Bacterial Classifications in the Genomic Era

by

Kevin Liang

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

Master of Science

in

Microbiology and Biotechnology

Department of Biological Sciences

University of Alberta

Abstract

Bacterial taxonomy is an integral part of all disciplines within the field of microbiology, as it allows researchers to communicate results efficiently, streamlining global collaboration. The ultimate goal of bacterial taxonomy is to create groups of organisms based not only on shared phenotypic and genomic traits, but also a common evolutionary history. To achieve this goal, the polyphasic approach, which examines phenotypic, genomic and phylogenetic data, is favored. Although the three major components of polyphasic taxonomy remain unchanged since it was first proposed in 1968, the methods in which we assess these aspects have improved significantly due to the abundance of whole genome sequences (WGS) available. In addition, WGS has also served as the basis for developing high-resolution subspecies level classification techniques. The research presented in this thesis therefore focuses on both applying modern techniques to the polyphasic approach to taxonomy and developing a standardized, easy-to-use high-resolution subspecies typing technique.

Traditionally, the 16S rRNA gene has been used to assess genomic and phylogenetic relationships for taxonomic purposes. Although it is now widely known that 16S rDNA is not suitable for species, genus or even family level taxonomic classifications, it is still commonly used to fulfill the phylogenetic aspect of polyphasic taxonomy within the family *Rhodobacteraceae*. Consequently, taxonomic inconsistencies have been a reoccurring problem since the conception of this group in 2005. To resolve taxonomic inconsistencies within this family, over 300 type strains with high-quality genomes were analyzed. As type strains are important reference material for classification, resolving taxonomic inconsistencies among these strains will ultimately help guide future taxonomic efforts and prevent the propagation of errors. Based on genomic and core-genome phylogenetic data, three species, and 25 genus level

misclassifications were identified. Combining a meta-analysis of phenotypes with genomic techniques, distinguishing phenotypic traits useful for family level classification were predicted. Furthermore, a general approach to taxonomy based on genomic and phylogenetic analyses is proposed, to validate taxonomic classifications but also highlight potential misclassifications.

Subspecies level classification is an integral part of epidemiological and clinical research, as it is important to differentiate between closely related pathogenic and non-pathogenic strains within the same species. A high-resolution subspecies level typing method, known as core-genome multilocus sequence typing (cgMLST) was developed for *Vibrio cholerae*, a bacterium best known as the causative agent of cholera. Traditionally, subspecies typing for *V. cholerae* was based on multilocus sequence typing (MLST), multilocus variable tandem repeats analysis (MLVA) or serotyping. These methods provided limited resolution, which restricted its use in an epidemiological setting. cgMLST, on the other hand, provides much greater resolution than any previously named method as it utilizes a significantly larger portion of the genome by analyzing all genes common to *V. cholerae*. An outbreak threshold capable of identifying outbreak related strains and potential sources of introduction is proposed. To help consolidate existing MLST information and also investigate large-scale ecological and epidemiological patterns, a sublineage threshold is defined which creates clusters similar to traditional MLST schemes. Using this threshold, a strong geographic signal is detected among environmental isolates not seen in clinical strains. This scheme, along with over 1,200 *V. cholerae* genomes and relevant provenance data, is currently available on PubMLST (https://pubmlst.org/vcholerae) for public access.

Research presented in this thesis demonstrates the importance of WGS-based analyses, not only for taxonomic classifications at the species level and above, but also at the subspecies

level. As next generation sequencing and bioinformatics techniques develop, WGS-based methods will inevitably become standard practices for bacterial classification.

Preface

Some research that forms part of this thesis results from collaborative work. Individual contributions for each of the chapters are listed below.

A version of chapter 2 will be submitted for publication as:

> "Liang, K.Y.H., Orata, F.D., Boucher, Y.F., and Case, R.J. Roseobacters in a sea of poly- and paraphyly: whole genome-based taxonomy of the family *Rhodobacteraceae* and the proposal for the split of the "roseobacter clade" into a novel family, *Roseobacteraceae* fam. nov."
>
> KYHL, YFB and RJC design the study and wrote the manuscript. FDO and KYHL performed bioinformatic analyses and collected data. YFB and RJC supervised the project.

A version of chapter 3 will be submitted for publication as:

> "Liang, K.Y.H., Orata, F.D., Islam, M.T., Nasreen, T., Alam, M., Tarr, C.L., Boucher, Y.F. A Vibrio cholerae Core-Genome Multilocus Sequence Typing Scheme to facilitate the epidemiological study of Cholera."
>
> KYHL and YFB design the study and wrote the manuscript. FDO and KYHL performed bioinformatic analyses and IMT helped with data analysis. CLT, TN and AM provided clinical and environmental strains. YFB supervised the project.

In addition to research presented here, I was also involved in other collaborative works. Appendix C provides a list of authored and coauthored publications.

Acknowledgements

I would like to thank everyone who has helped me throughout this process as none of this would have been possible. I would like to thank Dr. Yann Boucher for giving me the chance to join his lab and take part in his research. You were always willing to share your experience and give advice to help me deal with the stressful times and keeping me on track. You're a great PI and I only wished I joined your lab sooner. I would also like to thank to Dr. Rebecca Case. I could not have completed my *Rhodobacteraceae* taxonomy project without your assistance over the past two years. It has been a long process and I thank you for your patience.

I would like to thank my committee member Dr. Warren Gallin. I'm fortunate to have you as my undergrad program advisor and on my committee. Without your help during my undergrad, I would have had an impossibly difficult time getting all my courses organized. As my committee member you were always there to give valuable feedbacks pointing out possible gaps in my projects and/or explanations, which helped me a lot my preparation of my thesis and defense.

I would like to acknowledge Dr. Cheryl Tarr, Monica Im (Centers for Disease Control and Prevention) and Dr. Alam Munirul (International Centre for Diarrhoeal Disease Research, Bangladesh) whose contributions help shaped the thesis and projects that I am working on.

Last but not least, I would like to thank my lab-mates, Dr. Fabini Orata, Tareq Islam, and Nora Hussan. I have learned so much from each of you and I thank you all for making my grad school experience an enjoyable one.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

WGS	Whole Genome Sequences
cgMLST	Core genome multilocus sequence typing
MLST	Multilocus sequence typing
PG	Pandemic generating/phylcore lineage
NGS	Next generation sequencing
SNP	Single nucleotide polymorphism
ANI	Average nucleotide identity
MLSA	Multilocus sequence analysis
DDH	DNA-DNA hybridization
dDDH	*In-silico*/digital DNA-DNA hybridization
AAI	Average amino acid identity
PFGE	Pulse-field gel electrophoresis
MLVA	Multilocus variable number tandem repeat analysis
ST	Sequence type
VNTR	Variable number tandem repeat
PCR	Polymerase chain reactions
DMSP	Dimethylsulfoniopropionate
DMS	Dimethyl sulfide
MeSH	Methanethiol
PD	Patristic distance
AHL-QS	Acyl-homoserine lactone quorum sensing

GTR	General time reversible
SRA	Sequence read archives
cgST	cgMLST sequence type
DI	Dunn Index
ARI	Adjusted Rand Index
MST	Minimum spanning tree

Chapter 1: Introduction – application of whole genome sequencing to bacterial classification

Chapter 1

1.1 Bacterial classifications

A standardized bacterial classification system is important for microbial research as without it defining biological organisms would be nearly impossible, let alone communicating higher-level scientific concepts. However, taxonomy is more than just assigning names to novel isolates. It describes the process of systematically creating groups of organisms based not only on shared phenotypic and genomic traits but also a common evolutionary history.

Bacterial classification is a constantly evolving field. Each new technology allows us to study bacterial isolates in greater detail, making new ways of classifying organisms possible, in the hope of attaining a universal and stable method that can better achieve our goals in taxonomy. Unfortunately, with very few exceptions, there is rarely a set of well-defined standards that allows for consistent and stable classifications. There are multiple techniques that can be used for each level of classification, such as average amino acid identity, percentage of conserved protein, 16S-rRNA analyses, and DNA-DNA hybridization to name a few, but unfortunately, these techniques are not always consistent with one another. Despite this ambiguity, whole-genome sequences (WGS) and new bioinformatics techniques have provided some tangible criteria that can be used to make significant improvements to current practices.

1.1.1 Practical implications of bacterial taxonomy: from a clinical and an epidemiological perspective

Accurate bacterial classifications have important implications for both clinical and epidemiological practices. As is commonly the case, many species of bacteria known for their

pathogenicity harbour both dangerous and harmless strains. *Vibrio cholerae* for example, best known as the etiological agent of cholera, harbours a single lineage responsible for all major outbreaks since the early 19th century (Islam et al. 2017), the pandemic generating/phylcore (PG) lineage (Chun et al. 2009, Boucher 2016). Similarly, other pathogens such as *Escherichia coli*, *Salmonella enterica*, and *Vibrio parahaemolyticus,* also contain variants within the same species where some are pathogenic while others are not. *S. enterica* in particular contains six subspecies, but only one subspecies is responsible for nearly all infections in humans (Desai et al. 2013). Subspecies level classifications are therefore crucial to tracking the spread of illnesses by identifying pathogenic strains. The ability to identify pathogenic and non-pathogenic strains will also help us understand the emergence of pathogenic bacteria and their evolutionary history as comparative genomic and phylogenetic analyses can be performed on closely related strains to identify evolutionary events and/or genetic features that enabled some strains to become pathogenic while others to remain harmless or even beneficial (Wurtzel et al. 2012, Alavi et al. 2014, Cesbron et al. 2015).

In addition to the ability to differentiate between pathogenic and non-pathogenic strains, it is also important to distinguish among closely related pathogenic strains. For example, a majority of the illnesses caused by S. enterica, are due to a single subspecies, within which there exist over 2,000 serovars that differ in host and the illness that is caused (Porwollik et al. 2004). Similarly, only one lineage of *V. cholerae*, the PG lineage, is responsible for all major cholera outbreaks. As a result, outbreak strains are expected to be closely related and therefore, simply knowing that a particular isolate belongs to the pathogenic lineage will not provide any useful epidemiological information. This level of resolution is especially important for pandemic and epidemic diseases, as closely related pathogenic strains can rapidly spread to different regions,

causing many local outbreaks. In the past, epidemiological records based on patient travel, contacts and time of infection were all that could be used to deduce the origin of local outbreaks. However, genomic techniques that allow for subspecies level classifications will be able to quickly identify outbreak related strains and potential sources of introduction. This is because although pathogenic strains of a particular species are closely related, outbreak related strains are even more closely related. The ability to assess genomic similarity at this level is needed to identify whether isolates belong to the same outbreak or not. Comparing these isolates to a global reference database and identifying the next closest neighbours will provide an idea of likely sources of introductions (Katz et al. 2013).

1.1.2 Bacterial taxonomy as the fundamental component in understanding bacterial diversity and evolution

It is without question that bacterial taxonomy is important to understanding bacterial diversity. Bacteria have traditionally been one of the more difficult organisms to study, as unlike other macro-organisms, most are difficult to cultivate and observe in a laboratory setting (Stanier and Van Niel 1941). The importance of bacterial taxonomy can be seen in the changes in our understanding of microbial diversity and evolution throughout the years.

Early taxonomic efforts for bacteria were based on shape, motility, behaviour and habitat, as microscopy was all that could be used for classification purposes (Murray and Holt 1989). This approach grossly underestimated bacterial diversity as all isolates were grouped into merely six genera, distinguished only by shape (Cohn 1875). The development of cultivation-based methods to characterize bacteria greatly increased our understanding of bacterial diversity as it

allowed us to classify isolates base on a larger selection of phenotypic traits (Zengler et al. 2002). However, it is important to note that cultivation alone is not able to reveal the true diversity due to the difficulty in cultivating majority of the environmental isolates (Harwani 2013).

Gene sequencing, and later genome sequencing, has allowed us to gain a more comprehensive view of bacterial diversity (Kroes et al. 1999), as the cultivation of isolates is no longer required. The availability of DNA sequences has also allowed us to study bacterial evolution in greater detail and more accurately reconstruct phylogenetic relationships. 16S rRNA was one of the first molecules used to study bacterial diversity and evolutionary relationships (Woese 1987). Now with the decreasing cost of genome sequencing and the ease of genomic and metagenomic analyses through the use of readily available bioinformatics tools, whole genome sequences (WGS) are progressively replacing 16S rDNA in studying bacterial diversity and evolution.

1.2 Bacterial taxonomy at the species level and above

1.2.1 Polyphasic taxonomy

Bacterial taxonomy has long been a much-debated topic as we aim to identify universal and standardized classification techniques and to this day, many of the details (e.g., genus, family boundaries) remain ambiguous (Godreuil et al. 2005). However, it is commonly agreed that polyphasic taxonomy, which classifies bacteria based on phenotypic, genomic and phylogenetic traits, is the ideal approach (Vandamme et al. 1996, Thompson et al. 2015). Unlike other classification techniques, such as the genomospecies or taxospecies concepts, it does not rely on a single hypothesis or metric when determining taxonomic boundaries, but instead aims to integrate all information possible to achieve a stable taxonomic classification that reflects the

true ecological and evolutionary history of organisms (Vandamme et al. 1996). Since it was first proposed in 1968 (Colwell 1968) for bacterial classification from the species level and above, it remains an integral part of bacterial taxonomy. Although the concept of polyphasic taxonomy remains unchanged in that phenotypic, genomic and phylogenetic coherence are still the focus of this approach, techniques with which we assess these aspects have evolved significantly over the years.

1.2.2 Phenotypic aspects of the polyphasic approach

Phenotypic traits are the earliest criteria used for taxonomic classifications (Murray and Holt 1989) as they have been historically easier to evaluate than phylogenetic and genetic similarities, both of which require the ability to obtain genetic sequences. Phenotyping is important for a number of reasons. Firstly, accurate phenotypic descriptions at all levels of taxonomic classifications are required to understand the complex ecological roles different bacterial lineages play in the environment (Philippot et al. 2010). In addition, detail phenotypic descriptions of pathogenic strains highlighting important clinical characteristics such as antibiotic resistance and virulence phenotypes are also crucial for determining treatment strategies. Secondly, it is possible to correlate phenotypic data with genes presence, which may provide some insights into novel gene functions and annotations (Philippot et al. 2010). Lastly, phenotyping will allow us to provide accurate taxonomic descriptions of each bacterial lineage that highlights important unifying phenotypes that are also likely ancestral traits. The identification of ancestral characteristics will not only provide clues regarding the evolutionary paths taken, but also likely environmental conditions from which these isolates evolved. For

these reasons, although it is costly and time consuming, phenotyping is still an integral part of current taxonomic practice.

Ideally organisms belonging to the same groups, whether it be the same family, genus or species, will share some universal phenotypic traits, but this is not always the case and it is not uncommon for taxonomic descriptions to include phenotypes that are shared by some and not all (Wells et al. 1987, Garrity et al. 2015a, Orata et al. 2016). It is expected that as we examine higher -level taxonomic classifications, there will be fewer universally shared traits, as we are looking at a more diverse group of organisms. Therefore, instead of simply focusing on universal phenotypes, it is important to also identify unifying traits that are likely ancestral to the taxonomic group under study, as these will likely reflect a common evolutionary path and/or ecology (Philippot et al. 2010).

Despite the importance of phenotypic testing in bacterial taxonomy, it is not without limitations. Traditionally, 100 or more phenotypic tests are performed to identify a few defining characteristics. This unguided approach in phenotypic testing is not only costly, but also labour and time intensive. Although commercial microbial identification kits, such as the Biolog systems (https://www.biolog.com/products-portfolio-overview/microbial-identification/), are available allowing for more systematic and standardized phenotyping procedures, it is still difficult to scale up phenotypic testing to match the rate at which new genomes are sequenced. In addition, for some species with uncommon metabolisms or lifestyles, atypical phenotypic tests are required (Tindall et al. 2010), making it impossible to establish standardized phenotyping protocols. Another limitation to phenotyping in general is that it does not differentiate between ancestral and derived traits. Based on phenotypes alone, it is impossible to determine whether

isolates are phenotypically similar due to convergent evolution, lateral gene transfer or vertical descent.

Next generation sequencing (NGS) mitigated many of the limitations mentioned above. For one, using WGS, phenotyping no longer needs to be an unguided effort. There are now readily available tools that can annotate and predict potential phenotypes based only on WGS (Aziz et al. 2008, Goberna and Verdú 2016, Kanehisa et al. 2016). By first using *in-silico* methods to identify possible phenotypes, it can substantially reduce the number of phenotypic tests required to identify distinguishing traits.

1.2.3 Phylogenetic aspects of the polyphasic approach

As previously mentioned, some details regarding bacterial taxonomy remain ambiguous, but phylogenetic analyses do provide perhaps one of the few universally agreed upon rules for all levels of taxonomic classification: all taxonomic groups regardless of rank should be monophyletic (Rosselló-Móra and Amann 2015). Traditionally, 16S rDNA was used to reconstruct phylogenetic relationships. More recently, with the proliferation of WGS, higher resolution methods utilizing genome sequences, such as SNP-based and gene-by-gene-based methods for phylogenetic reconstruction, are becoming more common.

One of the key objectives of polyphasic taxonomy is to create clusters of isolates with a shared evolutionary history. As mentioned before, phenotyping is unable to differentiate between convergent evolution, vertical descent and lateral gene transfer; however, by reconstructing phylogenetic relationships using the numerous genes found in all the organisms under study, we can have more confidence that organisms with shared evolutionary history will be grouped

together. Ultimately, phylogenetic and phenotypic data should be consistent with one another (Murray et al. 1990).

The 16S rRNA gene was widely used for taxonomic purposes in the past, primarily because it is universally present in all isolates. This means that all unknown isolates can be analyzed using the same protocol to identify the general phylogenetic relationships with known isolates. This was extremely helpful as knowing even just the broader taxonomic classifications (i.e., class or order) can immediately narrow down the list of phenotypic tests that should be done and will help guide subsequent more in-depth phylogenetic and genomic analyses.

16S rRNA gene sequence analysis works best at the phylum level and becomes increasingly less reliable as we move towards lower taxonomic ranks (Poretsky et al. 2014, Ranjan et al. 2016). In addition to the lack of resolution, some isolates also contain multiple intragenomic copies with sequences different enough to be grouped as separate genera (Klappenbach et al. 2001, Acinas et al. 2004, Boucher et al. 2004, Case et al. 2007), adding yet another reason as to why the 16S rRNA gene alone is not suitable for determining taxonomic classifications. Therefore, although analysis of this gene may be the first indication that an isolate is a novel species or genus, taxonomic classifications at these levels should not rely solely on it (Tindall et al. 2010).

Progressively, the 16S rRNA gene is being replaced by WGS in phylogenetic analyses. A common high-resolution method that utilizes WGS is the SNP-based approach. One of the main advantages of the SNP-based method is that it can work with complete, draft genomes or even just NGS reads, making this approach less computationally intensive (Hall 2016). In addition, by eliminating the assembly and gene annotation steps, it also eliminates associated errors such as

annotation errors. As a result, SNP-based methods are often times faster than the gene-by-gene approach for phylogenetic reconstruction. There are generally two methods for SNP-based phylogenetic analysis: alignment-based and alignment-free methods.

A popular alignment based method is Parsnp (Treangen et al. 2014), which does require the assembly of NGS reads into draft or complete genomes. Highly similar genomes (≥97% average nucleotide identity (ANI) are used as input sequence and Parsnp will first produce a core-genome alignment and then identify potential SNPs. Although in some studies, core-genome alignments refer to the concatenation of multiple individual core gene alignments, here core genome simply refers to a collection of unique genomic sequences present in all organisms of interest. Once SNPs are identified, a series of quality filter parameters are required to remove SNPs present in recombinogenic regions, and poorly sequenced and/or assembled regions (Treangen et al. 2014). As alignment based methods for variant calling require the assembly of genomes, it is computationally more intensive than alignment free methods. However, alignment based methods often produce more reliable results as it is possible to easily identify indels and other structural mutations; therefore, are typically seen as the gold standard for SNP calling (Höhl and Ragan 2007, Treangen et al. 2014).

Although alignment-based methods produce more reliable results, because it is more computationally intensive, it is difficult to scale up for use in high-throughput analyses, which led to the development of numerous alignment free methods. A popular alignment-free method is the k-mer based approach where odd length short sequences, commonly known as k-mers, differing at only the central positions are identified among NGS reads of all isolates of interest (Gardner and Hall 2013). The flanking regions of these k-mers are then used to define the SNP locus (Gardner and Hall 2013). The advantage of this method is that the input files can be raw

reads, draft or complete genomes. This reduces the computational costs and can be used to effectively reconstruct the phylogenetic relationships of hundreds of isolates. Although some alignment free methods do require an assembled reference genome (Leekitcharoenphon et al. 2012), this is not always the case (Gardner et al. 2015)

Regardless of the variant identification method used for phylogenetic reconstruction, they share a common limitation, which is that they only work with closely related organisms. Parsnp, as previously mentioned, is recommended for organisms sharing 97% or more average nucleotide identity (ANI) (Treangen et al. 2014), which is above the commonly accepted species threshold of 95% (Goris et al. 2007). In addition to having closely related organisms for SNP-based phylogenetic analysis, there must also be stringent recombination detection parameters as single recombination events can result in numerous SNPs being created, which will impact the accuracy of phylogenetic reconstruction if not accounted for properly (Gardner and Hall 2013, Hall 2016).

Although a SNP-based phylogenetic approach can analyze hundreds of genomes with very minimal pre-processing (i.e., it does not require the assembly or annotation of genomes), it does only work with closely related organisms. To look at more distantly related organisms (i.e., within the same family), a gene-by-gene approach is required. Traditionally, only six to seven housekeeping genes are used for phylogenetic analyses as in multilocus sequence analysis (MLSA); however, advancements in NGS and bioinformatics tools have allowed for the inclusion of hundreds, if not thousands of genes, which can significantly improve the reliability and resolution of phylogenetic analyses. A common gene-by-gene approach is core-genome phylogeny, which in this context refers specifically to a collection of core genes.

Typically, next generation sequencing (NGS) reads are first assembled into draft or complete genomes before they are annotated. Core genes are then identified and individually aligned before concatenated into a single alignment that will be used for phylogenetic reconstruction (Chaudhari et al. 2016, Na et al. 2018). In addition to the ability of this method to work with more distantly related organisms, it is also possible to reconstruct phylogeny based on nucleotide or amino acid sequences depending on the group of organisms being studied. For more distantly related organisms, it is advisable to use amino acid sequences as it does not reach mutational saturation as quickly as nucleotides (Qin et al. 2014). On the other hand, nucleotide sequences are typically used for more closely related organisms, as it has a higher mutation rate providing greater level of resolution (Yamamoto and Harayama 1996). Although this core-genome based method can examine more distantly related organisms, it is important to note that the resolution will decrease as diversity increases, as fewer number of core genes will be identified (Chaudhari et al. 2016).

1.2.4 Genotyping in the polyphasic approach

Genomic similarities establish an efficient way to determine relatedness of hundreds of not thousands of isolates. In theory, genomically similar isolates should be descendants from the same ancestor and share similar phenotypic traits. Traditionally, genomic similarities can only be estimated through indirect methods, such as 16S rRNA gene sequence, single protein-coding gene, or changes in physical properties (e.g., DNA-DNA Hybridization and G+C content deviation). However, much like with phenotypic and phylogenetic analyses, whole-genome sequencing has allowed for the development of different direct and more accurate methods to

assess genomic similarities, including, among others, *in-silico* DNA-DNA hybridization (dDDH), average nucleotide identity (ANI), and average amino acid identity (AAI).

In terms of taxonomic classifications, DNA-DNA hybridization (DDH) is the gold standard for evaluating genomic relatedness and is commonly used to delineate novel species (Stackebrandt and Ebers 2006, Tindall et al. 2010). Based on empirical experiments, 70% at 5ºC melting temperature is suggested to serve as the species threshold where only isolates with 70% or more DDH values belong to the same species (Tindall et al. 2010). Although DDH is the gold standard for bacterial species taxonomy, it is a tedious process that is difficult to replicate, which makes it difficult to scale. This has led to the development of other metrics that aim to approximate or predict DDH experiment results.

One of the earliest methods to approximate DDH results is 16S rRNA gene sequence similarity. 16S rRNA genes, as previously mentioned, are conserved among all prokaryotes. Since this is also a relatively small molecule (~1.4kb) with conserved and variable regions, it can easily be partially or fully sequenced using traditional Sanger sequencing. It has been proposed that 97.5% (Stackebrandt and Goebel 1994), and later 98.7% sequence similarity (Stackebrandt and Ebers 2006), as the species cutoff; however, it should be noted that in all cases, 16S rDNA sequence similarity is not sufficient to confirm the identity of the species because even if isolates share high 16S rDNA sequence similarity (≥98.7%), it only suggests that there is a chance for DDH values to be above 70% (Tindall et al. 2010).

Both 16S rRNA gene sequence similarity and DDH experiments are indirect way of assessing genomic similarities. In order to directly determine genomic similarities, WGS must be used. Two commonly used WGS-based methods that can directly assess genomic similarities are

in-silico DNA-DNA Hybridization (dDDH) (Meier-Kolthoff et al. 2013) and average nucleotide identity (ANI) (Konstantinidis and Tiedje 2005a, Goris et al. 2007). Both methods utilize WGS to assess genomic similarity and predict traditional DDH results, as 70% DDH value is still the accepted species threshold. However, unlike traditional DDH experiment these WGS-based methods are easy to do and gives reproducible results allowing for the establishment of a more stable and practical species level taxonomic classification.

dDDH uses the same scale as traditional DDH experiment meaning that it retains the 70% threshold as the species cutoff (Meier-Kolthoff et al. 2013). On the other hand, ANI utilizes 95% as the species threshold as it was previously determined that isolates with 95% or more ANI will also have 70% or more DDH (Goris et al. 2007).

All previously mentioned methods, except for 16S rRNA sequencing, are primarily used for classification of species or lower taxonomic levels (Goris et al. 2007, Meier-Kolthoff et al. 2013); but there are also genomic metrics that can be used for genus and higher level classifications. A popular method is average amino acid identity (AAI). AAI is similar to ANI in theory, but instead of looking at nucleotide sequences, it utilizes amino acid sequences of conserved genes. Depending on the organisms of interest within genera comparisons can have values ranging from 60%-80% (Orata et al. 2018) and as a result there are no standardized AAI cutoffs that can be used for universal and systematic genus level classification. Although AAI is not able to definitively identify wrongly classified organisms, it is able to highlight potential misclassifications by highlighting unusual AAI values within a dataset (Qin et al. 2014, Orata et al. 2018).

1.3 Bacterial classification at the subspecies level

Subspecies level classification is useful to understanding biodiversity and evolution, but we will examine it from an epidemiological perspective, as it is becoming an integral part of public health and medical research. Numerous subspecies classification techniques have been applied to different pathogenic bacteria such as *Vibrio cholerae* (Katz et al. 2013, Kirchberger et al. 2016), *Listeria monocytogenes* (Moura et al. 2016), and *Salmonella enterica* (Brenner et al. 2000) in the aims of delineating between closely related bacterial strains.

Earlier methods for subspecies classification, such as serotyping or pulse-field gel-electrophoresis (PFGE), assess genomic relatedness through indirect means. The inability of these methods to resolve closely related strains make these techniques not suitable for large-scale global surveillance. Serotyping for example, is routinely used for typing *V. cholerae*; however, despite over 200 serotypes identified, only two (O1 and O139) are responsible for all major outbreaks (Islam et al. 2017). Needless to say, serotyping provided minimal information to the epidemiological surveillance of cholera.

Many of these methods are now replaced by sequence-dependent methods, such as multilocus sequence typing (MLST) and multilocus variable number tandem repeat analysis (MLVA), as these are usually less labour intensive and offer greater resolution. The ease of whole genome sequencing has caused yet another shift in the way we perform subspecies level classifications. A quick overview of sequence dependent subspecies classification techniques can, therefore, be largely split into the pre-genomic and the genomic era.

1.3.1 Subspecies level classifications in the pre-genomic era

Multilocus sequence typing (MLST) is a popular subspecies classification method that looks at internal fragments of housekeeping gene sequences to approximate genome level differences (Maiden et al. 1998). Currently, there are no set standards to the number of genes that are required and the number of genes can differ between studies. However, six to seven housekeeping genes are typically used, as it is a reasonable balance between the amount of resolution obtained and the cost of sequencing (Sabat et al. 2013, Dingle and MacCannell 2015). It works by first assigning a unique number to each unique gene sequence. The specific combination of numbers then constitutes a MLST profile, which is given an arbitrary sequence type (ST) designation. Similarity between isolates can then be evaluated by looking at how many loci are identical between strains. Strains differing at one or two loci are also typically assigned to clonal complexes, which can often be associated with particular pathovar or biotypes (Enright 2003, Forsythe et al. 2014). First proposed in 1998 (Maiden et al. 1998), MLST has now been applied to numerous bacterial species (Baldo and Werren 2007, Margos et al. 2008, Boonsilp et al. 2013, Octavia et al. 2013), most of which are publicly available on PubMLST (https://pubmlst.org).

There are a number of advantages to MLST that make it favourable for subspecies classification. One such benefit is the ease at which results can be easily communicated and standardized as it is based on sequences of well-defined housekeeping genes. As long as the genes are known, the same analysis can be easily repeated across laboratories. In addition, DNA sequences, allele profiles and ST designations can be easily stored on public servers such as PubMLST allowing researchers to analyze their dataset in a global context. Another advantage of

MLST is that any sequence level changes will be assigned a novel allele designation. As it is often difficult to determine whether changes are the result of recombination or a series of different mutations, by giving all new sequence a new allele designation, it reduces the effect of recombination events (Enright and Spratt 1999). By utilizing housekeeping genes, MLST also creates groups that are more likely to reflect the true evolutionary history as housekeeping genes are less likely to be horizontally transferred (Baldo and Werren 2007).

As with any typing methods, there are benefits and limitations. For MLST, one of the biggest limitations is that multiple individualized schemes may exist for different subgroups within the same species. This is because it is unreasonable to expect the same combination of six to seven housekeeping genes to resolve all subgroups within the same species equally well as different lineages may be under different selective pressure especially if the species contains pathogenic and non-pathogenic strains. As a result, some species such as *Staphylococcus epidermidis* (Wang et al. 2003, Wisplinghoff et al. 2003, Thomas et al. 2007) and *Vibrio cholerae* (Garg et al. 2003, Kotetishvili et al. 2003, Octavia et al. 2013) have more than one MLST scheme. This limits national and international collaboration, as results from different laboratories will not be comparable if they are using different schemes. Another important limitation of MLST is the lack of resolution due to the conservative nature in allele identification (i.e., any change whether it is 1 SNP or 100 SNPs is considered the same and given a new allele designation) means MLST cannot distinguish between closely related isolates, which limits its use in outbreak studies (Antwerpen et al. 2015). To exacerbate this problem, often times, only fragments of housekeeping genes are used and not the full length sequence (Maiden et al. 1998).

Multilocus variable number tandem repeats (VNTR) analysis (MLVA) is another pre-genomic era method that retains some of the benefits of MLST but provides greater resolution

than other commonly used methods such as MLST (Marsh et al. 2010, Lam et al. 2012, Chenal-Francisque et al. 2013, Martín et al. 2018) and PFGE (Keim et al. 2000). Generally, MLVA utilizes eight regions containing variable number tandem repeats (VNTR) (Keim et al. 2000). VNTR regions are PCR amplified and the number of repeats in each region is determined. A unique combination of repeats across all loci represents a MLVA type, which is used to distinguish between different strains.

MLVA was first proposed in 2000 to study *Bacillus anthracis* (Keim et al. 2000). Similar to MLST, it also provides reliable results that lend itself well to the establishment of global reference database and online tools (Guigon et al. 2008, Volpe Sperry et al. 2008, Grissa et al. 2008) because information can be easily coded and stored. MLVA can also be standardized by defining the specific VNTR loci used for particular schemes, and is one of the few relatively low cost methods that can be implemented globally without any specialized equipment (Martín et al. 2018). Together with the increase in resolution relative to previous methods (e.g., MLST and PFGE), it is frequently used in epidemiological research and was applied to numerous human pathogens such as *Vibrio cholerae* (Garrine et al. 2017, Bwire et al. 2018), *Salmonella* Typhimurium (Torpdahl et al. 2007), and *Clostridium difficile* (Marsh et al. 2010).

However, MLVA is not without limitations. The first major drawback is the detection of VNTRs. There are, in general, two ways to detect VNTRs; PCR-based methods and sequencing based methods. The easiest PCR-based method is perhaps multiplex PCR, where all loci can be analyzed at once (Sabat et al. 2013). The problem with this method is that it is impossible to determine which band corresponds to which locus. As a result, multiplex PCR will only produce a banding pattern that can be used to distinguish between strains, but will not produce a MLVA profile that can be easily stored and communicated. To circumvent this problem, it is possible to

perform single PCR reactions for each locus, which allows for the determination of the number of repeats based on band sizes. The problem with determining the number of repeats based only band size, whether it be through traditional gel electrophoresis or high-resolution capillary electrophoresis, is difficult and will produce variable results (Martín et al. 2018). In addition, there are also other mutations that can cause changes in band size without changing the number of repeats (Sabat et al. 2013). Ultimately, to obtain the true MLVA profile, we must sequence this region. Unfortunately, sequencing VNTR regions is notoriously difficult and prone to error whether through traditional Sanger sequencing or next generation sequencing techniques. In both cases, additional quality filter criteria are required to eliminate unreliable results (Kieleczawa 2006, Bartels et al. 2014). All PCR methods and also some sequencing methods also rely on the development of universal primers for these VNTR regions, which makes having a representative initial dataset crucial; however, often this is not possible resulting in primers that cannot be used when additional isolates are examined (Lindstedt 2005).

Another common limitation of MLVA is size homoplasy (Chenal-Francisque et al. 2013) where isolates share the same MLVA profile through convergent evolution and not by vertical descent. As a result, MLVA profiles may not reflect the true phylogenetic relationship and organisms may appear more closely related than they truly are. Although increasing the number of loci may partially compensate for this problem (Vergnaud and Pourcel 2009), this does not work in all cases and at times, accurate mutation rates are required to assess the quality of VNTR data (Kalvisa et al. 2016).

1.3.2 Bacterial subspecies level classification in the genomic era

Methods in the pre-genomic era were limited by sequencing technology, which is why both MLVA and MLST focused on only a limited number of loci. Just as the introduction of Sanger sequencing had a tremendous impact on bacterial typing, the introduction of next generation sequencing (NGS) also led to major innovations in subspecies typing techniques available for bacteria. Two primary methods are the gene-by-gene approach (Maiden et al. 2013) and single nucleotide polymorphism (SNP)-based method.

The gene-by-gene method is an extension of the traditional MLST approach. It works on the same concept as MLST, but instead of looking at six to seven genes, it utilizes thousands of genes. It retains the intuitive classification methods as traditional MLST; however, by increasing the number of genes considered, it has significantly increases the resolution. A commonly used gene-by-gene method is core-genome MLST (cgMLST).

cgMLST relies on the identification of a set of core genes, which are defined as genes shared by all isolates of interest. Typically, a diverse dataset of limited number of isolates are used to establish the scheme before evaluating its applicability on a larger sample set as it is computationally easier to do so. Although, strictly speaking, core genes must be present in all isolates within the data, due to limitations in sequencing technology, annotations and subsequent bioinformatic analyses, a more relaxed cutoff has sometimes been used for initial scheme development (Moura et al. 2016, Neumann et al. 2019), although not always (Ruppitsch et al. 2015, de Been et al. 2015). Since its introduction in 2013 (Maiden et al. 2013), cgMLST has been applied to numerous pathogenic bacteria highlighting its applicability in an epidemiological

setting (Antwerpen et al. 2015, Mellmann et al. 2016, Moura et al. 2016, Gonzalez-Escalona et al. 2017, Pearce et al. 2018).

One of the major advantages of cgMLST is the increased resolution allowing us to distinguish between closely related strains (Gonzalez-Escalona et al. 2017, Cody et al. 2017), while maintaining the intuitive classification system of traditional MLST. Although it has been known since MLST was initially introduced that the inclusion of greater number of genes will increase the resolution, it was not feasible due to time and cost constraints (Maiden et al. 1998). However, as whole-genome sequencing is becoming a routine practice in laboratories around the world and with readily available genome annotation software, this is no longer a limitation. In fact, cgMLST has been shown to be technically feasible and cost effective for real-time tracking of multi-drug resistant bacteria (Mellmann et al. 2016). Another important advantage of cgMLST is that it is compatible with traditional MLST methods. As MLST has been commonly used to study a wide range of organisms, ideally, new typing methods can be used to build on top of that information. Since MLST uses housekeeping genes, which are included in the core genome, any genes used in traditional MLST schemes should also be present in current cgMLST schemes. This means that by performing a single cgMLST analysis, it is possible to extract MLST data for all currently available MLST schemes. This increased in resolution, standardizability together with its compatibility with traditional MLST make cgMLST the preferred method for epidemiological research.

Despite its advantage, cgMLST also has important limitations. One of them is the reliance on the initial dataset for scheme development, sequencing technology and assembly software. As mentioned before, numerous cgMLST schemes were developed from an initial dataset containing a limited number of isolates. Although this approach reduces the

computational cost, it does require the initial dataset to be a representative sample as otherwise, it will result in a significant number of genes that are wrongly labeled as core genes, when in fact they are only present in a subset of the population. Sequence quality and assembly software also have a significant impact on core gene identification. Poorly sequenced and/or assembled genomes will result in gaps in the assemblies, which will reduce the number of core genes detected, lowering the resolution of the scheme. As a result, often there are quality filter criteria for genomes that can be used for cgMLST analysis (Moura et al. 2016). However, these limitations are primarily due to current limitations in technology whether at the sequencing step or any subsequent bioinformatics analyses. These are currently avoidable problems by ensuring schemes are developed and applied to reasonable quality data. In the future as these techniques mature, it is expected that these issues will become less significant.

