilucicola* NB-77
Sulfitobacter pontiacus 3SOLIMAR09	*Roseobacter pontiacus* 3SOLIMAR09
Sulfitobacter sp. CB2047	*Roseobacter pontiacus* CB2047
Sulfitobacter sp. EE-36	*Roseobacter pontiacus* EE-36
Sulfitobacter pontiacus EhN02	*Roseobacter pontiacus* EhN02
Sulfitobacter sp. NAS-14.1	*Roseobacter pontiacus* NAS-14.1
Sulfitobacter pseudonitzschiae H3	*Roseobacter pseudonitzschiae* H3
Sulfitobacter sp. 20_GPM-1509m	*Roseobacter* sp. 20_GPM-1509m
Tateyamaria sp. ANG-S1	*Roseobacter* sp. ANG-S1
Sulfitobacter sp. EhC04	*Roseobacter* sp. EhC04
Roseobacter sp. GAI101	*Roseobacter* sp. GAI101
Oceanibulbus sp. HI0021	*Roseobacter* sp. HI0021
Oceanibulbus sp. HI0023	*Roseobacter* sp. HI0023
Oceanibulbus sp. HI0027	*Roseobacter* sp. HI0027
Oceanibulbus sp. HI0040	*Roseobacter* sp. HI0040
Sulfitobacter sp. HI0054	*Roseobacter* sp. HI0054
Roseovarius atlanticus R12B	*Roseovarius atlanticus* R12B
Roseovarius indicus B108	*Roseovarius indicus* B108
Roseovarius mucosus DSM 17069	*Roseovarius mucosus* DSM 17069
Roseovarius nubinhibens ISM	*Roseovarius nubinhibens* ISM
Roseovarius sp. 217	*Roseovarius* sp. 217
Roseobacter sp. AzwK-3b	*Roseovarius* sp. AzwK-3b
Roseovarius sp. BRH_c41	*Roseovarius* sp. BRH_c41
Roseovarius sp. MCTG156(2b)	*Roseovarius* sp. MCTG156(2b)
Roseovarius sp. TM1035	*Roseovarius* sp. TM1035
Roseovarius tolerans EL-164	*Roseovarius tolerans* EL-164
Rubellimicrobium mesophilum DSM 19309	*Rubellimicrobium mesophilum* DSM 19309
Rubellimicrobium thermophilum DSM 16684	*Rubellimicrobium thermophilum* DSM 16684
Leisingera aquaemixtae CECT 8399	*Ruegeria aquaemixtae* CECT 8399
Leisingera aquimarina DSM 24565	*Ruegeria aquimarina* DSM 24565

Former species name and strain	New species name and strain
Family *Rhodobacteraceae*	
Pseudophaeobacter arcticus DSM 23566	*Ruegeria arcticus* DSM 23566
Ruegeria atlantica CECT 4292	*Ruegeria atlantica* CECT 4292
Aestuariivita boseongensis BS-B2	*Ruegeria boseongensis* BS-B2
Leisingera caerulea DSM 24564	*Ruegeria caerulea* DSM 24564
Ruegeria conchae TW15	*Ruegeria conchae* TW15
Leisingera daeponensis DSM 23529	*Ruegeria daeponensis* DSM 23529
Phaeobacter gallaeciensis DSM 26640	*Ruegeria gallaeciensis* DSM 26640
Ruegeria halocynthiae MOLA R1/13b	*Ruegeria halocynthiae* MOLA R1/13b
Phaeobacter gallaeciensis 2.10	*Ruegeria inhibens* 2.10
Phaeobacter inhibens DSM 16374	*Ruegeria inhibens* DSM 16374
Phaeobacter inhibens DSM 17395	*Ruegeria inhibens* DSM 17395
Phaeobacter inhibens S4Sm	*Ruegeria inhibens* S4Sm
Nautella italica CECT 7321	*Ruegeria italica* CECT 7321
Nautella italica CECT 7645	*Ruegeria italica* CECT 7645
Nautella sp. ECSMB14104	*Ruegeria italica* ECSMB14104
Ruegeria sp. R11	*Ruegeria italica* R11
Ruegeria lacuscaerulensis ITI-1157	*Ruegeria lacuscaerulensis* ITI-1157
Leisingera methylohalidivorans DSM 14336	*Ruegeria methylohalidivorans* DSM 14336
Ruegeria mobilis 270-3	*Ruegeria mobilis* 270-3
Ruegeria mobilis NBRC 101030	*Ruegeria mobilis* NBRC 101030
Ruegeria mobilis NBRC 102038	*Ruegeria mobilis* NBRC 102038
Ruegeria mobilis S1942	*Ruegeria mobilis* S1942
Ruegeria sp. TrichCH4B	*Ruegeria mobilis* TrichCH4B
Tropicibacter multivorans DSM 26470	*Ruegeria multivorans* DSM 26470
Sedimentitalea nanhaiensis DSM 24252	*Ruegeria nanhaiensis* DSM 24252
Ruegeria pomeroyi DSS-3	*Ruegeria pomeroyi* DSS-3
Phaeobacter sp. 11ANDIMAR09	*Ruegeria* sp. 11ANDIMAR09
Ruegeria sp. 6PALISEP08	*Ruegeria* sp. 6PALISEP08
Leisingera sp. ANG-DT	*Ruegeria* sp. ANG-DT
Leisingera sp. ANG-M1	*Ruegeria* sp. ANG-M1
Leisingera sp. ANG-M6	*Ruegeria* sp. ANG-M6
Leisingera sp. ANG-M7	*Ruegeria* sp. ANG-M7
Ruegeria sp. ANG-R	*Ruegeria* sp. ANG-R
Leisingera sp. ANG-S	*Ruegeria* sp. ANG-S
Leisingera sp. ANG-S3	*Ruegeria* sp. ANG-S3
Ruegeria sp. ANG-S4	*Ruegeria* sp. ANG-S4
Leisingera sp. ANG-S5	*Ruegeria* sp. ANG-S5
Leisingera sp. ANG-Vp	*Ruegeria* sp. ANG-Vp
Leisingera sp. ANG1	*Ruegeria* sp. ANG1
Ruegeria atlantica CECT 4293	*Ruegeria* sp. CECT 4293
Ruegeria sp. CECT 5091	*Ruegeria* sp. CECT 5091
Phaeobacter sp. CECT 5382	*Ruegeria* sp. CECT 5382
Leisingera sp. JC1	*Ruegeria* sp. JC1
Phaeobacter gallaeciensis JL2886	*Ruegeria* sp. JL2886
Rhodobacteraceae bacterium KLH11	*Ruegeria* sp. KLH11
Roseobacter sp. MED193	*Ruegeria* sp. MED193
Rhodobacteraceae bacterium O3.65	*Ruegeria* sp. O3.65
Phaeobacter sp. S26	*Ruegeria* sp. S26
Phaeobacter sp. S60	*Ruegeria* sp. S60
Roseobacter sp. SK209-2-6	*Ruegeria* sp. SK209-2-6
Ruegeria sp. TM1040	*Ruegeria* sp. TM1040
Ruegeria sp. ZGT108	*Ruegeria* sp. ZGT108
Ruegeria sp. ZGT118	*Ruegeria* sp. ZGT118
Pseudodonghicola xiamenensis DSM 18339	*Ruegeria xiamenensis* DSM 18339

Former species name and strain	New species name and strain
Family *Rhodobacteraceae*	
Ruegeria sp. PBVC088	*Sagittula alba* PBVC088
Mameliella alba UMTAT08	*Sagittula alba* UMTAT08
Pelagibaca bermudensis HTCC2601	*Sagittula bermudensis* HTCC2601
Salipiger mucosus DSM 16094	*Sagittula mucosus* DSM 16094
Tropicibacter naphthalenivorans DSM 19561	*Sagittula naphthalenivorans* DSM 19561
Citreicella sp. 357	*Sagittula* sp. 357
Yangia sp. CCB-MM3	*Sagittula* sp. CCB-MM3
Rhodobacteraceae bacterium PD-2	*Sagittula* sp. PD-2
Citreicella sp. SE45	*Sagittula* sp. SE45
Ponticoccus sp. SJ5A-1	*Sagittula* sp. SJ5A-1
Sagittula stellata E-37	*Sagittula stellata* E-37
Sediminimonas qiaohouensis DSM 21189	*Sediminimonas qiaohouensis* DSM 21189
Rhodobacteraceae bacterium EhC02	*Sediminimonas* sp. EhC02
Oceanicola sp. HL-35	*Sediminimonas* sp. HL-35
Rhodobacteraceae bacterium HLUCCO07	*Sediminimonas* sp. HLUCCO07
Shimia marina CECT 7688	*Shimia marina* CECT 7688
Thalassobius sp. CECT 5113	*Shimia* sp. CECT 5113
Thalassobius sp. CECT 5114	*Shimia* sp. CECT 5114
Phaeobacter sp. CECT 7735	*Shimia* sp. CECT 7735
Shimia sp. SK013	*Shimia* sp. SK013
Thalassobacter sp. 16PALIMAR09	*Thalassobacter stenotrophicus* 16PALIMAR09
Thalassobacter stenotrophicus 1CONIMAR09	*Thalassobacter stenotrophicus* 1CONIMAR09
Thalassobacter stenotrophicus CECT 5294	*Thalassobacter stenotrophicus* CECT 5294
Thalassobius gelatinovorus CECT 4357	*Thalassobius gelatinovorus* CECT 4357
Thalassobius mediterraneus CECT 5383	*Thalassobius mediterraneus* CECT 5383
Thalassobius mediterraneus CECT 5118	*Thalassobius* sp. CECT 5118
Thalassobius mediterraneus CECT 5120	*Thalassobius* sp. CECT 5120
Thioclava atlantica 13D2W-2	*Thioclava atlantica* 13D2W-2
Thioclava dalianensis DLFJ1-1	*Thioclava dalianensis* DLFJ1-1
Thioclava indica DT23-4	*Thioclava indica* DT23-4
Thioclava pacifica DSM 10166	*Thioclava pacifica* DSM 10166
Wenxinia marina DSM 24838	*Wenxinia marina* DSM 24838
Family *Stappiaceae* fam. nov.	
Pseudovibrio hongkongensis UST20140214-015B	*Falsivibrio hongkongensis* UST20140214-015B
Pseudovibrio stylochi UST20140214-052	*Falsivibrio stylochi* UST20140214-052
Labrenzia aggregata IAM 12614	*Labrenzia aggregata* IAM 12614
Labrenzia alba CECT 5094	*Labrenzia alba* CECT 5094
Labrenzia alba CECT 5095	*Labrenzia alba* CECT 5095
Labrenzia alba CECT 5096	*Labrenzia alba* CECT 5096
Labrenzia alexandrii CECT 5112	*Labrenzia alexandrii* CECT 5112
Labrenzia alexandrii DFL-11	*Labrenzia alexandrii* DFL-11
Labrenzia sp. C1B10	*Labrenzia* sp. C1B10
Labrenzia sp. C1B70	*Labrenzia* sp. C1B70
Labrenzia aggregata CECT 4801	*Labrenzia* sp. CECT 4801
Labrenzia alba CECT 7551	*Labrenzia* sp. CECT 7551
Labrenzia sp. CP4	*Labrenzia* sp. CP4
Labrenzia sp. DG1229	*Labrenzia* sp. DG1229
Labrenzia sp. OB1	*Labrenzia* sp. OB1
Roseibium sp. TrichSKD4	*Labrenzia* sp. TrichSKD4
Nesiotobacter exalbescens DSM 16456	*Nesiotobacter exalbescens* DSM 16456
Pannonibacter indicus DSM 23407	*Pannonibacter indicus* DSM 23407
Pannonibacter phragmitetus DSM 14782	*Pannonibacter phragmitetus* DSM 14782
*Pannonibacter phragmit*etus 31801	*Pannonibacter* sp. 31801

Former species name and strain	New species name and strain
Family *Stappiaceae* fam. nov.	
Pannonibacter phragmitetus CGMCC9175	*Pannonibacter* sp. CGMCC9175
Pseudovibrio axinellae Ad2	*Pseudovibrio axinellae* Ad2
Pseudovibrio denitrificans JCM 12308	*Pseudovibrio denitrificans* JCM 12308
Pseudovibrio sp. Ad13	*Pseudovibrio* sp. Ad13
Pseudovibrio sp. Ad14	*Pseudovibrio* sp. Ad14
Pseudovibrio sp. Ad26	*Pseudovibrio* sp. Ad26
Pseudovibrio sp. Ad37	*Pseudovibrio* sp. Ad37
Pseudovibrio sp. Ad46	*Pseudovibrio* sp. Ad46
Pseudovibrio sp. Ad5	*Pseudovibrio* sp. Ad5
Pseudovibrio sp. FO-BEG1	*Pseudovibrio* sp. FO-BEG1
Pseudovibrio sp. JE062	*Pseudovibrio* sp. JE062
Pseudovibrio sp. W64	*Pseudovibrio* sp. W64
Pseudovibrio sp. W74	*Pseudovibrio* sp. W74
Pseudovibrio sp. WM33	*Pseudovibrio* sp. WM33
Stappia indica EBBD 17.2	*Stappia indica* EBBD 17.2
Stappia indica SBBC 49	*Stappia indica* SBBC 49
Stappia stellulata DSM 5886	*Stappia stellulata* DSM 5886
Others	
Ahrensia marina LZD062	*Ahrensia marina* LZD062
Ahrensia sp. 13_GOM-1096m	*Ahrensia* sp. 13_GOM-1096m
Rhodovulum sp. PH10	*Rhodobacterales* bacterium PH10
Ahrensia sp. R2A130	*Rhodobacterales* bacterium R2A130