Various commercial software, such as BioNumerics (http://www.applied-maths.com/bionumerics) and SeqSphere^{+} (https://www.ridom.de/seqsphere/), and open-source software such as BIGSdb (Jolley and Maiden 2010) are available to implement cgMLST. All are capable of storing gene sequences, metadata, and identify allelic profiles. These tools will help standardize and automate the process of developing cgMLST schemes and further promote its use internationally. BIGSdb in particular is a web-application that is able to host many MLST, cgMLST and any variants of the gene-by-gene approach, allowing for the centralization of gene-based typing methods and the facilitation of global communication (Jolley and Maiden 2010). Open-source programs such as OrthoMCL, USearch and BPGA (Li et al. 2003, Edgar 2010, Chaudhari et al. 2016) can also help identify core genes with more customizable parameters. The increased flexibility in gene identifications provided by these tools will allow for the

development of more customized cgMLST schemes based on known ecology and evolutionary patterns.

One of the limitations of gene-by-gene method is the requirement for assembled genomes. SNP-based methods on the other hand, allow researchers to work with raw NGS reads thereby effectively reducing computational costs and associated errors. The SNP identification process is the same as to those used for reconstructing SNP-based phylogeny (section 1.2.3). Regardless of the variant calling method used, the post processing of SNPs is the same.

It has been shown that SNP-based methods have comparable resolution to gene-based methods (Chen et al. 2013, Qin et al. 2016). Development of online analysis tools have allowed researcher to process raw NGS reads and produce meaningful SNP data in as little as two hours making SNP-based method a feasible tool for routine clinical practice (Chen et al. 2013). SNP-based analysis have also been applied to various pathogens and have shown to be able to identify outbreaks (Chen et al. 2013, Katz et al. 2013, Kanagarajah et al. 2016, Qin et al. 2016). Another benefit of SNP based analysis is that the number of SNPs considered can be modified to achieve the best balance between cost and resolution. In *Salmonella enterica*, for example, as a little as 68 out of 22,000 SNPs can differentiate between major pathogenic strains (Wong et al. 2016), and in theory, one can choose any number between 68 and 22,000 SNPs to achieve the desired level of resolution.

A limitation to SNP-based methods, whether it is reference-free or reference-based, is that they are highly dependent on the programs used and the parameters chosen (Pightling et al. 2014, Schürch et al. 2018). Different programs and protocols have different quality filters and SNP calling criteria, which will produce different results; standardization is therefore required

for global participation. In comparison to gene-by-gene methods, SNP based methods are also more prone to biases caused by recombination and horizontal gene transfer as each event can create multiple SNPs and these must be dealt with accordingly (Schürch et al. 2018). Ultimately, this means that the number of SNP differences may not correlate to true evolutionary distance. Extra precaution is therefore required to avoid SNPs that are the result of recombination and in some organisms, anywhere from 30% to 97% of SNPs identified must be removed (Chen et al. 2013, Qin et al. 2016).

A number of tools are also available for SNP-based analyses. BioNumerics is a common commercial software for these analyses. It is convenient to use as all steps of the analysis can be done with a single package. As mentioned earlier, it is also capable of performing gene-by-gene analyses, so researchers can easily perform both SNP-based and gene-by-gene-based study for the same dataset using this tool. The major limitation of BioNumerics is the cost as not only must users purchase the software license, there are also an additional cost associated with performing subsequent analyses, such as SNP identification. Open-source alternatives, such as SAMtools, are available for performing SNP-based analysis using raw reads (Li et al. 2009). There are also other open-source tools to map reads onto reference genome for SNP identification, such as SMALT, BWA and MOSAIK, that allow for more customizability in the quality filter criteria (http://www.sanger.ac.uk/science/tools/smalt-0, Li and Durbin 2009, Lee et al. 2014). Different software have different advantages and excel at dealing with reads of varying quality and species of varying diversity; therefore, additional analyses should be perform to assess the suitability of each tool (Pightling et al. 2014).

1.4 Thesis objective and outline

1.4.1 Resolving taxonomic inconsistencies within the Rhodobacteraceae *family: proposal to move the roseobacter clade into* Roseobacteraceae *fam. nov and numerous genus and species level changes (Chapter 2)*

Rhodobacteraceae is a diverse family of bacteria found in many different environments (Simon et al. 2017). The diverse metabolic and phenotypic abilities mean members of this family play important environmental and ecological roles (Buchan et al. 2005, Moran et al. 2007). This family also includes one of the most abundant group of marine bacteria, the roseobacter clade, which were among the first group of marine bacteria to be cultivated in a laboratory setting; therefore, it is widely studied by oceanographers worldwide (Buchan et al. 2005). The cultivation of members of this family also corresponds with the rising popularity of 16S rRNA gene sequencing, and as such, taxonomic classifications relied heavily on this gene (Buchan et al. 2005). As previously discussed, although we aim to create monophyletic groups, 16S rRNA gene alone is not sufficient as it lacks resolution at the genus and species level. As a result, numerous taxonomic inconsistencies exist within this family.

Applying the polyphasic approach to taxonomy and utilizing the abundance of WGS now available, I aim to identify and resolve any taxonomic inconsistencies in this family. As type strains serve an important role as reference material for taxonomic purposes and have additional phenotypic information available, they were my focus in this study. I collected the genome sequences of over 300 type strains within the *Rhodobacteraceae* family and performed various phylogenetic and genomic similarity analyses to identify, family, genus and species level misclassifications based on currently accepted standards (e.g., ANI, AAI, dDDH). I identified 25

genera and 3 species that violated existing taxonomic standards. My analyses also showed strong evidence that the roseobacter clade, members of which are commonly found in the marine environment, should be moved to a novel family for which I propose the name *Roseobacteraceae* fam. nov.

1.4.2 Development of a cgMLST scheme for subspecies level classification of Vibrio cholerae and application in an epidemiological setting (Chapter 3)

Vibrio cholerae is best known as the causative agent of the acute diarrheal disease cholera. Cholera is a pandemic disease that has its origin in at least the early 19th century. Currently, it affects over 53 countries around the world causing millions of deaths annually (Ali et al. 2015). Epidemiological research and surveillance of a pandemic disease such as cholera requires a global effort. As such, a standardized, high-resolution typing scheme is required. Current methods for subspecies level identification of *V. cholerae*, such as SNP-based methods, MLST and MLVA, all suffer from important limitations that make them unsuitable for global surveillance effort. Using the readily available WGS of over 1,200 *V. cholerae* isolates, I have developed a core-genome MLST (cgMLST) scheme for typing *V. cholerae*. By utilizing over 2000 core genes, it provides greater resolution than any previously used methods. As cgMLST can be easily standardized and automated, it can analyze thousands of genomes efficiently. Utilizing PubMLST, cgMLST is able to collect and store data from *V. cholerae* genomes collected around the world allowing researchers to analyze their own datasets in a global context.

I evaluated the applicability of this scheme by applying it to the two best documented cholera outbreaks in modern history; Haiti and Yemen. The strength of this scheme is also

highlighted by a direct comparison with currently established methods such as SNP-based methods, MLST and MLVA.

Chapter 2: Roseobacters in a sea of poly- and paraphyly: whole genome-based taxonomy of the family *Rhodobacteraceae* and the proposal for the split of the "roseobacter clade" into a novel family, *Roseobacteraceae* fam. nov.

A version of chapter 2 will be submitted for publication as:

> "Liang, K.Y.H., Orata, F.D., Boucher, Y.F., and Case, R.J. Roseobacters in a sea of poly- and paraphyly: whole genome-based taxonomy of the family *Rhodobacteraceae* and the proposal for the split of the "roseobacter clade" into a novel family, *Roseobacteraceae* fam. nov."

KYHL, YFB and RJC design the study and wrote the manuscript. FDO and KYHL performed bioinformatic analyses and collected data. YFB and RJC supervised the project.

Chapter 2

2.1 Abstract

The *Rhodobacteraceae* family is a group of α-proteobacteria that is metabolically, phenotypically and ecologically diverse. It includes one of the most abundant group of marine bacteria, the roseobacter clade. The rapid pace of discovery of novel organisms in this clade in the last two decades meant that best practice for taxonomic classification, a polyphasic approach utilizing phenotypic, genotypic, and phylogenetic characteristics, was not always followed. Early efforts for the classification of these bacteria relied heavily on 16S rRNA gene sequence similarity and resulted in numerous taxonomic inconsistencies, with several poly- and paraphyletic genera within this family. Next-generation sequencing technologies have allowed whole-genome sequences to be obtained for most type species in this group, making a revision of its taxonomy possible. In this study, I performed whole-genome phylogenetic and genotypic analyses combined with a meta-analysis of phenotypic data to review taxonomic classifications of 331 type strains within the *Rhodobacteraceae* family. I identified three isolates that were misclassified as a novel species, and as such these were merged with existing species. In addition, I also identify seven paraphyletic genera and 17 polyphyletic genera, which were resolved by merging and splitting them as necessary. Members of the roseobacter clade not only have different environmental adaptions from other isolates within the family, but were also found to be different based on genomic, phylogenetic, and *in-silico* phenotypic data. As such, I propose to move this group of bacteria into a new family, *Roseobacteraceae* fam. nov. By resolving taxonomic inconsistencies of type strains within this family, I have established a set of coherent criteria based on whole-genome analysis that will help guide future taxonomic efforts and prevent the propagation of errors.

2.2 Introduction

Taxonomy is the science of characterizing, naming, and classifying organisms. Its aim is to group together organisms with shared evolutionary history from which they derived shared traits meaningful to their ecology and physiology (Wayne et al. 1987) and to this end, numerous criteria were established (Wayne et al. 1987, Konstantinidis and Tiedje 2005a, Mignard and Flandrois 2006, Meier-Kolthoff et al. 2013, Qin et al. 2014). These criteria were defined based on our understanding of prokaryotic diversity as well as the technology available. With each new method developed, we improve our understanding of bacterial diversity and evolution. As a result, many earlier taxonomic classifications are re-evaluated and modified (Parks et al. 2018, Wirth and Whitman 2018), as they are hypotheses that should be continuously verified when better techniques are available to reflect the current understanding (Garrity 2016).

Microbial taxonomy has changed substantially in the past few decades, embracing a polyphasic approach – phenotypic, genotypic, and phylogenetic – that considers a wide range of different traits for a systematic identification and description (Vandamme et al. 1996). Early phenotypic tests were primarily based on morphological characteristics; however, as morphological traits can vary within particular species, additional phenotypic tests were developed. This has led to what is known as numerical taxonomy, in which bacteria are represented by a long sequence of numbers, each representing individual phenotypic test (Sneath 2005). To make this process easier, we now have commercially available phenotypic tests, such as the Biolog system (https://www.biolog.com/products-portfolio-overview/microbial-identification/), that provides a systematic method to describe bacteria isolates base on a wide range of traits including, among others, carbon utilization, and antibiotic resistance. Despite the many standardized tests available, bacteria are rarely classified based on common phenotypic

tests. This is partly because bacteria are metabolically and phenotypically diverse and at times atypical phenotypic tests are required for species with specialized adaptive traits (Tindall et al. 2010). Labor and time constraints also dictate the number of phenotypic tests that can be done, and the number of isolates analyzed. Despite these limitations, phenotypic tests are still an invaluable part of every taxonomic classification regardless of ranks because it allows for detailed descriptions that highlight important ecological roles and/or clinical traits. Fortunately, next generation sequencing (NGS) has made it easier to determine phenotypic traits. This is because we are now able to extract phenotypic traits of medical, ecological, and physiological importance from WGS through *in-silico* means (Aziz et al. 2008, Kanehisa et al. 2016). Although *in-silico* phenotypic predictions are ultimately predictions that should be verified through experiments, these are based on extensive genomic and phylogenetic analyses and can still nonetheless help guide subsequent phenotypic test which will reduce both time and costs required to identify characteristic phenotypes useful for taxonomic purposes.

Genotypic and phylogenetic analyses are the other two important pillars of polyphasic taxonomy (Wayne et al. 1987). The earlier methods include G+C content deviation, 16S rRNA gene analyses, and DNA-DNA hybridization (DDH). Although these are still used today (Sasi Jyothsna et al. 2016, Rabus et al. 2019), they have important limitations.

DDH is proposed to be the gold standard for species delineation. It is widely accepted that 70% DDH at 5^{o}C melting temperature is the species cutoff, where isolates belong to the same species only if they meet or surpass this threshold (Wayne et al. 1987). DDH is notoriously difficult to reproduce and also time and labor intensive. In addition, only a limited number of laboratories are able to reliably perform DDH experiments (Gevers et al. 2005). Therefore, although this technique is seen as the gold standard for taxonomy, other methods are being

explored as a replacement for traditional DDH experiments or at least attempt to accurately predict traditional DDH results.

G+C content is one of the earlier methods that aim to predict DDH values. Conventional methods for measuring G+C content were indirect as they relied on variation in physical properties, such as melting temperature. These methods were later determined to be the primary sources of variations of past G+C content measurements (Klenk et al. 2014). Obviously, this limits the use of G+C content in bacterial taxonomy as it is impossible to attribute whether the variation seen is the result of the technique or actual differences. It is now possible to measure G+C content directly and more accurately using WGS and has been shown that within species variation of G+C content is usually less than 1% (Klenk et al. 2014). However, even with this more reliable method to measure G+C content, it alone cannot be used for species identification as even 0% G+C content deviation only corresponds to approximately 84% chance that traditional DDH values will be above the species cutoff of 70% (Klenk et al. 2014). In addition, isolates from different genera can have as little as 0.4% difference in their G+C content.

The most popular method used to predict DDH values is 16S rRNA gene analysis, as evident by the numerous 16S rRNA gene databases available such as Greengenes (https://greengenes.secondgenome.com), SILVA (https://www.arb-silva.de) and RDP (https://rdp.cme.msu.edu). It is used not only as a proxy for genomic similarities, but also to reconstruct phylogenetic trees. However, it should be noted that although 16S rDNA similarity values below 97.5% (Stackebrandt and Goebel 1994), and later 98.7% (Stackebrandt and Ebers 2006), means that high DDH values are unlikely, 16S rDNA identities above these thresholds do not necessarily guarantee high DDH values. This is in part because the full-length 16S rRNA gene, which is approximately 1.4kb long, is only a fraction of the size of a bacterial genome,

which can range from 130kb to 14Mb. It has been shown that isolates sharing up to 99% 16S rRNA gene sequence identity can still have DDH values as low as 23-50% (Wang et al. 2015) highlighting the discrepancy between 16S rRNA gene sequence identity and traditional DDH experiments. In addition to the lack of resolution, some isolates also contain multiple intragenomic copies with sequence different enough to be considered different genera based on phylogenetic analyses (Klappenbach et al. 2001, Acinas et al. 2004, Boucher et al. 2004, Case et al. 2007), adding yet another reason why the 16S rRNA gene alone is not suitable for determining taxonomic classification. 16S rDNA similarity should therefore only serve as a preliminary guide as to whether more in-depth genomic analyses are required (Tindall et al. 2010). Unfortunately, despite the increasing ease of bacterial genome sequencing, 16S rRNA gene-based phylogeny is still commonly used to fulfil the phylogenetic aspect of polyphasic taxonomy (Kim et al. 2010, Baek et al. 2015, Shin et al. 2017).

Fortunately, WGS has provided the basis to develop more accurate methods to assess genomic and phylogenetic relationships and has provided us with tangible standards for systematic classifications, more so for the species level than higher ranks. Many WGS-based methods developed have their basis in traditional DDH experiments, as it is still the gold standard for species delineation; however, unlike G+C content and 16S rRNA analysis mentioned above, these methods are much better able to predict traditional DDH experiments. One such method is average nucleotide identity (ANI), which is primarily used for species level delineation. It was determined that 95% ANI corresponds to 70% DDH values and is proposed as the species cutoff (Goris et al. 2007). Another method used for species classification that has its basis in traditional DDH is *in-silico* DDH (dDDH) calculated based on genome to genome distance and it retains the 70% species cutoff (Meier-Kolthoff et al. 2013). As ANI and dDDH

are reproducible and easily scaled to analyze hundreds of isolates, it is becoming standard practice for species delineation (Orata et al. 2016, 2018, Dees et al. 2017, Wirth and Whitman 2018).

Taxonomic standards become increasingly more ambiguous as we move to higher ranks. One method proposed was average amino acid identity (AAI), which is similar to ANI, but measures amino acid sequence identity instead. AAI is more suitable than ANI to assess higher taxonomic ranks, or more distantly related species, because amino acid does not reach mutational saturation as quickly as nucleotides (Qin et al. 2014). Although AAI can be used to resolve genus level relationships, there has yet to be an established genus level threshold as values for between genera comparisons can be anywhere from 60%-80% (Orata et al. 2018). Therefore, although AAI can be helpful when identifying comparisons that are different from expected and highlight potential misclassifications, unlike dDDH and ANI analyses, it will not provide a definitive answer. A polyphasic approach that includes detailed phylogenetic and genomic analyses supplemented with phenotypic data are still required for proper genus and higher-level classifications.

In this study I utilized the abundance of high-quality WGS data to perform a large-scale phylogenomic analyses on the type strains within the family *Rhodobacteraceae* to identify and resolve taxonomic inconsistencies. This family is the largest family within the order *Rhodobacterales* and is, according to the 2015 publication of the Bergey's manual, metabolically, phenotypically and genotypically diverse (Garrity et al. 2015b). *Rhodobacteraceae* was first proposed in 2005 based on 16S rRNA gene analysis and was named after the first described genus, *Rhodobacter* (Garrity et al. 2005). Because of its reliance on the 16S rRNA gene, taxonomic inconsistencies have been a reoccurring problem in this family since

it was first proposed. In fact, *Rhodobacteracea* was not even a legitimate name when it was proposed, as it included the genus *Hyphomonas*, the type genus of the family *Hyphomonadaceae* which violated rule 51b of the International Code of Nomenclature of Prokaryotes (Parker et al. 2019). It was only recognized as an official family within the order *Rhodobacterales* when *Hyphomonadaceae* was moved to the order *Caulobacterales* later that year based yet again on 16S rRNA gene-based phylogenetic analyses (Pujalte et al. 2007). However, the 2015 publication of the Bergey's manual placed *Hyphomonadacea* in the order *Rhodobacterales* again (Abraham and Rohde 2019), underlining once more the instability of taxonomic classifications at the family level based on the16S rRNA gene. This gene is therefore not suitable for family level analysis within this order let alone at the genus or species level.

Within the family *Rhodobacteraceae* is the roseobacter clade. Members of this group are historically important as it is one of the most readily cultivated groups of marine bacteria (Buchan et al. 2005). Members of the roseobacter clade can consists of up to 20% of coastal marine bacterial population making it one of the most abundant groups of marine bacteria (Moran et al. 2007). The roseobacter clade also plays important ecological and environmental roles. For example, it is one of two groups of bacteria that contains isolates capable of both pathways for dimethylsulfoniopropionate (DMSP) degradation; DMSP demethylation and DMSP cleavage (Luo and Moran 2014). These pathways utilize DMSP in different ways and are important for different reasons. The DMSP demethylation pathway converts DMSP into methanethiol (MeSH), which can be assimilated by marine bacteria (Reisch et al. 2011). The cleavage pathway, on the other hand, converts DMSP into DMS, which plays an important role in cloud formation and ultimately global climate regulation (Reisch et al. 2011, Moran et al. 2012). Much like the initial circumscription of the family *Rhodobacteraceae*, taxonomic

classifications within this group continues to rely heavily on 16S rRNA gene phylogeny. As a result, multiple studies have been published highlighting the numerous genus and species level misclassifications within this group (Breider et al. 2014, Wirth and Whitman 2018, Huang et al. 2018), but none has addressed it systematically.

To resolve taxonomic inconsistencies within this family, I focused on all type strains with draft or complete genomes available as of January 13th, 2019. By resolving taxonomic misclassifications among type strains, it will establish a set of taxonomically correct reference material that can help guide future taxonomic efforts and prevent the propagation of error. In addition, type strains also provide phenotypic data for a meta-analysis allowing us to more closely follow the polyphasic approach to taxonomy. A total of 25 genera and 3 species that violated existing taxonomic rules have been identified and must be addressed.

2.3 Results and Discussion

2.3.1 16S rRNA gene phylogeny provides little resolution within the **Rhodobacteraceae** *family relative to core genome analysis*

The 16S rRNA gene has played a major role in the taxonomic classification of many members within the *Rhodobacteraceae* family (Garrity et al. 2015b). The importance of 16S rDNA is further highlighted by the fact that the largest lineage within this family, the marine roseobacter clade, is defined by having members that share >89% 16S rRNA gene sequence similarity (Buchan et al. 2005). To determine the impact of using the 16S rRNA gene as the main molecular marker for naming new species and genera within this family, we reconstructed the phylogenetic tree of 331 type strains using full lengths 16S rRNA gene sequences, which are

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 (Fig 2.1). This is even more evident when nodes with less than 50% bootstrap support were collapsed, resulting in a poorly resolved tree backbone (Fig. 2.1). The inadequacy of 16S rRNA gene for use in genus-level classification is highlighted by the fact that only 23 genera in the entire family are monophyletic (Fig 2.1). In addition, genera such as *Yoonia*, *Loktanella* and *phaeobacter* which have previously been shown to be monophyletic based on a whole-genome approach (Breider et al. 2014, Wirth and Whitman 2018) are no longer monophyletic in the 16S rRNA tree (Fig 2.1, Fig 2.2).

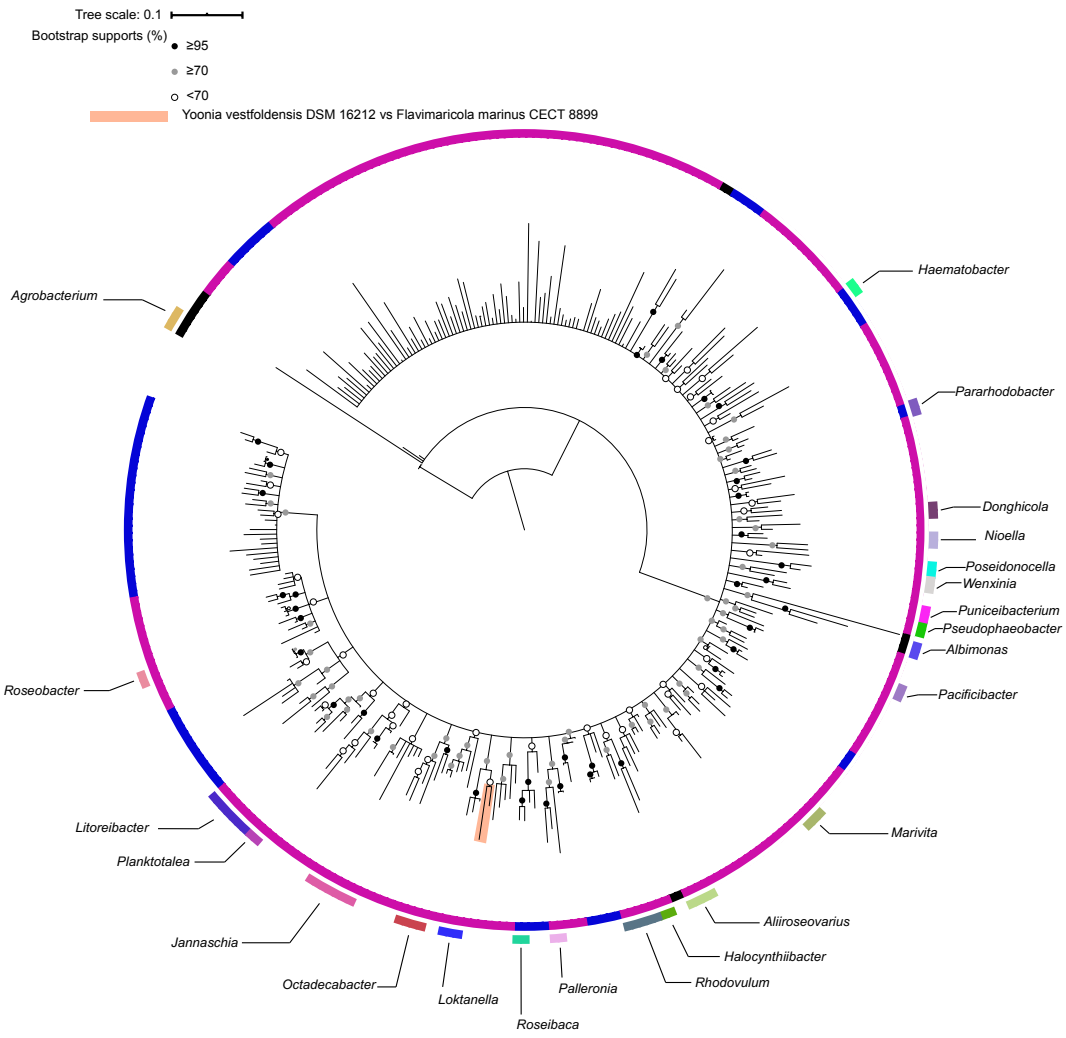
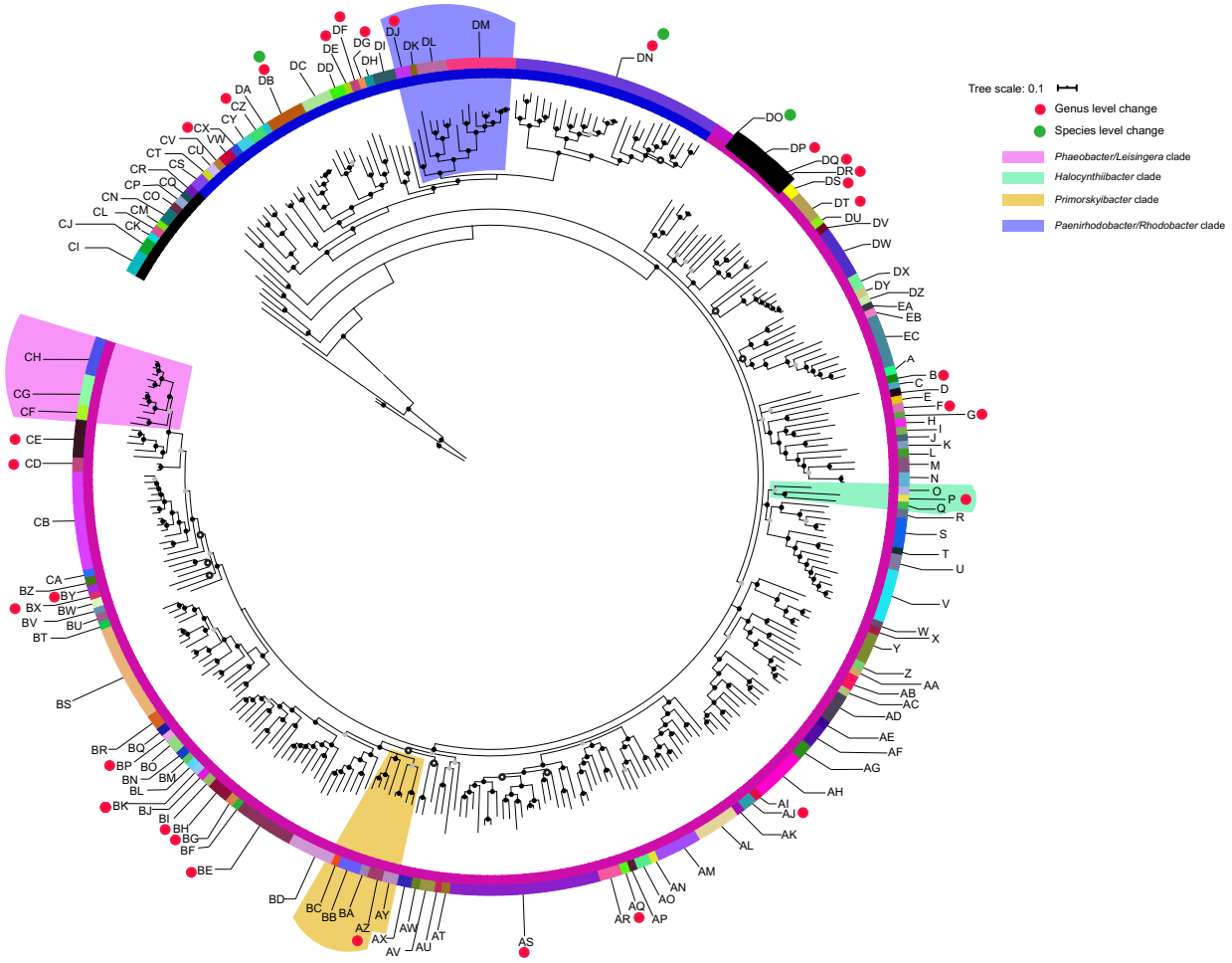


Fig 2.1: Phylogenetic tree of 331 *Rhodobacteraceae* type strains based on the full-length 16S rRNA gene. The tree was reconstructed using RAxML v8.2.11 with 1000 bootstrap replicates under the GTRGAMMA model and rooted with *Agrobacterium tumefaciens*. Bootstraps are indicated as black (≥95%), grey (≥70%) and white (<70%) circles. The inner ring represents the two major lineages within the family (Fig 2.2) and outer ring represents monophyletic genera. Blank regions of the outer ring represent polyphyletic genera based on this 16S rRNA gene tree.

A popular high-resolution method to reconstruct phylogenetic relationships is the core-genome approach. In this approach, phylogenetic trees are reconstructed with the concatenated alignments of core genes, which are defined as genes present in all organisms of interest. 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*, the phylogeny of the entire family was reconstructed in another study using 208 core genes from 106 strains (Simon et al. 2017). In my study, the core-genome phylogeny of the family was reconstructed using 140 core genes from a dataset of more than three times as large (331 type strains with the addition of three *Agrobacterium tumefaciens* outgroup; Table A1) as previously used (Simon et al. 2017), providing a more complete picture of the phylogenetic framework of the *Rhodobacteraceae* family (Fig 2.2). As expected, phylogenetic relationships are much better resolved in the core-genome phylogeny than with the 16S rRNA gene alone, with a well define backbone (Fig 2.2). Based on this core-genome tree, the *Rhodobacteraceae* family can be divided into two monophyletic lineages, one of which consists of the roseobacter clade and is composed primarily of organisms found in the marine environment (Buchan et al. 2005) (Fig 2.2 – inner ring, Table A1). It should be noted that the 16S rRNA gene phylogeny was unable to resolve these two lineages, meaning it would be difficult to determine even which lineage an isolate belongs to let alone which genus or species using this gene. In addition, genera that were not monophyletic in the 16S rRNA gene tree (e.g., *Yoonia*, *Leisingera*, and *Phaeobacter*) are now monophyletic with strong bootstrap support, consistent with prior studies (Wirth and Whitman 2018). However, several polyphyletic (*Albidovulum, Celeribacter, Defluviimonas, Gemmobacter, Lutimaribacter, Maribius, Oceanicola, Ponticoccus, Primorskyibacter, Pseudooceanicola, Pseudorhodobacter,*

Pseudoruegeria, Rhodobacter, Roseivivax, Ruegeria, Sulfitobacter, Thalassobius) and paraphyletic genera (*Paracoccus, Actibacterium, Tropicimonas, Roseovarius, Salipiger, Tropicibacter, Epibacterium*) remained (Table A2).



A: *Brevirhabdus*

B: *Pseudoalbidovulum*
Albidovulum xiamenense CGMCC 1.10789 → *Pseudoalbidovulum xiamenense* CGMCC 1.10789

C: *Hasllibacter* **D**: *Boseongicola* **E**: *Silicimonas*

F: *Roseivivax_A*
Roseivivax roseus DSM 23042 → *Roseivivax_A roseus* DSM 23042

G: *Maribius_A*
Maribius pontilimi GH1 23 1 → *Maribius_A pontilimi* GH1 23 1

H: *Tranquilimonas* **I**: *Oceaniglobus* **J**: *Kandeliimicrobium* **K**: *Oceaniovalibus*

L: *Hwanghaeicola* **M**: *Palleronia* **N**: *Maribius* **O**: *Pseudoruegeria*

P: *Pseudohalocynthiibacter*
Halocynthiibacter arcticus PAMC 20958 → *Pseudohalocynthiibacter arcticus* PAMC 20958

Q: *Halocynthiibacter* **R**: *Maritimibacter* **S**: *Aliiroseovarius* **T**: *Aquimixticola* **U**: *Pacificibacter*

V: *Celeribacter* **W**: *Marivivens* **X**: *Pseudooctadecabacter* **Y**: *Octadecabacter*

Z: *Oceanicola* **AA**: *Roseisalinus* **AB**: *Wenxinia* **AC**: *Pseudoroseicyclus* **AD**: *Limimaricola*

AE: *Flavimaricola* **AF**: *Loktanella* **AG**: *Cognatiyoonia* **AH**: *Yoonia* **AI**: *Salinihabitans*

AJ: *Cognatishimia*
Thalassobius activus CECT 5113 → *Cognatishimia activus* CECT 5113

AK: *Pseudopelagicola* **AL**: *Shimia* **AM**: *Pseudooceanicola* **AN**: *Sediminimonas*

AO: *Lutimaribacter* **AP**: *Litorimicrobium*

AQ: *Lutimaribacter_A*
Lutimaribacter litoralis KU5D5 → *Lutimaribacter_A litoralis* KU5D5

AR: *Thalassobius*

AS: *Roseovarius*
Pelagivirga sediminicola BH-SD19 → *Roseovarius sediminicola* BG-SD19
Pelagicola litorisediminis CECT 8287 → *Roseovarius litorisediminis* CECT 8287

AT: *Pseudaestuariivita* **AU**: *Rhodosalinus* **AV**: *Poseidonocella* **AW**: *Marinovum*

AX: *Donghicola* **AY**: *Puniceibacterium*

AZ: *Primorskyibacter*
Pseudooceanicola flagellatus CGMCC 1.12644
→ *Primorskyibacter flagellatus* CGMCC 1.12644

BA: *Primorskyibacter_A* **BB**: *Marivita* **BC**: *Thalassococcus* **BD**: *Roseivivax*

BE: *Salipiger*
Yangia pacifica CGMCC 1.3455 → *Salipiger pacifica* CGMCC 1.3455
Pelagibaca abyssi JLT2014 → *Salipiger abyssi* JLT2014

BF: *Citreimonas*

BG: *Roseivivax_B*
Roseivivax pacificus DSM 29329 → *Roseivivax_B pacificus* DSM 29329

BH: *Tropicibacter*
Pelagimonas varians DSM 23678 → *Tropicibacter varians* DSM 23678

BI: *Ponticoccus* **BJ**: *Sagittula*

BK: *Ponticoccus_A*
Ponticoccus marisrubri SJ5A-1 → *Ponticoccus_A marisrubri* SJ5A-1

BL: *Maliponia* **BM**: *Mameliella* **BN**: *Antarctobacter* **BO**: *Planktotalea*

BP: *Sulfitobacter_A*
Sulfitobacter pseudonitzschiae H3 → *Sulfitobacter_A pseudonitzschiae* H3

BQ: Ascidiaceihabitans **BR**: *Roseobacter* **BS**: *Sulfitobacter* **BT**: *Aestuariivita*

BU: *Pseudodonghicola* **BV**: *Marinibacterium* **BW**: *Sedimentitalea*

BX: *Ruegeria_A*
Ruegeria kandeliae J95 → *Ruegeria_A kandeliae* J95

BY: *Pseudooceanicola_A*
Pseudooceanicola lipolyticus 157 → *Pseudooceanicola_A lipolyticus* 157

BZ: *Jhaorihella* **CA**: *Cribrihabitans* **CB**: *Ruegeria*

CD: *Falsiruegeria*
Ruegeria litorea R37 CECT 7639 → *Falsiruegeria litorea* R37 CECT 7639
Ruegeria mediterranea M17 → *Falsiruegeria mediterranea* M17

CE: *Epibacterium*
Tritonibacter horizontis O3.65 → *Epibacterium horizontis* O3.65

CF: *Pseudophaeobacter* **CG**: *Phaeobacter*

CH: *Leisingera* **CI**: *Agrobacterium* **CJ**: *Ahrensia* **CK**: *Acuticoccus* **CL**: *Rubrimonas*

CM: *Oceanicella* **CN**: *Albimonas* **CO**: *Monaibacterium* **CP**: *Pontivivens* **CQ**: *Amylibacter*

CR: *Neptunicoccus* **CS**: *Parahodobacter* **CT**: *Roseicitreum* **CU**: *Roseinatronobacter*

CV: *Rhodobaca* **CW**: *Roseibaca*

CX: *Defluviimonas_A*
Defluviimonas indica DSM 24802 → *Defluviimonas_A indica* DSM 24802

CY: *Haematobacter*

CZ: *Gemmobacter_A*
Gemmobacter nectariphilus DSM 15620 → *Gemmobacter_A nectariphilus* DSM 15620
Gemmobacter megaterium DSM 26375 → *Gemmobacter_A megaterium* DSM 26375

DA: *Cereibacter*

DB: *Rhodobacter_A*
Rhodobacter megalophilus DSM 18937 → *Rhodobacter_A sphaeroides* DSM 18937
Rhodobacter sphaeroides 2.4.1 → *Rhodobacter_A sphaeroides* 2.4.1
Rhodobacter johrii JA192 → *Rhodobacter_A johrii* JA192
Rhodobacter azotoformans KA25 → *Rhodobacter_A azotoformans* KA25
Rhodobacter ovatus JA234 → *Rhodobacter_A ovatus* JA234

DC: *Pseudorhodobacter* **DD**: *Gemmobacter*

DE: *Pseudorhodobacter_A*
Pseudorhodobacter psychrotolerans PAMC 27389 → *Pseudorhodobacter_A psychrotolerans* PAMC 27389

DF: *Rhodobacter_B*
Rhodobacter blasticus DSM 2131 → *Rhodobacter_B blasticus* DSM 2131

DG: *Tabrizicola* **DH**: *Albidovulum* **DI**: *Defluviimonas*

DJ: *Rhodobacter_C*
Rhodobacter vinaykumarii JA123 → *Rhodobacter_C vinaykumarii* JA123
Rhodobacter veldkampii DSM 11550 → *Rhodobacter_C veldkampii* DSM 11550

DK: *Paenirhodobacter* **DL**: *Rhodobacter* **DM**: *Thioclava*

DN: *Paracoccus*
Paracoccus bengalensis DSM 582 → *Paracoccus versutus* DSM 582
Methylarcula marina VKM B-2159 → *Methylarcula marina* VKM B-2159

DO: *Rhodovulum*
Rhodovulum viride JA756 → *Rhodovulum kholense* JA756

DP: *Actibacterium*
Confluentimicrobium lipolyticum CECT 8621 → *Actibacterium lipolyticum* CECT 8621

DQ: *Oceanicola_A*
Oceanicola litoreus DSM 29440 → *Oceanicola_A litoreus* DSM 2944

DR: *Celeribacter_A*
Celeribacter manganoxidans DY25 → *Celeribacter_A manganoxidans* DY25

DS: *Tropicimonas*
Pseudoruegeria marinistellae SF-16 → *Tropicimonas marinistellae* SF-16

DT: *Pseudoruegeria_A*
Pseudoruegeria haliotis DSM 29328 → *Pseudoruegeria_A haliotis* DSM 29328
Pseudoruegeria lutimaris DSM 25294 → *Pseudoruegeria_A lutimaris* DSM 25294
Pseudoruegeria sabulilitoris GJMS-35 → *Pseudoruegeria_A sabulilitoris* GFMS-35

DU: *Planktomarina* **DV**: *Nereida* **DW**: *Litoreibacter* **DX**: *Nioella*

DY: *Roseicyclus* **DZ**: *Roseibacterium* **EA**: *Dinoroseobacter* **EB**: *Thalassobacter*

EC: *Jannaschia*

Fig 2.2: Phylogenetic tree of 331 *Rhodobacteraceae* type strains based on concatenated alignments of 140 core genes. The tree was reconstructed using RAxML v8.2.11 with 100 bootstrap replicates using the PROTGAMMAAUTO option for automatic model selection rooted with *Agrobacterium tumefaciens*. Bootstraps are indicated as black (≥95%), grey (≥70%) and white (<70%) circles. The inner ring represents the two major lineages within the family and outer rings represents monophyletic genera. Red and green dots represent where genus and species level changes respectively.