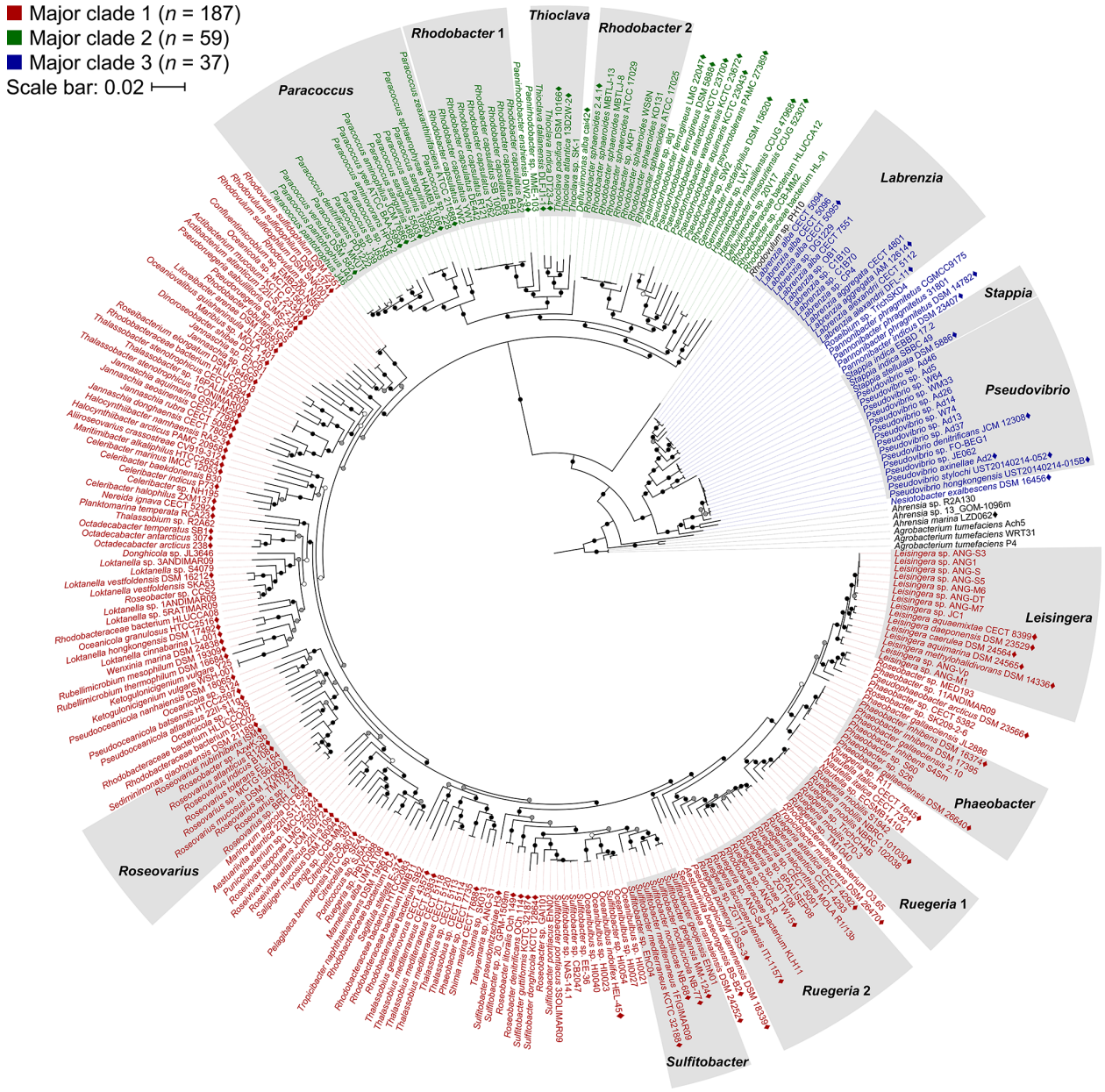

Figure C.1. The phylogenetic relationship of *Rhodobacteraceae* based on the core genome.

The tree was constructed from the concatenated alignment of 115 single-copy, protein-coding core genes (55,414 amino acid positions). Nodes with 50% or less bootstrap support are collapsed. Bootstrap support is indicated on the nodes as black circles (≥ 95%), gray circles (≥ 70%), or white circles (> 50%). Bar, 0.02 amino acid substitutions per site. The tree is split into three major clades: clade 1 (red), clade 2 (green), and clade 3 (blue). Smaller clades used in subsequent analyses are highlighted in gray. Diamonds after strain names indicate type strains.

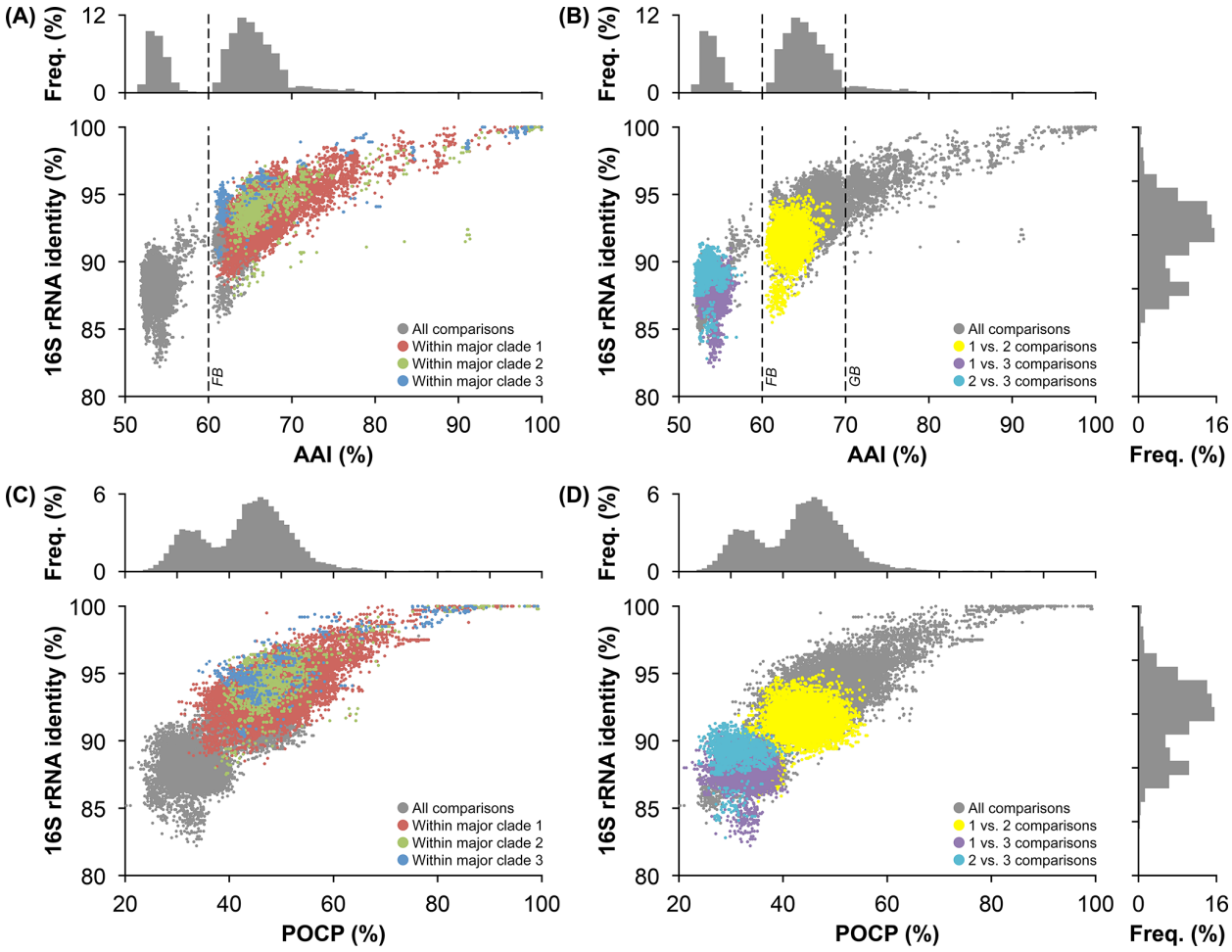

Figure C.2. Plot of 16S rRNA identity versus POCP or AAI comparisons within and between major clades of the core genome phylogenetic tree.

Plots were obtained from 41,905 pairwise comparisons (gray dots) and their frequencies (freq.) are indicated as bar graphs above (AAI or POCP) or on the right (16S rRNA identity) of each plot. (A, C) 16S rRNA identity vs. AAI intra-clade comparisons are plotted after the core genome phylogenetic tree was split into three major clades: clade 1, 187 members (red); clade 2, 59 members (green); clade 3, 37 members (blue). (B, D) Plot for 16S rRNA identity vs. POCP inter-clade comparisons: 1 vs. 2 (yellow), 1 vs. 3 (purple), 2 vs. 3 (blue). (A, B) Broken lines indicate the proposed AAI boundaries for family (FB, 60%) and genus (GB, 70%) delineations.

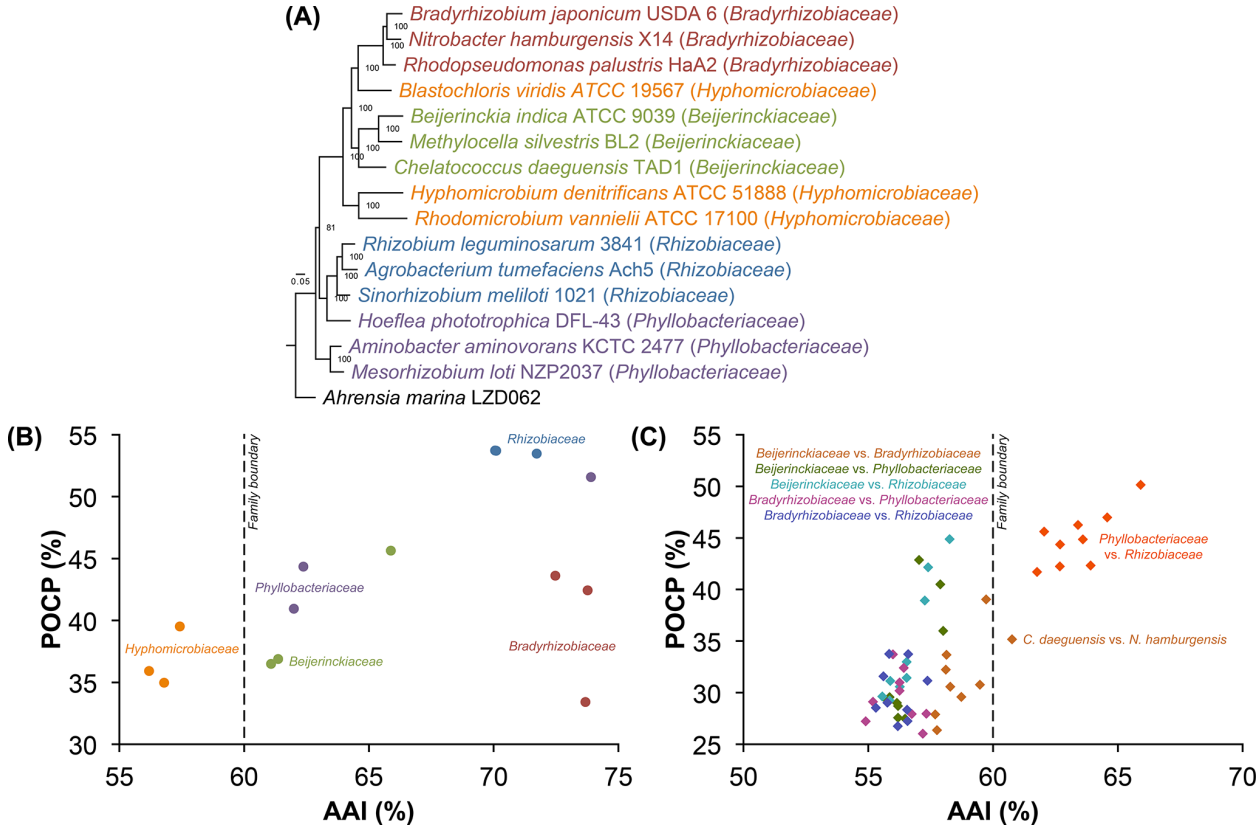

Figure C.3. Comparisons of whole-genome sequences of members of various families of the order *Rhizobiales*.

(A) The tree was constructed from the concatenated alignment of 461 single-copy, protein-coding core genes (155,842 amino acid positions) with *A. marina* LZD062^T as outgroup. Bootstrap support is indicated on the nodes. Bar, 0.05 amino acid substitutions per site. (B) POCP vs. AAI plot for intra-family comparisons. (C) POCP vs. AAI plot for inter-family comparisons (excluding comparisons with members of *Hyphomicrobiaceae*). Broken lines indicate the proposed AAI boundary for family delineation (60%).

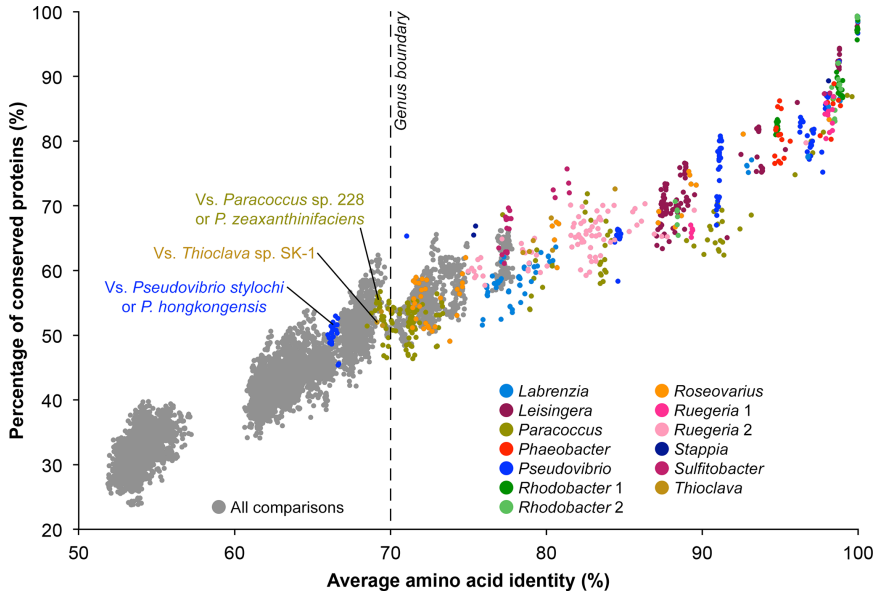

Figure C.4. Plot of POCP versus AAI comparisons within and between representative genera.