2.3.2 Evaluation of species designation within monophyletic genera

Unlike higher taxonomic ranks, there are clear genomic and phylogenetic criteria for species level delineation. dDDH and ANI are two common genomic metrics which uses 70% and 95% as the species threshold respectively (Richter and Rosselló-Móra 2009, Meier-Kolthoff et al. 2013). Phylogenetically, all isolates belonging to the same species must also be monophyletic (Rosselló-Móra and Amann 2015).

To identify possible taxonomic misclassifications at the species level, I calculated dDDH values for all species within currently named genera using the GGDC online tool (Meier-Kolthoff et al. 2013) (https://ggdc.dsmz.de). For polyphyletic genera, only comparisons within monophyletic clades are considered, as it is unlikely for two isolates that are not monophyletic to share more than 70% dDDH value. For cases where pairwise comparisons showed more than 70% dDDH, I then calculated ANI using JSpecies (Richter and Rosselló-Móra 2009) as a separate independent method of evaluating genomic relatedness. Only cases where dDDH, ANI and phylogenetic data support the merging of two species do I propose any taxonomic changes.

From all dDDH comparisons (Table A3), only three pairs of species had more than 70% dDDH value and 95% ANI (Fig 2.3, Table A3). The pairs of species were *Rhodobacter sphaeroides* 2.4.1^T (Imhoff et al. 1984) vs. *Rhodobacter megalophilus* DSM 18937^T (Arunasri et al. 2008), *Paracoccus bengalensis* DSM 17099^T (Ghosh et al. 2006) vs *Paracoccus versutus* DSM 582^T (Katayama et al. 1995), and *Rhodovulum viride* JA756^T (Srinivas et al. 2014) vs *Rhodovulum kholenses* DSM 19783^T (Kumar et al. 2008) with dDDH values of 82.9%, 81.9%, and 84.1 and the ANI values of 98.1%, 97.7% and 97.7% respectively (Fig 2.3, Table A3). The 16S rRNA gene sequence similarities (99.39%, 99.93%, 100% for *Rhodovulum*, *Paracoccus*, *Rhodobacter* respectively) were also higher than the 98.7% cutoff previously proposed

(Stackebrandt and Ebers 2006) (Table A4). In addition, all three pairs form monophyletic clades (Fig 2.2) with patristic distance (PD) ranging from 0.015 to 0.011 (Table A4). As PD is a measure of evolutionary distance based on our core-genome phylogenetic tree and the average within genera PD was 0.24 (Table A5), it shows that these pairs of isolates are not only monophyletic but are also closely related. Both phylogenetic and genomic evidence suggest these taxa should be renamed. In all cases, the species that was first identified will retain the species designation and all strain names are retained as per rule 38 of the code (Parker et al. 2019).

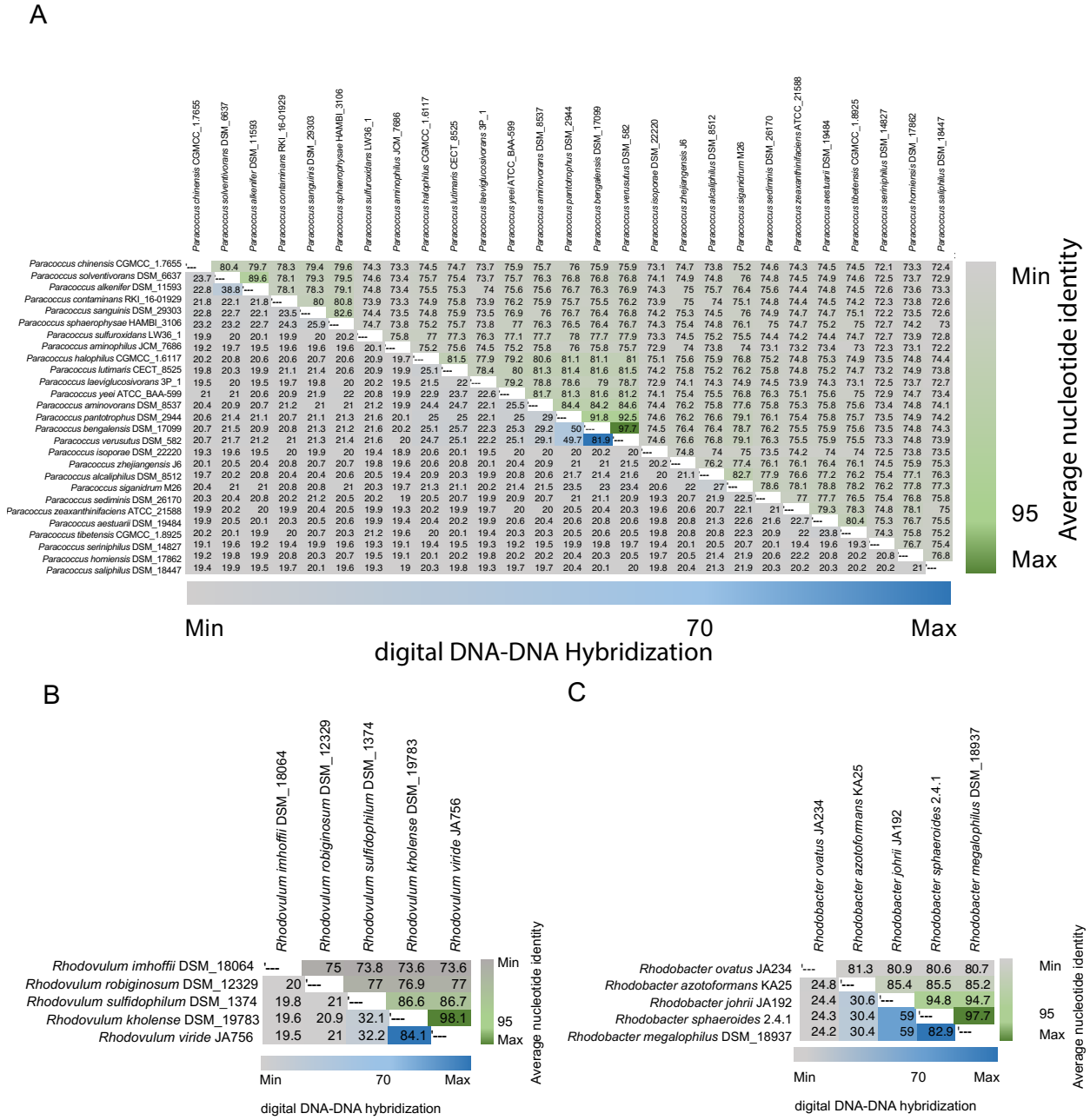


Fig 2.3: dDDH and ANI values showing the genera containing species level misclassifications. The species threshold (70% and 95% for dDDH and ANI respectively) are labeled on the X and Y axis.

Independent of my analyses, a recent publication specifically examined the taxonomic inconsistencies within the genus *Rhodobacter* using a concatenated phylogenetic tree based on 92 core genes (Suresh et al. 2019). In this study, the genus *Rhodobacter* was split into the same number of clades as our broader-scale core-genome phylogenetic analysis, highlighting the reliability and reproducibility of core-genome based phylogenies (Suresh et al. 2019) (Fig 2.2). This study, taken together with my analyses, provides strong evidence that *Rhodobacter* is indeed a polyphyletic genus that requires taxonomic reclassification. The same study also showed that *R. sphaeroides* 2.4.1^T and *R. megalophilus* DSM 18937^T are indeed the same species, which is consistent with my findings (Suresh et al. 2019) (Fig 2.3). As *R. megalophilus* DSM 18937^T differs from *R. sphaeroides* in a number of phenotypic traits including, among others, growth at 5^{o}C, vitamins required for growth and the ability to utilize citrate as carbon source, it was proposed that *R. megalophilus* should be considered a subspecies within *R. sphaeroides* 2.4.1^T (Suresh et al. 2019).

As for the other two pairs, I propose the following changes: *Rhodovulum viride* JA756^T (Srinivas et al. 2014) renamed as *Rhodovulum kholense* JA755^T comb. nov; and *Paracoccus bengalensis* DSM 17099^T (Ghosh et al. 2006) renamed as *Paracoccus versutus* DSM 17099^T comb. nov.

2.3.3 Genome guided genus level reclassifications supported by phylogenetic data.

Taxonomic classification at the genus level and higher is more difficult, as it lacks standardized metrics or guidelines. Although attempts were made to establish genomic standards for genus level classification, there has yet to be a consensus on what analyses to run and cutoff values to use (Orata et al. 2018). As a result, a polyphasic approach that also looks at phenotypic

and phylogenetic data is favored for assigning taxonomic classifications above the species level. However, it should be noted that although genome similarity analyses alone are not sufficient to justify genus level reclassifications, the relative ease in analyzing hundreds of isolates using a variety of metrics makes these methods effective initial approaches for sifting through a large quantity of data to identify potential misclassifications. These can be further examined from a phylogenetic and phenotypic perspective; both of which are more time-consuming and computationally intensive.

In the past, genus definition relied heavily on 16S rRNA gene sequence analyses. As such, genomically dissimilar organisms are sometimes grouped into the same genera because relatively distantly related organisms may still have similar 16S rRNA gene sequences. Take *Yoonia vestfoldensis* DSM 16212^T (Wirth and Whitman 2018) and *Flavimaricola marinus* CECT 8899^T (Wirth and Whitman 2018) for example. *Y. vestfoldensis* DSM 16212^T and *F. marinus* CECT 8899^T share a 96% 16S rDNA identity; however, *Y. vestfoldensis* DSM 16212^T has lower AAI, 1st, 2nd, and 3rd codon position similarity and higher PD when compared to *F. marinus* CECT 8899^T relative to other *Yoonia* species (Table A5). Therefore, if genus classifications were assigned based only on 16S rRNA gene sequence identity and 16S rRNA gene-based phylogeny, these two isolates would be grouped into the same genus despite being genomically dissimilar. This was indeed the case in the past (Van Trappen et al. 2004, Jung et al. 2016); however, genomic similarity analyses alone showed these isolates do not belong in the same genus and further phylogenetic and phenotypic analyses corroborated these results, leading to their reclassification in 2018 (Wirth and Whitman 2018). This highlights the importance of genomic similarity analyses as efficient methods for quickly identifying potential misclassifications that can help guide subsequent analyses.

Genomic metrics are therefore used to determine if there are any misclassifications among currently recognized monophyletic groups (Table A2, A5). It is clear that for all genomic metrics measured, that species within the same genus are more similar to each other than species between genera, as values for within genera comparisons are significantly different from between genera comparisons in all cases (Fig 2.4, Table A5). It is also worth noting that between and within genera comparisons always have some overlaps for all metrics considered (Fig 2.4). These overlaps are not unexpected, as even closely related genera can have different evolutionary rates due to differences in response to environmental factors (Ramette and Tiedje 2007), which means genera will contain species of varying degrees of diversity. This overlap is the primary reason why establishing a clear universal genus level boundary is difficult if not impossible.

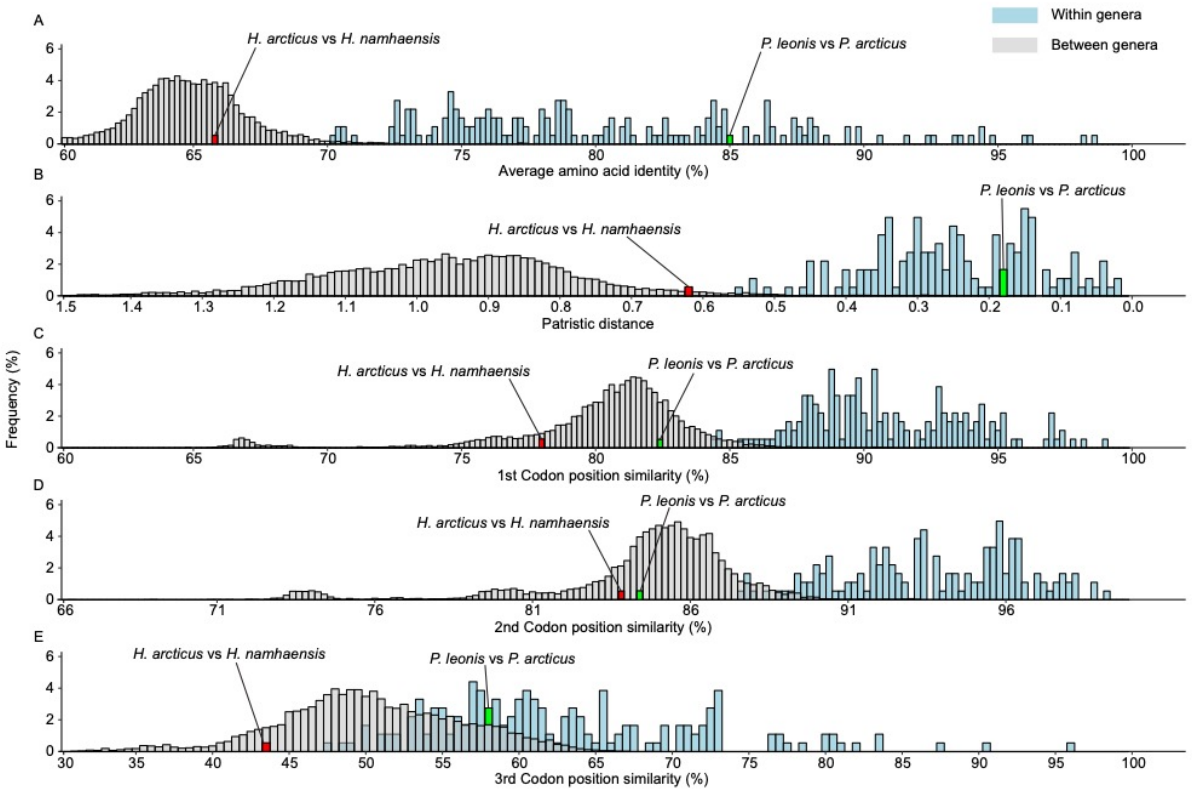


Fig 2.4: Genomic similarity within and between genera. Histogram of AAI, PD, 1st, 2nd and 3rd codon position similarity for all within (light blue) and between (grey) recognized monophyletic genera (Table A5). The distributions for within and between genera comparisons for all metrics are statistically significant ($p < 0.05$) based on Mann-Whitney U-test. The red and the green bars represent two atypical within genus comparisons.

All within genera comparisons have AAI values above 70% with only one exception; the comparison between *Halocynthiibacter arcticus* PAMC 20958^T (Baek et al. 2015) and *Halocynthiibacter namhaensis* RA2-3^T (Kim et al. 2014) at 65.8% (Fig 2.4); the only two named species within this genus as of November 28th, 2019. Other genomic metrics also show a similar pattern where these two isolates consistently have values more similar to those observed for between genera comparisons rather than within genus comparisons. As codon position similarities and AAI are measures of genomic similarities and PD is a measure of evolutionary distance, collectively these metrics show that *H. arcticus* PAMC 20958^T and *H. namhaensis* RA2-3^T are genomically and phylogenetically distinct, which is supported by the core-genome phylogeny as indicated by the relatively long branch length (Fig 2.2).

It is likely that *H. articus* PAMC 20958^T was misclassified, as it was originally identified to belong to the genus *Halocynthiibacter* based solely on 16S rRNA gene sequence identity and 16S rRNA gene-based phylogenetic analysis (Baek et al. 2015). Consistent with Baek et al, *H. articus* PAMC 20958^T does have the highest 16S rDNA identity with *H. namhaensis* RA2-3^T in at 96.6% (Table A5); however, based on my analysis, *H. articus* PAMC 20958^T shares a similarly high level of 16S rDNA identity with *Pseudopelagicola gijangensis* DSM 1005^T at 96.1% and *Thalassobius gelatinovorus* CECT 4357^T at 94.1% (Table A5). In addition, within the dataset used in this study, the ranges of 16S rDNA sequence identity for within and between genera comparisons are 93.3%-99.9% and 84.1%-97.9% respectively (Table A5). 16S rDNA identity of 96.6% is therefore not sufficient to support the placement of *H. arcticus* PAMC 20958^T with *H. namhaensis* RA2-3^T in the same genus. Based on other genomic and phylogenetic data (Fig 2.2, Fig 2.4), these isolates should, in fact, be considered as different genera.

The separation of these two isolates into different genera is also supported by differences in phenotypic traits previously identified (e.g., difference in temperature range, salt tolerance, pH tolerance, enzymatic activities and carbon metabolism) (Baek et al. 2015). As such, I proposed to move *H. arcticus* PAMC 20958^T to a new genera *Pseudohalocynthiibacter* gen. nov. This isolate, *Pseudohalocynthiibacter arcticus* PAMC 20958^T gen. nov comb. nov, will be the type species of the genus.

It is worth mentioning that although based on 1st and 2nd codon position similarity alone, *Pseudophaeobacter leonis* 306^T and *Pseudophaeobacter arcticus* DSM 23566^T also seem to belong to different genera, but unlike the *Halocynthiibacter* species, AAI, PD, and 3rd codon position similarity for these two *Pseudophaeobacter* species are within expected range (Fig 2.4). As genomic metrics are providing conflicting results for these two isolates, a definitive decision cannot be made until additional in-depth genomic, phylogenetic and phenotypic characterization is done.

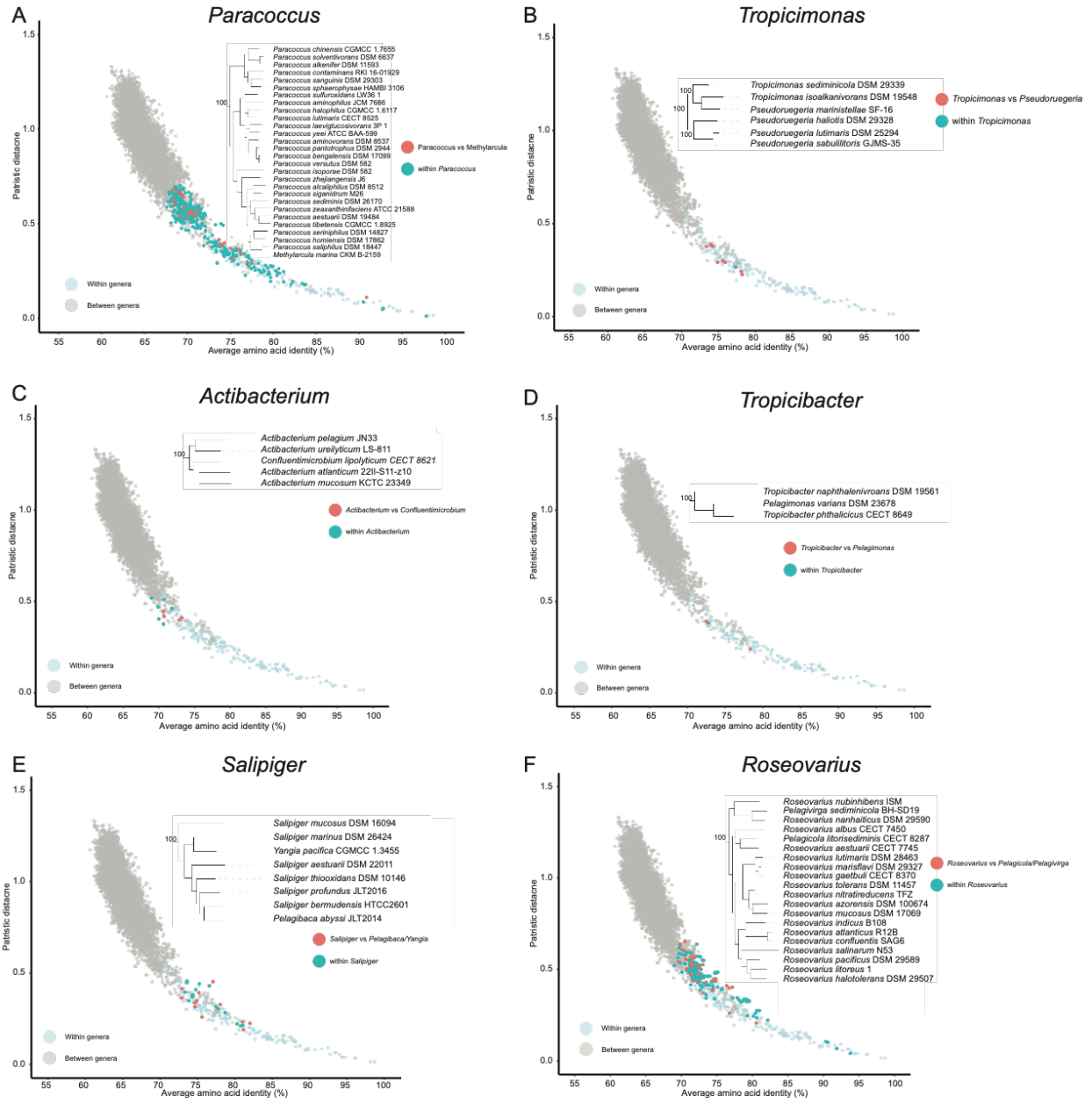
2.3.4 Phylogenetically guided genus level reclassifications

Monophyly is one of the few rules that can be universally applied to all levels of classifications (Parks et al. 2018), and16S rDNA-based phylogeny is typically used to determine whether this criterion is met. Taxonomic classification within the *Rhodobacteraceae* family is no exception, but as previously mentioned, the 16S rRNA gene is not able to resolve phylogenetic relationships among isolates in this order, resulting in numerous non-monophyletic genera (Table A2). One of the goals of this study is to resolve all non-monophyletic genera in this family by applying modern techniques to polyphasic taxonomy.

A total of seven paraphyletic genera were identified as they form monophyletic clades with one or more members of a different genus with strong bootstrap support (Fig 2.2, Table A2). There are seven monophyletic clades (one clade for each paraphyletic genus) consist of 16 genera in total, but ultimately only seven genera should retain their designation as all conflicting genera should be merged with existing genera. In all cases, the first described genus in each monophyletic clade, which I will refer to as the primary genus, will retain its genus designation, provided that the type species is available as per rule 38 and 42 of the code (Parker et al. 2019). For each monophyletic clade (which contains the primary genus and conflicting genera), PD and AAI comparisons are all within the range observed for typical within genus comparisons (Fig 2.5, Table A6), providing genomic support for the merging of these genera (Fig 2.2). As such we propose the following changes: transfer of *Confluentimicrobium lipolyticum* CECT 8621^T to the genus *Actibacterium*, *Tritonibacter horizontis* O3.65^T to the genus *Epibacterium*, *Pelagicola litorisediminis* CECT 8287^T and *Pelagivirga sediminicola* BH-SD19^T to the genus *Roseovarius*, *Yangia pacifica* CGMCC 1.3455^T and *Pelagibaca abyssi* JLT2014^T to the genus *Salipiger*, and *Pelagimonas varians* DSM 23678^T to the genus *Tropicibacter*.

The only paraphyletic genus that could not be resolved is *Paracoccus* (Davis et al. 1969), which forms a monophyletic clade with *Methylarcula marina* VKM B-2159^T (Trotsenko et al. 2000). As *Paracoccus* was described before *M. marina* VKM B-2159^T, the latter should be moved to the genus *Paracoccus*; however, because the type species of *Paracoccus* (*Paracoccus denitrificans* DSM 413) did not meet my minimum quality filter criteria of ≥95% complete and <5% contamination as determined by checkM (Parks et al. 2015), it was removed from my analyses. Therefore, a taxonomic reclassification cannot be proposed as it violates rule 42 of the code (Parker et al. 2019). Out of the seven paraphyletic genera identified, this is the only case in

which genome sequence quality impedes taxonomic resolution, highlighting the importance of not just having the whole genomes of type species, but that these genomes must also be of high quality as subsequent taxonomic classifications will rely heavily on these genomes.


A
Paracoccus
Paracoccus chinensis CGMCC 1.7655
Paracoccus solventivorans DSM 6637
Paracoccus alkenifer DSM 11593
Paracoccus contaminans RKI 16-01929
Paracoccus sanguinis DSM 29303
Paracoccus sphaerophysae HAMBI 3106
Paracoccus sulfuroxidans LW36 1
Paracoccus aminophilus JCM 7686
Paracoccus halophilus CGMCC 1.6117
Paracoccus lutimaris CECT 8525
Paracoccus laevigfucosivorans 3P 1
Paracoccus yeei ATCC BAA-599
Paracoccus aminovorans DSM 8537
Paracoccus pantotrophus DSM 2944
Paracoccus bengalensis DSM 17099
Paracoccus versutus DSM 582
Paracoccus isoporae DSM 582
Paracoccus zhejiangensis J6
Paracoccus alcaliphilus DSM 8512
Paracoccus siganidrum M26
Paracoccus sediminis DSM 26170
Paracoccus zeaxanthinifaciens ATCC 21588
Paracoccus aestuarii DSM 19484
Paracoccus tibetensis CGMCC 1.8925
Paracoccus seriniphilus DSM 14827
Paracoccus homiensis DSM 17862
Paracoccus saliphilus DSM 18447
Methylarcula marina CKM B-2159
Paracoccus vs Methylarcula
within Paracoccus
Within genera
Between genera
Patristic distacne
Average amino acid identity (%)
B
Tropicimonas
Tropicimonas sediminicola DSM 29339
Tropicimonas isoalkanivorans DSM 19548
Pseudoruegeria marinistellae SF-16
Pseudoruegeria haliotis DSM 29328
Pseudoruegeria lutimaris DSM 25294
Pseudoruegeria sabulilitoris GJMS-35
Tropicimonas vs Pseudoruegeria
within Tropicimonas
Within genera
Between genera
Patristic distacne
Average amino acid identity (%)
C
Actibacterium
Actibacterium pelagium JN33
Actibacterium ureilyticum LS-811
Confluentimicrobium lipolyticum CECT 8621
Actibacterium atlanticum 22II-S11-z10
Actibacterium mucosum KCTC 23349
Actibacterium vs Confluentimicrobium
within Actibacterium
Within genera
Between genera
Patristic distacne
Average amino acid identity (%)
D
Tropicibacter
Tropicibacter naphthalenivroans DSM 19561
Pelagimonas varians DSM 23678
Tropicibacter phthalicicus CECT 8649
Tropicibacter vs Pelagimonas
within Tropicibacter
Within genera
Between genera
Patristic distacne
Average amino acid identity (%)
E
Salipiger
Salipiger mucosus DSM 16094
Salipiger marinus DSM 26424
Yangia pacifica CGMCC 1.3455
Salipiger aestuarii DSM 22011
Salipiger thiooxidans DSM 10146
Salipiger profundus JLT2016
Salipiger bermudensis HTCC2601
Pelagibaca abyssi JLT2014
Salipiger vs Pelagibaca/Yangia
within Salipiger
Within genera
Between genera
Patristic distacne
Average amino acid identity (%)
F
Roseovarius
Roseovarius nubinhibens ISM
Pelagivirga sediminicola BH-SD19
Roseovarius nanhaticus DSM 29590
Roseovarius albus CECT 7450
Pelagicola litorisediminis CECT 8287
Roseovarius aestuarii CECT 7745
Roseovarius lutimaris DSM 28463
Roseovarius marisflavi DSM 29327
Roseovarius gaetbuli CECT 8370
Roseovarius tolerans DSM 11457
Roseovarius nitratireducens TFZ
Roseovarius azorensis DSM 100674
Roseovarius mucosus DSM 17069
Roseovarius indicus B108
Roseovarius atlanticus R12B
Roseovarius confluentis SAG6
Roseovarius salinarum N53
Roseovarius pacificus DSM 29589
Roseovarius litoreus 1
Roseovarius halotolerans DSM 29507
Roseovarius vs Pelagicola/Pelagivirga
within Roseovarius
Within genera
Between genera
Patristic distacne
Average amino acid identity (%)

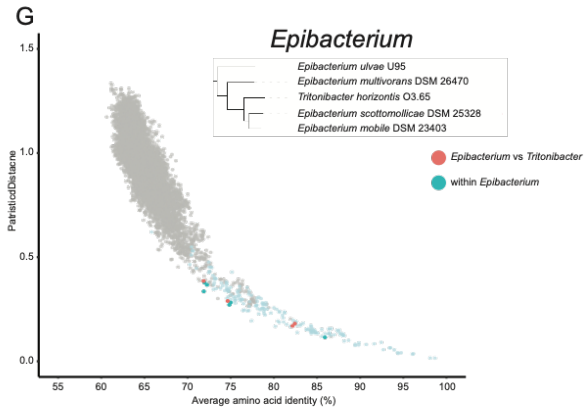
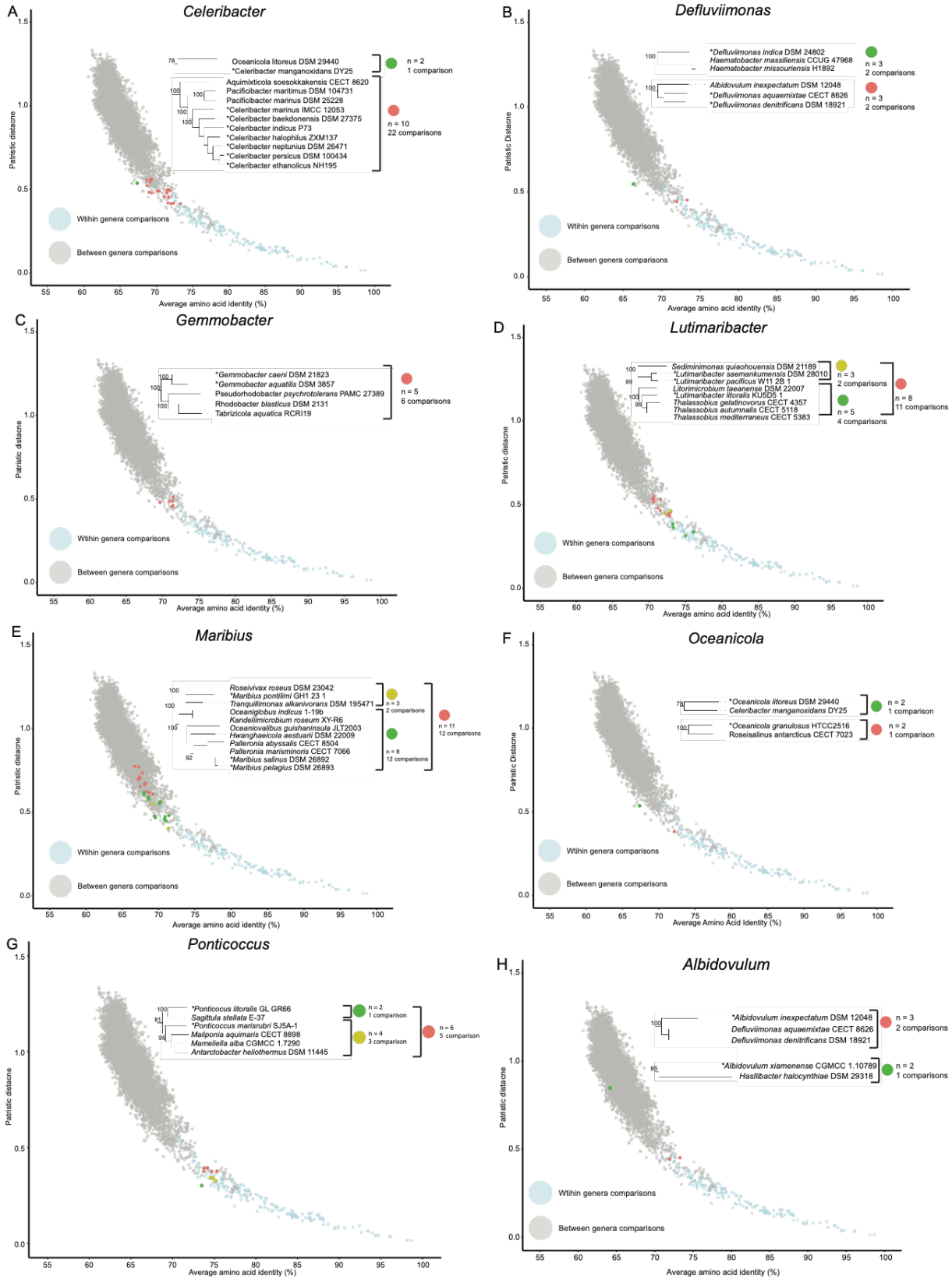


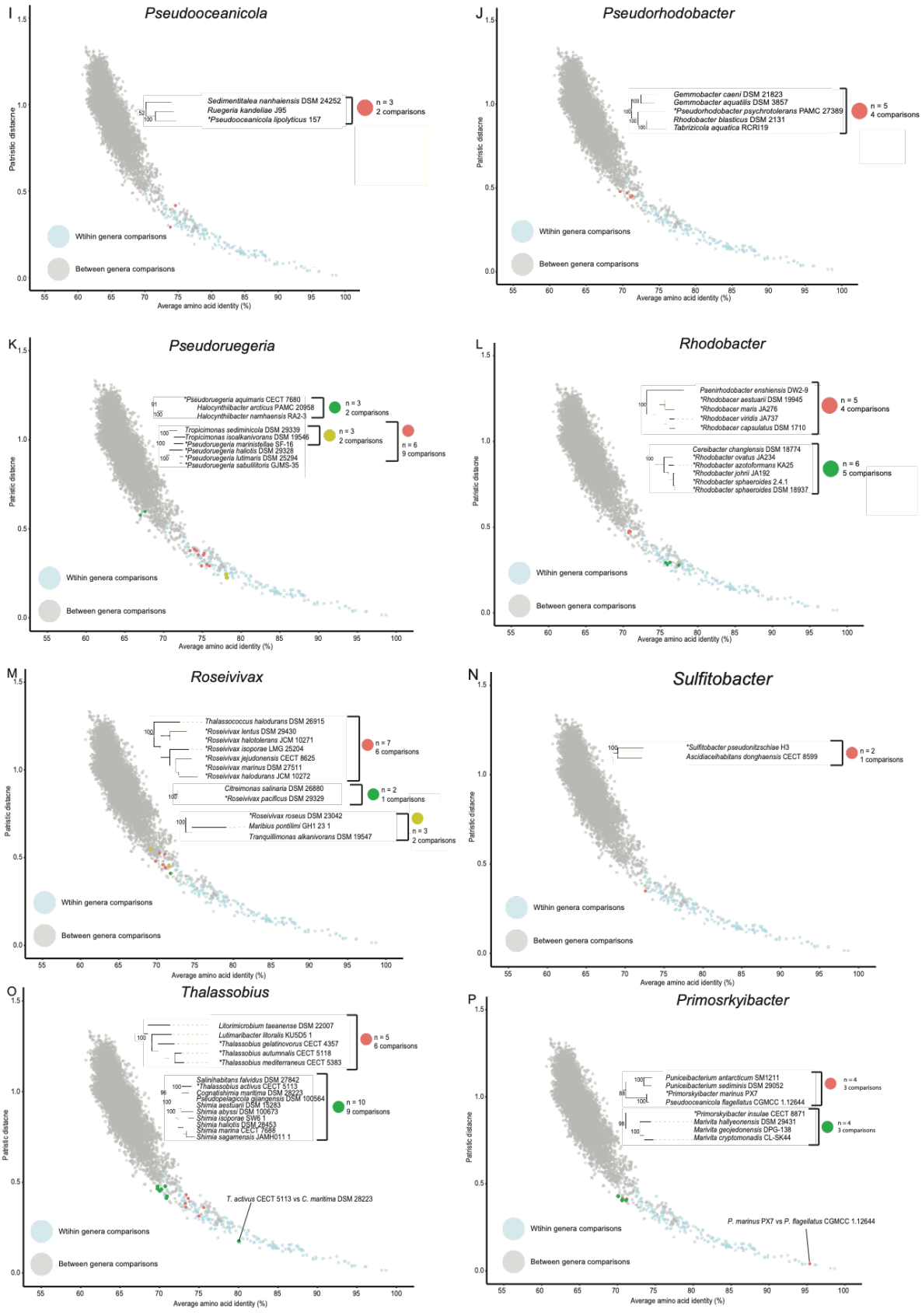
Fig 2.5: AAI and dDDH dot plots of all paraphyletic genera. Phylogenetic trees shown here are subsets of the core-genome tree (Fig 2.2).

Tropicimonas, unlike other paraphyletic genera identified, is monophyletic with four *Pseudoruegeria* strains, which themselves are a polyphyletic genus (Fig 2.5B). Genomic similarities between the two *Tropicimonas* species and the four *Pseudoruegeria* species are within the range expected for within genus comparisons (Fig 2.5B) and phylogenetically, it would resolve this paraphyletic genus to move the two *Tropicimonas* species into the genus *Pseudoruegeria*, as *Pseudoruegeria* was described before *Tropicimonas*. This, however, is not the most parsimonious solution as it results in two genus level changes. To respect the parsimony principle, only *Pseudoruegeria marinistellae* SF-16^T was moved into the genus *Tropicimonas*.