POCP vs. AAI was plotted for inter-genera comparisons (gray) and intra-genera comparisons (colored). *Ruegeria* and *Rhodobacter* are represented by two clades each (1 and 2). Broken lines indicate the proposed AAI boundary for genus delineation (70%).

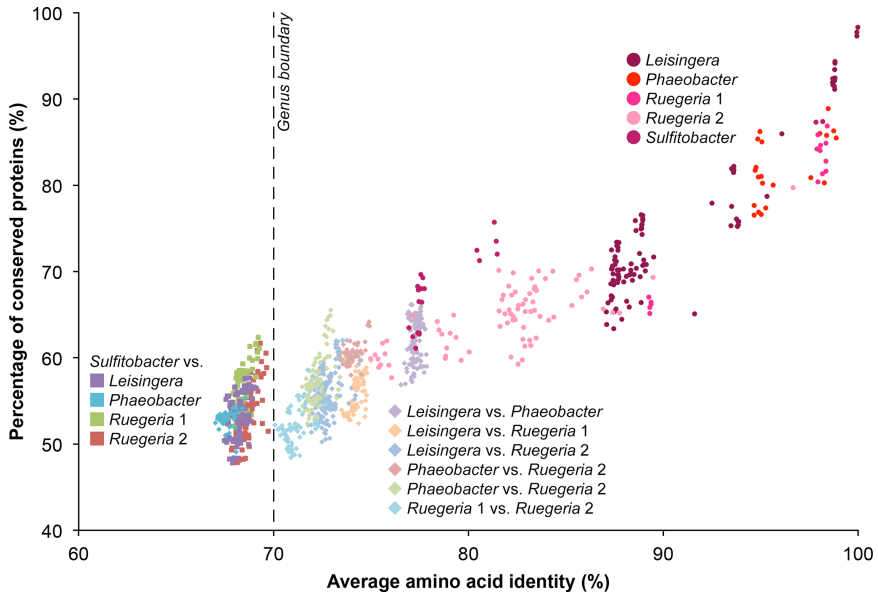

Figure C.5. Plot of POCP versus AAI comparisons within and between *Leisingera*, *Phaeobacter*, *Rugeria*, and *Sulfitobacter*.

POCP vs. AAI was plotted for inter-genera comparisons (diamonds and squares) and intra-genera comparisons (circles). *Ruegeria* is represented by two clades (1 and 2). Broken lines indicate the proposed AAI boundary for genus delineation (70%).

Appendix D: Supplementary Data for Chapter 4

Table D.1. Source and year of isolation of the *V. metoecus* and *V. cholerae* isolates used in this study.

Species and strain	Source	Geographical origin	Year of isolation	Accession number/s	Reference/s
Vibrio metoecus					
06-2478	Clinical (stool)	Mississippi, USA	2006	LCUD00000000	This study
07-2435	Clinical (leg wound)	North Carolina, USA	2007	LCUE00000000	This study
08-2459	Clinical (blood)	North Carolina, USA	2008	LCUF00000000	This study
2010V-1005	Clinical (ear)	USA	2010	LCUG00000000	This study
YB4D01	Water	Oyster Pond, MA, USA	2009	LBGO00000000	This study
YB5B04	Water	Oyster Pond, MA, USA	2009	LBGP00000000	This study
YB5B06	Water	Oyster Pond, MA, USA	2009	LBGQ00000000	This study
YB9D03	Water	Oyster Pond, MA, USA	2009	LBGR00000000	This study
OP3H	Water	Oyster Pond, MA, USA	2006	JJMN00000000	Kirchberger et al. (2014)
RC341	Water	Chesapeake Bay, MD, USA	1998	ACZT00000000	Haley et al. (2010a)
Vibrio cholerae					
YB1A01	Water	Oyster Pond, MA, USA	2009	LBCL00000000	This study
YB1G06	Water	Oyster Pond, MA, USA	2009	LBFV00000000	This study
YB2A05	Water	Oyster Pond, MA, USA	2009	LBFW00000000	This study
YB2A06	Water	Oyster Pond, MA, USA	2009	LBFX00000000	This study
YB2G01	Water	Oyster Pond, MA, USA	2009	LBFY00000000	This study
YB2G05	Water	Oyster Pond, MA, USA	2009	LBFZ00000000	This study
YB2G07	Water	Oyster Pond, MA, USA	2009	LBGA00000000	This study
YB3B05	Water	Oyster Pond, MA, USA	2009	LBGB00000000	This study
YB3G04	Water	Oyster Pond, MA, USA	2009	LBGC00000000	This study
YB4B03	Water	Oyster Pond, MA, USA	2009	LBGD00000000	This study
YB4C07	Water	Oyster Pond, MA, USA	2009	LBGE00000000	This study
YB4F05	Water	Oyster Pond, MA, USA	2009	LBGF00000000	This study
YB4G05	Water	Oyster Pond, MA, USA	2009	LBGG00000000	This study
YB4G06	Water	Oyster Pond, MA, USA	2009	LBGH00000000	This study
YB4H02	Water	Oyster Pond, MA, USA	2009	LBGI00000000	This study
YB5A06	Water	Oyster Pond, MA, USA	2009	LBGJ00000000	This study
YB6A06	Water	Oyster Pond, MA, USA	2009	LBGK00000000	This study
YB7A06	Water	Oyster Pond, MA, USA	2009	LBGL00000000	This study
YB7A09	Water	Oyster Pond, MA, USA	2009	LBGM00000000	This study
YB8E08	Water	Oyster Pond, MA, USA	2009	LBGN00000000	This study
12129(1)	Water	Australia	1985	ACFQ00000000	Chun et al. (2009)
1587	Clinical	Lima, Peru	1994	AAUR00000000	Chun et al. (2009)
2010EL-1786	Clinical	Artibonite, Haiti	2010	CP003069; CP003070	Reimer et al. (2011)
2740-80	Water	Gulf Coast, USA	1980	AAUT00000000	Chun et al. (2009)
623-39	Water	Bangladesh	2002	AAWG00000000	Chun et al. (2009)
877-163	Water	Bangladesh	2002	LBNV00000000	This study
AM-19226	Clinical	Bangladesh	2001	AATY00000000	Chun et al. (2009)
BX 330286	Water	Australia	1986	ACIA00000000	Chun et al. (2009); Haley et al. (2010b)
CIRS101	Clinical	Dhaka, Bangladesh	2002	ACVW00000000	Chun et al. (2009); Grim et al. (2010)
MAK757	Clinical	Sulawesi, Indonesia	1937	AAUS00000000	Chun et al. (2009)
MJ-1236	Clinical	Matlab, Bangladesh	1994	CP001485; CP001486	Chun et al. (2009); Grim et al. (2010)
MO10	Clinical	Madras, India	1992	AAKF00000000	Chun et al. (2009)
MZO-2	Clinical	Bangladesh	2001	AAWF00000000	Chun et al. (2009)
MZO-3	Clinical	Bangladesh	2001	AAUU00000000	Chun et al. (2009)
N16961	Clinical	Bangladesh	1971	AE003852; AE003853	Heidelberg et al. (2000)
O395	Clinical	India	1965	CP001235; CP001236	Feng et al. (2008)
RC385	Plankton	Chesapeake Bay, MD, USA	1998	AAKH00000000	Chun et al. (2009)
TM 11079-80	Sewage	Brazil	1980	ACHW00000000	Chun et al. (2009)
TMA21	Seawater	Brazil	1982	ACHY00000000	Chun et al. (2009)

Species and strain	Source	Geographical origin	Year of isolation	Accession number/s	Reference/s
Vibrio cholerae					
V51	Clinical	USA	1987	AAKI00000000	Chun et al. (2009)
V52	Clinical	Sudan	1968	AAKJ00000000	Chun et al. (2009)
VL426	Water	Maidstone, Kent, UK	Unknown	ACHV00000000	Chun et al. (2009)

Table D.2. Integron regions extracted from the *V. metoecus* and *V. cholerae* genomes.

Species and strain	Size (bp)	GC (%)	Total ORFs	Hypothetical ORFs (%)
Vibrio metoecus				
06-2478	91,284	41.1	115	52.2
07-2435	101,971	41.9	147	57.8
08-2459	140,984	41.3	208	65.9
2010V-1005	121,408	41.7	156	62.2
OP3H	117,537	41.4	156	64.1
RC341	71,256	42.0	85	56.5
YB4D01	128,849	40.7	179	69.3
YB5B04	86,814	41.5	121	71.9
YB5B06	111,519	41.0	147	72.1
YB9D03	117,178	41.3	188	69.7
Vibrio cholerae				
MZO-3	132,885	42.5	214	48.1
N16961	125,669	42.2	187	75.4
O395	116,524	41.4	199	80.4
RC385	125,710	42.0	219	42.9
TM 11079-80	62,779	42.9	106	38.7
VL426	56,384	42.0	57	47.4
YB1A01	90,204	41.4	109	64.2
YB1G06	71,516	42.0	90	50.0
YB2A06	116,704	41.6	164	64.0
YB4B03	95,111	41.7	119	56.3
YB4C07	67,955	42.4	97	48.5

Table D.3. Count of HGT events for *V. metoecus* and *V. cholerae* strains based on 376 single-copy core genes with inferred HGT.

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, $P = 0.0053$. Multiple *V. cholerae* YB strains from Oyster Pond belong to the same clonal complex (CC), and only one strain from each clonal complex was included.

Species and strain	Classification	HGT count	% of total
Vibrio metoecus			
07-2435	Clinical	151	23.05
OP3H	Environmental	55	8.40
06-2478	Clinical	48	7.33
YB4D01	Environmental	43	6.56
08-2459	Clinical	38	5.80
YB5B06	Environmental	37	5.65
2010V-1005	Clinical	31	4.73
YB5B04	Environmental	30	4.58
RC341	Environmental	29	4.43
YB9D03	Environmental	27	4.12
Vibrio cholerae			
V51	Non-O1/O139 clinical	21	3.21
RC385	Environmental	17	2.60
TMA21	Environmental	16	2.44
YB2G01 (CC 5)	Environmental	16	2.44
MZO-2	Non-O1/O139 clinical	10	1.53
YB4F05 (CC 3)	Environmental	9	1.37
1587	Non-O1/O139 clinical	8	1.22
877-163	Environmental	7	1.07
AM-19226	Non-O1/O139 clinical	6	0.92
MAK757	O1/O139 clinical	6	0.92
2740-80	Environmental	5	0.76
BX 330286	Environmental	4	0.61
O395	O1/O139 clinical	4	0.61
TM 11079-80	Environmental	4	0.61
YB4B03 (CC 2)	Environmental	4	0.61
12129(1)	Environmental	3	0.46
2010EL-1786	O1/O139 clinical	3	0.46
CIRS101	O1/O139 clinical	3	0.46
MJ-1236	O1/O139 clinical	3	0.46
MO10	O1/O139 clinical	3	0.46
N16961	O1/O139 clinical	3	0.46
623-39	Environmental	2	0.31
MZO-3	Non-O1/O139 clinical	2	0.31
V52	Non-O1/O139 clinical	2	0.31
VL426	Environmental	2	0.31
YB7A06 (CC 4)	Environmental	2	0.31
YB3B05 (CC 1)	Environmental	1	0.15

Classification	HGT count	% of total
Vibrio metoecus		
Clinical	268	40.92
Environmental	221	33.74
Vibrio metoecus total	489	74.66
Vibrio cholerae		
Environmental	92	14.05
Non-O1/O139 clinical	49	7.48
O1/O139 clinical	25	3.82
Vibrio cholerae total	166	25.34
Total	655	100.00

Table D.4. ANI of the core genome of the *V. cholerae* isolates from Oyster Pond.

ANI was determined from the pairwise comparison of 1,560 single-copy core genes (≈ 1.42 mbp). Members of the different clonal complexes (CC) are indicated. Highlighted in gray are pairwise ANI comparisons within clonal complexes.

	YB3B05 (CC 1)	**YB4C07 (CC 1)**	**YB8E08 (CC 1)**	**YB2G07 (CC 2)**	**YB4B03 (CC 2)**	**YB4G05 (CC 2)**	**YB1G06 (CC 3)**	**YB2A05 (CC 3)**	**YB2G05 (CC 3)**	**YB4F05 (CC 3)**	**YB6A06 (CC 3)**	**YB7A09 (CC 3)**	**YB1A01 (CC 4)**	**YB5A06 (CC 4)**	**YB7A06 (CC 4)**	**YB2A06 (CC 5)**	**YB2G01 (CC 5)**	**YB3G04 (CC 5)**	**YB4G06 (CC 5)**	**YB4H02 (CC 5)**
YB3B05		100	100	99	99	99	98	98	98	98	98	98	98	98	98	98	98	98	98	98
YB4C07	100		100	99	99	99	98	98	98	98	98	98	98	98	98	98	98	98	98	98
YB8E08	100	100		99	99	99	98	98	98	98	98	98	98	98	98	98	98	98	98	98
YB2G07	99	99	99		100	100	98	98	98	98	98	98	98	98	98	98	98	98	98	98
YB4B03	99	99	99	100		100	98	98	98	98	98	98	98	98	98	98	98	98	98	98
YB4G05	99	99	99	100	100		98	98	98	98	98	98	98	98	98	98	98	98	98	98
YB1G06	98	98	98	98	98	98		100	100	100	100	100	99	99	99	98	98	98	98	98
YB2A05	98	98	98	98	98	98	100		100	100	100	100	99	99	99	98	98	98	98	98
YB2G05	98	98	98	98	98	98	100	100		100	100	100	99	99	99	98	98	98	98	98
YB4F05	98	98	98	98	98	98	100	100	100		100	100	99	99	99	98	98	98	98	98
YB6A06	98	98	98	98	98	98	100	100	100	100		100	99	99	99	98	98	98	98	98
YB7A09	98	98	98	98	98	98	100	100	100	100	100		99	99	99	98	98	98	98	98
YB1A01	98	98	98	98	98	98	99	99	99	99	99	99		100	100	98	98	98	98	98
YB5A06	98	98	98	98	98	98	99	99	99	99	99	99	100		100	98	98	98	98	98
YB7A06	98	98	98	98	98	98	99	99	99	99	99	99	100	100		98	98	98	98	98
YB2A06	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98		100	100	100	100
YB2G01	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	100		100	100	100
YB3G04	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	100	100		100	100
YB4G06	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	100	100	100		100
YB4H02	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	100	100	100	100	

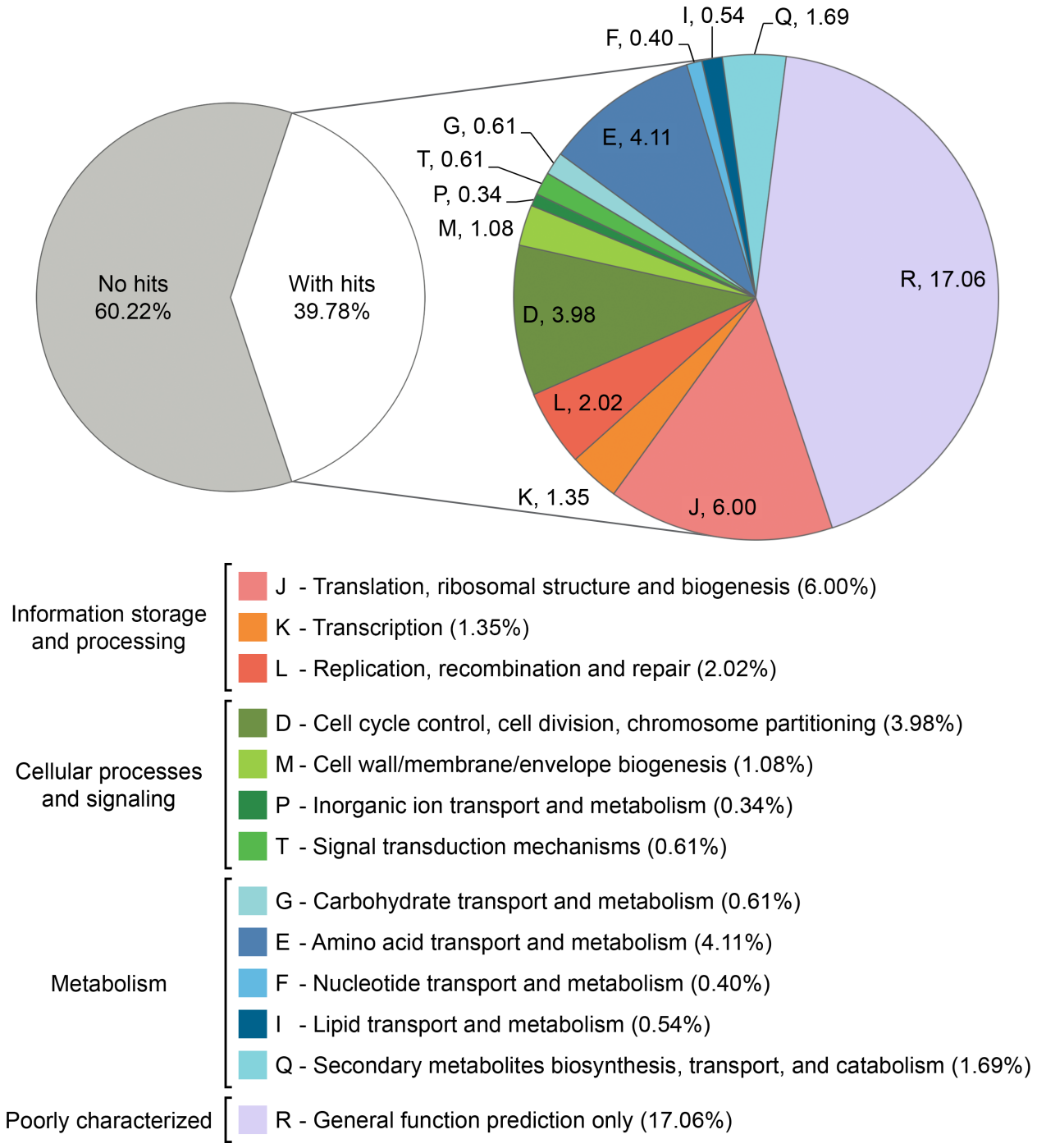

Figure D.1. Predicted functional roles of the 1,452 *V. metoecus* and *V. cholerae* chromosomal integron gene cassettes.

Functions predicted from 116 gene families based on the COG database. The numbers on the pie chart represent percentage of genes in a category. Genes categorized as "function unknown" or not assigned into a COG category because there is no similarity to the database are combined ("no hits," gray), whereas the genes with assigned COG categories ("with hits," white) belong to four major groups: information storage and processing (reds), cellular processes and signaling (greens), metabolism (blues), and poorly characterized genes with general function prediction only (purple).

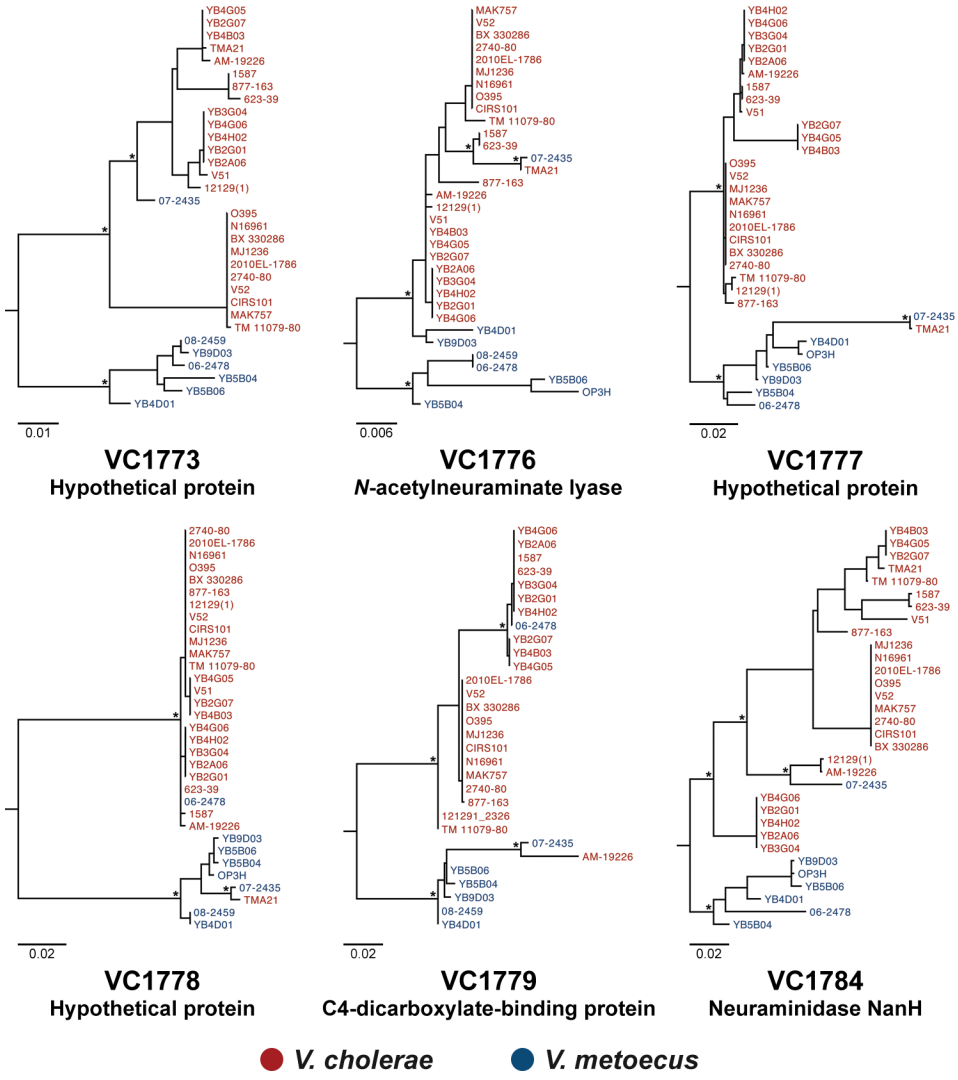

Figure D.2. Phylogenetic trees of orthologous gene families of islet B (*nan-nag* cluster, VC1773-VC1784) of the VPI-2.

Six out of eleven gene families have reliable bootstrap support. Relevant bootstrap support (≥ 70%) is indicated with *. Bar, nucleotide substitutions per site.

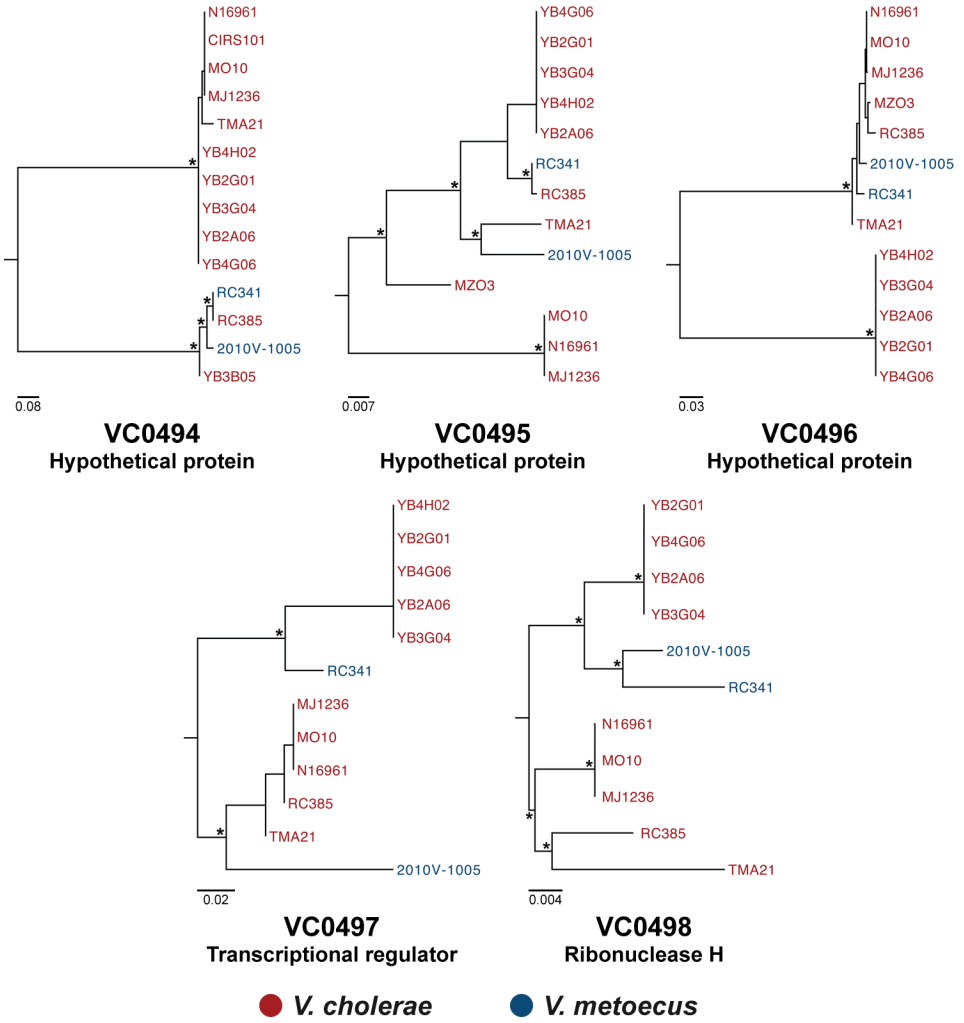

Figure D.3. Phylogenetic trees of orthologous gene families of the core VSP-2 (5' end, VC0494-VC0498).

Five out of five gene families have reliable bootstrap support. Relevant bootstrap support (≥ 70%) is indicated with *. Bar, nucleotide substitutions per site. The 3' end (VC0504-VC0510) did not produce trees with reliable bootstrap support.

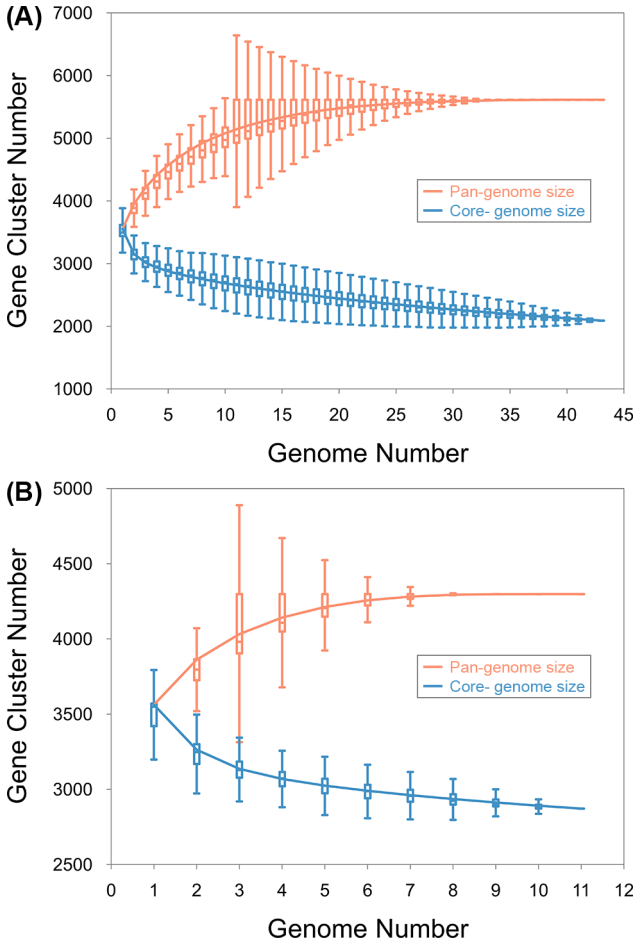

Figure D.4. The pan-genome and core genome profiles for (A) *V. cholerae* and (B) *V. metoecus*.