2.3.5 Reclassification at the genus level: addressing polyphyletic genera

Based on my core-genome phylogenetic analysis, I have identified 17 polyphyletic genera (Fig 2.2, Table A2). For each polyphyletic genus, the clade containing the type species, which I will call the primary clade, retains the genus designation as per rule 39a of the code (Parker et al. 2019). All isolates that are part of a polyphyletic genus but are not part of the primary clade will be given a new genus designation or merged with other genera. For each clade where genus level reclassification is required, within and between genera comparisons for all relevant genera were performed (Fig 2.6, Table A6).


A
Celeribacter
Oceanicola litoreus DSM 29440
*Celeribacter manganoxidans DY25
n = 2
1 comparison
Aquimixticola soesokkakensis CECT 8620
Pacificibacter maritimus DSM 104731
Pacificibacter marinus DSM 25228
*Celeribacter marinus IMCC 12053
*Celeribacter baekdonensis DSM 27375
*Celeribacter indicus P73
*Celeribacter halophilus ZXM137
*Celeribacter neptunius DSM 26471
*Celeribacter persicus DSM 100434
*Celeribacter ethanolicus NH195
n = 10
22 comparisons
Wtihin genera comparisons
Between genera comparisons
Patristic distacne
Average amino acid identity (%)
B
Defluviimonas
*Defluviimonas indica DSM 24802
Haematobacter massiliensis CCUG 47968
Haematobacter missouriensis H1892
n = 3
2 comparisons
Albidovulum inexpectatum DSM 12048
*Defluviimonas aquaemixtae CECT 8626
*Defluviimonas denitrificans DSM 18921
n = 3
2 comparisons
Wtihin genera comparisons
Between genera comparisons
Patristic Distacne
Average amino acid identity (%)
C
Gemmobacter
*Gemmobacter caeni DSM 21823
*Gemmobacter aquatilis DSM 3857
Pseudorhodobacter psychrotolerans PAMC 27389
Rhodobacter blasticus DSM 2131
Tabrizicola aquatica RCRI19
n = 5
6 comparisons
Wtihin genera comparisons
Between genera comparisons
Patristic distacne
Average amino acid identity (%)
D
Lutimaribacter
Sediminimonas qiaohouensis DSM 21189
*Lutimaribacter saemankumensis DSM 28010
n = 3
2 comparisons
*Lutimaribacter pacificus W11 2B 1
Litorimicrobium taeanense DSM 22007
*Lutimaribacter litoralis KUSD5 1
Thalassobius gelatinovorus CECT 4357
Thalassobius autumnalis CECT 5118
Thalassobius mediterraneus CECT 5383
n = 5
4 comparisons
n = 8
11 comparisons
Wtihin genera comparisons
Between genera comparisons
Patristic distacne
Average amino acid identity (%)
E
Maribius
Roseivivax roseus DSM 23042
*Maribius pontilimi GH1 23 1
Tranquillimonas alkanivorans DSM 195471
n = 3
2 comparisons
Oceaniglobus indicus 1-19b
Kandeliimicrobium roseum XY-R6
Oceanivalibus guishaninsula JLT2003
Hwanghaeicola aestuarii DSM 22009
Palleronia abyssalis CECT 8504
Palleronia marisminoris CECT 7066
*Maribius salinus DSM 26892
*Maribius pelagius DSM 26893
n = 8
12 comparisons
n = 11
12 comparisons
Wtihin genera comparisons
Between genera comparisons
Patristic distacne
Average amino acid identity (%)
F
Oceanicola
*Oceanicola litoreus DSM 29440
Celeribacter manganoxidans DY25
n = 2
1 comparison
*Oceanicola granulosus HTCC2516
Roseisalinus antarcticus CECT 7023
n = 2
1 comparison
Wtihin genera comparisons
Between genera comparisons
Patristic Distacne
Average Amino Acid Identity (%)
G
Ponticoccus
*Ponticocus litoralis GL GR66
Sagittula stellata E-37
*Ponticoccus marisrubri SJ5A-1
n = 2
1 comparison
Maliponia aquimaris CECT 8898
Mameliella alba CGMCC 1.7290
Antarctobacter heliothermus DSM 11445
n = 4
3 comparison
n = 6
5 comparison
Patristic distacne
Average amino acid identity (%)
H
Albidovulum
*Albidovulum inexpectatum DSM 12048
Defluviimonas aquaemixtae CECT 8626
Defluviimonas denitrificans DSM 18921
n = 3
2 comparisons
*Albidovulum xiamenense CGMCC 1.10789
Hasllibacter halocynthiae DSM 29318
n = 2
1 comparisons
Wtihin genera comparisons
Between genera comparisons
Patristic distacne
Average amino acid identity (%)


I
Pseudooceanicola
Sedimentitalea nanhaiensis DSM 24252
Ruegeria kandeliae J95
*Pseudooceanicola lipolyticus 157
n = 3
2 comparisons
Wtihin genera comparisons
Between genera comparisons
Patristic distacne
Average amino acid identity (%)
J
Pseudorhodobacter
Gemmobacter caeni DSM 21823
Gemmobacter aquatilis DSM 3857
*Pseudorhodobacter psychrotolerans PAMC 27389
Rhodobacter blasticus DSM 2131
Tabrizicola aquatica RCRI19
n = 5
4 comparisons
K
Pseudoruegeria
*Pseudoruegeria aquimaris CECT 7680
Halocynthiibacter arcticus PAMC 20958
Halocynthiibacter namhaensis RA2-3
n = 3
2 comparisons
Tropicimonas sediminicola DSM 29339
Tropicimonas isoalkanivorans DSM 19546
*Pseudoruegeria marinistellae SF-16
*Pseudoruegeria halotis DSM 29328
*Pseudoruegeria lutimaris DSM 25294
*Pseudoruegeria sabulilitoris GJMS-35
n = 3
2 comparisons
n = 6
9 comparisons
L
Rhodobacter
Paenirhodobacter enshiensis DW2-9
*Rhodobacer aestuarii DSM 19945
*Rhodobacer maris JA276
*Rhodobacer viridis JA737
*Rhodobacer capsulatus DSM 1710
n = 5
4 comparisons
Cereibacter changlensis DSM 18774
*Rhodobacter ovatus JA234
*Rhodobacter azotoformans KA25
*Rhodobacter johrii JA192
*Rhodobacter sphaeroides 2.4.1
*Rhodobacer sphaeroides DSM 18937
n = 6
5 comparisons
M
Roseivivax
Thalassococcus halodurans DSM 26915
*Roseivivax lentus DSM 29430
*Roseivivax halotolerans JCM 10271
*Roseivivax isoporae LMG 25204
*Roseivivax jejudonensis CECT 8625
*Roseivivax marinus DSM 27511
*Roseivivax halodurans JCM 10272
n = 7
6 comparisons
Citreimonas salinaria DSM 26880
*Roseivivax pacificus DSM 29329
n = 2
1 comparisons
*Roseivivax roseus DSM 23042
Maribius pontilimi GH1 23 1
Tranquillimonas alkanivorans DSM 19547
n = 3
2 comparisons
N
Sulfitobacter
*Sulfitobacter pseudonitzschiae H3
Ascidiaceihabitans donghaensis CECT 8599
n = 2
1 comparisons
O
Thalassobius
Litorimicrobium taeanense DSM 22007
Lutimaribacter litoralis KU5D5 1
*Thalassobius gelatinovorus CECT 4357
*Thalassobius autumnalis CECT 5118
*Thalassobius mediterraneus CECT 5383
n = 5
6 comparisons
Salinihabitans falvidus DSM 27842
*Thalassobius activus CECT 5113
Cognatishimia maritima DSM 28223
Pseudopelagicola gijangensis DSM 100564
Shimia aestuarii DSM 15283
Shimia abyssi DSM 100673
Shimia isoporae SW6 1
Shimia haliotis DSM 28453
Shimia marina CECT 7696
Shimia sagamiensis JAMH011 1
n = 10
9 comparisons
T. activus CECT 5113 vs C. maritima DSM 28223
P
Primosrkyibacter
Puniceibacterium antarcticum SM1211
Puniceibacterium sediminis DSM 29052
*Primorskyibacter marinus PX7
Pseudooceanicola flagellatus CGMCC 1.12644
n = 4
3 comparisons
*Primorskyibacter insulae CECT 8871
Marivita hallyeonensis DSM 29431
Marivita geojedonensis DPG-138
Marivita cryptomonadis CL-SK44
n = 4
3 comparisons
P. marinus PX7 vs P. flagellatus CGMCC 1.12644

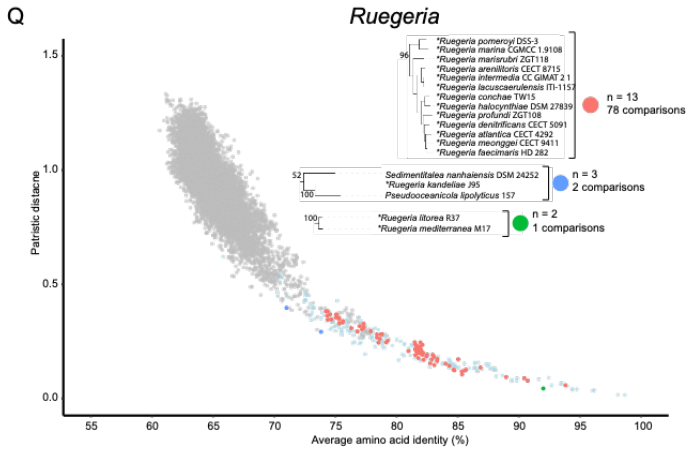


Fig 2.6: AAI and dDDH dot plots of all polyphyletic genera. Phylogenetic trees shown here are subsets of the core-genome tree (Fig 2.2).

A majority of the comparisons between the polyphyletic genera and their closest neighbors resulted in borderline AAI and PD values, where they fall in the overlap region of between and within genus comparisons (Fig 2.6). Besides the lack of resolution from these genomic similarity indicators, there are several reasons why it is more difficult to make decisions on taxonomic classifications for polyphyletic as opposed to paraphyletic genera. Unlike the latter, where conflicting genera must be merged with existing ones to achieve resolution, polyphyetic genera can be resolved by either merging conflicting genera with existing ones or giving them novel genus designations. Therefore, although we know that these isolates must be renamed as they are currently polyphyletic, how they should be renamed remains unclear for some. Following the polyphasic approach in bacterial taxonomy, if genomic similarity and phylogenetic analyses are inconclusive, the decision must then rely on phenotypic traits. As we are using type species to determine genus boundaries, this information is generally available from original publications.

Although majority of the cases are ambiguous (Fig 2.6), there are two polyphyletic genera that can be partially resolved and two that can be fully resolved based on phylogenetic and genomic data. One genus that can be partially resolved is the genus *Pseudoruegeria*. It is currently split into three clades (Fig 2.6K). One clade contains *Pseudoruegeria marinistellae* SF-16^T, which will be transferred to the genus *Tropicimonas*, which as previously mentioned, is the most parsimonious solution. Another clade contains the type species, *P. aquimaris* CECT 7680^T with two *Halocynthibacter* species. Genomic comparisons between *P. aquimaris* CECT 7680^T and the *Halocynthibacter* species clearly falls within the range of between genera comparisons (Fig 2.6K, Table A6) and therefore, both genera will retain their designations. The last clade contains exclusively of *Pseudoruegeria haliotis* DSM 29328^T, *P. lutimaris* DSM 25294^T, *P.*

sabulilitoris GJMS-35^T. The closest relative of these three isolates is the genus *Tropicimonas* (Fig 2.6K). Genomic comparisons between these three *Pseudoruegeria* and the *Tropicimonas* isolates resulted in inconclusive values for both AAI and PD (Fig 2.6K). To determine whether these three *Pseudoruegeria* species should merged with *Tropicimonas* or be given a new genus designation will require additional phenotypic data.

Another polyphyletic genus that can be partially resolved is *Ruegeria*. It is currently split into three monophyletic clades (2.6Q). *Ruegeria atlantica* CECT 4292^T is the type species and it is in a monophyletic group with only *Ruegeria* isolates (2.6Q). This clade will therefore retain the genus designation. Another clade contains *Ruegeria kandeliae* J95^T, *Sedimentalea nanhaiensis* DSM 24252^T and *Pseudooceanicola lipolyticus* 157^T. Since genomic metrics between these three species (*S. nanhaiensis* DSM 24252^T, *P. lipolyticus* 157^T and *R. kandeliae* J95^T) are inconclusive, additional phenotypic data is required (Fig 2.6Q, Table A6). The last clade contains exclusively of *R. litorea* R37^T and *R. mediterranea* M17^T and will be given a new genus designation, for which I propose the name *Falsiruegeria* and designate *Falsiruegeria litorea* R37^T gen. nov comb. nov as the type species of the genus.

Albidovulum is one of the two genera that can be fully resolved (Fig 2.6H). Currently, this genus is split into two separate monophyletic clades; *A. xiamenense* CGMCC 1.10789^T is grouped together with *Hasllibacter halocynthiae* DSM 29318^T and *A. inexpectatum* DSM 12048^T is grouped together with *Defluviimonas aquaemixtae* CECT 8626^T and *D. denitrificans* DSM 18921^T (Fig 2.6H). The type species for the genus *Albidovulum* is *A. inexpectatum*. Currently, *A. inexpectatum* DSM 12048^T is basal to the two *Delufviimonas* species, and as such it will retain its genus designation as this will not result in any para- or polyphyletic genus. The genomic comparisons between *A. xiamenense* CGMCC 1.10789^T and *H. halocynthiae* DSM

29318^T is more straightforward, as these values clearly fall within the range for between genera comparisons (Fig 2.6H, Table A6) and therefore these two isolates will remain as separate genera. As a result, *A. xiamenense* CGMCC 1.10789^T will require a new genus designation, which I propose to be *Falsialbidovulum* gen. nov and *F. xiamenense* CGMCC 1.10789^T gen. nov comb. nov will the type species of this genus.

The other polyphyletic genus that can be fully resolved is *Thalassobius* (Fig 2.6O). This genus is currently split into two monophyletic clades. The first clade contains *T. gelatinovorus* CECT 4357^T, *T. autumnalis* CECT 5118^T, *T. mediterraneus* CECT 5383^T, *Litorimicrobium taeanense* DSM 22007^T and *Lutimaribacter litoralis* KU5D5 1^T. The second clade contains *T. activus* CECT 5113^T with *Salinihabitans falvidus* DSM 27842^T, *Cognatishimia maritima* DSM 28223^T, *Pseudopelagicola gijangensis* DSM 100564^{T}and the genus *Shimia* (Fig 2.6O). All comparisons between *Thalassobius* with other genera in the first clade resulted in borderline values. As *Thalassobius* is monophyletic and contains the type species of the genus, this group will retain its genus designation. *T. activus* CECT 5113^{T}will therefore require a new genus designation or be merged with existing genera. AAI and PD values between *T. activus* CECT 5113^T and *Cognatishimia maritima* DSM 28223^T fall within range expected for within genus comparison (Fig 2.6O); therefore, I propose to rename *T. activus* CECT 5113^T as *Cognatishimia activus* CECT 5113^T comb. nov.

2.3.6 Workflow for the incorporation of new genomes for consistent taxonomic classifications.

As it is not practical to reconstruct core-genome phylogenetic trees of all type strains each time new genomes are available, there needs to be a way to quickly and accurately identify

the phylogenetic relationships of unknown isolates to known isolates. The 16S rRNA gene was previously used for this purpose, but it is at best able to identify the family to which an isolate belongs, which will still contain hundreds if not thousands of isolates.

An efficient way to narrow down a list of close relatives and was used in this study for the incorporation of new genomes, is through the use AAI. Ideally the 10 closest relatives can be determined based on pairwise AAI comparisons between the unknown isolate and all type strains. This effectively reduces the dataset to only 11 isolates (10 closest relative and the unknown isolate), for which in-depth phylogenomic analyses, such as the reconstruction of a core-genome phylogenetic tree, can easily be done

I collected two additional type strains (*Primorskyibacter sedentarius* DSM 104836^T and *Phaeobacter piscinae* P-14^T) and one novel genus (*Sinirhodobacter*) as examples for which I can apply this approach. The identity of these two species were confirmed as not only do they form a strongly supported monophyletic clade with their proposed genera, the overall structures of the smaller phylogenetic trees are similar to the core-genome tree that was reconstructed with a more extensive dataset (Fig 2.2, 2.7AB).

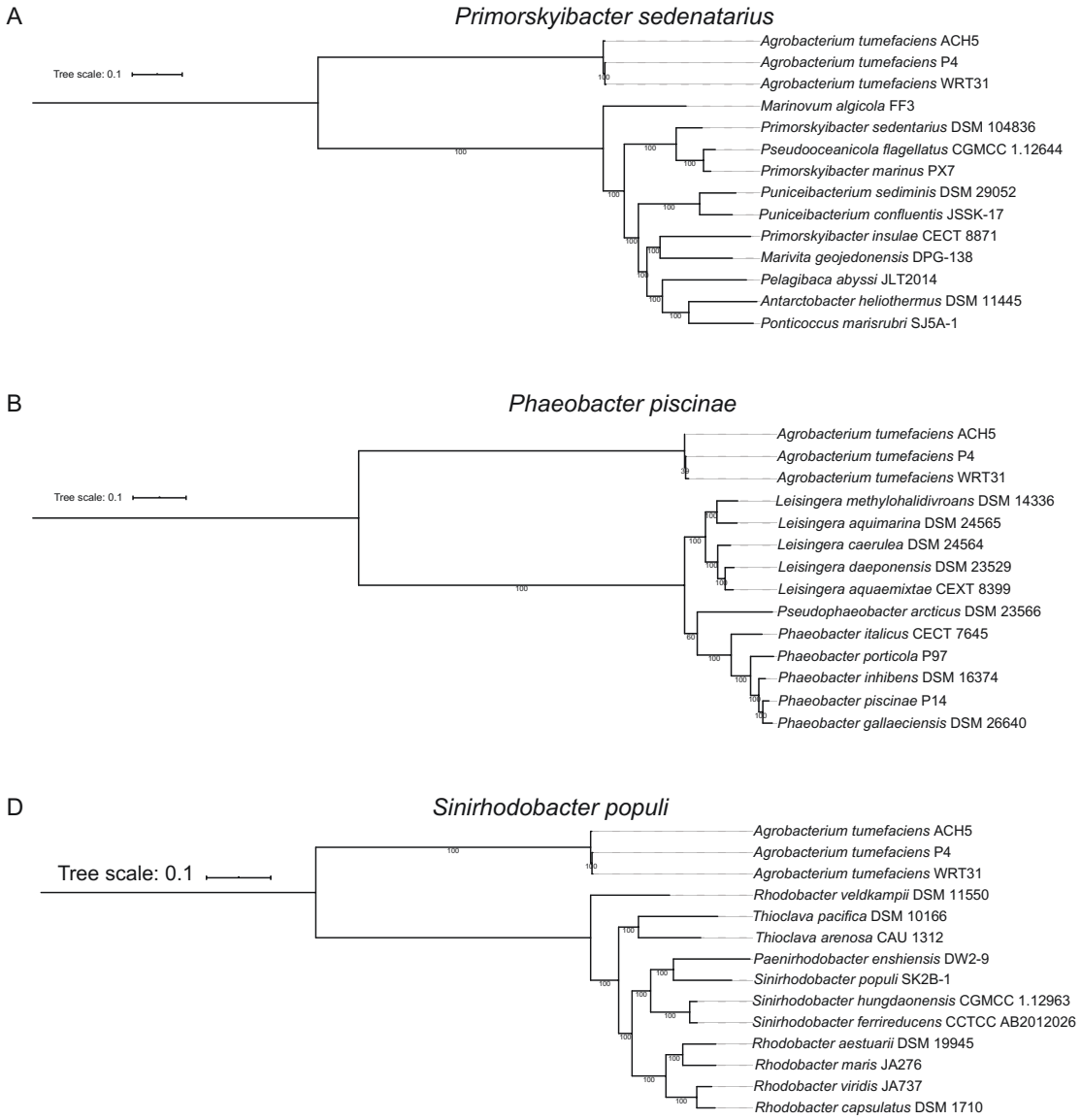


Fig 2.7: Core-genome phylogenetic trees of new genomes added after the creation of the phylogenomic framework. Each phylogenetic tree contains three *Agrobacterium tumefaciens* strains as outgroup, the newly uploaded genomes and its 10 closest relatives determined based on AAI. Core genes are identified for each group separately. Each tree is built using RAxML v8.2.12 with PROTGAMMAAUTO option for automatic model selection. Branch supports is evaluated with 100 bootstrap replicates.

As *Primorskyibacter* is currently a polyphyletic genus (Fig 2.6P), confirming that the type species *P. sedentarius* DSM 104836^T forms a monophyletic cluster with *P. marnius* PX7^T allows us to partially resolve this genus, which was previously not possible (Fig 2.6P). *Primorskyibacter* is currently split into two monophyletic clades. *P. insulae* CECT 8871^T is grouped together with the genus *Marivita* and *Primorksyibacter marinus* PX7^T is grouped together with *Pseudooceanicola flagellatus* CGMCC 1.12644^T (Fig 2.6P). AAIs between *Primorskyibacter marinus* PX7^T and the two *Puniceibacterium* species are 71% whereas comparison with *Pseudooceanicola flagellatus* CGMCC 1.12644^T is clearly different from the rest at 95% (Fig 2.6P, Table A6). PD shows a similar pattern where comparisons between *Primorksyibacter marinus* PX7^T and the two *Puneiceibacterium* species are 0.41, but 0.042 when compared with *Pseudooceanicola flagellatus* CGMCC 1.12644^T.This extremely high AAI value suggest that there is even the possibility that these two isolates (*Primorskyibacter marinus* PX7^T and *Pseudooceanicola flagellatus* CGMCC 1.12644^T) belong to the same species (Konstantinidis and Tiedje 2005b); but a dDDH value of 52.4% clearly shows that these are different species. Taking AAI, PD and dDDH together, these two isolates should be classified as distinct species within the same genus. As *Pseudooceanicola flagellatus* CGMCC 1.12644^T (Huang et al. 2018) was described after *Primorskyibacter* (Romanenko et al. 2011), it will be transferred to this genus. Consequently, *P. insulae* CECT 8871^T will either have to be merged with existing genera or given a novel genus designation.

This approach was also able to highlight potential misclassification of a novel genus. *Sinirhodobacter* is a novel genus proposed in 2013 as the close relative of *Rhodobacter* (Yang et al. 2013). This study has shown that members of *Sinirhodobacter* is the sister taxa of the genus *Rhodobacter* and *Thioclava* is basal to both (Yang et al. 2013). This relationship, where the type

species of the genus *Sinirhodobacter* (*Sinirhodobacter ferrireducens* CCTCC AB2012026^T) is the sister taxa to the genus *Rhodobacter*, was confirmed by my core-genome phylogenetic analysis (Fig 2.7C). However, different from the previous study, my analysis shows that *Paenirhodobacter* is a closer relative to *Sinirhodobacter* than *Rhodobacter* (Fig 2.7). In addition, *Paenirhodobacter* forms a monophyletic clade with *Sinirhodobacter populi* SK2B-1^T (Fig 2.7C) making the genus *Sinirhodobacter* a paraphyletic genus. *Paenirhodobacter* is likely misclassified as not only are both *Paenirhodobacter* and *Sinirhodobacter* differentiated from *Rhodobacter* by their lack of phototrophic abilities, the initial analyses describing *Paenirhodobacter* did not include any *Sinirhodobacter* strains (Yang et al. 2013, Wang et al. 2014). In addition, *Paenirhodobacter* also shares a number of phenotypic traits with *Sinirhodobacter* as both are positive for urease activity, arginine dihydrolase and utilization of maltose and negative for indole production. Since *Paenirhodobacter* was described in 2014 (Wang et al. 2014) after *Sinirhodobacter* was described (Yang et al. 2013), I propose to rename *Paenirhodobacter enshiensis* DW2-9^T (currently the only named species within this genus) as *Sinirhodobacter enshiensis* DW2-9^T comb. nov.

The three examples presented highlight the benefits of this approach as an efficient first step in determining taxonomic classifications of novel genomes as it can provide validation to the proposed taxonomic classification or point out potential misclassifications. It should be noted that although phylogenetic data and AAI alone may not be sufficient to justify all taxonomic classifications, this approach can still serve to guide subsequent in-depth genomic, phylogenetic and phenotypic analyses that involves an even larger dataset of closely related strains.

Although this approach was used in this study, as the number of genomes increases it will become increasingly less feasible to compute all pairwise AAI in a timely manner without

sufficient computing power. In such cases, close relatives can still be identified using tools such as GTDB-tk (Chaumeil et al. 2019) to identify the general phylogenetic placements of unknown isolates among known type strains. A shortlist of close relatives can then be selected for more in-depth analyses as before.

2.3.7 Proposal to move the roseobacter clade into the new family **Roseobacteraceae** *fam. nov.*

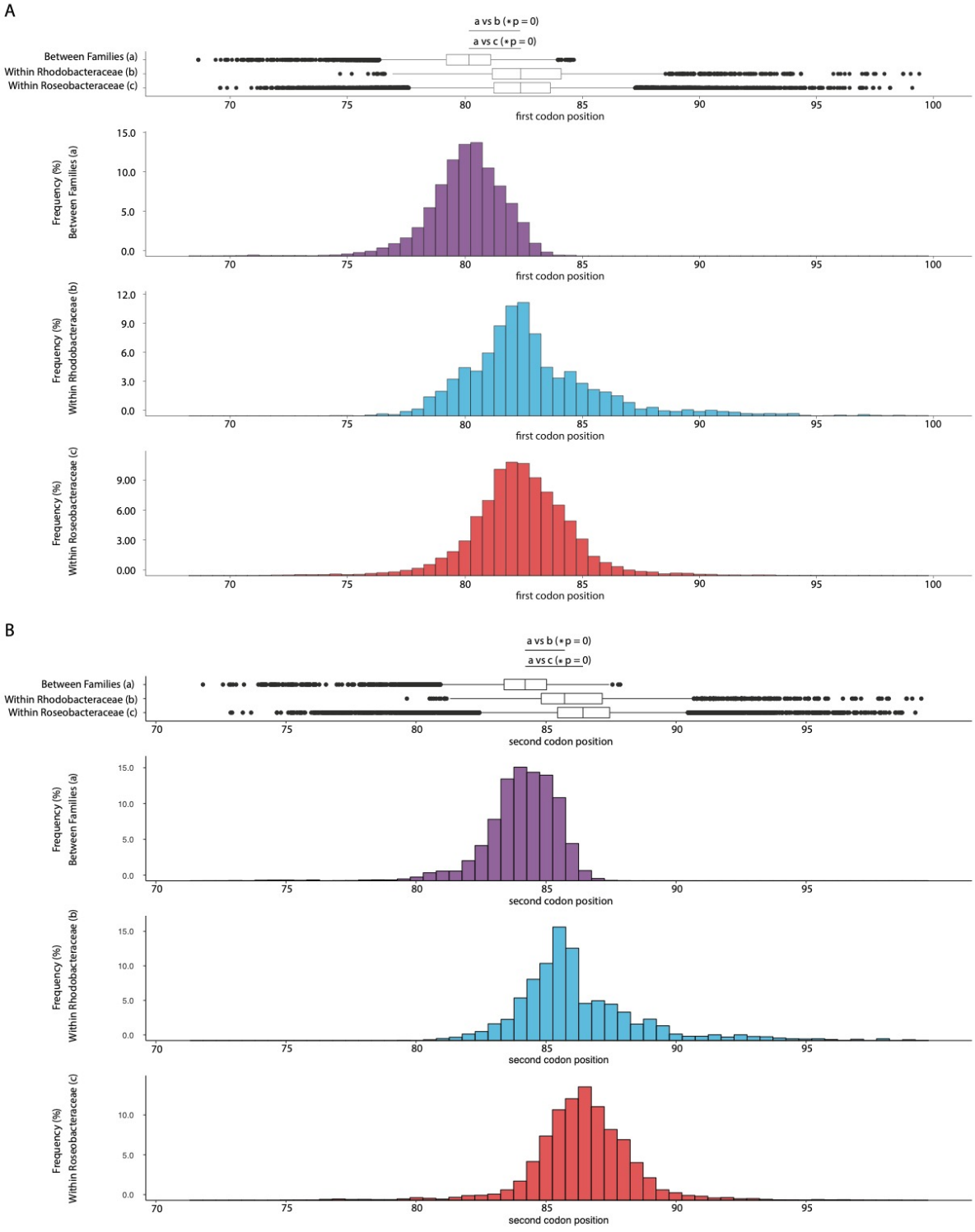
As previously mentioned, the roseobacter clade has an important historical role in the field of oceanography (Buchan et al. 2005, Moran et al. 2007) and is studied by many worldwide. Members of this clade also play important roles in regulating biogeochemical cycles and climate conditions (Buchan et al. 2005, Brinkhoff et al. 2008). Despite its importance, there is no standardized terminology to refer to this clade. It was previously suggested to refer to this group as the marine roseobacter clade based on marine and non-marine adaptions (Simon et al. 2017); however, as not all members of the roseobacter clade live in marine environments and not all isolates outside of the roseobacter clade live in non-marine environments (Table A1), this term does not refer to the roseobacter clade specifically, but rather to a polyphyletic group within the family.

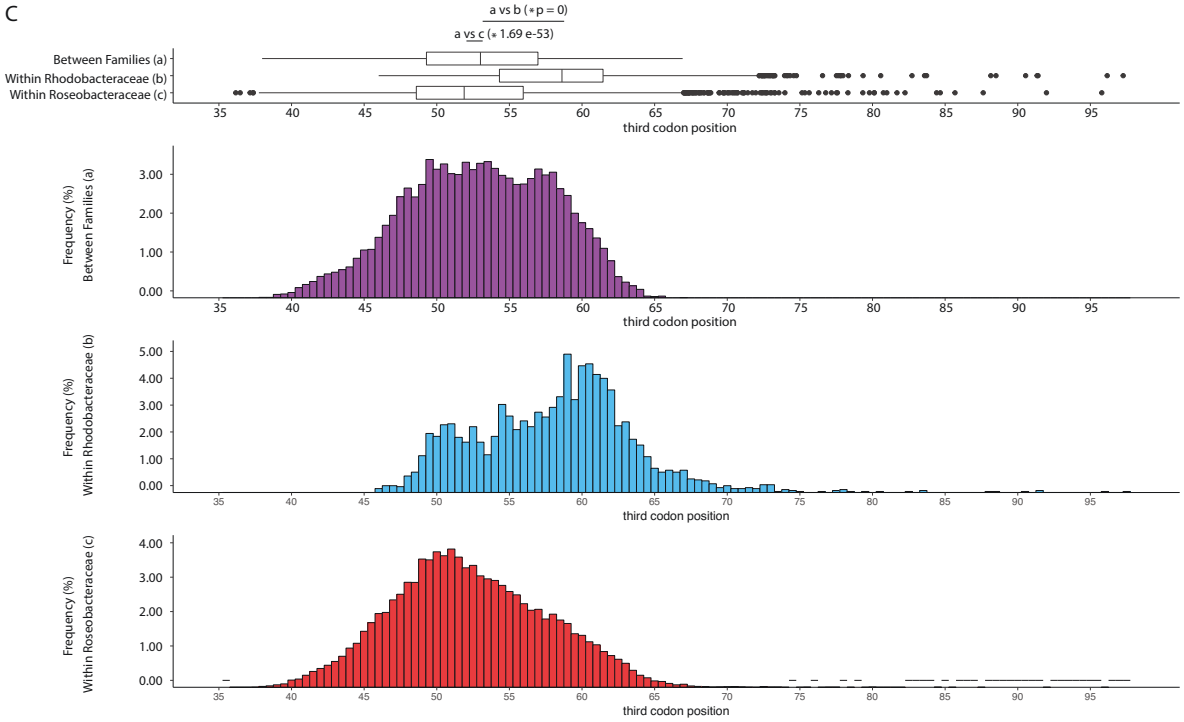
To establish a phylogenetically coherent classification for the roseobacter clade, I performed a meta-analysis of phenotypic traits as well as comprehensive genomics and phylogenomic analyses looking at similarities and differences between the roseobacter clade and its closest relatives. I identified a number of genomic and likely phenotypic differences between the roseobacter clade and other members of the *Rhodobacteraceae* family. As such I propose to move this clade to a new family, *Roseobacteraceae* fam. nov, named based on the first described

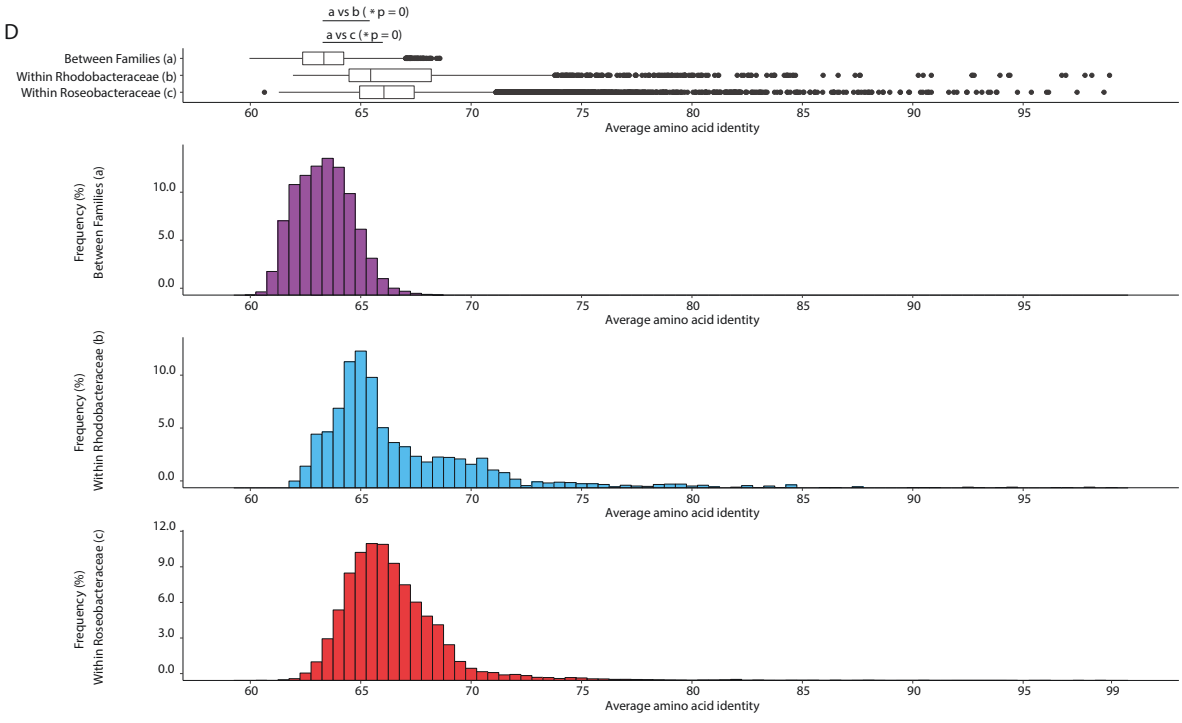
genus, *Roseobacter* (Shiba 1991). All members outside of this clade will retain the *Rhodobacteraceae* family designation.

2.3.7.1 Phylogenetic and genomic analysis shows a clear distinction between the **Roseobacteraceae** *family and the* **Rhodobacteraceae** *families.*

Phylogenetically, the *Roseobacteracae* family forms a monophyletic clade, with 100% bootstrap support, clearly separating it from other members of the *Rhodobacteracea* family (Fig. 2.2 – inner ring). This is consistent with multiple studies that the roseobacter clade is monophyletic and distinct from the rest of the family (Simon et al. 2017, Parks et al. 2018). Values of all genomic metrics (AAI, and codon position similarities) for within family comparisons are significantly higher than between family comparisons ($p = 0$, Fig 2.8A-D). PD, which is a measure of evolutionary distance, for within family comparisons is also significantly smaller than between family comparisons ($p = 0$, Fig. 2.8E).