Profiles are based on (A) 42 *V. cholerae* genomes and (B) 10 *V. metoecus* genomes. Graphs show the number of gene clusters (families) with every addition of a genome into the gene pool.

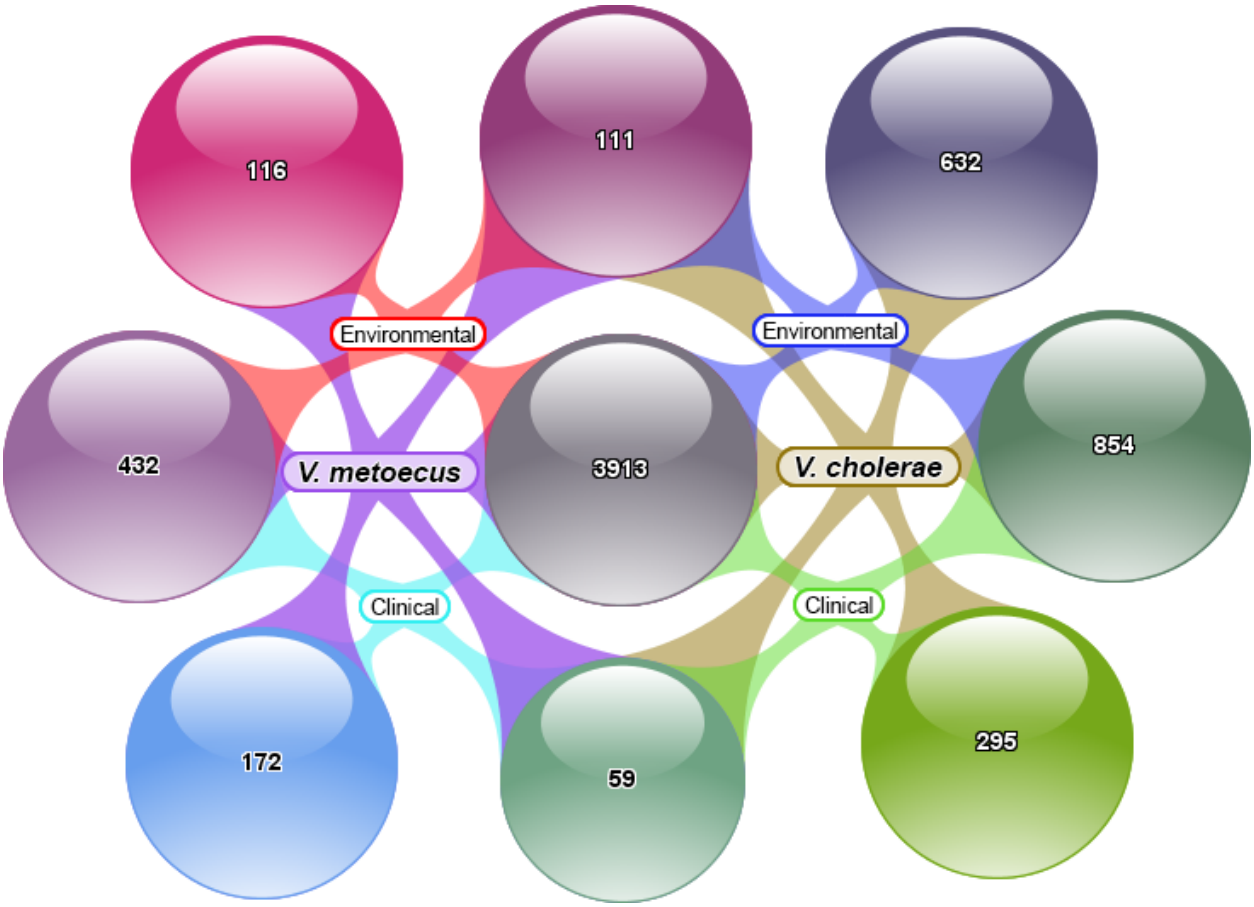

Figure D.5. Cluster diagram of all orthologous gene families of the *V. metoecus*–*V. cholerae* pan-genome.

Numbers indicate gene families containing members of various groups of strains. Clockwise from top left bubble: environmental *V. metoecus* only (n = 116), environmental *V. metoecus* and *V. cholerae* (n = 111), environmental *V. cholerae* only (n = 632), environmental and clinical *V. cholerae* (n = 854), clinical *V. cholerae* only (n = 295), clinical *V. metoecus* and *V. cholerae* (n = 59), clinical *V. metoecus* only (n = 172), environmental and clinical *V. metoecus* (n = 432). The gene families in the middle bubble contain environmental and clinical *V. metoecus* and *V. cholerae* members (n = 3,913).

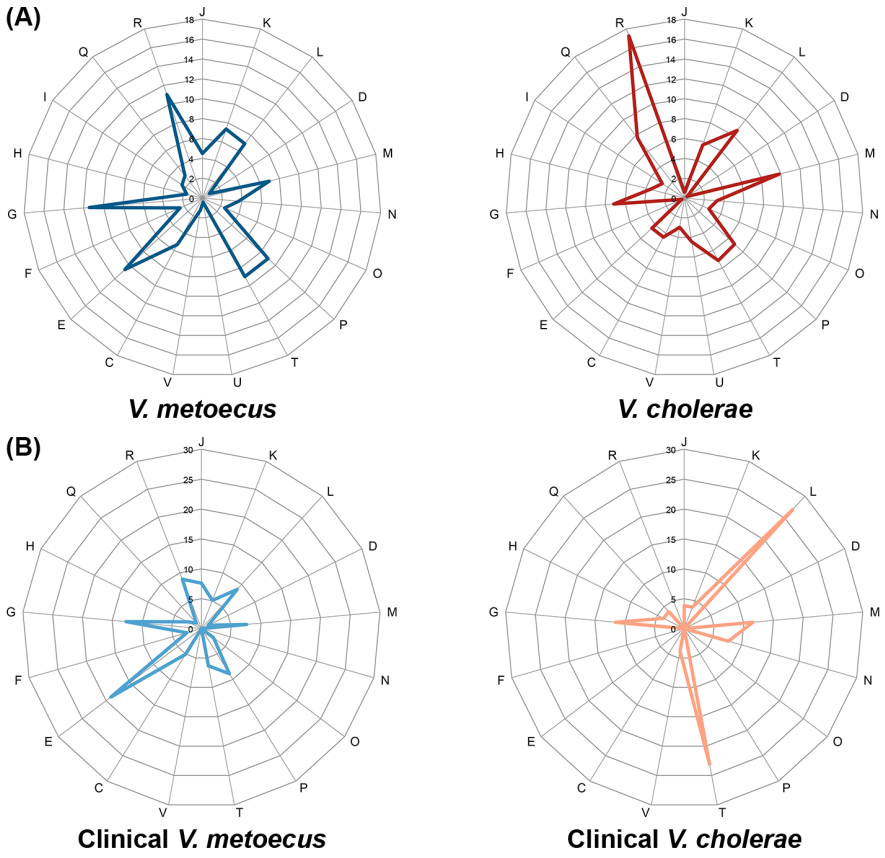

Figure D.6. Comparison of orthologous gene families and their abundance, uniquely found in (A) *V. metoecus* or *V. cholerae* and (B) clinical *V. metoecus* or clinical *V. cholerae*.

Functional roles for genes were predicted based on the COG database. The numbers on the charts represent percentage of genes in a COG category. Genes categorized as "function unknown" or not assigned into a COG category because there is no similarity to the database are not included.

References

Chun, J., Grim, C.J., Hasan, N.A., Lee, J.H., Choi, S.Y., Haley, B.J., et al. (2009). Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 106(36), 15442-15447.

Feng, L., Reeves, P.R., Lan, R., Ren, Y., Gao, C., Zhou, Z., et al. (2008). A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS One* 3(12), e4053.

Grim, C.J., Hasan, N.A., Taviani, E., Haley, B., Chun, J., Brettin, T.S., et al. (2010). Genome sequence of hybrid *Vibrio cholerae* O1 MJ-1236, B-33, and CIRS101 and comparative genomics with *V. cholerae*. *Journal of Bacteriology* 192(13), 3524-3533.

Haley, B.J., Grim, C.J., Hasan, N.A., Choi, S.Y., Chun, J., Brettin, T.S., et al. (2010a). Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. *BMC Microbiology* 10, 154.

Haley, B.J., Grim, C.J., Hasan, N.A., Taviani, E., Chun, J., Brettin, T.S., et al. (2010b). The pre-seventh pandemic *Vibrio cholerae* BX 330286 El Tor genome: evidence for the environment as a genome reservoir. *Environmental Microbiology Reports* 2(1), 208-216.

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., et al. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406(6795), 477-483.

Kirchberger, P.C., Turnsek, M., Hunt, D.E., Haley, B.J., Colwell, R.R., Polz, M.F., et al. (2014). *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. *International Journal of Systematic and Evolutionary Microbiology* 64(9), 3208-3214.

Reimer, A.R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., et al. (2011). Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerging Infectious Diseases* 17(11), 2113-2121.

Appendix E: Supplementary Data for Chapter 5

Table E.1. Core housekeeping genes used to determine genome completeness.

Completeness is the percentage of these genes present in each *V. cholerae* or *V. metoecus* genome. The accession numbers of genes are from the reference genome *V. cholerae* N16961.

Gene name	Gene description	Accession no.
adk	Adenylate kinase	NP_230632
alas	Alanyl-tRNA synthetase	NP_230196
argS	Arginyl-tRNA synthetase	NP_231706
aroA	5-Enolpyruvylshikimate-3-phosphate synthetase	NP_231368
aroC	Chorismate synthase	NP_231747
aspS	Aspartyl-tRNA synthetase	NP_230811
birA	Biotin-[acetylCoA carboxylase] holoenzyme synthetase	NP_229973
coaE	Dephospho-CoA kinase	NP_232057
cysS	Cysteinyl-tRNA synthetase	NP_231482
dfp	Fused 4'-phosphopantothenolcysteine decarboxylase	NP_229872
dnaE	DNA polymerase III, alpha subunit	NP_231876
dnaN	DNA polymerase III, beta subunit	NP_062597
efp	Elongation factor EF-P	NP_230854
eno	Enolase	NP_232076
ffh	Signal recognition particle (SRP) component with 4.5S RNA	NP_230211
folP	7,8-Dihydropteroate synthase	NP_230287
frr	Ribosome recycling factor	NP_231888
ftsY	Fused signal recognition particle (SRP) receptor	NP_229805
fusA	Protein chain elongation factor EF-G, GTP-binding	NP_230015
gyrB	DNA gyrase subunit B	NP_229675
ileS	Isoleucyl-tRNA synthetase	NP_230331
infB	Fused protein chain initiation factor 2, IF-2	NP_230292
infC	Protein chain initiation factor, IF-3	NP_232684
lepA	GTP-binding membrane protein	NP_232092
leuS	Leucyl-tRNA synthetase	NP_230603
mdh	Malate dehydrogenase	NP_230086
metG	Methionyl-tRNA synthetase	NP_230681
ndk	Nucleoside diphosphate kinase	NP_230405
nusB	Transcription antitermination protein	NP_231898
pgi	Glucose-6-phosphate isomerase	NP_230028
pheS	Phenylalanine tRNA synthetase, alpha subunit	NP_230864
pheT	Phenylalanine tRNA synthetase, beta subunit	NP_230865
pncB	Nicotinate phosphoribosyltransferase	NP_232499
pnp	Polynucleotide phosphorylase/polyadenylase	NP_230296
prfA	Peptide chain release factor RF-1	NP_231810
priA	Primosome factor n' (replication factor Y)	NP_232306
pros	Prolyl-tRNA hydrolase	NP_230522
pstA	Phosphate transporter subunit	NP_232473
pth	Peptidyl-tRNA hydrolase	NP_231815
purA	Adenylosuccinate synthetase	NP_232230
pyrB	Aspartate carbamoyltransferase, catalytic subunit	NP_232139
pyrF	Orotidine-5'-phosphate decarboxylase	NP_231545
pyrG	CTP synthetase	NP_232077
pyrH	Uridylate kinase	NP_231889
recA	DNA strand exchange and recombination protein	NP_230194
recG	ATP-dependent DNA helicase	NP_232338

Gene name	Gene description	Accession no.
purC	Phosphoribosylaminoimidazole-succinocarboxamide synthetase	NP_230835
rnc	RNase III	NP_232090
rnhB	Ribonuclease HII	NP_231877
rplA	50S ribosomal subunit protein L1	NP_229979
rplB	50S ribosomal subunit protein L2	NP_232221
rplC	50S ribosomal subunit protein L3	NP_232224
rplD	50S ribosomal subunit protein L4	NP_232223
rplE	50S ribosomal subunit protein L5	NP_232212
rplF	50S ribosomal subunit protein L6	NP_232209
rplK	50S ribosomal subunit protein L11	NP_229978
rplL	50S ribosomal subunit protein L7/12	NP_229981
rplM	50S ribosomal subunit protein L13	NP_230221
rplN	50S ribosomal subunit protein L14	NP_232214
rplO	50S ribosomal subunit protein L15	NP_232205
rplP	50S ribosomal subunit protein L16	NP_232217
rplQ	50S ribosomal subunit protein L17	NP_232198
rplR	50S ribosomal subunit protein L18	NP_232208
rplS	50S ribosomal subunit protein L19	NP_230215
rplT	50S ribosomal subunit protein L20	NP_232686
rplV	50S ribosomal subunit protein L22	NP_232219
rplX	50S ribosomal subunit protein L24	NP_232213
rpoA	RNA polymerase, alpha subunit	NP_232199
rpoB	RNA polymerase, beta subunit	NP_229982
rpsB	30S ribosomal subunit protein S2	NP_231891
rpsC	30S ribosomal subunit protein S3	NP_232218
rpsD	30S ribosomal subunit protein S4	NP_232200
rpsE	30S ribosomal subunit protein S5	NP_232207
rpsG	30S ribosomal subunit protein S7	NP_230014
rpsH	30S ribosomal subunit protein S8	NP_232210
rpsI	30S ribosomal subunit protein S9	NP_230222
rpsJ	30S ribosomal subunit protein S10	NP_232225
rpsK	30S ribosomal subunit protein S11	NP_232201
rpsL	30S ribosomal subunit protein S12	NP_230013
rpsM	30S ribosomal subunit protein S13	NP_232202
ruvA	Component of RuvABC resolvasome, regulatory subunit	NP_231480
ruvB	DNA helicase, component of RuvABC resolvasome	NP_231479
secA	Preprotein translocase subunit, ATPase	NP_232024
secY	Preprotein translocase membrane subunit	NP_232204
serS	Seryl-tRNA synthetase	NP_230755
smpB	Trans-translation protein	NP_230495
tgt	tRNA-guanine transglycosylase	NP_230390
thrS	Threonyl-tRNA synthetase	NP_232683
tig	Peptidyl-prolyl cis/trans isomerase A	NP_231930
toxR	Cholera toxin transcriptional activator	NP_230630
tpiA	Triosephophate isomerase	NP_232298
trmD	tRNA (guanine-1-) -methyltransferase	NP_230214
trpA	Tryptophan synthase, alpha subunit	NP_230814
trpC	Indole-3-glycerol phosphate synthase	NP_230816
trpS	Tryptophanyl-tRNA synthetase	NP_232251
truB	tRNA pseudouridine synthase B	NP_230294
tsf	Protein chain elongation factor EF-Ts	NP_231890
tyrS	Tyrosyl-tRNA synthetase	NP_230119
uvrA	ATPase and DNA damage recognition protein	NP_230048
uvrB	Exonuclease of nucleotide excision repair	NP_230664
uvrB	Exonuclease of nucleotide excision repair	NP_230664

Gene name	Gene description	Accession no.
valS	Valyl-tRNA synthetase	NP_232132
yabC	S-Adenosyl-dependent methyltransferase activity	NP_232039
ybeY	Conserved protein	NP_230607
yidC	Cytoplasmic insertase, Sec system	NP_062588

Table E.2. ANI (%) from the pairwise comparisons* of the *V. cholerae* (Vc) and *V. metoecus* (Vm) whole genome sequences.

	Vc N16961	Vc OYP1G01	Vc OYP2A12	Vc OYP2E01	Vc OYP3B05	Vc OYP3F10	Vc OYP4B01	Vc OYP4C07	Vc OYP4G08	Vc OYP4H06	Vc OYP4H11	Vc OYP6D06	Vc OYP6E07	Vc OYP6F08	Vc OYP6F10	Vc OYP7C09	Vc OYP8C06	Vc OYP8F12
Vc N16961		98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Vc OYP1G01	98		98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Vc OYP2A12	98	98		98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Vc OYP2E01	98	98	98		98	98	98	98	98	98	100	98	98	100	98	98	98	98
Vc OYP3B05	98	98	98	98		98	98	100	98	98	98	98	98	98	98	98	98	98
Vc OYP3F10	98	98	98	98	98		98	98	98	98	98	98	98	98	98	98	98	98
Vc OYP4B01	98	98	98	98	98	98		98	98	100	98	98	98	98	98	98	98	98
Vc OYP4C07	98	98	98	98	100	98	98		98	98	98	98	98	98	98	98	98	98
Vc OYP4G08	98	98	98	98	98	98	98	98		98	98	98	98	98	98	98	98	98
Vc OYP4H06	98	98	98	98	98	98	100	98	98		98	98	98	98	98	98	98	98
Vc OYP4H11	98	98	98	100	98	98	98	98	98	98		98	98	100	98	98	98	98
Vc OYP6D06	98	98	98	98	98	98	98	98	98	98	98		98	98	98	98	98	98
Vc OYP6E07	98	98	98	98	98	98	98	98	98	98	98	98		98	98	98	98	98
Vc OYP6F08	98	98	98	100	98	98	98	98	98	98	100	98	98		98	98	98	98
Vc OYP6F10	98	98	98	98	98	98	98	98	98	98	98	98	98	98		100	98	98
Vc OYP7C09	98	98	98	98	98	98	98	98	98	98	98	98	98	98	100		98	98
Vc OYP8C06	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98		100
Vc OYP8F12	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	100	
Vm OP3H	86	86	86	86	86	87	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP4D01	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP4E03	86	86	86	86	86	87	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP5B04	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP5B06	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP5H08	86	86	86	86	86	87	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP8G05	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP8G09	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP8G12	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP8H05	86	86	86	86	86	87	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP9B03	86	86	86	86	86	87	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP9B09	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP9C12	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP9D03	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP9D09	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP9E03	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vm OYP9E10	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86

*Intra-species pairwise comparisons average = 98.24% (*V. cholerae*), 97.91% (*V. metoecus*); inter-species pairwise comparisons average = 86.23%.

Table E.3. ANI (%) from the pairwise comparisons* of the *V. metoecus* (Vm) and *V. cholerae* (Vc) whole genome sequences.

	Vm OP3H	Vm OYP4D01	Vm OYP4E03	Vm OYP5B04	Vm OYP5B06	Vm OYP5H08	Vm OYP8G05	Vm OYP8G09	Vm OYP8G12	Vm OYP8H05	Vm OYP9B03	Vm OYP9B09	Vm OYP9C12	Vm OYP9D03	Vm OYP9D09	Vm OYP9E03	Vm OYP9E10
Vm OP3H		98	97	97	98	98	97	98	97	97	97	97	97	97	97	97	97
Vm OYP4D01	98		98	98	98	100	98	98	98	98	98	98	98	98	97	98	98
Vm OYP4E03	97	97		97	97	97	98	97	98	98	98	98	98	98	100	98	98
Vm OYP5B04	97	97	97		98	98	98	98	98	98	98	98	98	98	97	98	98
Vm OYP5B06	98	98	98	98		98	98	100	98	98	98	98	98	98	98	98	98
Vm OYP5H08	98	100	97	98	98		98	98	98	98	98	98	98	98	97	98	98
Vm OYP8G05	97	98	98	98	98	98		98	98	98	98	98	99	98	98	98	98
Vm OYP8G09	98	98	97	98	100	98	98		98	98	98	98	98	98	97	98	98
Vm OYP8G12	97	98	97	98	98	98	98	98		98	98	98	98	98	97	98	98
Vm OYP8H05	97	98	98	98	98	98	98	98	98		100	98	98	98	98	98	98
Vm OYP9B03	97	97	98	97	98	97	98	98	97	100		98	98	98	98	98	98
Vm OYP9B09	97	98	97	98	98	98	98	98	98	98	98		98	98	97	98	98
Vm OYP9C12	97	98	98	98	98	98	99	98	98	98	98	98		98	98	98	98
Vm OYP9D03	97	98	98	98	98	98	98	98	98	98	98	98	98		98	100	100
Vm OYP9D09	97	98	100	97	98	98	98	98	98	98	98	98	98	98		98	98
Vm OYP9E03	97	98	98	98	98	98	98	98	98	98	98	98	98	100	98		100
Vm OYP9E10	97	98	98	98	98	98	98	98	98	98	98	98	98	100	98	100	
Vc N16961	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP1G01	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP2A12	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP2E01	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP3B05	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP3F10	87	87	87	86	86	87	86	86	86	87	87	86	86	86	87	86	86
Vc OYP4B01	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP4C07	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP4G08	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP4H06	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP4H11	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP6D06	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP6E07	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP6F08	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP6F10	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP7C09	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP8C06	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
Vc OYP8F12	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86

*Intra-species pairwise comparisons average = 97.91% (*V. metoecus*), 98.24% (*V. cholerae*); inter-species pairwise comparisons average = 86.23%.

Table E.4. Count of horizontal gene transfer (HGT) events in the core genome of *V. cholerae* and *V. metoecus*.

Transfer events were counted from 554 single-copy core gene phylogenetic trees, where *V. cholerae* and *V. metoecus* members cannot be partitioned into two perfect clades. The values indicate the number of times a strain is a recipient of a gene transfer event, where that strain clustered with members of the other species (donor) in a tree with reliable bootstrap support (≥ 70%).

Species and strain	HGT count	Percent of total
Vibrio cholerae		
OYP6E07	42	3.1
OYP8C06	32	2.3
OYP8F12	32	2.3
OYP4G08	27	2.0
OYP1G01	24	1.8
OYP7C09	18	1.3
OYP3F10	17	1.2
OYP4H11	17	1.2
OYP6F10	17	1.2
OYP2A12	16	1.2
OYP4B01	16	1.2
OYP6D06	16	1.2
OYP4H06	15	1.1
OYP2E01	14	1.0
OYP4C07	14	1.0
OYP3B05	13	1.0
OYP6F08	11	0.8
Vibrio cholerae total	341	24.9
Vibrio metoecus		
OP3H	85	6.2
OYP5H08	73	5.3
OYP4D01	72	5.3
OYP4E03	67	4.9
OYP9D09	67	4.9
OYP8H05	66	4.8
OYP9B03	64	4.7
OYP5B06	63	4.6
OYP8G09	63	4.6
OYP8G12	63	4.6
OYP9B09	59	4.3
OYP5B04	58	4.2
OYP8G05	53	3.9
OYP9C12	47	3.4
OYP9D03	43	3.1
OYP9E10	43	3.1
OYP9E03	41	3.0
Vibrio metoecus total	1,027	75.1
Total	1,368	100.0

Table E.5. Count of horizontal gene transfer (HGT) events in the accessory genome of *V. cholerae* and *V. metoecus*.

Transfer events were counted from 175 phylogenetic trees, where *V. cholerae* and *V. metoecus* members cannot be partitioned into two perfect clades. The values indicate the number of times a strain is a recipient of a gene transfer event, where that strain clustered with members of the other species (donor) in a tree with reliable bootstrap support (≥ 70%).

Species and strain	HGT count	Percent of total
Vibrio cholerae		
OYP3F10	43	5.5
OYP8C06	30	3.8
OYP8F12	28	3.6
OYP6E07	25	3.2
OYP1G01	24	3.1
OYP3B05	22	2.8
OYP4G08	22	2.8
OYP4C07	20	2.6
OYP6F08	19	2.4
OYP7C09	18	2.3
OYP2A12	17	2.2
OYP6F10	16	2.0
OYP2E01	15	1.9
OYP4B01	15	1.9
OYP4H06	14	1.8
OYP4H11	14	1.8
OYP6D06	12	1.5
Vibrio cholerae total	354	45.2
Vibrio metoecus		
OYP9B09	48	6.1
OP3H	32	4.1
OYP5B04	32	4.1
OYP4E03	30	3.8
OYP9D09	30	3.8
OYP8G12	28	3.6
OYP9C12	26	3.3
OYP8H05	25	3.2
OYP9B03	24	3.1
OYP5H08	23	2.9
OYP8G05	23	2.9
OYP8G09	22	2.8
OYP4D01	21	2.7
OYP9E03	20	2.6
OYP5B06	17	2.2
OYP9E10	15	1.9
OYP9D03	13	1.7
Vibrio metoecus total	429	54.8
Total	783	100.0

Table E.6. Result of the transformation experiment to determine competency of *V. cholerae* (Vc) and *V. metoecus* (Vm).

All samples were done in duplicate. Naturally competent *V. cholerae* A1552 with and without added genomic DNA served as controls. Colony counts are reported in colony forming units (CFU) per mL of culture. Transformation frequency (TF) was calculated by dividing the number of transformants by the total number of bacteria.