A
a vs b (* p = 0)
a vs c (* p = 0)
Between Families (a)
Within Rhodobacteraceae (b)
Within Roseobacteraceae (c)
first codon position
Frequency (%) Between Families (a)
Frequency (%) Within Rhodobacteraceae (b)
Frequency (%) Within Roseobacteraceae (c)
B
second codon position


C
a vs b (∗p = 0)
a vs c (∗1.69 e-53)
Between Families (a)
Within Rhodobacteraceae (b)
Within Roseobacteraceae (c)
third codon position
Frequency (%)
Between Families (a)
Within Rhodobacteraceae (b)
Within Roseobacteraceae (c)


D
a vs b (∗p = 0)
a vs c (∗p = 0)
Between Families (a)
Within Rhodobacteraceae (b)
Within Roseobacteraceae (c)
Average amino acid identity
Frequency (%)
Between Families (a)
Within Rhodobacteraceae (b)
Within Roseobacteraceae (c)

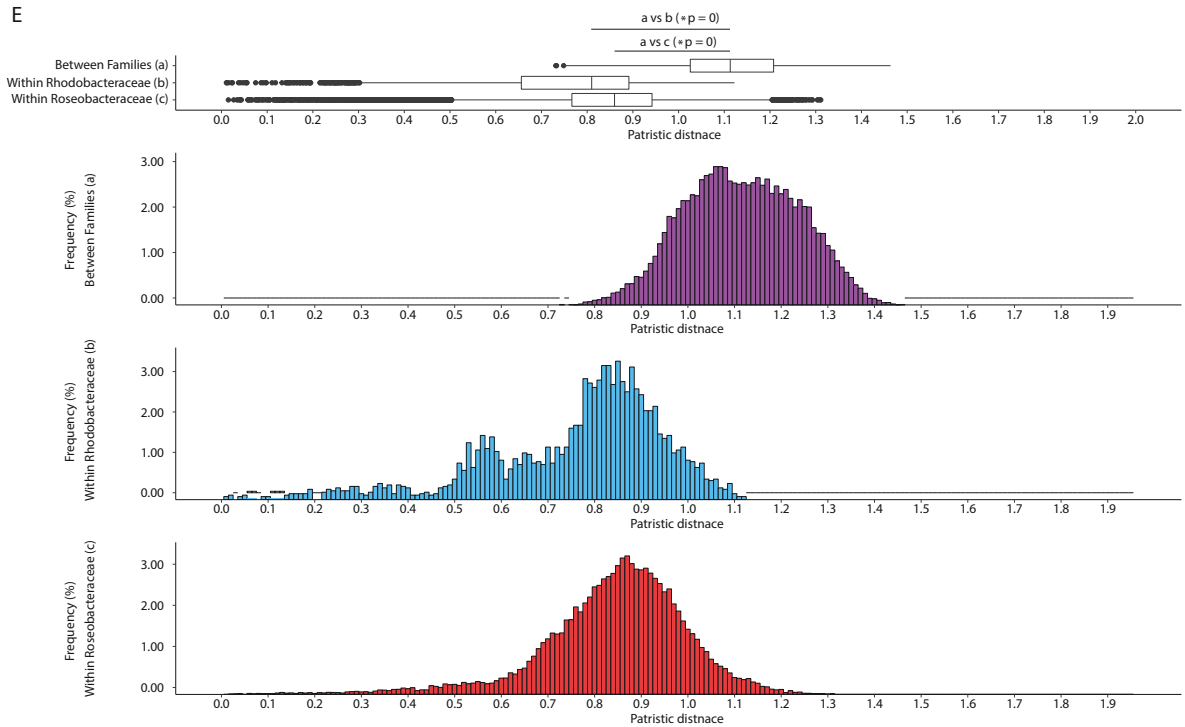
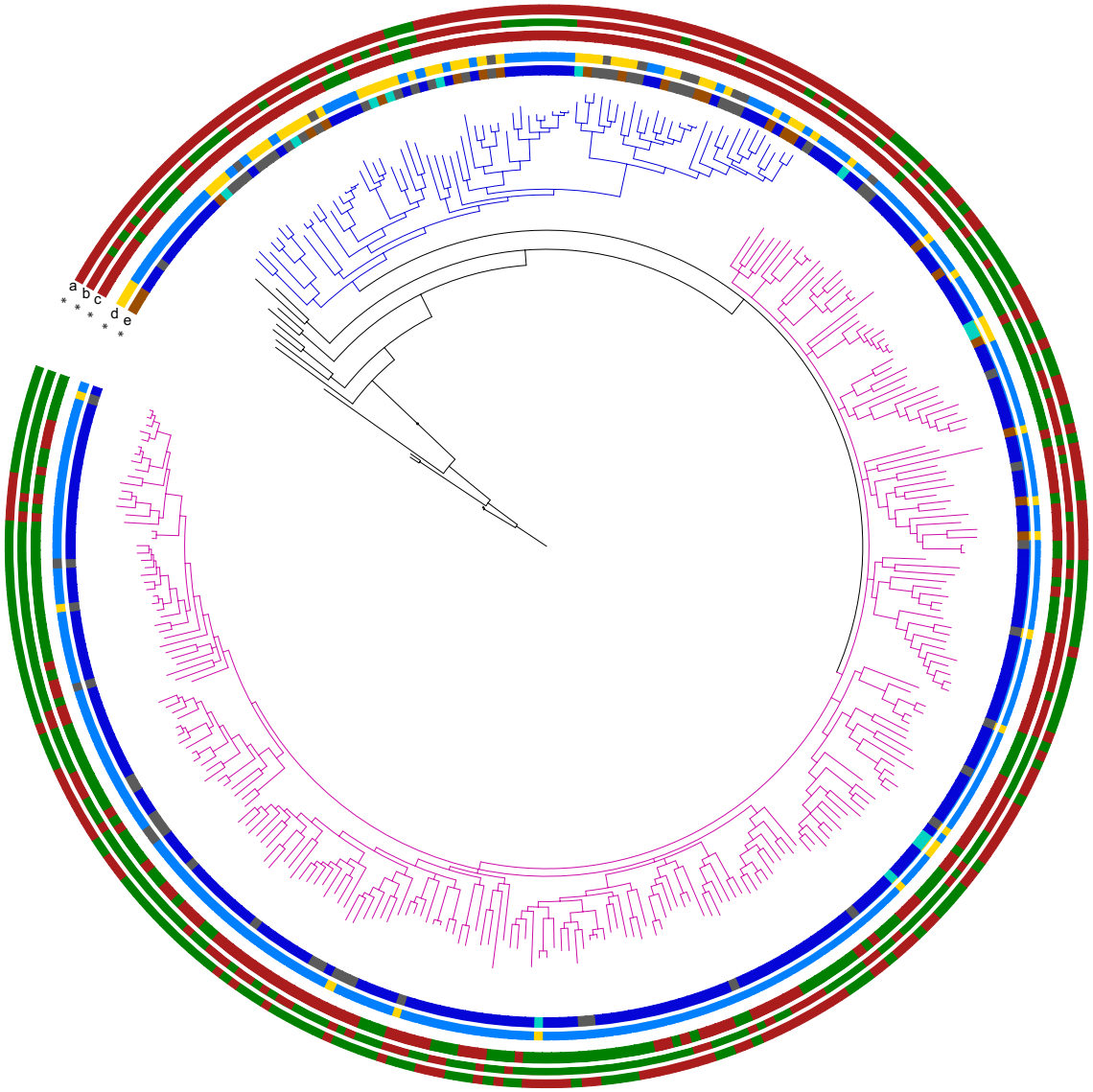


Fig 2.8: Histogram of AAI, codon position similarities and PD for within *Roseobacteracea* (Red), within *Rhodobacteraceae* (blue) and between family (purple) comparisons. Boxplots show 1.5 interquartile range, 25th, 50th and 75th percentile. Significant differences between distributions ($p < 0.05$) are represented by a *.

Taken together, there is strong evidence that the *Roseobacteraceae* family is phylogenetically and genomically distinct from the *Rhodobacteraceae* family and should be considered a novel family.

2.3.7.2 Predicted phenotypic characteristic shows difference in adaptive traits between the **Roseobacteraceae** *family and the* **Rhodobacteraceae** *family.*

Environmentally, the two families are different where 89% of the *Roseobacterace* are isolated in environments with high salt content (defined here as environments with ≥3.5% w/v NaCl concentration; the average NaCl concentration of sea water) (Fig. 2.9). On the other hand, only 39% of the *Rhodobacteraceae* family are isolated from such environments.



Tree scale: 0.1

Rhodobacteraceae

Roseobacteraceae

* Significant difference in proportion between lineages (p<0.05)

a AHL - Quorum Sensing (p < 0.00001)
Present 56% Roseobacteraceae, 4% Rhodobacteraceae
Absent

b DMSP Cleavage (p < 0.00001)
Present 64% Roseobacteraceae, 37% Rhodobacteraceae
Absent

c DMSP Demethylation (p < 0.00001)
Present 52% Roseobacteraceae, 7% Rhodobacteraceae
Absent

d Salinity level (p < 0.00001)
Low 9% Roseobacteraceae, 49% Rhodobacteraceae
High 89% Roseobacteraceae, 39% Rhodobacteraceae

e Environments (p < 0.00001 proportion of isolates found in Marine environment)
Marine 82% Roseobacteraceae, 37% Rhodobacteraceae
Terrestrial 2% Roseobacteraceae, 20% Rhodobacteraceae
Freshwater 3% Roseobacteraceae, 8% Rhodobacteraceae
Other 13% Roseobacteraceae, 35% Rhodobacteraceae

Fig 2.9: Same core-genome phylogenetic tree as before (Fig 2.2), but colored based on phenotypic traits and environment of isolation. Branches of the *Roseobacteraceae* and *Rhodobacteraceae* families are highlighted in pink and blue respectively. Rings represent presents (green)/absents (red) of AHL-quorum sensing (a), DMSP cleavage pathway (b), DMPS demethylation pathway (b), together with the environment of isolation (c) (marine (blue), terrestrial (brown), freshwater (light blue) and other (grey)) and salinity levels (d) (high: ≥3.5% NaCl, light blue; low: <3.5%, yellow). Significant difference in proportion between the two families is ($p < 0.05$ according to proportional Z test) is marked by a *.

Different environments also lead to different adaptions. Three pathways characteristics of the family *Roseobacteraceae* were identified by combining a meta-analysis of phenotypic traits with comprehensive genomic similarity analyses. For each pathway, I chose a number of functional marker genes, based on current literature, as indication of present/absent of each pathway (Fig. 2.9, Table A7, A8).

2.3.7.2.1 Sulfur metabolism: DMSP demethylation and DMSP cleavage pathways

Dimethylsulfoniopropionate (DMSP) is a ubiquitous sulfur containing compound found in the ocean produced by many marine phytoplankton and macroalgae, which can serve as an osmoprotectant (Moran et al. 2012), antioxidant (Sunda et al. 2002) or as a defense mechanism against grazing (Strom et al. 2003). As DMSP is also a source of carbon and sulfur for marine bacteria, it is a known chemoattractant (Seymour et al. 2010). Marine bacteria can utilize DMSP in two ways (Moran et al. 2012): the demethylation pathway, which produces methanethiol (MeSH), and the cleavage pathway, which produces DMS (Moran et al. 2003, Reisch et al. 2011). MeSH is an important source of cellular sulfur and it has long been known that bacterial can incorporate MeSH directly into sulfur containing amino acids (Visscher et al. 1992, González et al. 1999). The second pathway cleaves DMSP into DMS, a volatile sulfur compound that plays an important role in global climate regulation (Lovelock et al. 1972, Charlson et al. 1987, Vallina and Simó 2007, Moran et al. 2012) and is an important part of the sulfur cycle. Members of the *Roseobacteraceae* family are one of the few that are known to contain both pathways (Moran et al. 2003), suggesting the importance of DMSP to this family bacteria.

All isolates capable of the DMSP demethylation pathway have at least one homolog each of the *dmdABC* genes (Moran et al. 2012). Therefore, in my dataset, these three genes (*dmdABC*) are used as functional markers for the demethylation pathways where only isolates with at least one homolog of each are considered to be capable of DMSP demethylation. The cleavage pathway is more complicated as there are six homologous DMSP lyases (*dddWPQDLY*) and not only can isolate contain multiple copies of each gene, it is also not necessary to have all six homologs (Moran et al. 2012). I, therefore, used all six DMSP lyase genes as functional markers for the DMSP cleavage pathways, but isolates that contain any number of the six genes are considered to be able to cleave DMSP. Overall, 64% of the *Roseobacteraceae* are capable of the cleavage pathway which is significantly higher than the 38% of the *Rhodobacteraceae* that are able to do so (Fig. 2.9, Table A7). The demethylation pathway shows a similar pattern where 52% of the *Roseobacteraceae* are likely able to perform DMSP demethylation compared to only 7% of the *Rhodobacteraceae* (Fig 2.9, Table A7). DMSP cleavage and DMSP demethylation is present in majority of the *Roseobacteraceae* family suggest these two pathways are likely ancestral traits within this family and was subsequently lost by some.

The importance of DMSP to the *Roseobacterace* family is further highlighted by the fact that 40% of the *Rosoebacteraceae* are capable of performing both pathways whereas only 3% of the *Rhodobacteraceae* can. The difference in proportion of isolates capable DMSP degradation between these two families is consistent with the fact that majority of the *Roseobacteracae* family are found in the marine environment in association with marine algae blooms where DMSP is commonly found (Buchan et al. 2005).

2.3.8.2.2 Quorum sensing: Acyl-homoserine lactone production and response

Marine bacteria can be broadly classified as free-living, those that can thrive on minimal nutrient, or patch-associated, those that are able to exploit small nutrient rich patches (Luo and Moran 2015). Patch-associated bacteria, such as the members of the *Roseobacterace* family, generally have a larger genome size encoding a variety of genes that allow these bacteria to respond quickly to changes in the environment (Luo and Moran 2015). One of the adaptations that members of the *Roseobacteraceae* family is quorum sensing. Quorum sensing is an important behavioral modulation mechanism that regulates many phenotypes that requires coordinated behavior, such as biofilm formation and pathogenicity (Case et al. 2008). This mechanism allows bacteria to quickly respond to different environmental cues and effectively cope with the changes in their environments.

Acyl-homoserine lactone-based quorum sensing (AHL-QS) is the most commonly described QS mechanism in *Proteobacteria* (Case et al. 2008) and is highly conserved within the *Roseobacteraceae* family (Cude and Buchan 2013). A complete AHL-QS circuit consists of two genes, *luxRI* (Case et al. 2011). The LuxR protein is the response protein. It mediates gene expression of other proteins in the cell and also activates the *luxI* gene. The LuxI protein is responsible for the synthesis of AHL. Not only can a single organism have more than one copies of the *LuxRI* genes, there can also be more copies of one than the other (Case et al. 2008). In this study, isolates that contain at least one copy each of the *luxRI* genes are considered likely capable of AHL-QS. I found that 56% of the *Roseobacteraceae* family is capable of AHL-QS which is significantly higher than the 4% of the *Rhodobacteraceae* family (Fig. 2.9, Table A8). AHL-QS, therefore, seems to be a trait that is more prominent in the *Roseobacteraceae* family

than the *Rhodoabacteraceae* family which again fits with known ecological data as majority of the *Roseobacteraceae* family are marine bacteria.

2.4 Conclusion

This study highlights several issues with the taxonomic classifications within the *Rhodobacteraceae* family of the order *Rhodobacterales*. Overall, I identified 17 polyphyletic and seven paraphyletic genera. All paraphyletic genera, with the exception of *Paracoccus*, were resolved by merging conflicting genera with existing ones. When possible, polyphyletic genera were resolved by splitting or merging isolates based on genomic and phylogenetic data. Genomic metrics (i.e, AAI, codon position similarities), phylogenetic data, and PD show that the genus *Halocynthiibacter*, which currently contains only two species, should be split into separate genera. Three species level misclassifications were also identified and resolved based on dDDH, ANI and phylogenetic analysis. Lastly, I also proposed to move the roseobacter clade into a new family, *Roseobacteraceae* fam. nov. based on phylogenetic, genomic and *in-silico* phenotypic analysis.

2.5 Materials and Method

2.5.1 Dataset descriptions

As of January 13th, 2019, WGS of 342 type strains within the *Rhodobacteraceae* family were available on GenBank (National Center for Biotechnology Information) (Table A9). In addition, three *Agrobacterium tumefaciens* strains were used as the outgroup for all phylogenetic analyses. Plasmid sequences were excluded from analyses where possible.

2.5.2 Genome annotation and core gene identification

An important limitation of core-genome phylogeny is the quality of the assembled genomes used. Low sequence quality or poorly assembled genomes will affect gene annotations and the number of core genes identified (Moura et al. 2016) which will ultimately affect the reconstruction of the phylogeny. I addressed this issue by ensuring our genome sequences are complete or nearly complete (i.e., ≥95% complete) with low levels of contamination (>5%) as outlined in CheckM (Parks et al. 2015), which assess these criteria based on the present and the number of copies of a set of well-defined core genes. As a result, we excluded 11 genomes from our initial dataset of 342 genome sequences obtained from GenBank leaving us with 331 genomes (Table A10).

All 331 high-quality *Rhodobacteraceae* genomes that meet my quality filter criteria (≥95% complete and <5% contamination) together with three *Agrobacterium tumefaciens* strains were annotated using RAST 2.0 (Aziz et al. 2008) or Prodigal 2.6.3 (https://github.com/hyattpd/Prodigal). Core genes, as defined as genes present in all organisms of interest, are then identified using Bacterial Pan Genome Analysis (BPGA) (Chaudhari et al. 2016) based on USearch clustering algorithm (Edgar 2010).

2.5.3 16S rRNA phylogenetic analysis

16S rDNA-based phylogeny for the full dataset of 333 genomes (331 type strains and three *Agrobacterium tumefaciens* strains) (Table A1) was reconstructed to highlight the impact of 16S rDNA-based phylogeny has on taxonomic classification within this group. A single copy

of the full length 16S rRNA gene was extracted from all 333 organisms. These sequences were then aligned with muscle with default parameters (Edgar 2004). The final alignment (with 1,628 nucleotide positions) was used to reconstruct a maximum-likelihood phylogenetic tree using RAxML 8.2.12 (Stamatakis 2014). The GTR (general time reversible) nucleotide substitution model and gamma model of rate heterogeneity was used and robustness of branches was estimated with 1000 bootstrap replicates.

2.5.4 core-genome phylogenetic analysis

There were 140 core genes identified from the 331 genomes in the initial dataset. For every core gene, the amino acid sequences were aligned with Muscle (Edgar 2004) with default parameters. The core gene alignments were then concatenated using Geneious 8.1.8 (Kearse et al. 2012), resulting in final alignments with 71,480 amino acid positions. This alignment was used to reconstruct a core-genome phylogenetic tree of the 333 genomes RAxML 8.2.11 (Stamatakis 2014) with PROTGAMMAAUTO option for automatic model selection. Robustness of branches was estimated with 100 bootstrap replicates.

2.5.5 Species delineations

There are established genomic and phylogenetic criteria for species level designation. Phylogenetically, the minimum requirement for a set of isolates to be considered as members of the same species is that they must form monophyletic clades, for which I assessed using core-genome phylogenetic trees. Average nucleotide identity (ANI) and *in-silico* DNA-DNA Hybridization (dDDH) are the two genomic metrics used for species delineation. Based on

previous studies, 70% dDDH and 95% ANI were used as the species level threshold (Goris et al. 2007, Richter and Rosselló-Móra 2009, Meier-Kolthoff et al. 2013). ANI was calculated using JSpecies (Richter and Rosselló-Móra 2009) and dDDH was calculated with the online GGDC calculator tool (Meier-Kolthoff et al. 2013) with default parameters for both.

To identify any species level misclassifications, dDDH was calculated for isolates belonging to the same genus. For polyphyletic genera, only isolates within the same monophyletic clades are considered as it is impossible for isolates to share more than 70% dDDH values if they are not monophyletic with each. For any genera where species level misclassifications was identified, ANI was calculated for those comparisons. Taxonomic classifications were only proposed for isolates that meet or surpass both the dDDH and ANI species threshold and are also monophyletic.

2.5.6 Assessing genomic level similarities for genus and family level

AAI and codon position similarities were used to assess genus and family level genomic similarities. Evolutionary distance based on the core-genome phylogenetic tree was quantified using patristic distance (PD). AAI was calculated using compareM (https://github.com/dparks1134/CompareM), and codon position similarities were calculated using Geneious 8.1.8 (https://www.geneious.com) and translatorX (Abascal et al. 2010) respectively.

2.5.7 Genus level delineation based on genomic and phylogenetic data

To assess if any currently recognized genera are misclassified, genomic metrics were calculated for all within and between monophyletic genera comparisons, which excludes any poly- and paraphyletic genera. Mann-Whitney U-test were used to assess the significance of the difference between the two distributions. Poly- and paraphyletic genera are identified based on the core-genome phylogenetic tree (Fig 2.2).

For paraphyletic genera, the first described genus within the clade, which I will refer to as the primary genus, will retain the genus designation. Other conflicting genera within that clade will be merged with the primary genus as per rule 38 of the code (Parker et al. 2019). All genomic metrics were calculated for within and between genera comparisons within these clades. These values were compared to those obtained from within and between recognized monophyletic genera comparisons (Table A5) to determine whether genomic similarities among genera I am merging falls within the expected range of within genus comparisons.

For polyphyletic genera, the clade containing the type species of the genus, which I will refer to as the primary clade, will retain the genus designation as per rule 39a of the code (Parker et al. 2019). All other clades will require a novel genus level designation or merged with existing genera. Similar to paraphyletic genera, for all clades where genus level reclassifications are required, genomic metrics are used to determine whether these genera should be merged with existing genera or given novel genus designations.

2.5.8 Family level delineation based on genomic, phenotypic and phylogenetic data

Genomic similarities at the family level were assessed based on AAI and codon position similarities. Core-genome phylogeny was used to assess phylogenetic relationships. Environment of isolation and salinity level were collected from original publications. Phenotypic traits characteristics of the *Roseobacteracae* family and marker genes used to assess present/absent of these traits were identified from current literature. Using Rast annotations (Aziz et al. 2008), present/absent of major pathways were assessed for all 245 members of the *Roseobacteraceae* family and 75 members of the *Rhodobacteraceae* family. Significance of the difference in proportion of these pathways between the two families was assessed using proportion Z-test.

2.6 Data availability

All genomes used in this study were retrieved from NCBI Genbank database. The list of accessions numbers is listed in table A9.

2.7 Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (to YFB and RJC), the Integrated Microbial Biodiversity program of the Canadian Institute for Advanced Research (to YFB), the graduate student scholarships from Alberta Innovates – Technology Futures (to KYHL and FDO), Canada Graduate Scholarship Master's program from Natural Sciences and Research Council of Canada (to KYHL) and the Bank of Montréal Financial Group (to FDO). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Chapter 3: A *Vibrio Cholerae* Core-Genome Multilocus Sequence Typing Scheme to Facilitate the Epidemiological Study of Cholera.

A version of chapter 3 will be submitted for publication as:

> "Liang, K.Y.H., Orata, F.D., Islam, M.T., Nasreen, T., Alam, M., Tarr, C.L., Boucher, Y.F. A Vibrio cholerae Core-Genome Multilocus Sequence Typing Scheme to facilitate the epidemiological study of Cholera."

KYHL and YFB design the study and wrote the manuscript. FDO and KYHL performed bioinformatic analyses and IMT helped with data analysis. CLT, TN and AM provided clinical and environmental strains. YFB supervised the project.

Chapter 3

3.1 Abstract

Core genome multilocus sequence typing (cgMLST) has gained popularity in recent years in epidemiological research and subspecies level classification. cgMLST retains the intuitive nature of traditional MLST but offers much greater resolution by utilizing significantly larger portions of the genome. Here, I introduce a cgMLST scheme for *Vibrio cholerae*, a bacterium abundant in marine and freshwater environments and the etiologic agent of cholera. A set of 2,443 core genes ubiquitous in *V. cholerae* were used to analyze a comprehensive dataset of 1,262 clinical and environmental strains collected from 52 countries, including 65 genomes newly sequenced in this study. I established a sublineage threshold, based on 133 allelic differences that creates clusters nearly identical to traditional MLST types providing context and backwards compatibility to new cgMLST classifications. I also defined an outbreak threshold, based on seven allelic differences, that is capable of identifying strains that are part of the same outbreak and closely related isolates which could give clues on its origin. Using this scheme, I confirmed the South Asian origin of modern epidemics and identified a strong geographic affinity among sublineages of environmental isolates. Advantages of cgMLST are highlighted by a direct comparison with existing classification methods such as MLST, and single nucleotide polymorphism-based methods. cgMLST outperforms all existing methods in terms of resolution, standardization, and ease-of-use. I anticipate this scheme will serve as a basis for a universally applicable and standardized classification system for *V. cholerae* research and epidemiological surveillance in the future. This cgMLST scheme is publicly available on PubMLST (https://pubmlst.org/vcholerae/).

3.2 Introduction

Vibrio cholerae is the causative agent of cholera, an acute diarrheal disease. Cholera is transmitted in a fecal-oral route mostly by contaminated food or water (Jahan 2016, Momba and Azab El-Liethy 2017). The case fatality rate (CFR) of cholera can be up to 50% without treatment, but with proper medical care, CFR is usually less than 1% (Clemens et al. 2017, Momba and Azab El-Liethy 2017). In developed countries, with proper water treatment facilities, cholera is practically non-existent aside from imported cases. Unfortunately, this cannot be said for many developing countries lacking this infrastructure, where cholera has been endemic for centuries such as in parts of South Asia (Kaper et al. 1995). As it is also difficult to eradicate cholera (Islam et al. 2017), this disease often becomes endemic in regions where it has been introduced, for example in Latin America in 1991 (Dalsgaard et al. 1997, Choi et al. 2016), Haiti in 2010 (Orata et al. 2014), and Yemen in 2016 (Weill et al. 2018). It is estimated that there are over a million cholera cases each year resulting in tens of thousands of deaths worldwide (Ali et al. 2015). Being an indicator of healthcare and socio-economic disparities (Mintz et al. 2013, Legros 2018), this disease is often under-reported due to its negative influence on tourism as it implies poor water quality (Sack et al. 2006). In addition, it should be noted that only one lineage of *V. cholerae* is responsible for all documented pandemics and major outbreaks with numerous harmless environmental isolates found worldwide (Islam et al. 2017). Together with the lack of a universally applicable and standardized classification method, outbreak surveillance and source attribution is often challenging (Orata et al. 2014, Jahan 2016). The Haiti outbreak for example, due to these limitations, required continuous extensive genomic and epidemiological research since the beginning of the outbreak to determine the source of introduction, which was not

confirmed until August 2011 even though it broke out in July 2010 (Hendriksen et al. 2011, Frerichs et al. 2012, Katz et al. 2013, Orata et al. 2014, Frerichs 2016).

A typing method for use in global surveillance of pandemic causing pathogens such as *V. cholerae* should be efficient and easy to use, with the potential to be applied to all *V. cholerae* strains around the world. Therefore, it must have the capacity to analyze thousands of genomes efficiently and new genomes should be easily typed as they get sequenced. As all cholera outbreaks are caused by a single lineage of *V. cholerae*, the pandemic generating/phylocore genome (PG) lineage, which includes 7th pandemic El Tor, El Tor sister, El Tor progenitor, classical and classical sister clade (Chun et al. 2009, Boucher 2016, Islam et al. 2017), this method should also be able to differentiate isolates at a fine scale and separate outbreaks caused by genetically similar strains. Such a method will help create a comprehensive database with detailed epidemiological data that will allow for the analysis of future outbreak strains in a global context and guide subsequent epidemiological analyses. Different methods for subspecies level classification and outbreak surveillance have been developed for *V. cholerae*. These methods include serotyping, multilocus sequence typing (MLST) (Octavia et al. 2013, Kirchberger et al. 2016), multilocus variable number of tandem repeats (VNTR) analysis (MLVA) (Garrine et al. 2017, Bwire et al. 2018), and single nucleotide polymorphism (SNP)-based approaches (Katz et al. 2013). Despite the popularity of these methods, there are important limitations to each.

Serotyping based on the presence of cell surface O-antigens is one of the earliest attempts at subspecies level classification of *V. cholerae*. There are now over 200 serogroups of *V. cholerae* identified; however, only two serogroups, O1 and O139, have been found to be responsible for all major documented epidemics and pandemics (Safa et al. 2010, Boucher et al. 2015). Serogroup O1 can be further divided into two biotypes (El Tor and Classical) and three

serotypes (Inaba, Hikojima, and Ogawa) (Momba and Azab El-Liethy 2017). The lack of resolution within the epidemic strains and the possibility of serogroup conversion (Mandal et al. 2011) limits the use of serotyping in epidemiological studies.

MLST provides a standardized classification method that is based on a collection of six to seven well-defined housekeeping genes (Maiden et al. 1998). MLST was used to study a number of cholera outbreaks and allowed the descriptions of its general population structure (Horwood et al. 2011, Luo et al. 2013). It is reproducible and provides reliable results; however, it is unable to differentiate between closely related strains which limits its use in outbreak surveillance (Gonzalez-Escalona et al. 2008, Maiden et al. 2013). In addition, direct comparisons between different MLST schemes are difficult, as different schemes utilize different housekeeping genes.

MLVA utilizes VNTR regions, which are under less selective pressure than housekeeping genes. This method therefore provides greater resolution than MLST for some bacterial species (Lam et al. 2012, Chenal-Francisque et al. 2013). However, due to their rapid mutation rate, VNTR regions are more affected by homoplasy where two isolates may share the same MLVA profile due to convergent mutation and not by vertical descent (Vogler et al. 2011). As a result, MLVA may produce clusters that do not necessarily reflect phylogenetic relationships (Struelens and Brisse 2013). Two common PCR-based methods exist for the typing of VNTR regions, but each have significant limitations (Sabat et al. 2013). The first method is multiplex PCR which can analyze all loci at once, but it is impossible to determine which bands correspond to which loci; therefore, this method only produces a banding pattern for strain identification, which makes it difficult to standardize and communicate results. The second method is the separate amplification of VNTR regions but determining the number of repeats based on amplicon size information alone is difficult if the difference in size is not large enough. In addition, different

types of mutations that do not necessarily change the number of repeats, can cause a change in amplicon size. Sequencing is needed to confirm MLVA profiles, but repeat regions increase the chances of sequencing errors (Klassen and Currie 2012). Due to these limitations, stringent quality control is required for reliable MLVA analysis (Danin-Poleg et al. 2007).

SNP-based analysis is one of the most common whole-genome-based methods currently being used and was applied to various outbreaks (Katz et al. 2013, Leekitcharoenphon et al. 2014, Wong et al. 2016). It relies on the identification of conserved SNPs in strains of interest using next-generation sequence reads or assembled genomes. The number of SNPs can then be related to the evolutionary distance between isolates. SNP-based analysis provides reliable results with sufficient resolution for epidemiological studies, but it is sensitive to horizontal gene transfer and recombination events, as each event will result in many SNPs being created. The number of SNPs between two strains, therefore, does not necessarily reflect the true phylogenetic relationship. SNPs found in recombinogenic regions should therefore be removed which, depending on the organism of interest, can be anywhere from 30% to 97% of SNPs identified (Chen et al. 2013, Qin et al. 2016). Since recombination and horizontal gene transfer events are common in *V. cholerae* (Meibom et al. 2005, Borgeaud et al. 2015, Wang et al. 2016), SNP-based methods, although suitable in individual epidemiological studies, will be difficult to serve as a universal classification method for *V. cholerae*.

Core-genome MLST (cgMLST), also known as the gene-by-gene approach, overcomes the various limitations of previously mentioned subtyping methods and was established to serve as a universally applicable standardized typing scheme. Similar to MLST, cgMLST relies on individual gene sequences to differentiate between closely related strains; however, instead of using six to seven housekeeping genes, cgMLST utilizes hundreds to thousands of core genes,

which are genes commonly found in all strains of a species. By utilizing a much larger portion of the genome, cgMLST provides superior resolution compared to traditional MLST schemes. By combining the expandable and standardized classification method that made traditional MLST favourable with the resolution of whole-genome-based methods, cgMLST is becoming more popular in epidemiological and ecological studies (de Been et al. 2015, Moura et al. 2016, Cody et al. 2017, Janowicz et al. 2018, Bletz et al. 2018, Neumann et al. 2019, Jones et al. 2019). This method has the added advantage of backwards compatibility with all MLST schemes, meaning it is possible to determine MLST profiles of any isolates based on their cgMLST profiles, since cgMLST, by definition, would include all housekeeping genes. This allows for a 1:1 mapping of any previously established MLST scheme to the cgMLST scheme, helping consolidate existing classification information.

Another major benefit of cgMLST is that, much like traditional MLST methods, it is possible to establish different clustering thresholds to define important groups. Clonal complexes are examples of clustering thresholds established by MLST schemes, where each clonal complex corresponds to a cluster of isolates that share at most one allelic difference across all seven genes sequenced. Some important clonal complexes were shown to correspond to either groups established by a previous typing method (Sails et al. 2003) or major outbreak strains (Leavis et al. 2006). However, cgMLST offers even greater flexibility than MLST in this regard, given the number of loci considered. With small clustering thresholds where groups are created based on the sharing of a large number of alleles, it is possible to identify closely related strains useful in epidemiological studies. On the other hand, with larger clustering thresholds, it is possible to identify lineage- or even sublineage-level differences to study large scale patterns and answer broader ecological questions. The benefits of defining clustering thresholds with cgMLST

schemes have already been demonstrated in other human pathogens, such as *Brucella melitensis* (Janowicz et al. 2018), *Campylobacter jejuni* (Cody et al. 2017) , *Clostridium difficile* (Bletz et al. 2018), *Enterococcus faecium* (de Been et al. 2015), and *Listeria monocytogenes* (Moura et al. 2016).

In this study, I introduce a cgMLST scheme for the genome-wide typing of *V. cholerae* and demonstrate its universality and efficacy by applying it to known cholera outbreaks around the world. The advantages of cgMLST are presented by comparing the scheme with previously established classification methods. Additionally, I have produced a 1:1 mapping of the cgMLST scheme against two MLST schemes for *V. cholerae* (Octavia et al. 2013, Kirchberger et al. 2016), allowing for the consolidation of existing classification information. The cgMLST scheme, genome sequences used in this study, and relevant epidemiological information are publicly available on PubMLST (https://pubmlst.org/vcholerae/), which allows for the automatic annotation and subsequent analyses of hundreds of newly uploaded *V. cholerae* genomes in a global context. This increase in efficiency, standardizability, and resolution compared to current methods makes cgMLST the most suitable classification scheme for large scale *V. cholerae* surveillance. By applying this scheme to over 1,200 isolates collected around the world, it was possible to establish outbreak and sublineage thresholds which not only allowed us to validate the South Asian origin of many modern epidemics as proposed in previous studies (Reimer et al. 2011, Islam et al. 2017, Weill et al. 2017) but also identified a strong geographic signal among environmental strains where isolates from the same sublineage are also from the same geographic region; a pattern that is not seen in clinical isolates.

3.3 Materials and Method

3.3.1 Dataset description

On November 6th 2018, 1,172 *V. cholerae* genomes consisting of 800 draft and complete genomes and 372 sequence read archives (SRAs) were available from both publicly available databases and private collections and were selected as our dataset. One hundred sixteen SRAs from a recent study on the Yemen cholera outbreak (Weill et al. 2018) were subsequently added as an independent evaluation of the cgMLST scheme (Table B1). The 488 SRAs were assembled using skesa (Souvorov et al. 2018) or the CLC Genomics Workbench 7 (https://www.qiagenbioinformatics.com) using default parameters. This total dataset of 1,288 included twenty-six genomes with less than 90% of the core genes, which were identified using USearch (Edgar 2010) based on RAST (Aziz et al. 2008) annotations. These twenty-six genomes were removed from subsequent analyses resulting in a final dataset of 1,262 genomes collected from 52 countries and spanning 82 years from 1937 to 2018 (Table B2). These include a historical collection from the 6th cholera pandemic, clinical isolates from outbreaks in various countries (e.g., Bangladesh, India, Haiti, Yemen, the Democratic Republic of Congo, and Russia), and environmental isolates from different parts of the world (e.g., USA, Mexico, Australia, etc).

3.3.2 Gene identification and allele assignments

Instead of using the full dataset of 1,288 genomes for core gene identification, I selected a subset of high-quality genomes for this purpose. The reason is because core gene identification is highly dependent on the initial dataset and the inclusion of poorly assembled and/or sequenced data will reduce the number of core genes identified (Moura et al. 2016). First, 800 already

assembled draft or complete genomes were selected for core gene identification. Low-quality assemblies were then eliminated by removing genomes with less than 40× coverage and/or N50 values less than 40 kb. From a previously established cgMLST scheme for *L. monocytogenes*, 40× coverage and 20kb N50 value were used as cutoff thresholds, as genomes that do not meet these criteria resulted in a low proportion of loci being called (Moura et al. 2016). The 40× threshold was adopted in this study; however, because the average *V. cholerae* genome size (~ 4 Mb) is larger than the average *L. monocytogenes* genome (~ 3 Mb), 40 kb was instead selected as the N50 cutoff. The use of these cutoffs resulted in the removal of 82 genomes.

The remaining 718 genomes were annotated using RAST (Aziz et al. 2008) and USearch (Edgar 2010) and a tentative set of core genes that were on average present in 99% of the genomes were selected. An additional 13 genomes were removed, as they lacked more than 90% for the core genes (Table B3), resulting in a final dataset of 705 high-quality genomes. However, an additional 26 genomes were subsequently removed for the core gene analysis as it has been previously suggested that they form a highly divergent lineage within the *V. cholerae* (Liang et al. 2017, 2019, Islam et al. 2018), ensuring that the dataset used for core gene identification consists only of unambiguously *V. cholerae* isolates (also as verified by average nucleotide identity (Goris et al. 2007) and digital DNA-DNA hybridization (Meier-Kolthoff et al. 2013) between genomes). Completeness and potential contamination of all remaining 679 genomes were also independently evaluated by checkM, which estimates these values based on the presence and number of copies of a set of pre-defined single copy marker genes (Parks et al. 2015) (Table B4). All genomes were, according to the criteria established by checkM, nearly complete (≥97%) with medium to low levels of contamination (<7%) (Parks et al. 2015).

Each orthologous gene was compared against the *V. cholerae* N16961 reference genome using BLASTN (Altschul et al. 1990) to determine gene function. Any gene family with no homolog in N16961 or classified as pseudogenes on the NCBI GenBank database were removed, meaning N16961 was 100% complete for the cgMLST scheme. Any gene that was present in more than one copy in any of the initial 679 genomes was also removed, as they were considered paralogous. Thus, in this context, core genes are defined as being present in at least 90% of the 679 high-quality assembled genomes in a single copy. By choosing a relaxed cutoff of 90% completeness, we accounted for missing genes due to sequencing, annotation, or assembly errors while ensuring there is sufficient resolution to differentiate between closely related strains, with at least 2,199 loci remaining for classification purposes. The final cgMLST scheme utilizes a set of 2,443 single-copy core gene loci, which is 2,425,296 bp in size and covering approximately 61% of the genome. The list of core genes is available on PubMLST (https://pubmlst.org/vcholerae/).

Automated scripts in BIGSdb (Jolley and Maiden 2010) were used to perform allele calls and assignments for all 1,264 isolates. Allele calls were made only for complete coding sequences with minimum of 70% similarity and 70% length coverage at the nucleotide level, as previously described (Moura et al. 2016). Default settings were used for all other parameters.