Species and strain	Transformants 1 (CFU/mL)	Transformants 2 (CFU/mL)	Total 1 (CFU/mL)	Total 2 (CFU/mL)	TF 1	TF 2	Average TF	Standard deviation
Vc OYP1G01	1.60E+05	1.60E+05	3.80E+08	3.20E+08	4.21E–04	5.00E–04	4.61E–04	5.58E–05
Vc OYP2A12	1.00E+04	8.00E+03	2.60E+08	3.00E+08	3.85E–05	2.67E–05	3.26E–05	8.34E–06
Vc OYP2E01	0	0	1.00E+08	1.20E+08	0	0	0	0
Vc OYP3B05	0	0	3.40E+08	3.20E+08	0	0	0	0
Vc OYP3F10	0	0	2.80E+08	3.20E+08	0	0	0	0
Vc OYP4B01	0	0	2.60E+08	2.40E+08	0	0	0	0
Vc OYP4C07	0	0	2.80E+08	3.80E+08	0	0	0	0
Vc OYP4G08	0	0	2.40E+08	3.80E+08	0	0	0	0
Vc OYP4H11	0	0	2.60E+08	3.40E+08	0	0	0	0
Vc OYP6D06	0	0	2.60E+08	2.60E+08	0	0	0	0
Vc OYP6F08	0	0	3.00E+08	3.40E+08	0	0	0	0
Vc OYP6F10	0	0	2.80E+08	3.00E+08	0	0	0	0
Vc OYP7C09	0	0	3.80E+08	4.00E+08	0	0	0	0
Vc OYP8C06	0	0	1.80E+08	2.20E+08	0	0	0	0
Vc OYP8F12	0	0	3.80E+08	4.00E+08	0	0	0	0
Vm OYP4D01	0	0	3.40E+08	3.80E+08	0	0	0	0
Vm OYP4E03	0	0	3.80E+08	3.20E+08	0	0	0	0
Vm OYP5B04	0	0	4.00E+08	4.20E+08	0	0	0	0
Vm OYP5B06	0	0	3.00E+08	3.40E+08	0	0	0	0
Vm OYP8G05	0	0	3.40E+08	4.40E+08	0	0	0	0
Vm OYP8G09	0	0	3.40E+08	3.20E+08	0	0	0	0
Vm OYP8G12	0	0	4.40E+08	4.20E+08	0	0	0	0
Vm OYP8H05	0	0	4.00E+08	3.60E+08	0	0	0	0
Vm OYP9B03	0	0	1.80E+08	2.40E+08	0	0	0	0
Vm OYP9B09	0	0	3.20E+08	3.00E+08	0	0	0	0
Vm OYP9C12	0	0	3.80E+08	3.40E+08	0	0	0	0
Vm OYP9D09	0	0	2.80E+08	2.60E+08	0	0	0	0
Vm OYP9E03	0	0	3.80E+08	3.40E+08	0	0	0	0
Vm OYP9E10	0	0	2.00E+08	2.80E+08	0	0	0	0
Vc A1552	2.40E+05	3.60E+05	5.60E+08	4.80E+08	4.29E–04	7.50E–04	5.89E–04	2.27E–04
Vc A1552 (No DNA)	0	0	4.60E+08	5.00E+08	0	0	0	0

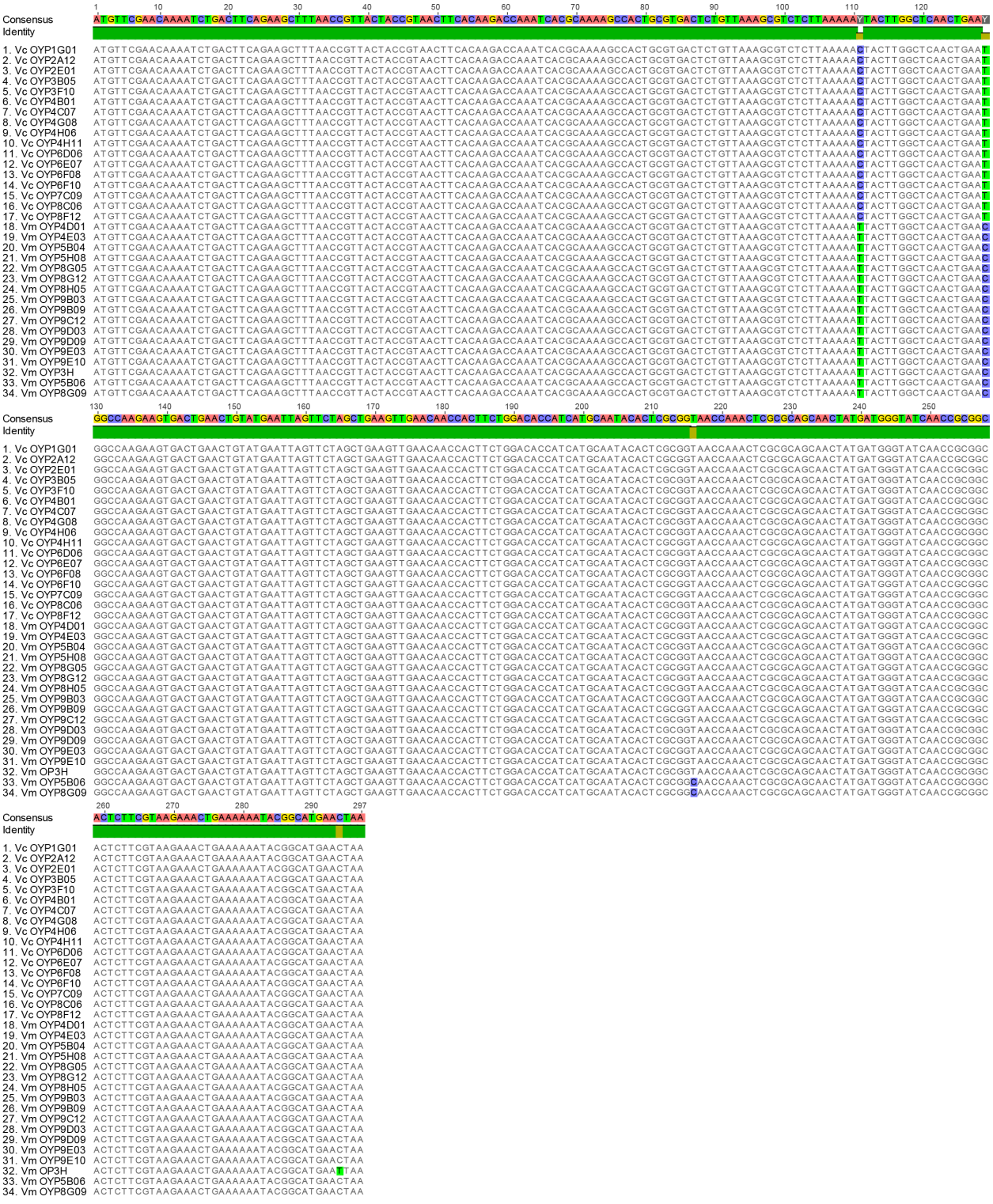

Figure E.1. Alignment of the VC0290 gene sequences (297 bp) from *V. cholerae* (Vc) and *V. metoecus* (Vm).

For each sequence, colored nucleotides are those that do not match the consensus (majority) or no specific base call can be made for that position (e.g., 50/50 C and T). Green bars represent regions in the alignment with 100% identity (yellow for < 100% identity).

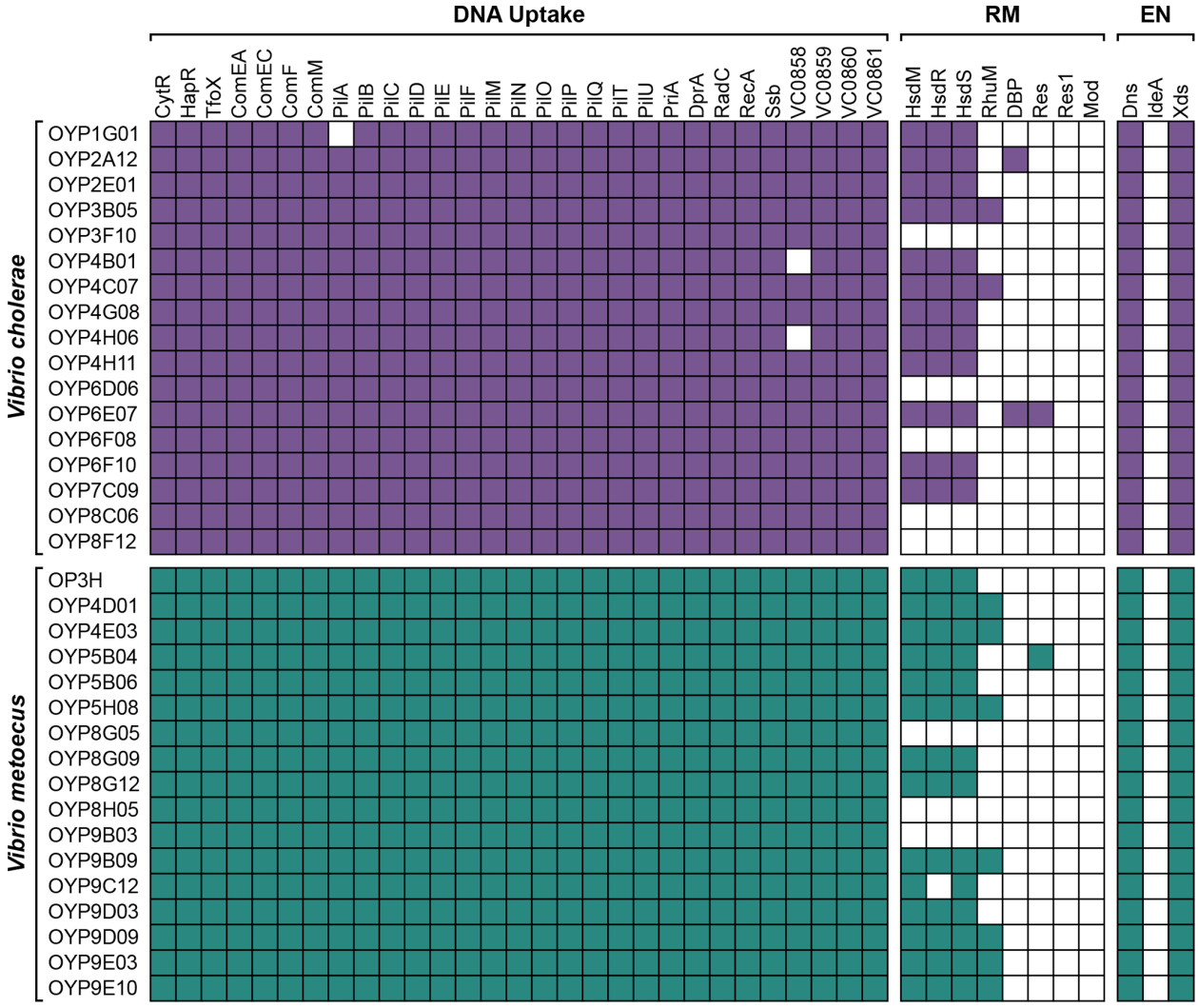

Figure E.2. Presence/absence map of genes involved in DNA uptake and restriction-modification (RM), as well as endo- and exonucleases (EN) in the *V. cholerae* and *V. metoecus* genomes.

Colored squares represent the presence of a homologue based on BLAST sequence similarity search; white squares represent the absence of genes.

Appendix F: Supplementary Data for Chapter 6

Table F.1. A chronological timeline of events involving the cholera outbreak in Haiti (2010–2014).

Highlighted in gray are significant studies on determining the origin of the cholera outbreak.

2010	
July 28 – August 14	1,400 cases of cholera were reported in Nepalgunj, Nepal, a city close to the Nepal-India border (Integrated Regional Information Networks, 2014).
September – October	Nepalese MINUSTAH troops underwent medical examination. None of the troops showed symptoms of severe diarrhea and dehydration (Lantagne et al., 2014).
September 23	A surge in cholera cases occurred in Kathmandu, Nepal, where the MINUSTAH troops trained shortly prior to deployment to Haiti (Maharjan, 2010).
October 8-24	MINUSTAH troops arrived in Haiti from Nepal. They set up camp in Meille, Artibonite Department, Haiti (Lantagne et al., 2014).
October 12	The first case of cholera was reported – a man who bathed and drank water from the Latem River in Mirebalais, a town two kilometers from Meille (Ivers and Walton, 2012).
October 19	The Haitian Ministry of Public Health and Population (MSPP) was alerted of an unusually high number of cholera cases from the Artibonite and Centre Departments (Centers for Disease Control and Prevention, 2010a).
October 20	St. Marc in the Artibonite River Delta reported an explosive outbreak of cholera cases (Lantagne et al., 2014).
October 21-22	The Haiti National Public Health Laboratory (LNSP) identified *Vibrio cholerae* El Tor O1 in stool samples of patients. Samples were sent to the Centers for Disease Control and Prevention (CDC) for confirmation (Centers for Disease Control and Prevention, 2010a).
October 23	First cholera outbreak in Haiti was officially announced by the MSPP (Centers for Disease Control and Prevention, 2011).
October 27	News journalists started to point blame for the cholera outbreak at the MINUSTAH troops after their investigations showed improper sewage waste disposal in the camp (Al Jazeera English, 2010; Katz, 2010).
November 7-27	French epidemiologist, Dr. Renaud Piarroux, and colleagues conducted epidemiological investigations to determine the source of the cholera outbreak (Piarroux et al., 2011).
November 13	The CDC also identified *V. cholerae* El Tor O1 from stool samples received from the LNSP. Results indicated that a single strain caused illness among the patients, which suggested that *V. cholerae* was likely introduced to Haiti in one event (Centers for Disease Control and Prevention, 2010a).
November 17	Haitians started riots denouncing the MINUSTAH troops and demanding they leave the country (BBC News, 2010; Desvarieux, 2010).
December 3	MSPP reported that cholera had spread to all ten departments of Haiti (Centers for Disease Control and Prevention, 2010b).