3.3.3 Core-genome sequence type (cgST) assignment

cgST, which was defined as a unique combination of alleles of all loci included in the scheme, was assigned for all isolates, excluding those from the Yemen outbreak study (8), with an in-house script, as previously described (Garg et al. 2003). Briefly, missing loci were replaced

with the most common allele when assigning cgSTs, allowing for a conservative estimate of diversity. The 116 isolates from the Yemen cholera outbreak study (Weill et al. 2018) were annotated automatically by uploading them to PubMLST. PubMLST treated missing alleles as 'N'. cgSTs were assigned to each allele profile, treating 'N' as a regular allele designation. However, different from typical allele designations, 'N's can represent any allelic sequence; therefore, some isolates may contain multiple cgST designations, all of which are possibly true cgSTs. For isolates with more than one cgST suggested by PubMLST, postprocessing was done using an in-house script to identify the most likely cgST, which was determined by assuming missing loci contained the most common allele (Table B5). It is expected that as genome sequencing becomes more reliable, higher quality genomes will be available and any missing data can be updated as needed.

3.3.4 MLST scheme and sequence type (ST) assignments

Two MLST schemes developed for *V. cholerae* were mapped to this cgMLST scheme. The first MLST scheme developed in 2013 by Octavia and colleagues (Octavia et al. 2013) was used to study the global population structure of non-O1/non-O139 *V. cholerae* and is currently hosted on PubMLST. All isolates uploaded to PubMLST were automatically annotated with this scheme. Any missing data in this scheme was ignored and no ST designation was assigned. The second MLST scheme developed in 2016 by Kirchberger and collegaues (Kirchberger et al. 2016) was used to study the population structure of environmental *V. cholerae* in a region on the US East Coast. The second MLST scheme is not currently hosted on PubMLST, but because the housekeeping genes in this scheme are also found in the cgMLST scheme, a similar in-house script used in cgST assignments was used to assign ST designations. Therefore, all isolates in

this study, when possible, were assigned three designations – two ST designations based on the two previous MLST schemes (Octavia et al. 2013, Kirchberger et al. 2016) and one cgST designation based on the cgMLST scheme from this study.

3.3.5 Outbreak and sublineage clustering thresholds

A clustering threshold was defined as the maximum number of allelic differences found within a cluster. All clusters were produced based on the single-linkage clustering method, which meant an isolate belonged to a cluster if it linked with any isolate within that cluster. Two metrics were used as general guidelines for determining clustering thresholds. The first metric used was the Dunn Index (DI), which measured clustering efficiencies (Dunn 1974). Briefly, the DI was highest for a network (i.e., the network has the best clustering efficiency) when the intra-cluster distances were minimized, and the inter-cluster distances were maximized. Since isolate distances were measured based on allelic differences, a high DI resulted in clusters where isolates were more closely related to those found within the same cluster than those found in a different cluster. The DI was calculated using the R package 'clvalid' and 'boot' with 100 bootstrap replicates for each threshold and graphed using the R package 'ggplot2' (Brock et al. 2008, Wickham 2009, Canty et al. 2017, R Core Team 2017).

The second metric used was the Adjusted Rand Index (ARI), which measured the level of similarity between two networks when clustering the same set of isolates by measuring the amount of agreements (i.e., the number of pairs that were grouped either as being in the same cluster or different cluster in both networks) and disagreements (i.e., the number of pairs that were grouped together in one network but grouped separately in another) (Hubert and Arabie

1985). The values ranged from -1 (i.e, two networks are exactly opposite) to 1 (i.e, two networks are identical). ARI was used to determine the level of similarity between various clustering thresholds and the MLST schemes. ARI was calculated using the R package 'clues' and graphed using 'ggplot2' (Wickham 2009, Chang et al. 2010, R Core Team 2017).

3.3.6 Minimum spanning tree (MST)

All MSTs, unless otherwise specified, were constructed using GrapeTree MSTv2 (Zhou et al. 2018). Loci with missing data were included in the profile as "–". GrapeTree provided a novel algorithm that accounted for missing data when constructing an MST, an important feature since missing data is common in whole and core genome-based analyses. GrapeTree is currently integrated within PubMLST, which allows for quick visualization of the dataset with any provenance data.

3.3.7 Phylogenetic analysis

Parsnp v1.2 (Treangen et al. 2014) was used to reconstruct the phylogenetic tree using *V. cholerae* N16961 as the reference genome. The -x flag was used to enable filtering of SNPs in recombinogenic regions as identified by PhiPack (Bruen et al. 2006). Default settings were used for all other parameters. The phylogenetic tree included 1,146 genomes (all genomes except for the 116 isolates from the recent Yemen cholera outbreak study (Weill et al. 2018)). Since all isolates sequenced in that study belonged to the 7th pandemic El Tor lineage, it would have had limited impact on the overall structure of the tree. The phylogeny was visualized and annotated using iTOL (Letunic and Bork 2007).

3.3.8 Biogeographical analysis of environmental isolates

All isolates that were not part of the PG lineages (Chun et al. 2009, Boucher 2016) were first clustered based on the sublineage threshold using the python package networkX (Hagberg et al. 2008). Missing alleles were replaced with the most common allelic designation when calculating pairwise differences to establish a more conservative estimate of diversity. The network was then visualized using Cytoscape (Shannon et al. 2003).

3.3.9 Data availability

All previously sequenced *V. cholerae* genomes and the additional 65 genomes sequenced in this study will be deposited on NCBI GenBank database and is currently publicly available on PubMLST (https://pubmlst.org/vcholerae). Table B6 lists the accession numbers and links for all the genomes used in this study. In addition, all genome sequences, allelic profiles, cgST designations, ST designations, and relevant epidemiological data are publicly available on PubMLST (https://pubmlst.org/vcholerae/).

3.4 Results and Discussion

3.4.1 A high-resolution typing scheme for pandemic V. cholerae

The highest level of resolution of any cgMLST scheme is defined by cgSTs, where a unique cgST represents a unique allelic profile. Isolates that belong to the same cgST are expected to be phylogenetically very closely related, as although they may not have the exact genomic sequence, they do have the same sequence for all 2,443 core gene loci. I identified a

total of 1,026 cgSTs from 1,262 genomes collected from 52 countries. Even with this extensive dataset, I have yet to sample anywhere close to the total predicted cgST diversity for the global *V. cholerae* population (Fig. 3.1). All isolates were given at least one cgST designation and up to two MLST ST designations based on two previously established MLST schemes (Octavia et al. 2013, Kirchberger et al. 2016) (Table B2). MLST STs are defined based on the unique combination of all loci of a particular MLST scheme, which ideally uses six to seven well-defined housekeeping genes. Only 12 STs are exclusively present in the 7th pandemic El Tor lineage using the traditional MLST schemes (Octavia et al. 2013, Kirchberger et al. 2016), while 560 cgSTs are uniquely present in this group based on the cgMLST scheme (Table B2).

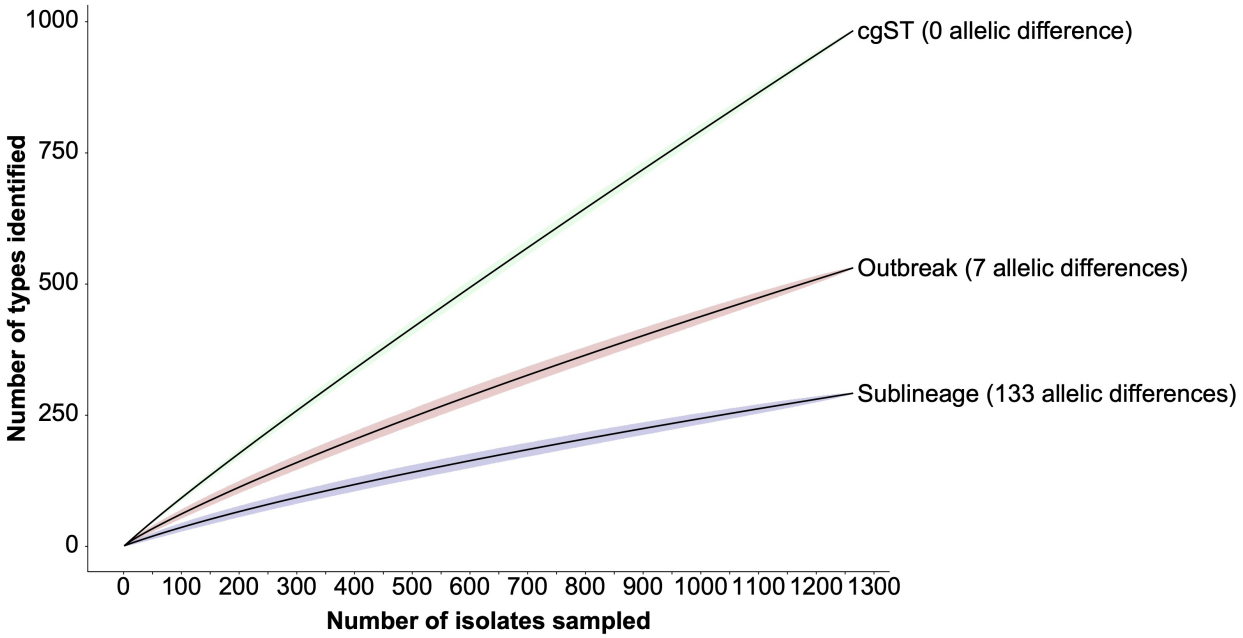


Fig 3.1: Rarefaction curve for cgST, outbreak threshold (7 allelic difference) and the sublineage threshold (133 allelic difference) computed using mothur (P. D. Schloss et al. 2009) with default parameters.

As the El Tor lineage is responsible for most cholera outbreaks around the world since the beginning of the 7th pandemic (Salim et al. 2005), this superior ability to resolve between closely related strains in the 7th pandemic El Tor lineages makes cgMLST more suitable in outbreak surveillance than traditional MLST schemes.

3.4.2 Backwards compatibility with previous subspecies classification methods

Much like how cgSTs are important in studying closely related strains typical in outbreaks, it is also important in establishing a standardized nomenclature at a higher level to answer broader ecological questions. Here, I propose a sublineage definition for *V. cholerae* based on our cgMLST scheme.

Pairwise allelic differences calculated between all isolates shows three major peaks (Fig. 3.2A). The first peak ends at 40 allelic differences, and the second peak ends at 133 allelic differences (Fig. 3.2B). The last peak begins at 2,200 allelic differences (Fig. 3.2A), which is expected due to mutational saturation (i.e. every single allele in the scheme is different between the two distantly related strains being compared). Both breaks (i.e., 40 or 133 allelic differences) could represent a potential sublineage delineation. To choose between the two thresholds, the clustering efficiency is measured by calculating the Dunn Index (DI) (Dunn 1974). Since cluster distances are measured by allelic differences, the network with the best clustering efficiency (i.e., the highest DI) will also produce clusters that best represent biological relationships, as isolates are more closely related to themselves than to isolates from other clusters. A DI was calculated for each clustering threshold in the range of 1 to 1,000 allelic differences with 100 bootstrap replicates (Fig. 3.3). As the clustering threshold defines the maximum number of allelic

differences within a cluster, the smaller the threshold, the more closely related the isolates will be within a cluster. It is clear that DIs in the range of 0 to 50 allelic differences are significantly lower than the DIs in the range of 100 to 350 allelic differences, with 133 being a clear local maximum. Since 133 allelic differences has the best clustering efficiency and it also represents a natural break where most isolate pairs have either less than or much greater number of allelic differences (Fig. 3.2B), it was chosen as the sublineage threshold.

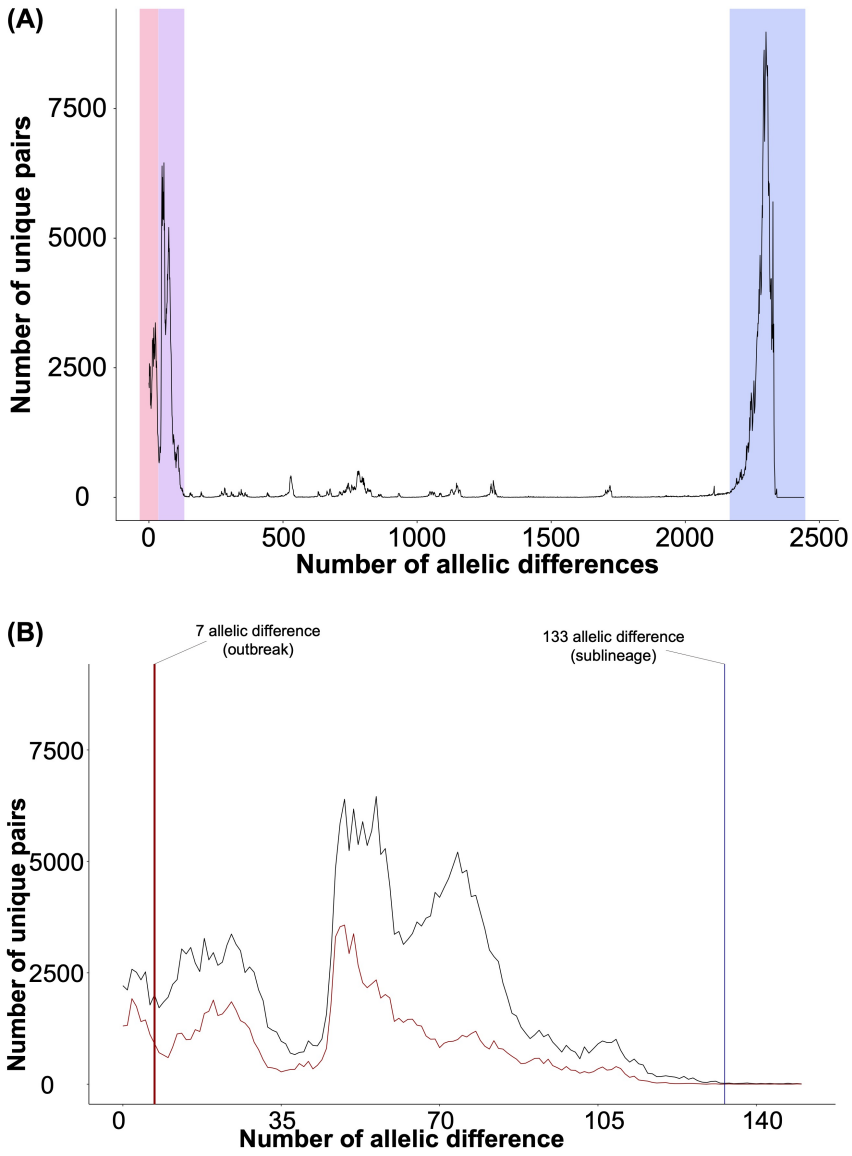


Fig 3.2: Pairwise allelic differences based on the cgMLST scheme for all isolates used in this study. Both plots show the frequency of allelic mismatches in pairwise comparisons. A) Pairwise comparisons of up to 2,443 allelic differences are shown. Major peaks are shaded. B) Comparisons with up to 150 allelic differences are shown. Vertical lines indicate the outbreak threshold (red) and sublineage threshold (blue). Pairwise comparisons of only clinical isolates are shown in red.

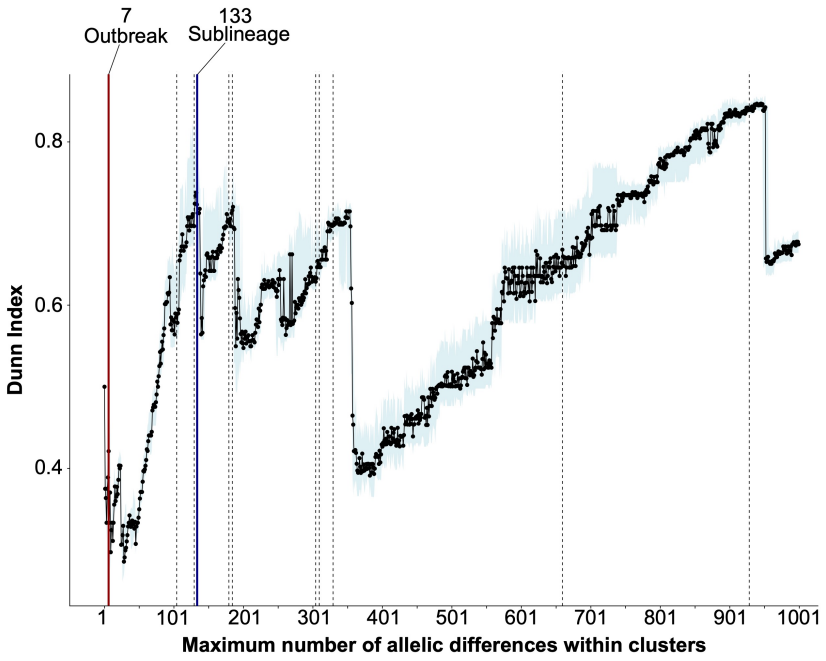


Fig 3.3: Plot showing the Dunn Index for clustering thresholds ranging from 1 to 1,000 allelic differences. Each clustering threshold is bootstrapped 100 times. The median, plotted with the light blue shade, indicates the 25th to 75th percentile range. Dark blue and dark red vertical lines indicate the sublineage and outbreak thresholds, respectively. The dotted lines represent other clustering thresholds used in the ARI calculations (Fig. 3.4B, 3.6).

Because cgMLST includes all housekeeping genes, information from the two MLST schemes previously developed for *V. cholerae* (Octavia et al. 2013, Kirchberger et al. 2016) can now be consolidated with the cgMLST scheme by creating a 1:1 cgMLST to MLST map. To evaluate the similarities between the sublineage threshold and the MLST schemes, I created a minimum spanning tree (MST) for all Bangladesh isolates (n = 255) showing only edges with 133 allelic differences or fewer (Fig. 3.4A and Fig. 3.5). Each cluster therefore represents a single sublineage. Bangladesh was chosen to compare cgMLST and MLST as it is the most extensively sampled country both in terms of clinical and environmental isolates in our dataset. The sublineage threshold produces clusters that closely resembles traditional MLST STs. Based on the 2013 MLST scheme (Octavia et al. 2013), each sublineage corresponds to exactly one ST (Fig. 3.5), whereas there is only one sublineage that contains two STs based on the 2016 MLST scheme (Kirchberger et al. 2016) (Fig. 3.4A). All but two isolates (*V. cholerae* strains N16961 and A19) belong to ST1. A19 and N16961 belong to ST290, which differs from ST1 at only one of seven MLST loci (Table B5). The reason these two isolates are of a different MLST ST could only be partly explained; they were isolated at an earlier time point (1970s near the start of the 7th pandemic (Mutreja et al. 2011)) than most of the remaining isolates, which were isolated from 1991 onwards (Table B2).

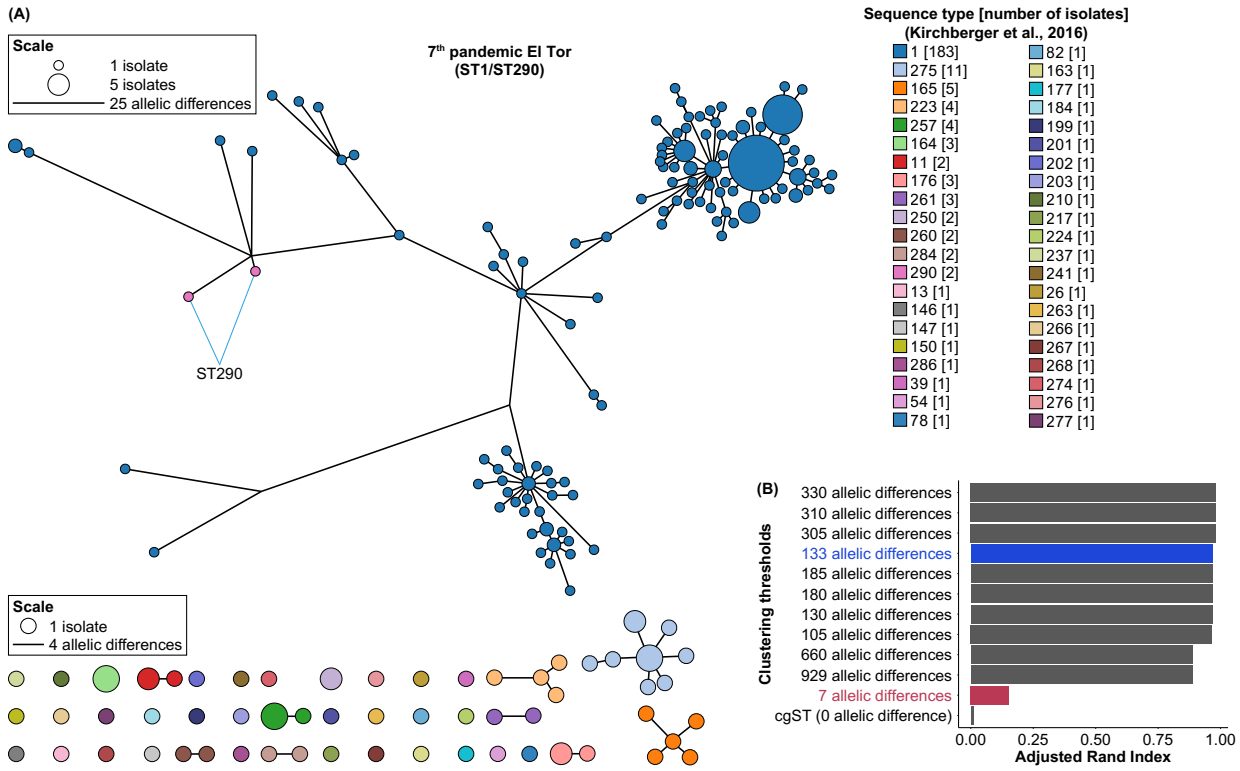


Fig 3.4: Evaluation of network similarities between cgMLST sublineage threshold (133 allelic differences) and the 2013 MLST ST (Kirchberger et al. 2016). A) Networks of all sublineages identified using only *V. cholerae* isolates from Bangladesh (*n* = 255). Each cluster represents a sublineage and includes isolates with less than or equal to 133 allelic differences with each other. Each node represents a cgST and is colored by sequence type based on the 2016 MLST scheme (Kirchberger et al. 2016). Size of the nodes are proportional to the number of isolates. The length of the connecting lines within a cluster is proportional to the number of allelic differences. B) Adjusted Rand Index for individual pairwise comparisons between predefined clustering thresholds (Fig. 3.3) and the 2016 MLST scheme (Kirchberger et al. 2016). The sublineage clustering threshold (i.e., 133 allelic difference) and outbreak threshold (i.e., 7 allelic difference) are indicated in blue and red bars, respectively.

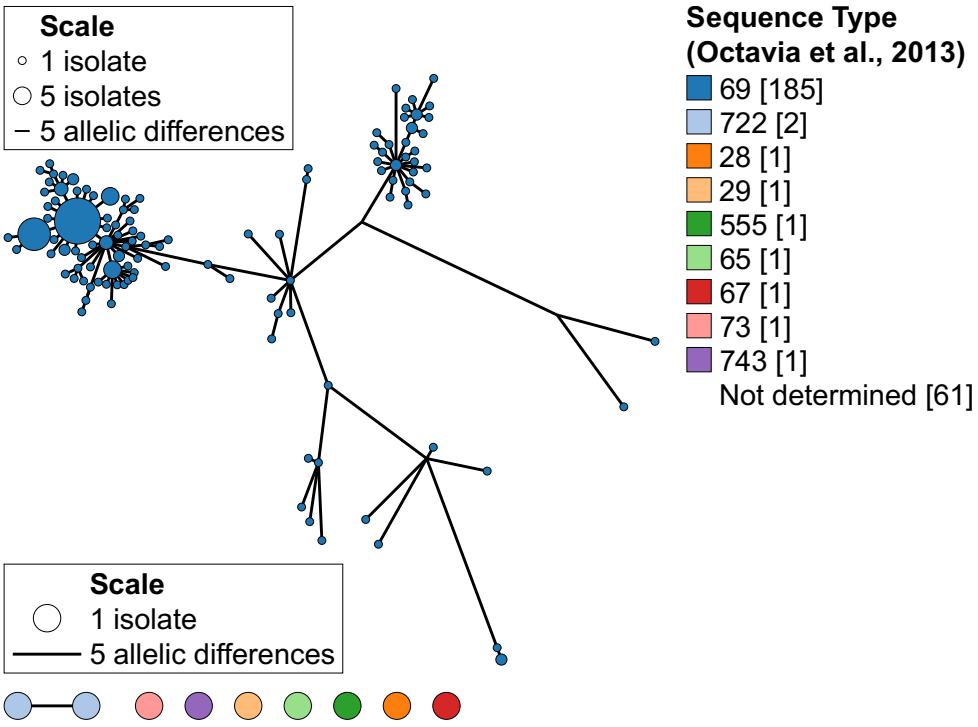


Fig 3.5: Evaluation of network similarities between cgMLST sublineage threshold (133 allelic differences) and the 2013 MLST scheme (Octavia et al. 2013). Networks of all sublineages identified using only *V. cholerae* isolates from Bangladesh (n = 255). Each cluster represents a sublineage and includes isolates with less than or equal to 133 allelic differences with each other. Each node represents a cgST and is colored by sequence type based on the 2013 MLST scheme (Octavia et al. 2013). Size of the nodes are proportional to the number of isolates. The length of the connecting lines within a cluster is proportional to the number of allelic differences.

It is impossible to visually evaluate similarities between two MSTs with over 1,200 nodes each simply due to the sheer volume of data. ARI was therefore used as a metric to determine network similarities (Hubert and Arabie 1985). In order to determine whether the sublineage threshold (i.e., 133 allelic differences) is indeed the best match to traditional MLST schemes, I chose 11 clustering thresholds distributed across the range of 1 to 1,000 allelic differences (Fig. 3.3) to compare with the MLST schemes. These additional thresholds are chosen as they have a relatively high DI compared to their immediate neighbours. More data points were chosen in the range of 105 to 330 allelic differences, as it was expected thresholds in this range will best match the traditional MLST schemes. Interestingly, all thresholds in that range had comparable ARIs regardless of the MLST schemes in question (Fig. 3.3B, Fig. 3.6), indicating that all of them, including the sublineage threshold, produces clusters similar to the MLST scheme. This would suggest that there can be a large range of diversity within a single MLST ST where isolates can have anywhere from 0 (i.e., have the same cgST) to 330 allelic differences. Although clustering thresholds between 105 to 330 allelic differences produce similar clusters to a traditional MLST scheme, 133 allelic difference was chosen as the sublineage threshold as it has the best clustering efficiency (Fig. 3.3) and it represents a natural breakpoint in the currently sampled population (Fig. 3.2B).

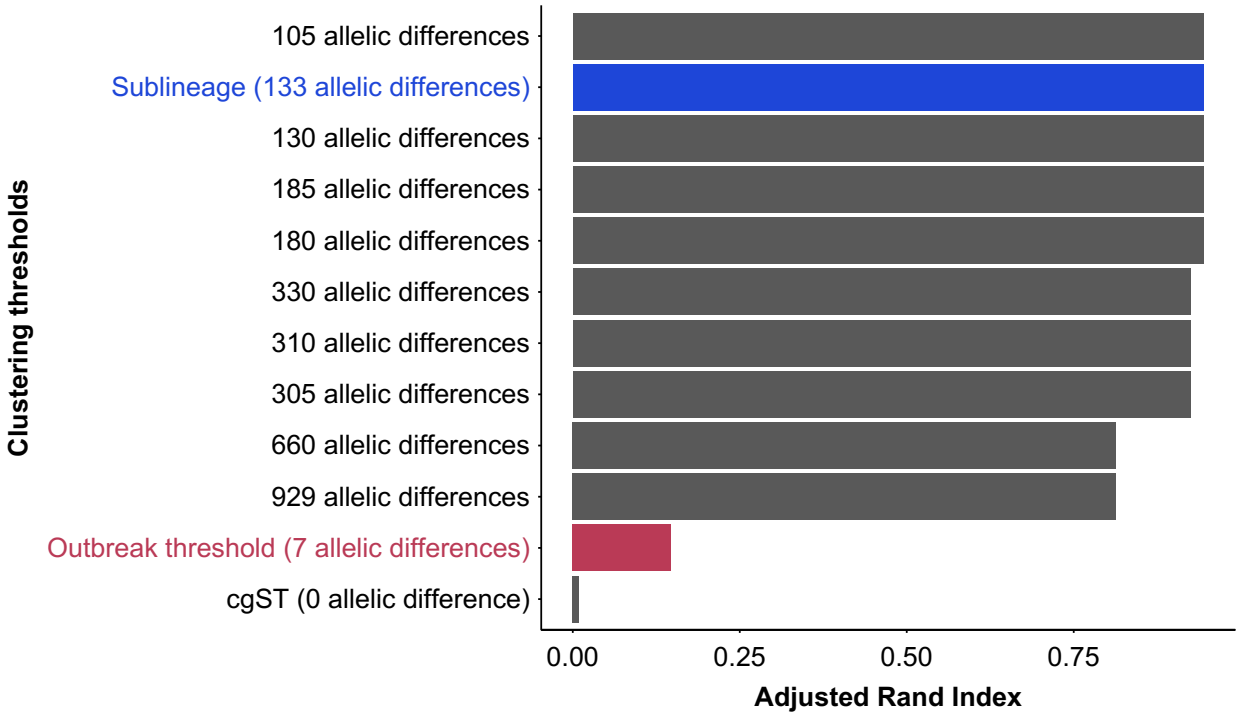


Fig 3.6: Adjusted rand index calculated with the same method as before (Fig 3.4), but compared with the 2013 MLST scheme (Octavia et al. 2013). The same allelic thresholds were chosen and the outbreak and sublineage thresholds are highlighted in red and blue respectively

A phylogenetic tree of 1,146 isolates was used to assess the phylogenetic support of the sublineage threshold across different *V. cholerae* strains (Fig. 3.7). This tree includes all *V. cholerae* isolates within my dataset with the exception of the 116 clinical isolates published recently from the Yemen cholera outbreak study (Weill et al. 2018), which all belong to the 7th pandemic El Tor lineage. The strains within the PG lineage, which include the 7th pandemic El Tor, El Tor progenitor, El Tor sister, Classical, and Classical sister groups (Chun et al. 2009, Boucher 2016), are closely related with little genetic variation. These lineages are therefore collapsed in the phylogenetic tree as the relationships between them are not well resolved. All sublineages formed monophyletic clades, although in some cases the most basal branch is of a different sublineage (e.g., *V. cholerae* strains T5 or 506315) creating paraphyletic clades. Ideally, each sublineage would correspond to exactly one full monophyletic clade. The reason this is not seen is likely the lack of sampling, leading to the grouping of relatively distantly related isolates together in the same clade. Further sampling will likely resolve these cases into two separate monophyletic clades. Out of 1,262 isolates, we identified 291 sublineages, and 19 of which belong exclusively to the PG lineage. Of the 292 sublineages, 223 are singletons. Based on the rarefaction curve, much like cgSTs, the total sublineage diversity of *V. cholerae* is far from being sampled (Fig. 3.1).

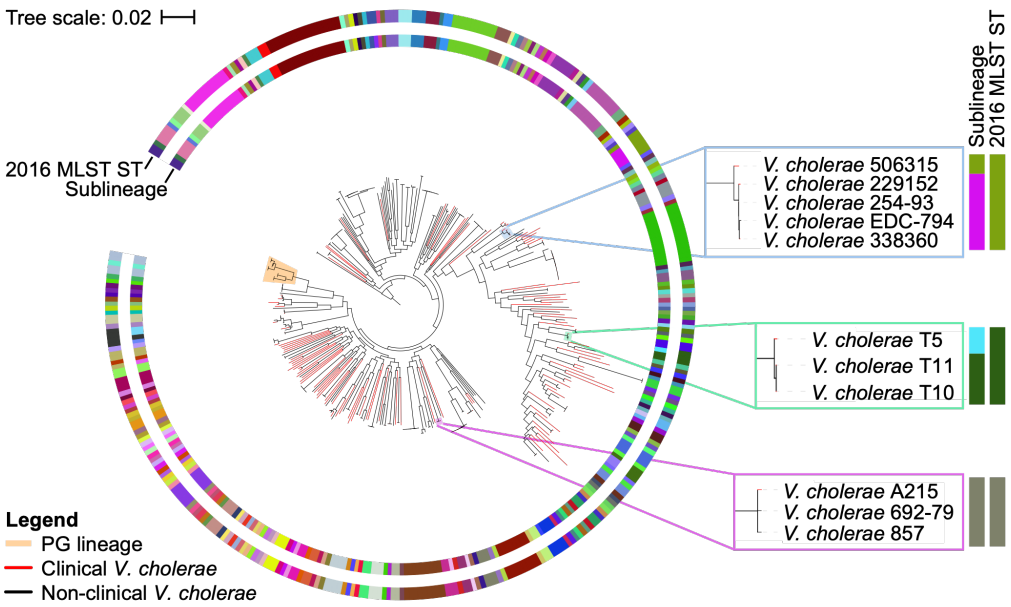


Fig 3.7: Phylogenetic tree of 1,148 *V. cholerae* isolates (excluding the 116 isolates from the Yemen outbreak study (Weill et al. 2018)) reconstructed using Parsnp v1.2 (Treangen et al. 2014). All group inside the PG lineage (7th pandemic El Tor, El Tor progenitor, El Tor sister, Classical, and Classical sister) are collapsed. Outer rings represents clustering by sequence type based on the 2016 MLST scheme by Kirchberger et al. (Kirchberger et al. 2016), whereas the inner ring represents clustering based on the sublineage threshold (i.e., 133 allelic differences). Branches of clinical strains are colored in red. The phylogenetic tree is rooted with a basal lineage to *V. cholerae* (collapsed) (Islam et al. 2018, Liang et al. 2019)

The sublineage concept has been applied to numerous pathogens and as such were each defined differently depending on the pathogen in question. Some have defined sublineages based on natural breaks in genetic similarities (Moura et al. 2016), while others may use sublineage to refer specifically to traditional MLST STs (Lucidarme et al. 2015) or even finer level of resolution below the MLST ST level based on whole genome analyses (Royer et al. 2019). There is, however, one unifying feature of all sublineage definitions - that they all refer to monophyletic clades. Sublineages are defined in this study based on natural breaks in allelic differences calculated from cgMLST profiles and were put into context by comparing with two traditional MLST schemes. I have shown that our definition of sublineage not only form monophyletic clades (Fig 3.7), but also corresponds to any traditional MLST ST designation. This sublineage definition will therefore play a crucial role in consolidating information from all previous MLST analyses.

3.4.3 A universal south Asian origin for modern cholera outbreaks

With the continual improvements of next-generation sequencing techniques, whole genome sequencing is expected to become a standard practice or even the first identification tool used in clinical and epidemiological studies. It is therefore critical to develop rapid typing scheme for genome sequence data that had the power to inform us about the relationship of a novel isolate with known strains. This is done here by defining what I term an 'outbreak threshold' based on the cgMLST scheme, which can identify outbreak related strains and potential sources of introduction. The outbreak threshold is expected to be less than 40 allelic differences as outbreak isolates are very closely related. There is a minor discontinuity at seven allelic differences where isolate pairs have either less than or more than this number of allelic

differences (Fig. 3.2B). Looking at the DI, the local maximum in the range of 0 to 50 occurs at seven allelic differences as well (Fig. 3.3) making this cutoff a likely candidate for an outbreak threshold. When applying the outbreak threshold to the full datasets containing all sequenced *V. cholerae* genomes meeting the minimum quality threshold, major clusters were examined to evaluate the ability of cgMLST in identifying strains that are part of the same outbreak.

One of the major outbreak clusters identified, with no prior information required, contains the Haiti and the Yemen outbreaks, which are the two best documented cholera outbreaks in modern history (Frerichs et al. 2012, Katz et al. 2013, Orata et al. 2014, Eppinger et al. 2014, Weill et al. 2018). Isolates collected from these outbreaks form a single cluster with Dominican Republic, Eurasian (India, Russia, Nepal, and Ukraine), and African (Tanzania, Kenya, and Somalia) isolates (Fig. 3.8A). The Dominican Republic isolates are closely related to the Haiti outbreak strains. Given the close proximity of the two countries, co-located on the island of Hispaniola, it was not surprising that isolates from Haiti would eventually spread to the Dominican Republic (Katz et al. 2013). The 7th pandemic El Tor lineage spread across the world from South Asia in three separate waves (Mutreja et al. 2011). The third wave, being the most recent distribution event, has been claimed to be responsible for the outbreaks in Haiti and Yemen (Weill et al. 2018). It is therefore not surprising to see Haiti and Yemen form a single cluster with India at its center. Nepal is the known source of introduction for the Haiti outbreak in 2010 (Frerichs et al. 2012), and comparisons with over 1,200 *V. cholerae* isolates from all over the world still show the Nepalese isolates are indeed the closest relatives to the Haitian isolates (Fig. 3.8A).

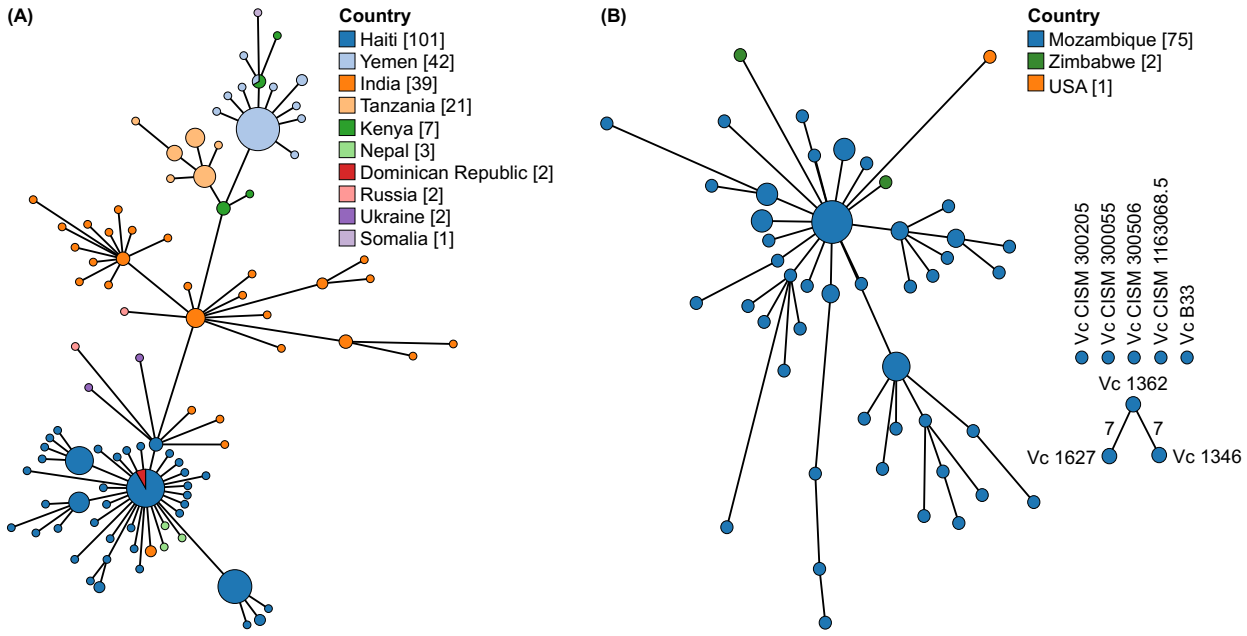


Fig 3.8: Minimum spanning trees isolated when the outbreak threshold (7 allelic differences) was applied to the complete dataset of 1,264 isolates. A) All isolates which clustered together with the isolates from Haiti (blue) and Yemen (light blue) based on the clustering threshold of seven allelic differences. B) All isolates clustered with the Mozambique isolates based on the clustering threshold of seven allelic differences. Additional Mozambique isolates that are not part of the same outbreak cluster are also shown. Three isolates, two from Zimbabwe (green) and one from the USA (orange), are connected as they share seven or fewer allelic differences with the Mozambique isolates. In both panels, the size of the nodes is proportional to the number of isolates. Length of the lines is proportional to the number of allelic differences and all connections have less than or equal to seven allelic differences.

Cholera is still endemic in Africa (Ali et al. 2015), and caused several major reported outbreaks in different countries over last few decades (World Health Organization 2017) including Mozambique (Cavailler et al. 2006, Garrine et al. 2017) and Zimbabwe (Sardar et al. 2013). Another major cluster groups most of the Mozambique isolates together with two Zimbabwe isolates (strains CP1038(11) and 2011EL-1137) and one USA isolate (2009V-1116) (Fig. 3.8B). Based on cgMLST analysis, it is evident that these two Zimbabwe isolates are closely related to the Mozambique isolates differing at four or less alleles. The close proximity of the two countries suggests these are likely travel-associated cases. Although outbreaks involving Mozambique isolates (Garrine et al. 2017), and the Zimbabwe isolates ((Reimer et al. 2011, Hasan et al. 2012)) have been independently studied, the link between these isolates have not been shown before. Global cgMLST analysis is therefore an invaluable tool as it allows for the identification of links between independent studies. However, with only two Zimbabwe isolates in the dataset, additional sampling in this region is required to understand the epidemiology of this outbreak. According to the NCBI BioSample database, strain 2009V-1116 was collected by the Centers for Disease Control and Prevention in 2009 and is associated with travel to Pakistan. Since the 7th pandemic El Tor lineage has been circulating in Asian and Middle Eastern countries for a long time (Hu et al. 2016), it is possible that, at least within our dataset, the Mozambique isolates are the closest relative to this specific Pakistan strain.

3.4.4 Confirmation of an African connection for the Yemen outbreak

The Yemen cholera outbreak began in October 2016 with 11 confirmed cases (http://www.emro.who.int/pandemic-epidemic-diseases/cholera/cholera-cases-in-yemen.html). By January 2017, there were already over 10,000 cholera cases with 99 associated deaths

(http://www.emro.who.int/pandemic-epidemic-diseases/cholera/weekly-update-cholera-cases-in-yemen-15-jan-2017.html). By the end of that year, there were over 900,000 cholera cases (http://www.emro.who.int/pandemic-epidemic-diseases/cholera/outbreak-update-cholera-in-yemen-19-december-2017.html). The Yemen cholera outbreak continues on today as the largest cholera outbreak in modern history. As isolates from this outbreak were only recently made available (Weill et al. 2018), they were not part of the initial dataset for the cgMLST scheme development. These isolates were added and analyzed on PubMLST after the scheme had been established. This set of isolates therefore serves as an independent test of the universality and applicability of the cgMLST scheme. To determine the potential origin of the Yemen outbreak and its phylogenetic relationships with existing *V. cholerae* strains, the Yemen isolates were compared with other 7th pandemic El Tor isolates from Asian and African countries (Table B1). All allele designations and cgST assignments were done automatically on PubMLST. MST was built using these isolates and all connections with seven and fewer allelic differences are represented as solid lines (Fig. 3.9). Isolates connected by solid lines therefore belong in the same outbreak cluster as defined by the outbreak threshold of seven allelic differences. Isolates from Yemen, Kenya, and Haiti all cluster with the central Indian isolates with seven or fewer allelic differences; however, the closest relatives to the Yemen isolates are those from Kenya with four or fewer allelic differences (Fig. 3.9). The Indian isolates are the next closest connection but there is no direct linkage between these and the Yemen isolate. This pattern is consistent with the work of Weill and colleagues (Weill et al. 2018), where they suggested that the Yemen outbreak strains may have come from East Africa which itself came from South Asia based on SNP-base phylogenetic analysis and Bayesian evolutionary analysis (Weill et al. 2018)

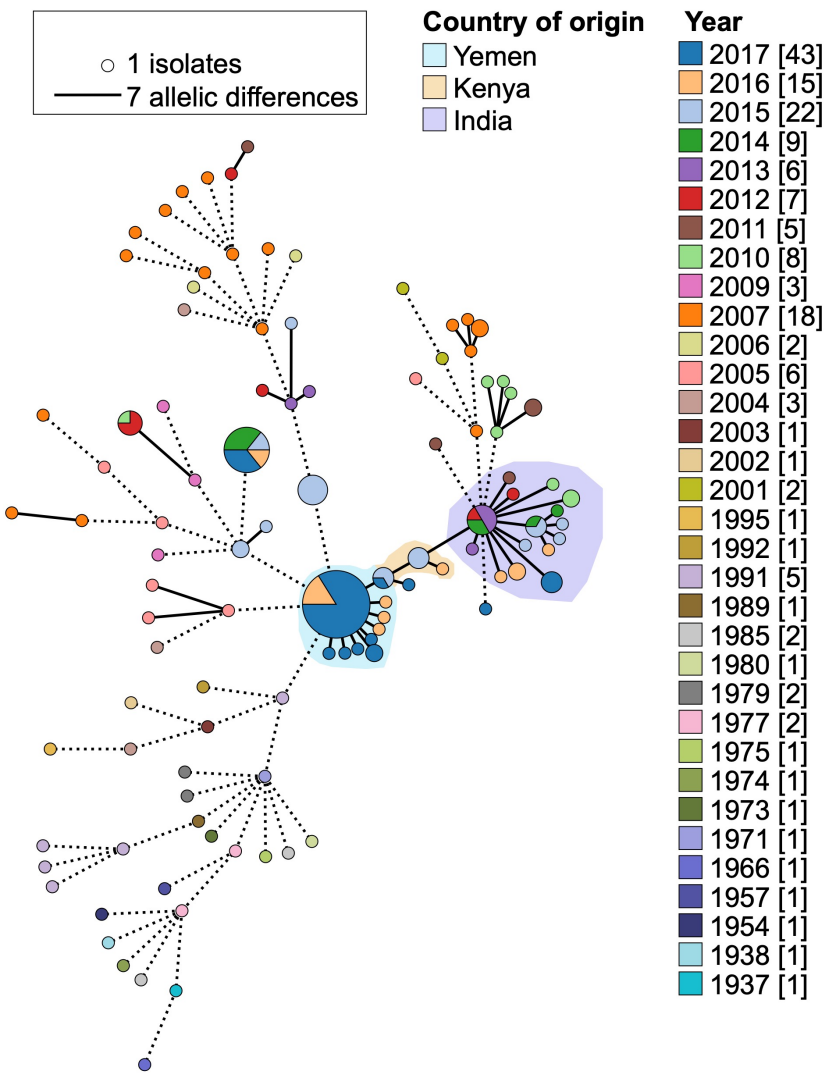


Fig 3.9: cgMLST Minimum Spanning Tree of all Yemen isolates and representative 7th pandemic El Tor strains (Table B1). All isolates connected by dotted lines share eight or more allelic differences (not drawn to scale). All isolates connected with solid lines share seven or fewer allelic differences (i.e., they belong to the same outbreak cluster; drawn to scale). Each node represents a cgST that is colored by year of collection. The outbreak clusters are shaded by country.

Unlike the limited samples available from African cholera outbreaks, the Haiti and Yemen outbreaks are significant cases for epidemiological investigations because *V. cholerae* has been heavily sampled from these countries as well as surrounding regions. Two major limitations in genomic epidemiology have been the lack of a universal classification scheme and a comprehensive database; however, this is no longer the case in the genomic era as sequencing technology is becoming increasingly more accessible (Orata et al. 2014). A genomic approach, as shown here, is able to produce accurate predictions of potential origins of outbreaks and provides us with sufficient resolution to accurately track the spread of the disease. Therefore, genomic analysis should be the first step in any epidemiological study as not only will it help guide subsequent analyses and investigations, but consistently sequencing new genomes will also help expand and refine the current global *V. cholerae* genome database.

3.4.5 Increased resolution for the history of cholera in Mozambique: comparing cgMLST to MLVA

The 7th pandemic reached Africa in 1970 and cholera appeared in Mozambique at roughly the same time (Weill et al. 2017). Since its introduction, cholera has been endemic in that country and has continued to cause multiple outbreaks (Garrine et al. 2017). A popular tool for outbreak investigation is MLVA (Danin-Poleg et al. 2007, Lam et al. 2012), which was recently used to study *V. cholerae* strains collected in Mozambique over multiple years (Garrine et al. 2017). MLVA is a subspecies typing method similar to MLST in concept; however, it utilizes variable number tandem repeats (VNTR) instead of gene sequences. As VNTR mutates at a faster rate than conserved genes, it has been shown that MLVA provides greater resolution than MLST for some species (Lam et al. 2012, Chenal-Francisque et al. 2013). To establish a

direct comparison between my cgMLST scheme and this MLVA scheme, I examined the MSTs created by both methods focusing on only shared isolates (Fig. 3.10). The MLVA identified 26 profiles forming two clonal complexes and four singletons (Fig. 3.10A) (Garrine et al. 2017). A similar population structure is seen with the cgMLST analysis (Fig. 3.10B), including the four singletons identified in the MLVA. The central node in the cgMLST MST consists mostly of isolates with MLVA profile '8,4,6,18,21' similar to the central node in the MLVA MST (Garrine et al. 2017). The two clonal complexes (CC) identified in the MLVA MST are also identified in cgMLST MST with the smaller CC2 being at least four allelic differences away from the larger CC1.

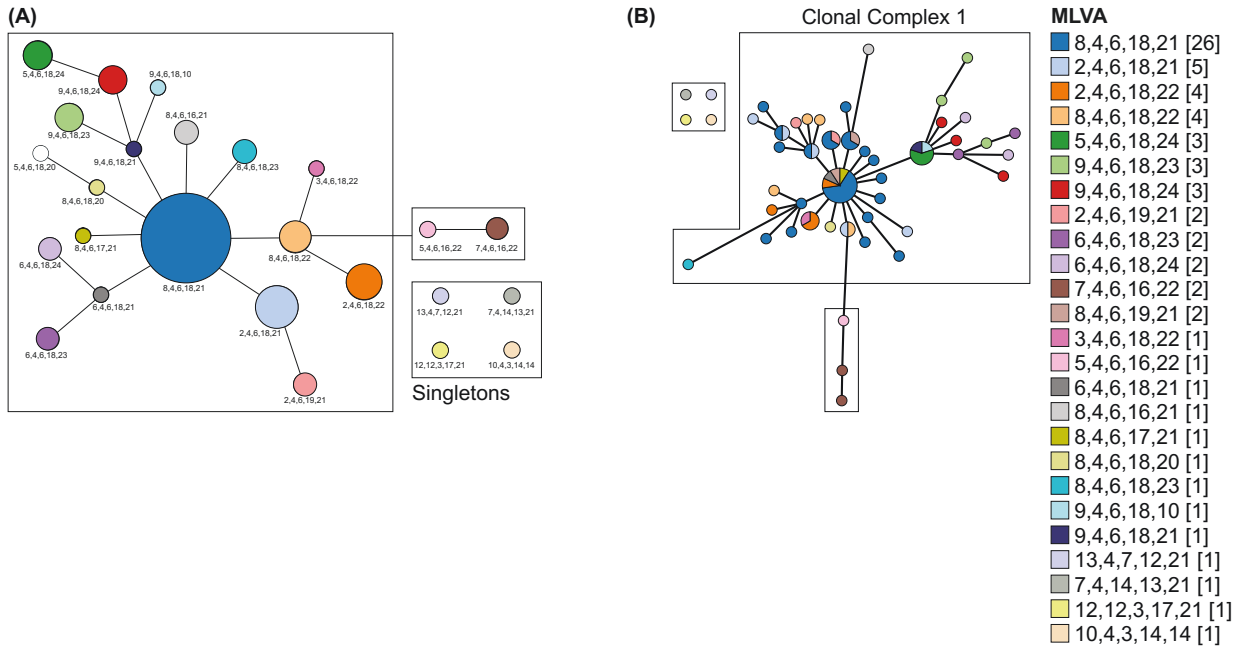


Fig 3.10: Comparing MLVA with cgMLST analysis focusing on only Mozambique isolates. A) Minimum spanning tree (MST) based on MLVA adapted from a previous study (Garrine et al. 2017). Each node represents a MLVA type and is assigned a unique color. B) MST based on cgMLST analysis of isolates used in the previous study. Each node represents a unique cgST but colored based on MLVA types.

Although there are a few MLVA types that were grouped into a single cgST, such as cgST1 and gST114, indicating cgMLST was unable to resolve the differences in these MLVA types, there are many MLVA types such as profile '2,4,6,18,21', profile '7,4,6,16,22', profile '9,4,6,18,24', and profile '8,4,6,18,22', that were split into multiple cgSTs. Overall there are 26 MLVA types as opposed to 48 cgST types, showing that the latter provides better resolution overall than the former. The cgMLST analysis overlaid with isolation dates shows that in Mozambique, *V. cholerae* strains are highly clonal and strains from the same cgST can cause outbreaks over multiple years (e.g. cgST114 and cgST94) (Fig. 3.11), which corroborates the claim made in the initial MLVA study (Garrine et al. 2017), where the same MLVA type can be seen over multiple years.

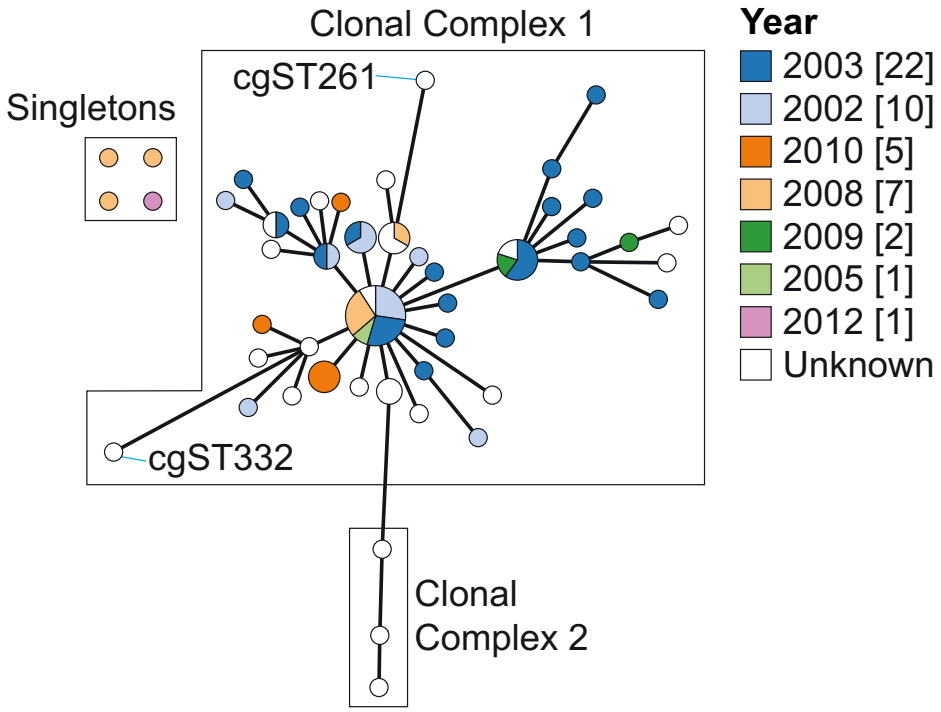


Fig 3.11: cgMLST analysis of Mozambique outbreaks isolates. Each node represents a unique cgST. Size of nodes and length of connecting lines are proportional to the number of isolates and allelic difference respectively. Each node is colored based on year of isolation.

In addition to increased resolution, cgMLST also produces more reliable and reproducible results than MLVA, as it eliminates errors associated with the detection of VNTR regions using PCR or sequencing-based methods. For the same reason that MLST is less affected by convergent evolution compared to MLVA (Struelens and Brisse 2013), cgMLST is also less affected by convergent evolution.

3.4.6 Standardizing the genotypes responsible for the Haiti 2010 cholera outbreak: comparing cgMLST and SNP-based analyses

One of the largest cholera outbreaks in modern history occurred in Haiti following the devastating earthquake in 2010 (Orata et al. 2014, Guillaume et al. 2018). Prior to this outbreak, there were no documented cholera cases in Haiti (Katz et al. 2013, Boucher 2016). Since the initial introduction, *V. cholerae* now remains endemic in Haiti and is responsible for thousands of cholera cases annually (Guillaume et al. 2018). Multiple studies have strongly suggested that the Haitian strains are in fact imported from Nepal (by the UN Nepalese troops) and the outbreak occurred as a result of both inappropriate sanitary practice and the lack of screening of UN troops upon their arrival in Haiti (Frerichs et al. 2012, Orata et al. 2014, Frerichs 2016, Guillaume et al. 2018).

A SNP-based approach was used to study the evolutionary dynamics of *V. cholerae* in Haiti (Katz et al. 2013). This technique relies on the identification of SNPs in draft or closed genomes. The primary benefit of this method is that assembly and annotation are not required. It is also capable of resolving closely related strains using whole-genome data. However, SNP-

based methods are highly influenced by recombination events and quality filter parameters chosen (Pightling et al. 2014).

To establish a direct comparison between the cgMLST scheme and SNP-based analysis, we focused on MSTs of only Haitian outbreak isolates (Fig. 3.12). All Haitian isolates are closely related according to the cgMLST scheme, sharing at most four allelic differences with each other (Fig. 3.12A). The Haitian and Nepalese isolates, therefore, also belong to the same sublineage (SL6) which is consistent with the fact that these isolates belong to the same MLST ST (either ST1 or ST69 based on the 2016 and 2013 MLST scheme respectively) (Octavia et al. 2013, Kirchberger et al. 2016) (Table B2). The overall population structure is similar between the two methods where we have SNP ST1 as the center of the MST and ST2 and ST3 extending from that likely ancestral genotype (Fig. 3.12). SNP ST1, ST2, and ST3 can be split into 11, 2, and 3 different cgSTs, respectively (Fig. 3.12A). There is only one case, cgST66, where it contains isolates from SNP ST1 and ST3. Overall, cgMLST was able to differentiate 39% of the isolates while SNP-based analyses can differentiate 35%, showing comparable level of resolution. As expected, both the cgMLST and the SNP-based analyses showed that the Haiti outbreak is highly clonal where most isolates belong to the same cgST or SNP ST (Katz et al. 2013). However, one important advantage of cgMLST over SNP-based analysis is that, the former can be easily standardized because it relies on a predefined set of core genes. Based on these standardized genes, we can establish a systematic nomenclature system. This makes cgMLST more suitable than SNP-based method as a universally applicable classification system for epidemiological studies and research worldwide.

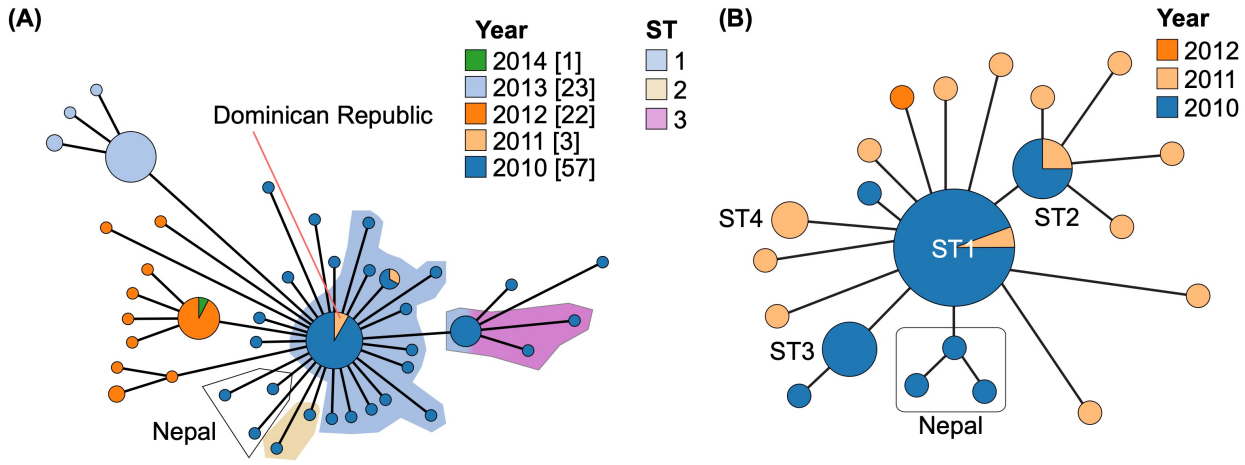


Fig 3.12: Comparison between cgMLST and SNP based analysis focusing on Haiti outbreak related strains. A) Minimum spanning tree (MST) constructed based on cgMLST analysis. Each node represents a unique cgST and is colored by year of isolation. All isolates not from Haiti are labeled. Background shading represents SNP ST. B) MST constructed based on SNP analysis. Each node is colored based on year of isolation and SNP ST and isolates not from Haiti are labeled. In all panels, size of nodes and length of connecting lines are proportional to number of isolates and allelic differences respectively.

3.4.7 Environmental isolates differ from clinical strains by their diversity and their associations with specific geographical locations

To look at the geographic signal of *V. cholerae*, we eliminated all clinical isolates and those that belong to the PG lineages (Chun et al. 2009, Boucher 2016). This is because the geographic signal of clinical strains can be skewed, as pathogenic strains can travel long distances in a short period of time through association with human hosts. The geographical analysis was therefore performed only with environmental isolates.

Along with all the publicly available environmental strains that are not part of the PG lineages, there are a total of 195 isolates spanning 9 countries. After grouping isolate at the sublineage level (i.e., each cluster have at most 133 allelic differences), it could be noted that all isolates from the same sublineage also shared a country of origin, with the exception of strains 692-79 and 857 (Fig. 3.13), which are from the USA and Bangladesh, respectively. Phylogenetic analysis shows these isolates to be closely related to strain A215, a clinical isolate from the USA (Fig. 3.7). All three strains contain the *toxR* gene, a toxin transcriptional regulator common in pathogenic *V. cholerae* (Childers and Klose 2007), and genes encoding for the Mannose-sensitive hemagglutinin pilus, the RTX toxin, and hemolysin (*hlyA*), all of which are putative virulence factors for *V. cholerae*. In addition, strains A215 and 857 also have the zona occludens toxin gene. Similar toxin gene contents among these three isolates and close phylogenetic relationships suggest that strains 692-79 and 857 may also be pathogenic or at least associated with a pathogenic strain and are capable of surviving inside a human host. This provides evidence that although clinical isolates can spread across the world rapidly and closely related isolates can be from very different parts of the world, environmental isolates from the same geographic origin share an affinity among each other at least at the sublineage level. It is

important to note that my dataset contains a relatively small number of environmental isolates that are not part of the PG lineages. Therefore, this distinct distribution pattern based on geographic origin may be a result of currently insufficient sampling of environmental *V. cholerae* worldwide. With large-scale environmental sampling, it will be possible to determine with greater accuracy the evolutionary rate and distribution pattern of *V. cholerae* in the environment using cgMLST. In addition, this method will become an invaluable tool in dealing with these big datasets as it provides an efficient and standardized method of classification.

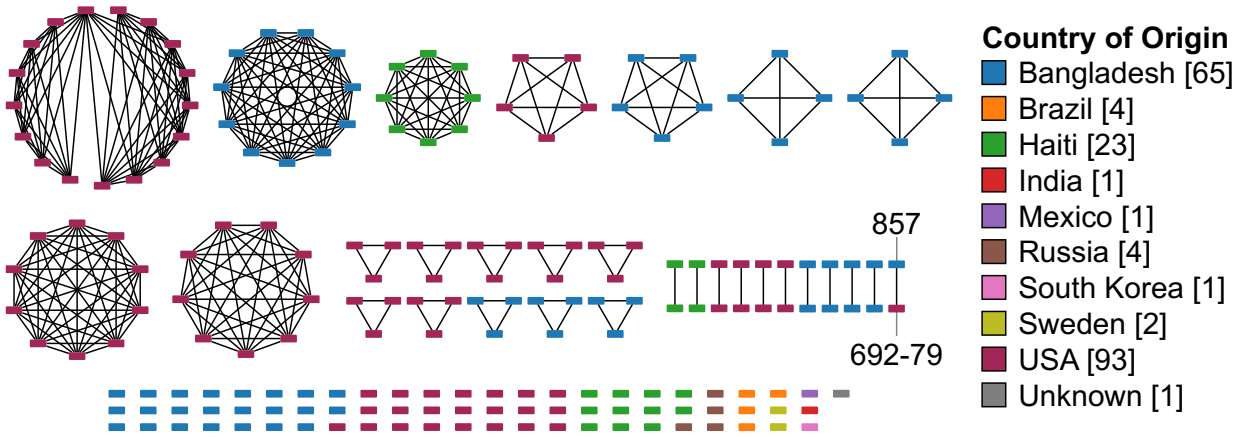


Fig 3.13: Sublineage clusters of non-clinical environmental isolates that are not part of the PG lineages. Each node represents an isolate and are colored by country of isolation.

3.5 Conclusion

With an extensive collection of over 1,200 *V. cholerae* isolates, I developed a cgMLST scheme based on 2,443 core genes. I established a sublineage-level definition based on 133 allelic differences as part of our standardized classification scheme. It was determined by comparisons with previous MLST schemes that the cgMLST sublineage classification can be used as a proxy for traditional MLST. Additionally, the universality and applicability of the scheme have been tested by looking at various cholera outbreak cases. I also determined an outbreak threshold based on seven allelic differences that groups outbreak isolates together with strains from potential source of introduction. This threshold creates clusters that are consistent with known epidemiological data when applied to the Haiti and Yemen cholera outbreaks, two of the best-documented cholera outbreaks in modern history. Also, I was able to confirm the South Asian origin of modern cholera outbreaks. Furthermore, although sampling is limited, a geographic signal at the sublineage leel not seen in clinical strains could be identified among environmental isolates that are not part of the PG lineage (Chun et al. 2009, Boucher 2016). Lastly, this scheme is fully implemented on PubMLST (https://pubmlst.org/vcholerae/) for public access. All newly available genomes uploaded to PubMLST will be annotated automatically and a cgST designation will be assigned to isolates with less than 100 missing loci. Relevant epidemiological data and the variety of analytical and visualization tools are all integrated on PubMLST, allowing for a quick analysis of any newly sequenced genome in a global context. This scheme will be an important tool for future large-scale epidemiological and biogeographical research.

3.6 Acknowledgements

I acknowledge the helpful feedback provided by Dr. Keith Jolley (University of Oxford) regarding the development of the cgMLST scheme, as well as the implementation of this scheme on PubMLST. I also thank Monica Im (Centers for Disease Control and Prevention) for assistance with obtaining whole genome sequences. MA acknowledges the governments of Bangladesh, Canada, Sweden, and United Kingdom for providing core/unrestricted support.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (to YFB); the Integrated Microbial Biodiversity program of the Canadian Institute for Advanced Research (to YFB); federal appropriations to the Centers for Disease Control and Prevention through the Advanced Molecular Detection Initiative (to CLT); the graduate student scholarships from Alberta Innovates – Technology Futures (to KYHL, FDO, MTI and TN); the NSERC Canada Graduate Scholarship – Master's Program (to KYHL); the University of Alberta Faculty of Graduate Studies and Research (Queen Elizabeth II Graduate Scholarship to KYHL and TN), and the Bank of Montréal Financial Group (to FDO). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Chapter 4: Concluding remarks – Moving forward: polyphasic taxonomy in the genomic era

Chapter 4

4.1 The future of polyphasic taxonomy

4.1.1 Application of genomic metrics in bacterial taxonomy

16S rRNA gene sequence analysis has played a significant role in bacterial taxonomy, especially in the *Rhodobacteraceae* and *Roseobacteraceae* families as the cultivation of members within these families corresponded to the raising popularity of the use of 16S rRNA gene for taxonomic classifications (Buchan et al. 2005). However, I have shown that the 16S rRNA gene does not have sufficient resolution to confidently resolve family level relationships within the order *Rhodobacterales*, let alone genus and species level (section 2.2, 2.3.1). In addition, it is now widely known that organisms may contain multiple intragenomic copies of 16S rDNA with sequences different enough to be considered as different genera (Klappenbach et al. 2001, Acinas et al. 2004, Boucher et al. 2004, Case et al. 2007). This fact, combined with the small size and the low mutation rate of the 16S rRNA gene, means it cannot be used as a reliable proxy for genomic similarities. Taxonomic inconsistencies, therefore, has been a reoccurring issue within this family since its conception in 2005 (Garrity et al. 2005).

Rather than relying on indirect methods, such as 16S rRNA gene sequence identity, to assess genomic similarities, it is now possible to directly measure it using WGS. Not only are WGS-based methods more accurate and reproducible, they can also easily be scaled to analyze hundreds of isolates efficiently. Popular WGS-based methods include dDDH, ANI, and AAI, which have already provided tangible taxonomic classification standards, more so for the species level than higher ranks. 70% dDDH and 95% ANI are currently regularly used for species level delineations (Baek et al. 2015, Orata et al. 2018, Rabus et al. 2019). Although there are yet to be universally accepted standards for taxonomic classifications at the genus rank and above, it can

nonetheless be used to identify potential misclassifications. With the abundance of WGS and the numerous readily available bioinformatics tools to assess genomic relatedness, it is reasonable to expect that WGS-based methods will become standard practices in the field of taxonomy.

Using this genomic approach, it was possible to identify species, genus and family level misclassifications within the *Rhodobacteraceae* and *Roseobacteraceae* families (chapter 2). Out of 53 monophyletic genera consisting of 265 species, I identified three genera which contained species level misclassifications based on the accepted species threshold of 70% and 95% for dDDH and ANI respectively (Goris et al. 2007, Meier-Kolthoff et al. 2013) (section 2.3.2). At the genus level, AAI, 1st, 2nd and 3rd codon position similarities also showed that the genus *Halocynthiibacter* should be split into two genera; a conclusion supported by both phylogenetic and phenotypic data (section 2.3.3). At the family level, these genomic metrics also showed that the *Roseobacteraceae* family is genomically distinct from the *Rhodobacteraceae* family (section 2.3.7.1). This exemplifies the importance of genomic similarity analyses in establishing stable taxonomic classifications as it is the only high-throughput method that can be used for large-scale analyses to ensure taxonomic consistency in all taxonomic levels.

4.1.2 Changes in phenotypic characterizations

Traditional bacterial classifications require extensive phenotypic characterization; however, it is not realistic to expect this approach to keep up with the rate at which genomes are sequenced. This problem becomes apparent when looking at the large discrepancy in the number of described species between NCBI and the Bergey's manual of Systematics of Archaea and Bacteria. The family *Rhodobacteraceae* for example have only 30 described genera in the 2015

publication of the Bergey's manual (Garrity et al. 2015b); however, as of November 22nd, 2019 there were 153 genera with genomes available.

Although phenotypic testing is a costly and time-consuming process, it is an irreplaceable part of taxonomic classifications as accurate phenotypic descriptions are required to understand the complex ecological and environmental roles bacteria play in the environments. Furthermore, such descriptions highlighting relevant virulence and antibiotic resistance traits are also crucial for clinical and epidemiological practices. To keep up with genomic progress, it is expected that phenotypic analyses will require genomic guidance in the future (Rosselló-Móra and Amann 2015).

Through a meta-analysis of adaptive traits in the novel *Roseobacteraceae* family, I was able to identify a number of important phenotypes (e.g., quorum sensing, carbon monoxide oxidation, and sulfur metabolism) and marker genes ancestral to this lineage. A comprehensive genomic approach then allowed me to analyze and predict presence/absence of these phenotypic traits for hundreds of isolates, effectively extending these comprehensive phenotypic experiments of a few selected strains to hundreds of isolates (chapter 2.3.7.2). Although *in-silico* phenotypic analyses are ultimately predictions that must be verified experimentally, it is impossible to exhaustively test all possible phenotypes for all isolates. However, through this approach, I was able to reduce the number of phenotypes that should be tested by identifying traits that are likely able to differentiate between the groups in question. As bioinformatics tools mature, together with the ever-expanding database of WGS, *in-silico* phenotypic analyses will likely provide more accurate and comprehensive phenotypic predictions in the future not unlike how dDDH is currently able to accurately predict traditional DDH experiment results.

Three pathways that are characteristics of the *Roseobacteraceae* family were identified and they are DMSP demethylation, DMSP cleavage and quorum sensing. These pathways were significantly more prominent in the *Roseobacteraceae* fam. nov than in the *Rhodobacteraceae* family, which is supported by current understanding of their ecology (section 2.3.7.2). DMSP for example, commonly produced by phytoplanktons, has long been known to be an important chemoattractant for marine bacteria (Moran et al. 2012) (section 2.3.7.2.1). Similarly, acyl-homoserine lactone-based quorum sensing have also been shown to be a common feature among members of the *Roseobacteraceae* family (Cude and Buchan 2013, Zan et al. 2014) (section 2.3.7.2.2).

4.1.3 Importance of genome-scale phylogenetic analysis in the aim to establish stable taxonomic classifications

Phylogeny is one of the three key factors of polyphasic taxonomy. It is crucial for the development of a stable set of classification standards as monophyly is one of the few, if not only, rule that can be universally applied to all levels of classification (Rosselló-Móra and Amann 2015, Parks et al. 2018). It is expected that WGS-based phylogenetic analyses will play an increasingly bigger role in taxonomic classification as it can ensure that only organisms with a shared evolutionary history are grouped together avoiding potential misclassifications due to convergent evolution. Through a core-genome phylogenetic analysis, I was able to identify seven paraphyletic and 17 polyphyletic genera that must be addressed (section 2.3.4). It also highlighted two major lineages that are likely different families which later genomic and *in-silico* phenotypic analysis confirmed (section 2.3.7.1, 2.3.7.2).

Although in my studies, I looked no further than the family level, there are numerous studies highlighting the use of genomic similarity analyses in higher taxonomic classifications ranging from the order to the phylum level (Sen et al. 2014, Waite et al. 2017).

4.1.3 Genomic and phylogenetic analyses as a guide to ensure stable taxonomic classifications

Through genomic and phylogenetic analyses, I was able to identify and resolve numerous taxonomic inconsistencies among type strains, which should help guide future taxonomic classifications. In addition to resolving current taxonomic misclassifications, it is equally important to establish a workflow to ensure future taxonomic classifications are stable. As it is impossible to reconstruct core-genome phylogenetic trees with hundreds, if not thousands, of type strains each time a new genome is added, there needs to be a way to quickly and accurately identify close relatives of unknown isolates among known type strains. One way this can be achieved is through a genomic approach, which as previously mentioned is suitable for high-throughput analysis. By calculating pairwise AAI values between newly added genomes with all known type strains, I was able to quickly identify the closest relatives of the new genomes (section 2.3.6). Using this much smaller dataset of just the novel genome and a few close relatives, I was able to perform in-depth phylogenetic analysis. This can then serve as a starting point to guide subsequent phenotypic and additional genomic analyses. Through this approach, I confirmed the taxonomic classifications of two type strains with genomes recently made available and identified a genus level misclassification (section 2.3.6).

Aside from using genomic similarity analyses, there are also online tools available that can help identify closely related strains to guide subsequent analyses. One such tool is GTDB-tk,

which places any bacterial and archaeal genomes into a reference tree based on 120 core genes (Chaumeil et al. 2019). This allows researchers to identify, with confidence, the general phylogenetic placements of their genomes. From that, a shortlist of genomes can be selected for more in-depth analyses.

4.2 Future of the *Vibrio cholerae* cgMLST scheme

The current cgMLST scheme for *V. cholerae* is based on a collection of 2,443 core genes, Based on this scheme, I have also established an outbreak threshold capable of identifying outbreak related strains and potential sources of introductions. Using this threshold, I was also able to confirm the Nepalese and East African origin of the Haiti and Yemen outbreak respectively (Orata et al. 2014, Weill et al. 2017) (section 3.4.4 and 3.4.6). In addition, I also proposed a sublineage threshold based on 133 allelic differences that creates clusters similar to any traditional MLST ST which will help consolidate information from existing studies. Applying this threshold, I have identified a strong geographic signal among environmental isolates not seen in clinical strains (section 3.4.7). This pattern is consistent with our understanding that clinical *V. cholerae* strains can spread through asymptomatic carriers (Ackers et al. 1997).

Currently, PubMLST only hosts the 2013 MLST scheme (Octavia et al. 2013), but not the 2016 MLST scheme (Kirchberger et al. 2016). In addition, we have previously identified a novel gene marker (*viuB*) suitable for subspecies level typing of *Vibrio cholerae* (Kirchberger 2017). Future progress will incorporate other molecular typing methods outside of cgMLST and MLST schemes, such as *viuB* and 16S rRNA gene sequence on the same platform. This will

allow researchers to easily analyze their isolates in a global context with all commonly used sequence-based typing methods.

To encourage the use of cgMLST, there are two limitations that should be addressed. The first limitation is that cgMLST only works with assembled genomes, which requires users to have at least an introductory knowledge in bioinformatics and genome assemblies. The second limitation is that genomes uploaded to NCBI must be uploaded separately to PubMLST and therefore, requires periodic update to ensure the cgMLST scheme is current.