December 7	The epidemiological report of Dr. Piarroux and colleagues leaked online. Their report indicated that cholera was imported, which started in the MINUSTAH camp and spread throughout Haiti through the contamination of the Artibonite River (Doland and Lederer, 2010).
December 9	First molecular study on the origin of cholera in Haiti was published. The study suggested a South Asian origin (Bangladesh) of *V. cholerae* and was introduced to Haiti through human transmission (Chin et al., 2011).
December 17	The United Nations (UN) Secretary-General Ban Ki-Moon established the Independent Panel of Experts ("UN Panel of Experts") to conduct independent investigations to determine the source of cholera (Lantagne et al., 2014).
	2011
February 14-18	The UN Panel of Experts conducted investigations in Haiti, including epidemiological, water and sanitation, and molecular analysis investigations (Lantagne et al., 2014).
May 4	The UN Panel of Experts published the results of their investigations. They endorsed the human transmission of cholera to Haiti. However, *V. cholerae* strains from Nepal were not available at that time, thus, a direct comparison between Haitian and Nepalese strains was not performed (Lantagne et al., 2014).
July 7	The study conducted by Dr. Piarroux and colleagues (from November 2010) was officially published (Piarroux et al., 2011).
August 23	First paper to include Nepalese *V. cholerae* strains was published. Strains from Nepal and Haiti were almost identical based on genomic analysis, reinforcing Nepal as the origin of the Haitian outbreak (Hendriksen et al., 2011).
September 29	Shanchol, a vaccine for cholera, by the International Vaccine Institute (Seoul, South Korea) and Shantha Biotechnics (Hyderabad, India) was approved for worldwide use by the World Health Organization (Butler, 2011).
October 14	The UN withdrew 2,750 troops from Haiti (Charbonneau, 2011).
November 3	The human rights groups, Institute of Justice and Democracy in Haiti (IJDH) and the Bureau des Avocats Internationaux, filed a case on behalf of 5,000 Haitian cholera victims against the UN and the MINUSTAH for inadequate screening and sanitation and demanded installation of national water and sanitation system, compensation for the victims, and a public apology (Institute for Justice and Democracy in Haiti, 2011).
December 21	The UN acknowledged receipt of the petition and promised a response "in due course" (Institute for Justice and Democracy in Haiti, 2014a).
	2012
April 12	Partners in Health (PIH) and the Haitian Group for the Study of Kaposi's Sarcoma and Opportunistic Infections vaccinated 100,000 people against cholera with the Shanchol vaccine. Two weeks later, a second dose was given to the 91% of those who received the first dose. This was to demonstrate that, until new water sanitation systems can be constructed, vaccination can serve as a temporary substitute to prevent the spread of cholera (Adams, 2013).
June 18	A study supporting the climatic hypothesis was published, suggesting that environmental strains of *V. cholerae* may have resided undetected in Haitian waterways and rapidly

	multiplied due to favourable environmental conditions (Hasan et al., 2012).
November 20	Mekalanos and colleagues refuted the Hasan study (from June 2012), mainly questioning the unreliable culture methods of the study and also stressed the convincing scientific evidence supporting the human transmission of cholera to Haiti (Mekalanos et al., 2012). Frerichs and colleagues also refuted the Hasan study citing, among other things, that the study "offered no evidence that *V. cholerae* non-O1/O139 played a notable role in the cholera epidemic origin" (Frerichs et al., 2012).
December 11	The UN contributed $23.5 million to Haiti, 1% of the $2.2 billion plan by PIH and the MSPP to eliminate cholera by 2022, primarily through investments in clean water and proper sanitation (UN News Service, 2012).
	2013
February 21	The UN Secretary-General rejected the IJDH case (from November 3, 2011) stating that the claims were not receivable because of the UN's diplomatic immunity (Institute for Justice and Democracy in Haiti, 2014a).
May 7	IJDH responded to the UN arguing that the international law requires the UN to "consider and settle claims filed by third parties for injury illness and death attributable to the UN or its peacekeeping forces." It also gave notice to the UN that the lawyers would pursue a lawsuit if an appropriate response is not received in 60 days (Institute for Justice and Democracy in Haiti, 2013b).
May 22	Due to the availability of additional scientific evidence, the UN Panel of Experts released an update of their report, concluding that the MINUSTAH camp was the most likely source of introduction of cholera into Haiti (Lantagne et al., 2014).
May 30	Eighteen members of the US Congress led by Congresswoman Maxine Waters wrote to the UN Secretary-General calling for him to use his powers to ensure that the UN takes responsibility for the cholera outbreak (Veal, 2013).
July 2	A study was published, which stated that the *V. cholerae* strains have not gained new genetic material since the introduction to Haiti. Also, the strains have the limited ability to acquire genetic material from the environment (Katz et al., 2013).
July 5	The UN Legal Counsel responded to the letter from the lawyers of the cholera victims (from May 7, 2011), saying that the claims are not receivable (Institute for Justice and Democracy in Haiti, 2013c). The UN Secretary-General responded to the letter from the US Congress (from May 30, 2011), saying that the claims of the cholera victims are not receivable. He also reiterated the report by the UN Panel of Experts, that lack of access to clean water and sanitation infrastructure is the key factor that facilitated the epidemic (Institute for Justice and Democracy in Haiti, 2013d).
August 6	The Yale University Law School report held the UN responsible for the cholera outbreak in Haiti, stating that the UN is legally and morally obliged to provide remedy to the cholera victims despite the claim of the UN for diplomatic immunity (Transnational Development Clinic et al., 2013).
October 9	IJDH filed a lawsuit in the Federal District Court in Manhattan, New York against the UN on behalf of the Haitian victims of the cholera epidemic (Institute for Justice and Democracy in Haiti, 2013a).

2014	
January 12	The fourth year anniversary since the January 2010 earthquake (United States Geological Survey, 2010).
February 28	Gustavo Gallon, UN-appointed expert on human rights in Haiti, called for the compensation of the victims and those responsible for the outbreak should be punished (Institute for Justice and Democracy in Haiti, 2014b).
March 11	A new lawsuit against the UN was filed by IJDH in the Federal Court of Brooklyn, New York on behalf of 1,500 Haitians (Associated Press, 2014).
March 12	The Obama administration showed support to the UN, stating the UN has diplomatic immunity and is shielded from liability for the cholera outbreak in Haiti (Gersham, 2014).
March 25	Sandra Honoré, UN Special Representative and head of the MINUSTAH, told the UN Security Council that Haiti still has the highest number of cholera cases in the world (Spielmann, 2014).

References

Adams, P. (2013). Cholera in Haiti takes a turn for the worse. *Lancet* 381(9874), 1264.

Al Jazeera English (2010). *UN investigates Haiti outbreak* [Online]. Doha, Qatar: Al Jazeera Satellite Network. Available: http://www.aljazeera.com/news/africa/2010/10/201010284141214196 7.html [Accessed 25 March 2014].

Associated Press (2014). *Haitians sue U.N. over cholera epidemic* [Online]. New York, NY, USA: Time, Inc. Available: http://time.com/20177/haitians-sue-u-n-over-cholera-epidemic [Accessed 25 March 2014].

BBC News (2010). *UN appeals for calm after cholera riots in Haiti* [Online]. London, UK: British Broadcasting Corporation. Available: http://www.bbc.co.uk/news/world-latin-america-11772283 [Accessed 25 March 2014].

Butler, D. (2011). No quick fix for Haiti cholera. *Nature* 478(7369), 295-296.

Centers for Disease Control and Prevention (2010a). Update: cholera outbreak — Haiti, 2010. *MMWR: Morbidity and Mortality Weekly Report* 59(45), 1473-1479.

Centers for Disease Control and Prevention (2010b). Update: outbreak of cholera — Haiti, 2010. *MMWR: Morbidity and Mortality Weekly Report* 59(48), 1586-1590.

Centers for Disease Control and Prevention (2011). *Cholera in Haiti: one year later* [Online]. Atlanta, GA, USA: Centers for Disease Control and Prevention. Available: http://www.cdc.gov/haiticholera/haiti_cholera.htm [Accessed 25 March 2014].

Charbonneau, L. (2011). *U.N. council approves withdrawal of some Haiti troops* [Online]. New York, NY, USA: Thomson Reuters. Available: http://www.reuters.com/article/2011/10/14/us-haiti-un-idUSTRE79D53V20111014 [Accessed 25 March 2014].

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., et al. (2011). The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 364(1), 33-42.

Desvarieux, J. (2010). *At the heart of Haiti's cholera riots, anger at the U.N.* [Online]. New York, NY, USA: Time, Inc. Available: http://content.time.com/time/world/article/0,8599,2032437,00.html [Accessed 25 March 2014].

Doland, A., and Lederer, E. (2010). *Haiti cholera likely from UN troops, expert says* [Online]. New York, NY, USA: FOX News Network, LLC. Available: http://www.foxnews.com/world/2010/12/07/haiti-cholera-likely-troops-expert-says [Accessed 25 March 2014].

Frerichs, R.R., Boncy, J., Barrais, R., Keim, P.S., and Piarroux, R. (2012). Source attribution of 2010 cholera epidemic in Haiti. *Proceedings of the National Academy of Sciences of the United States of America* 109(47), E3208.

Gersham, J. (2014). *Obama administration backs U.N. in disputes over Haiti cholera epidemic* [Online]. New York, NY, USA: Dow Jones & Company, Inc. Available: http://blogs.wsj.com/law/2014/03/12/obama-administration-backs-u-n-in-disputes-over-haiti-cholera-epidemic [Accessed 25 March 2014].

Hasan, N.A., Choi, S.Y., Eppinger, M., Clark, P.W., Chen, A., Alam, M., et al. (2012). Genomic diversity of 2010 Haitian cholera outbreak strains. *Proceedings of the National Academy of Sciences of the United States of America* 109(29), E2010-E2017.

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., et al. (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio 2(4), e00157-11.

Institute for Justice and Democracy in Haiti (2011). *Petition for relief: English (IJDH-BAI)* [Online]. Boston, MA, USA: Institute for Justice and Democracy in Haiti. Available: http://www.ijdh.org/2011/11/topics/law-justice/chief-claims-unit-minustah-log-base-room-no-25a-boule-toussaint-louverture-clercine-18-tabarre-haiti-ijdh-bai/#.UzOwsa1dVX8 [Accessed 25 March 2014].

Institute for Justice and Democracy in Haiti (2013a). *Cholera complaint against the UN* [Online]. Boston, MA, USA: Institute for Justice and Democracy in Haiti. Available: http://www.ijdh.org/2013/10/topics/health/cholera-complaint-against-the-un/#.UzNwpK1dU4R [Accessed 25 March 2014].

Institute for Justice and Democracy in Haiti (2013b). *Cholera victims' response to UN letter* [Online]. Boston, MA, USA: Institute for Justice and Democracy in Haiti. Available: http://www.ijdh.org/2013/05/topics/law-justice/cholera-victims-response-to-un-letter/#.UzO1FK1dVX- [Accessed 25 March 2014].

Institute for Justice and Democracy in Haiti (2013c). *Letter from the United Nations Legal Counsel to cholera victims' lawyers* [Online]. Boston, MA, USA: Institute for Justice and Democracy in Haiti. Available: http://www.ijdh.org/2013/07/topics/health/letter-from-the-united-nations-legal-counsel-to-mr-brian-concannon-institute-for-justice-democracy-in-haiti/#.UzO40K1dVX8 [Accessed 25 March 2014].

Institute for Justice and Democracy in Haiti (2013d). *UN Secretary General Ban Ki-Moon's letter to US Congress* [Online]. Boston, MA, USA: Institute for Justice and Democracy in Haiti. Available: http://www.ijdh.org/2013/07/topics/health/unsg-letter-to-rep-maxine-waters/#.UzO5Oa1dVX8 [Accessed 25 March 2014].

Institute for Justice and Democracy in Haiti (2014a). *Cholera litigation* [Online]. Boston, MA, USA: Institute for Justice and Democracy in Haiti. Available: http://www.ijdh.org/cholera/cholera-litigation/#.UzOzIK1dVX9 [Accessed 25 March 2014].

Institute for Justice and Democracy in Haiti (2014b). *UN expert calls for accountability as organization ducks cholera suit* [Online]. Boston, MA, USA: Institute for Justice and Democracy in Haiti. Available: http://www.ijdh.org/2014/02/topics/health/un-expert-calls-for-accountability-as-organization-ducks-cholera-suit/#.UzPGqK1dVX8 [Accessed 25 March 2014].

Integrated Regional Information Networks (2014). *In brief: cholera outbreak in Nepal under control, says WHO* [Online]. Nairobi, Kenya: Integrated Regional Information Networks. Available: http://www.irinnews.org/report/90231/in-brief-cholera-outbreak-in-nepal-under-control-says-who [Accessed 25 March 2014].

Ivers, L.C., and Walton, D.A. (2012). The "first" case of cholera in Haiti: lessons for global health. *American Journal of Tropical Medicine and Hygiene* 86(1), 36-38.

Katz, J.M. (2010). *UN probes base as source of Haiti cholera outbreak* [Online]. Washington, DC, USA: The Washington Times, LLC. Available: http://www.washingtontimes.com/news/2010/oct/27/un-probes-base-as-source-of-haiti-cholera-outbreak [Accessed 25 March 2014].

Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., et al. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4(4), e00398-13.

Lantagne, D., Nair, G.B., Lanata, C.F., and Cravioto, A. (2014). "The cholera outbreak in Haiti: where and how did it begin?," in *Cholera Outbreaks: Current Topics in Microbiology and Immunology,* eds. B.G. Nair & Y. Takeda. (Berlin, Germany: Springer-Verlag), 145-164.

Maharjan, L. (2010). *Cholera outbreak looms over capital* [Online]. Kathmandu, Nepal: International Media Network Nepal Pvt. Ltd. Available: http://www.thehimalayantimes.com/fullNews.php?headline=Cholera+outbreak+looms+over+capital&NewsID=258974 [Accessed 25 March 2014].

Mekalanos, J.J., Robins, W., Ussery, D.W., Davis, B.M., Schadt, E., and Waldor, M.K. (2012). Non-O1 *Vibrio cholerae* unlinked to cholera in Haiti. *Proceedings of the National Academy of Sciences of the United States of America* 109(47), E3206.

Piarroux, R., Barrais, R., Faucher, B., Haus, R., Piarroux, M., Gaudart, J., et al. (2011). Understanding the cholera epidemic, Haiti. *Emerging Infectious Diseases* 17(7), 1161-1168.

Spielmann, P.J. (2014). *UN: Haiti has more cholera than any other nation* [Online]. New York, NY, USA: ABC News Network. Available: http://abcnews.go.com/International/wireStory/haiti-cholera-nation-23041690 [Accessed 25 March 2014].

Transnational Development Clinic, Global Health Justice Partnership, and Association Haitienne de Droit de L'Environnment (2013). "Peacekeeping without accountability: The United Nations' responsibility for the Haitian cholera epidemic". (New Haven, CT, USA: Yale Law School).

UN News Service (2012). *UN launches new initiative to eliminate cholera in Haiti and Dominican Republic* [Online]. New York, NY, USA: United Nations. Available: http://www.un.org/apps/news/story.asp?NewsID=43743&Cr=cholera#.UzO9La1dVX9 [Accessed 25 March 2014].

United States Geological Survey (2010). *Magnitude 7.0 - Haiti region* [Online]. Reston, VA, USA: United States Geological Survey. Available: http://earthquake.usgs.gov/earthquakes/eqinthenews/2010/us2010rja6/us2010rja6.php [Accessed 25 March 2014].

Veal, L. (2013). *Congresswoman Waters urges United Nations to commit resources to eradicate cholera in Haiti* [Online]. Washinton, DC, USA: United States Congress. Available: http://waters.house.gov/news/documentsingle.aspx?DocumentID=336057 [Accessed 25 March 2014].