To address the first concern, it would be ideal to integrate assembly tools, such as skesa (Souvorov et al. 2018), to enable automatic assembly of NGS reads. This will reduce the technical challenge of utilizing this scheme allowing users to upload draft/complete genomes or NGS reads. The second limitation can be addressed by developing back-end tools that will periodically retrieve newly uploaded genomes and NGS reads from NCBI then perform the proper quality filter and incorporate them into the scheme, which will ensure the cgMLST scheme is always up-to-date.

One of the challenges in studying *V. cholerae* is that many local outbreaks are not well documented and/or reported due to the potential of negative impact on tourism as it implies poor water quality (Ali et al. 2015). As cgMLST relies on a comprehensive database to identify global and local patterns, this makes it difficult to make detailed predictions of the distribution patterns of localized outbreaks and track the spread of *V. cholerae*. Our collaboration with the CDC and ICDDR,B will hopefully encourage the use of the cgMLST analysis as standard protocols for epidemiological studies around the world, as well as the sequencing of clinical and environmental isolates. With sufficient global sampling, it is our goal to produce detailed

epidemiological data of cholera outbreaks around the world, similar to what was previously done with the Yemen outbreak (Weill et al. 2017), with significantly fewer steps and lower technical threshold to complete these analyses.

The ultimate goal is to develop an easy-to-use online tool that is not only rooted in a comprehensive database, but will also allow users to easily visualize genomic and phylogenetic relationships with biogeographical (e.g., origin of introduction) and epidemiological data (e.g., virulent genes, antibiotic resistances, etc) through the integration of other readily available bioinformatics tools such as GrapeTree (Zhou et al. 2018), Phyloviz (Ribeiro-Gonçalves et al. 2016) and GenGis (Parks et al. 2009). This will allow epidemiologists to quickly analyze new *V. cholerae* isolates in a global context and predict important clinically relevant traits. Such a tool will become an invaluable resource for epidemiological research worldwide.

References

References

Abascal, F., Zardoya, R., and Telford, M.J. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. **38**(suppl_2): W7–W13. doi:10.1093/nar/gkq291.

Abraham, W., and Rohde, M. 2019. *Hyphomonadaceae*. *In* Bergey's Manual of Systematics of Archaea and Bacteria. Wiley. pp. 1–10. doi:10.1002/9781118960608.fbm00349.

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. J. Bacteriol. **186**(9): 2629–2635. Available from https://www.ncbi.nlm.nih.gov/pubmed/15090503.

Ackers, M., Pagaduan, R., Hart, G., Greene, K.D., Abbott, S., Mintz, E., and Tauxe, R. V. 1997. Cholera and sliced fruit: Probable secondary transmission from an asymptomatic carrier in the United States. Int. J. Infect. Dis. **1**(4): 212–214. doi:10.1016/S1201-9712(97)90039-4.

Alavi, P., Starcher, M.R., Thallinger, G.G., Zachow, C., Müller, H., and Berg, G. 2014. *Stenotrophomonas* comparative genomics reveals genes and functions that differentiate beneficial and pathogenic bacteria. BMC Genomics **15**(1): 482. doi:10.1186/1471-2164-15-482.

Ali, M., Nelson, A.R., Lopez, A.L., and Sack, D.A. 2015. Updated global burden of cholera in endemic countries. PLoS Negl. Trop. Dis. **9**(6): e0003832. doi:10.1371/journal.pntd.0003832.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment

search tool. J. Mol. Biol. **215**(3): 403–410. doi:10.1016/S0022-2836(05)80360-2.

Antwerpen, M.H.H., Prior, K., Mellmann, A., Höppner, S., Splettstoesser, W.D.D., and Harmsen, D. 2015. Rapid high resolution genotyping of *Francisella tularensis* by whole genome sequence comparison of annotated genes ("MLST+"). PLoS One **10**(4): e0123298. doi:10.1371/journal.pone.0123298.

Arunasri, K., Venkata Ramana, V., Sproer, C., Sasikala, C., and Ramana, C. V. 2008. *Rhodobacter megalophilus* sp. nov., a phototroph from the Indian Himalayas possessing a wide temperature range for growth. Int. J. Syst. Evol. Microbiol. **58**(8): 1792–1796. doi:10.1099/ijs.0.65642-0.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O., Formsma, K., Kubal, M., Vonstein, V., Stevens, R., McNeil, L.K., Edwards, R.A., Pusch, G.D., Reich, C., Glass, E.M., Olsen, G.J., Paczian, T., Overbeek, R.A., Meyer, F., Vassieva, O., DeJongh, M., Osterman, A.L., Disz, T., Best, A.A., Gerdes, S., Parrello, B., Bartels, D., Olson, R., and Paarmann, D. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**(1): 75. doi:10.1186/1471-2164-9-75.

Baek, K., Lee, Y.M., Shin, S.C., Hwang, K., Hwang, C.Y., Hong, S.G., and Lee, H.K. 2015. *Halocynthiibacter arcticus* sp. nov., isolated from Arctic marine sediment. Int. J. Syst. Evol. Microbiol. **65**(11): 3861–3865. doi:10.1099/ijsem.0.000507.

Baldo, L., and Werren, J.H. 2007. Revisiting *Wolbachia* supergroup typing based on WSP: Spurious lineages and discordance with MLST. Curr. Microbiol. **55**(1): 81–87. doi:10.1007/s00284-007-0055-8.

Bartels, M.D., Petersen, A., Worning, P., Nielsen, J.B., Larner-Svensson, H., Johansen, H.K., Andersen, L.P., Jarløv, J.O., Boye, K., Larsen, A.R., and Westh, H. 2014. Comparing whole-genome sequencing with sanger sequencing for spa typing of methicillin-resistant *staphylococcus aureus*. J. Clin. Microbiol. **52**(12): 4305–4308. doi:10.1128/JCM.01979-14.

de Been, M., Pinholt, M., Top, J., Bletz, S., Mellmann, A., van Schaik, W., Brouwer, E., Rogers, M., Kraat, Y., Bonten, M., Corander, J., Westh, H., Harmsen, D., and Willems, R.J.L. 2015. Core Genome Multilocus Sequence Typing Scheme for High-Resolution Typing of *Enterococcus faecium*. J. Clin. Microbiol. **53**(12): 3788–3797. doi:10.1128/JCM.01946-15.

Bletz, S., Janezic, S., Harmsen, D., Rupnik, M., and Mellmann, A. 2018. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Genome-Wide Typing of *Clostridium difficile*. J. Clin. Microbiol. **56**(6): e01987-17. doi:10.1128/JCM.01987-17.

Boonsilp, S., Thaipadungpanit, J., Amornchai, P., Wuthiekanun, V., Bailey, M.S., Holden, M.T.G., Zhang, C., Jiang, X., Koizumi, N., Taylor, K., Galloway, R., Hoffmaster, A.R., Craig, S., Smythe, L.D., Hartskeerl, R.A., Day, N.P., Chantratita, N., Feil, E.J., Aanensen, D.M., Spratt, B.G., and Peacock, S.J. 2013. A Single Multilocus Sequence Typing (MLST) Scheme for Seven Pathogenic *Leptospira* Species. PLoS Negl. Trop. Dis. **7**(1): e1954. doi:10.1371/journal.pntd.0001954.

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): 63–68.

Boucher, Y. 2016. Sustained Local Diversity of *Vibrio cholerae* O1 Biotypes in a Previously Cholera-Free Country. MBio **7**(3): e00570-16. doi:10.1128/mBio.00570-16.Copyright.

Boucher, Y., Douady, C.J., Sharma, A.K., Kamekura, M., and Doolittle, W.F. 2004. Intragenomic Heterogeneity and Intergenomic Recombination among Haloarchaeal rRNA Genes. J. Bacteriol. **186**(12): 3980–3990. doi:10.1128/JB.186.12.3980-3990.2004.

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. Front. Microbiol. **6**: 1120. doi:10.3389/fmicb.2015.01120.

Breider, S., Scheuner, C., Schumann, P., Fiebig, A., Petersen, J.J., Pradella, S., Klenk, H.-P.P., Brinkhoff, T., GÃ¶ker, M., and Göker, M. 2014. Genome-scale data suggest reclassifications in the Leisingera-Phaeobacter cluster including proposals for *Sedimentitalea* gen. nov. and *Pseudophaeobacter* gen. nov. Front. Microbiol. **5**: 416. doi:10.3389/fmicb.2014.00416.

Brenner, F.W.W., Villar, R.G.G., Angulo, F.J.J., Tauxe, R., and Swaminathan, B. 2000. *Salmonella* nomenclature. J. Clin. Microbiol. **38**(7): 2465–2467. Am Soc Microbiol.

Brinkhoff, T., Giebel, H.A., and Simon, M. 2008. Diversity, ecology, and genomics of the Roseobacter clade: a short overview. Arch. Microbiol. **189**(6): 531–539. doi:10.1007/s00203-008-0353-y.

Brock, G., Pihur, V., Datta, S., and Datta, S. 2008. clValid: An R Package for Cluster Validation.

J. Stat. Softw. **25**(4): 1–22. Available from http://www.jstatsoft.org/v25/i04/.

Bruen, T.C., Philippe, H., and Bryant, D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics **172**(4): 2665–2681. doi:10.1534/genetics.105.048975.

Buchan, A., González, J.M., and Moran, M.A. 2005. Overview of the marine Roseobacter lineage. Appl. Environ. Microbiol. **71**(10): 5665–5677. doi:10.1128/AEM.71.10.5665-5677.2005.

Bwire, G., Sack, D.A., Almeida, M., Li, S., Voeglein, J.B., Debes, A.K., Kagirita, A., Buyinza, A.W., Orach, C.G., and Stine, O.C. 2018. Molecular characterization of *Vibrio cholerae* responsible for cholera epidemics in Uganda by PCR, MLVA and WGS. PLoS Negl. Trop. Dis. **12**(6): e0006492. doi:10.1371/journal.pntd.0006492.

Canty, A., Ripley, B.D., and others. 2017. boot: Bootstrap R (S-Plus) Functions. R Packag. version **1**(7).

Case, R.J., Boucher, Y., Dahllof, I., Holmstrom, C., Doolittle, W.F., and Kjelleberg, S. 2007. Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. Appl. Environ. Microbiol. **73**(1): 278–288. doi:10.1128/AEM.01177-06.

Case, R.J., Labbate, M., and Kjelleberg, S. 2008. AHL-driven quorum-sensing circuits: their frequency and function among the Proteobacteria. ISME J. **2**(4): 345. Nature Publishing Group.

Case, R.J., Longford, S.R., Campbell, A.H., Low, A., Tujula, N., Steinberg, P.D., and Kjelleberg,

S. 2011. Temperature induced bacterial virulence and bleaching disease in a chemically defended marine macroalga. Environ. Microbiol. **13**(2): 529–537. doi:10.1111/j.1462-2920.2010.02356.x.

Cavailler, P., Lucas, M., Perroud, V., McChesney, M., Ampuero, S., Guérin, P.J., Legros, D., Nierle, T., Mahoudeau, C., Lab, B., Kahozi, P., Deen, J.L., von Seidlein, L., Wang, X.Y., Puri, M., Ali, M., Clemens, J.D., Songane, F., Baptista, A., Ismael, F., Barreto, A., and Chaignat, C.L. 2006. Feasibility of a mass vaccination campaign using a two-dose oral cholera vaccine in an urban cholera-endemic setting in Mozambique. Vaccine **24**(22): 4890–4895. doi:10.1016/j.vaccine.2005.10.006.

Cesbron, S., Briand, M., Essakhi, S., Gironde, S., Boureau, T., Manceau, C., Fischer-Le Saux, M., and Jacques, M.-A. 2015. Comparative Genomics of Pathogenic and Nonpathogenic Strains of *Xanthomonas arboricola* Unveil Molecular and Evolutionary Events Linked to Pathoadaptation. Front. Plant Sci. **6**: 1126. doi:10.3389/fpls.2015.01126.

Chang, F., Qiu, W., Zamar, R.H., Lazarus, R., and Wang, X. 2010. clues: An R Package for Nonparametric Clustering Based on Local Shrinking. J. Stat. Softw. **33**(4): 1–16. Available from http://www.jstatsoft.org/v33/i04/.

Charlson, R.J., Lovelock, J.E., Andreae, M.O., and Warren, S.G. 1987. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. Nature **326**(6114): 655. Nature Publishing Group.

Chaudhari, N.M., Gupta, V.K., and Dutta, C. 2016. BPGA-an ultra-fast pan-genome analysis pipeline. Sci. Rep. **6**(March): 24373. Nature Publishing Group. doi:10.1038/srep24373.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. doi:10.1093/bioinformatics/btz848.

Chen, C., Zhang, W., Zheng, H., Lan, R., Wang, H., Du, P., Bai, X., Ji, S., Meng, Q., Jin, D., Liu, K., Jing, H., Ye, C., Gao, G.F., Wang, L., Gottschalk, M., and Xu, J. 2013. Minimum core genome sequence typing of bacterial pathogens: A unified approach for clinical and public health microbiology. J. Clin. Microbiol. **51**(8): 2582–2591. doi:10.1128/JCM.00535-13.

Chenal-Francisque, V., Passet, V., Brisse, S., Cantinelli, T., Diancourt, L., Pourcel, C., Lecuit, M., Leclercq, A., Tran-Hykes, C., and Bracq-Dieye, H. 2013. Optimized Multilocus Variable-Number Tandem-Repeat Analysis Assay and Its Complementarity with Pulsed-Field Gel Electrophoresis and Multilocus Sequence Typing for *Listeria monocytogenes* Clone Identification and Surveillance. J. Clin. Microbiol. **51**(6): 1868–1880. doi:10.1128/jcm.00606-13.

Childers, B.M., and Klose, K.E. 2007. Regulation of virulence in Vibrio cholerae: The ToxR regulon. Future Microbiol. **2**(3): 335–344. doi:10.2217/17460913.2.3.335.

Choi, S.Y., Rashed, S.M., Hasan, N.A., Alam, M., Islam, T., Sadique, A., Johura, F.-T., Eppinger, M., Ravel, J., Huq, A., Cravioto, A., and Colwell, R.R. 2016. Phylogenetic Diversity of *Vibrio cholerae* Associated with Endemic Cholera in Mexico from 1991 to 2008. MBio **7**(2): e02160-15. doi:10.1128/mBio.02160-15.

Chun, J., Grim, C.J., Hasan, N.A., Lee, J.H., Choi, S.Y., Haley, B.J., Taviani, E., Jeon, Y.-S.,

Kim, D.W., Lee, J.-H., Brettin, T.S., Bruce, D.C., Challacombe, J.F., Detter, J.C., Han, C.S., Munk, A.C., Chertkov, O., Meincke, L., Saunders, E., Walters, R.A., Huq, A., Nair, G.B., and Colwell, R.R. 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. Proc. Natl. Acad. Sci. **106**(36): 15442–15447. doi:10.1073/pnas.0907787106.

Clemens, J.D., Nair, G.B., Ahmed, T., Qadri, F., and Holmgren, J. 2017. Cholera. Lancet **390**(10101): 1539–1549. doi:10.1016/S0140-6736(17)30559-7.

Cody, A.J., Bray, J.E., Jolley, K.A., McCarthy, N.D., and Maiden, M.C.J. 2017. Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of *Campylobacter jejuni* and *C. coli* Human Disease Isolates. J. Clin. Microbiol. **55**(7): 2086–2097. doi:10.1128/JCM.00080-17.

Cohn, F. 1875. Studies on bacteria. Contrib. to Biol. ofplants (in Ger. **1**: 127–222.

Colwell, R.R. 1968. Polyphasic taxonomy of bacteria. Symposium on Taxonomic Studies of Microorganisms by Instrumental Analysis of their Components and Metabolites. *In* International Conference on Culture Collections, Tokyo, Japan. ICRO/UNRSCO. University Park Press, Baltimore, Md.

Cude, W.N., and Buchan, A. 2013. Acyl-homoserine lactone-based quorum sensing in the Roseobacter clade: complex cell-to-cell communication controls multiple physiologies. Front. Microbiol. **4**: 336. Frontiers.

Dalsgaard, A., Skov, M.N., Serichantalergs, O., Echeverria, P., Meza, R., and Taylor, D.N. 1997. Molecular evolution of *Vibrio cholerae* O1 strains isolated in Lima, Peru, from 1991 to

1995. J. Clin. Microbiol. **35**(5): 1151–1156. Am Soc Microbiol.

Danin-Poleg, Y., Cohen, L.A., Gancz, H., Broza, Y.Y., Goldshmidt, H., Malul, E., Valinsky, L., Lerner, L., Broza, M., and Kashi, Y. 2007. *Vibrio cholerae* Strain Typing and Phylogeny Study Based on Simple Sequence Repeats. J. Clin. Microbiol. **45**(3): 736–746. doi:10.1128/JCM.01895-06.

Davis, D.H., Doudoroff, M., Stainer, R.Y., and Mandel, M. 1969. Proposal to reject the genus *Hydrogenomonas*: Taxonomic implications. Int. J. Syst. Bacteriol. **19**(4): 375–390. doi:10.1099/00207713-19-4-375.

Dees, M.W., Lysøe, E., Rossmann, S., Perminow, J., and Brurberg, M.B. 2017. *Pectobacterium polaris* sp. nov., isolated from potato (*Solanum tuberosum*). Int. J. Syst. Evol. Microbiol. **67**(12): 5222–5229. doi:10.1099/ijsem.0.002448.

Desai, P., T., Porwollik, S., Long, F., Cheng, P., Wollam, A., Clifton, S.W.W., Weinstock, G., M., and McClelland, M. 2013. Evolutionary Genomics of *Salmonella enterica* Subspecies. MBio **4**(2): e00579-12. doi:10.1128/mBio.00579-12.

Dingle, T.C., and MacCannell, D.R. 2015. Molecular Strain Typing and Characterisation of Toxigenic *Clostridium difficile*. *In* Methods in Microbiology. Elsevier. pp. 329–357.

Dunn, J.C. 1974. Well-separated clusters and optimal fuzzy partitions. J. Cybern. **4**(1): 95–104. doi:10.1080/01969727408546059.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5): 1792–1797. doi:10.1093/nar/gkh340.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**(19): 2460–2461. doi:10.1093/bioinformatics/btq461.

Enright, M. 2003. The evolution of a resistant pathogen – the case of MRSA. Curr. Opin. Pharmacol. **3**(5): 474–479. doi:10.1016/S1471-4892(03)00109-7.

Enright, M.C., and Spratt, B.G. 1999. Multilocus sequence typing. Trends Microbiol. **7**(12): 482–487. Elsevier. doi:10.1007/978-1-60327-999-4_11.

Eppinger, M., Pearson, T., Koenig, S.S.K., Pearson, O., Hicks, N., Agrawal, S., Sanjar, F., Galens, K., Daugherty, S., Crabtree, J., Hendriksen, R.S., Price, L.B., Upadhyay, B.P., Shakya, G., Fraser, C.M., Ravel, J., and Keim, P.S. 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. doi:10.1128/mBio.01721-14.

Forsythe, S.J., Dickins, B., and Jolley, K.A. 2014. *Cronobacter*, the emergent bacterial pathogen *Enterobacter sakazakii* comes of age; MLST and whole genome sequence analysis. BMC Genomics **15**(1): 1121. doi:10.1186/1471-2164-15-1121.

Frerichs, R.R. 2016. Deadly river: Cholera and cover-up in post-earthquake Haiti. Cornell University Press, Ithaca.

Frerichs, R.R., Keim, P.S., Barrais, R., and Piarroux, R. 2012. Nepalese origin of cholera epidemic in Haiti. Clin. Microbiol. Infect. **18**(6): E158–E163. doi:10.1111/j.1469-0691.2012.03841.x.

Gardner, S.N., and Hall, B.G. 2013. When Whole-Genome Alignments Just Won't Work: kSNP v2 Software for Alignment-Free SNP Discovery and Phylogenetics of Hundreds of Microbial Genomes. PLoS One **8**(12): e81760. doi:10.1371/journal.pone.0081760.

Gardner, S.N., Slezak, T., and Hall, B.G. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics **31**(17): 2877–2878. doi:10.1093/bioinformatics/btv271.

Garg, P., Aydanian, A., Smith, D., Morris, J.G.J.G., Nair, G.B., Stine, O.C., Antonia, A., Smith, D., Morris, J.G.J.G., Nair, G.B., and Stine, O.C. 2003. Molecular Epidemiology of O139 *Vibrio cholerae*: Mutation, Lateral Gene Transfer, and Founder Flush. Emerg. Infect. Dis. **9**(7): 810–814. doi:10.3201/eid0907.020760.

Garrine, M., Mandomando, I., Vubil, D., Nhampossa, T., Acacio, S., Li, S., Paulson, J.N., Almeida, M., Domman, D., Thomson, N.R., Alonso, P., and Stine, O.C. 2017. Minimal genetic change in *Vibrio cholerae* in Mozambique over time: Multilocus variable number tandem repeat analysis and whole genome sequencing. PLoS Negl. Trop. Dis. **11**(6): e0005671. doi:10.1371/journal.pntd.0005671.

Garrity, G.M. 2016. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? J. Clin. Microbiol. **54**(8): 1956–1963. doi:10.1128/JCM.00200-16.

Garrity, G.M., Bell, J.A., and Lilburn, T. 2005. Family I. *Rhodobacteraceae* fam. nov. Bergey's Man. Syst. Bacteriol. **2**(Part C): 161.

Garrity, G.M., Bell, J.A., and Lilburn, T. 2015a. *Helicobacteraceae* fam. nov. Bergey's Man. Syst. Archaea Bact.: 1–1. doi:10.1002/9781118960608.fbm00211.

Garrity, G.M., Bell, J.A., and Lilburn, T. 2015b. *Rhodobacteraceae* fam. nov. *In* Bergey's Manual of Systematics of Archaea and Bacteria. John Wiley & Sons, Ltd, Chichester, UK. pp. 1–2. doi:10.1002/9781118960608.fbm00173.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Peer, Y. Van De, Vandamme, P., Thompson, F.L., and Swings, J. 2005. Re-evaluating prokaryotic species. **3**(September): 733–739.

Ghosh, W., Mandal, S., and Roy, P. 2006. *Paracoccus bengalensis* sp. nov., a novel sulfur-oxidizing chemolithoautotroph from the rhizospheric soil of an Indian tropical leguminous plant. Syst. Appl. Microbiol. **29**(5): 396–403. doi:10.1016/j.syapm.2005.10.004.

Goberna, M., and Verdú, M. 2016. Predicting microbial traits with phylogenies. ISME J. **10**(4): 959–967. doi:10.1038/ismej.2015.171.

Godreuil, S., Cohan, F.M., Shah, H., and Tibayrenc, M. 2005. Which species concept for bacteria?—An E-debate. Infect. Genet. Evol. **5**: 375.

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*. J. Clin. Microbiol. **55**(6): 1682–1697. doi:10.1128/JCM.00227-17.

Gonzalez-Escalona, N., Martinez-Urtaza, J., Romero, J., Espejo T, R., Jaykus, L.-A., and DePaola, A. 2008. Determination of Molecular Phylogenetics of *Vibrio parahaemolyticus* Strains by Multilocus Sequence Typing. J. Bacteriol. **190**(8): 2831–2840. doi:10.1128/JB.01808-07.

González, J.M., Kiene, R.P., and Moran, M.A. 1999. Transformation of Sulfur Compounds by an Abundant Lineage of Marine Bacteria in the $α$-Subclass of the ClassProteobacteria. Appl. Environ. Microbiol. **65**(9): 3810–3819. Am Soc Microbiol.

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. Int. J. Syst. Evol. Microbiol. **57**(1): 81–91. doi:10.1099/ijs.0.64483-0.

Grissa, I., Bouchon, P., Pourcel, C., and Vergnaud, G. 2008. On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. Biochimie **90**(4): 660–668. doi:10.1016/j.biochi.2007.07.014.

Guigon, G., Cheval, J., Cahuzac, R., and Brisse, S. 2008. MLVA-NET - A standard Web Database for Bacterial Genotyping and Surveillance. Eurosurveillance **13**(4): 18863.

Guillaume, Y., Ternier, R., Vissieres, K., Casseus, A., Chery, M.J., and Ivers, L.C. 2018. Responding to cholera in Haiti: Implications for the national plan to eliminate cholera by 2022. J. Infect. Dis. **218**(Suppl 3): S167–S170. doi:10.1093/infdis/jiy491.

Hagberg, A., Swart, P., and S Chult, D. 2008. Exploring network structure, dynamics, and function using NetworkX. *In* Varoquaux G, Vaught T, Millman J (ed), Proceedings of the 7th Python in Science Conference (SciPy2008). pp. 11–16.

Hall, B.G. 2016. Effects of sequence diversity and recombination on the accuracy of phylogenetic trees estimated by kSNP. Cladistics **32**(1): 90–99. doi:10.1111/cla.12113.

Harwani, D. 2013. The great plate count anomaly and the unculturable bacteria. Microbiology

2(9): 350–351.

Hasan, N.A., Choi, S.Y., Eppinger, M., Clark, P.W., Chen, A., Alam, M., Haley, B.J., Taviani, E., Hine, E., Su, Q., and others. 2012. Genomic diversity of 2010 Haitian cholera outbreak strains. Proc. Natl. Acad. Sci. **109**(29): E2010--E2017. National Acad Sciences.

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., Bortolaia, V., Pearson, T., Waters, A.E., Prasad Upadhyay, B., Devi Shrestha, S., Adhikari, S., Shakya, G., Keim, P.S., and Aarestrup, F.M. 2011. Population Genetics of *Vibrio cholerae* from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak. MBio **2**(4): e00157-11. doi:10.1128/mBio.00157-11.

Höhl, M., and Ragan, M.A. 2007. Is Multiple-Sequence Alignment Required for Accurate Inference of Phylogeny? Syst. Biol. **56**(2): 206–221. doi:10.1080/10635150701294741.

Horwood, P., Collins, D., Jonduo, M., Rosewell, A., Dutta, S., Dagina, R., Ropa, B., Siba, P., and Greenhill, A. 2011. Clonal Origins of *Vibrio cholerae* O1 El Tor Strains, Papua New Guinea, 2009–2011. Emerg. Infect. Dis. **17**(11): 2063. doi:10.3201/eid1711.110782.

Hu, D., Liu, B., Feng, L., Ding, P., Guo, X., Wang, M., Cao, B., Reeves, P.R., and Wang, L. 2016. Origins of the current seventh cholera pandemic. Proc. Natl. Acad. Sci. **113**(48): E7730–E7739. doi:10.1073/pnas.1608732113.

Huang, M.-M., Guo, L.-L., Wu, Y.-H., Lai, Q.-L., Shao, Z.-Z., Wang, C.-S., Wu, M., and Xu, X.-W. 2018. *Pseudooceanicola lipolyticus* sp. nov., a marine alphaproteobacterium, reclassification of *Oceanicola flagellatus* as *Pseudooceanicola flagellatus* comb. nov. and emended description of the genus *Pseudooceanicola*. Int. J. Syst. Evol. Microbiol. **68**(1):

409–415. doi:10.1099/ijsem.0.002521.

Hubert, L., and Arabie, P. 1985. Comparing partitions. J. Classif. **2**(1): 193–218. doi:10.1007/BF01908075.

Imhoff, J.F., Trüper, H.G., and Pfennig, N. 1984. Rearrangement of the Species and Genera of the Phototrophic "Purple Nonsulfur Bacteria." Int. J. Syst. Evol. Microbiol. **34**(3): 340–343. doi:10.1099/00207713-34-3-340.

Islam, M.T., Alam, M., and Boucher, Y. 2017. Emergence, ecology and dispersal of the pandemic generating Vibrio cholerae lineage. Int. Microbiol. **20**(3): 106–115. doi:10.2436/20.1501.01.291.

Islam, M.T., Liang, K., Im, M.S., Winkjer, J., Busby, S., Tarr, C.L., and Boucher, Y. 2018. Draft Genome Sequences of Nine *Vibrio* sp. Isolates from across the United States Closely Related to Vibrio cholerae. Microbiol. Resour. Announc. **7**(21): e00965-18. doi:10.1128/MRA.00965-18.

Jahan, S. 2016. Cholera - epidemiology, prevention and control. *In* Significance, Prevention and Control of Food Related Diseases. Croatia: InTech. *Edited by* H.A. Makun. InTechOpen, Rijeka, Croatia. pp. 145–157. doi:10.5772/63358.

Janowicz, A., De Massis, F., Ancora, M., Camma, C., Patavino, C., Battisti, A., Prior, K., Harmsen, D., Scholz, H., Zilli, K., Sacchini, L., Di Giannatale, E., and Garofolo, G. 2018. Core Genome Multilocus Sequence Typing and Single Nucleotide Polymorphism Analysis in the Epidemiology of *Brucella melitensis* Infections. J. Clin. Microbiol. **56**(9): e00517-18.

Jolley, K.A., and Maiden, M.C.J. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11**(1): 595.

Jones, R.C., Harris, L.G., Morgan, S., Ruddy, M.C., Perry, M., Williams, R., Humphrey, T., Temple, M., and Davies, A.P. 2019. Phylogenetic Analysis of *Mycobacterium tuberculosis* Strains in Wales by Use of Core Genome Multilocus Sequence Typing To Analyze Whole-Genome Sequencing Data. J. Clin. Microbiol. **57**(6): e02025-18. doi:10.1128/JCM.02025-18.

Jung, Y.-T., Park, S., Lee, J.-S., and Yoon, J.-H. 2016. *Loktanella marina* sp. nov., isolated from seawater. Int. J. Syst. Evol. Microbiol. **66**(7): 2528–2533. doi:10.1099/ijsem.0.001084.

Kalvisa, A., Tsirogiannis, C., Silamikelis, I., Skenders, G., Broka, L., Zirnitis, A., Jansone, I., and Ranka, R. 2016. MIRU-VNTR genotype diversity and indications of homoplasy in *M. avium* strains isolated from humans and slaughter pigs in Latvia. Infect. Genet. Evol. **43**: 15–21. Elsevier B.V. doi:10.1016/j.meegid.2016.05.013.

Kanagarajah, S., Waldram, A., Dolan, G., Jenkins, C., Ashton, P.M., Carrion Martin, A.I., Davies, R., Frost, A., Dallman, T.J., De Pinna, E.M., Hawker, J.I., Grant, K.A., and Elson, R. 2016. Whole genome sequencing reveals an outbreak of *Salmonella Enteritidis* associated with reptile feeder mice in the United Kingdom, 2012-2015. Food Microbiol. **71**(April 2015): 32–38. Elsevier Ltd. doi:10.1016/j.fm.2017.04.005.

Kanehisa, M., Sato, Y., and Morishima, K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J. Mol. Biol. **428**(4): 726–731. doi:10.1016/j.jmb.2015.11.006.

Kaper, J.B., Morris, J.G., and Levine, M.M. 1995. Cholera. Clin. Microbiol. Rev. **8**(1): 48–86.

Katayama, Y., Hiraishi, A., and Kuraishi, H. 1995. *Paracoccus thiocyanatus* sp. nov., a new species of thiocyanate-utilizing facultative chemolithotroph, and transfer of *Thiobacillus versutus* to the genus *Paracoccus* as *Paracoccus versutus* comb. nov. with emendation of the genus. Microbiology **141**(6): 1469–1477. doi:10.1099/13500872-141-6-1469.

Katz, L.S.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S.S., Turnsek, M.A.A., Guo, Y., Wang, S., Paxinos, E.E.E., Orata, F., Gladney, L.M.M., Stroika, S., Folster, J.P.P., Rowe, L., Freeman, M.M.M., Knox, N., Frace, M., Boncy, J., Graham, M., Hammer, B.K.K., Boucher, Y., Bashir, A., Hanage, W.P.P., Domselaar, G.V. Van, Tarr, L., Van Domselaar, G., Tarr, C.L.L., and Domselaar, G.V. Van. 2013. Evolutionary Dynamics of *Vibrio cholerae* O1 following a Single-Source Introduction to Haiti. MBio **4**(4): e00398-13. Am Soc Microbiol. doi:10.1128/mBio.00398-13.Editor.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**(12): 1647–1649. doi:10.1093/bioinformatics/bts199.

Keim, P., Price, L.B., Klevytska, A.M., Smith, K.L., Schupp, J.M., Okinaka, R., Jackson, P.J., and Hugh-Jones, M.E. 2000. Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J. Bacteriol. nicht bekannt **182**(10): 2928–2936.

Kieleczawa, J. 2006. Fundamentals of sequencing of difficult templates-An overview. J. Biomol. Tech. **17**(3): 207–217.

Kim, K.K., Lee, J.-S., Lee, K.C., Oh, H.-M., and Kim, S.-G. 2010. *Pontibaca methylaminivorans* gen. nov., sp. nov., a member of the family *Rhodobacteraceae*. Int. J. Syst. Evol. Microbiol. **60**(9): 2170–2175. doi:10.1099/ijs.0.020172-0.

Kim, Y.-O., Park, S., Kim, H., Park, D.-S., Nam, B.-H., Kim, D.-G., and Yoon, J.-H. 2014. *Halocynthiibacter namhaensis* gen. nov., sp. nov., a novel *alphaproteobacterium* isolated from sea squirt *Halocynthia roretzi*. Antonie Van Leeuwenhoek **105**(5): 881–889. doi:10.1007/s10482-014-0142-3.

Kirchberger, P.C. 2017. Population dynamics of *Vibrio cholerae* and its close relative Vibrio metoecus in an aquatic ecosystem. University of Alberta. doi:https://doi.org/10.7939/R36M33H7D.

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. Appl. Environ. Microbiol. **82**(18): 5576–5586. doi:10.1128/AEM.01177-16.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. 2001. rrndb: the Ribosomal RNA Operon Copy Number Database. Nucleic Acids Res. **29**(1): 181–184. Available from https://www.ncbi.nlm.nih.gov/pubmed/11125085.

Klassen, J.L., and Currie, C.R. 2012. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. BMC Genomics **13**(1): 14. doi:10.1186/1471-2164-13-14.

Klenk, H.-P., Meier-Kolthoff, J.P., and Göker, M. 2014. Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. Int. J. Syst. Evol. Microbiol. **64**(2): 352–356. doi:10.1099/ijs.0.056994-0.

Konstantinidis, K.T., and Tiedje, J.M. 2005a. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. **102**(7): 2567–2572. doi:10.1073/pnas.0409727102.

Konstantinidis, K.T., and Tiedje, J.M. 2005b. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. **187**(18): 6258–6264. doi:10.1128/JB.187.18.6258-6264.2005.

Kotetishvili, M., Stine, O.C., Chen, Y., Kreger, A., Sulakvelidze, A., Sozhamannan, S., and Morris, J.G. 2003. Multilocus sequence typing has better discriminatory ability for typing *Vibrio cholerae* than does pulsed-field gel electrophoresis and provides a measure of phylogenetic relatedness. J. Clin. Microbiol. **41**(5): 2191–2196. doi:10.1128/JCM.41.5.2191-2196.2003.

Kroes, I., Lepp, P.W., and Relman, D.A. 1999. Bacterial diversity within the human subgingival crevice. Proc. Natl. Acad. Sci. **96**(25): 14547–14552. doi:10.1073/pnas.96.25.14547.

Kumar, P.A., Aparna, P., Srinivas, T.N.R., Sasikala, C., and Ramana, C. V. 2008. *Rhodovulum kholense* sp. nov. Int. J. Syst. Evol. Microbiol. **58**(7): 1723–1726. doi:10.1099/ijs.0.65620-0.

Lam, C., Octavia, S., Reeves, P.R., and Lan, R. 2012. Multi-locus variable number tandem repeat analysis of 7th pandemic *Vibrio cholerae*. BMC Microbiol. **12**(1): 82. doi:10.1186/1471-2180-12-82.

Leavis, H.L., Bonten, M.J., and Willems, R.J. 2006. Identification of high-risk enterococcal clonal complexes: global dispersion and antibiotic resistance. Curr. Opin. Microbiol. **9**(5): 454–460. doi:10.1016/j.mib.2006.07.001.

Lee, W.P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., and Marth, G.T. 2014. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One **9**(3): e90581. doi:10.1371/journal.pone.0090581.

Leekitcharoenphon, P., Kaas, R.S., Thomsen, M.C.F., Friis, C., Rasmussen, S., and Aarestrup, F.M. 2012. snpTree--a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics **13**: S6. doi:10.1186/1471-2164-13-s7-s6.

Leekitcharoenphon, P., Nielsen, E.M., Kaas, R.S., Lund, O., and Aarestrup, F.M. 2014. Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*. PLoS One **9**(2): e87991. doi:10.1371/journal.pone.0087991.

Legros, D. 2018. Global cholera epidemiology: Opportunities to reduce the burden of cholera by 2030. J. Infect. Dis. **218**(Suppl 3): S137–S140. doi:10.1093/infdis/jiy486.

Letunic, I., and Bork, P. 2007. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics **23**(1): 127–128. doi:10.1093/bioinformatics/btl529.

Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14): 1754–1760. doi:10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,

and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16): 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, L., Stoeckert, C.J.J., and Roos, D.S.S. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. **13**(9): 2178–2189. doi:10.1101/gr.1224503.candidates.

Liang, K., Islam, M.T., Hussain, N., Winkjer, N.S., Im, M.S., Rowe, L.A., Tarr, C.L., and Boucher, Y. 2019. Draft Genome Sequences of Eight Vibrio sp. Clinical Isolates from across the United States That Form a Basal Sister Clade to *Vibrio cholerae*. Microbiol. Resour. Announc. **8**(3): e01473-18. doi:10.1128/MRA.01473-18.

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 Announc. **5**(35): e00905-17. doi:10.1128/genomeA.00905-17.

Lindstedt, B.A. 2005. Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. Electrophoresis **26**(13): 2567–2582. doi:10.1002/elps.200500096.

Lovelock, J.E., Maggs, R.J., and Rasmussen, R.A. 1972. Atmospheric dimethyl sulphide and the natural sulphur cycle. Nature **237**(5356): 452. Nature Publishing Group.

Lucidarme, J., Hill, D.M.C., Bratcher, H.B., Gray, S.J., du Plessis, M., Tsang, R.S.W., Vazquez, J.A., Taha, M.-K., Ceyhan, M., Efron, A.M., Gorla, M.C., Findlow, J., Jolley, K.A., Maiden, M.C.J., and Borrow, R. 2015. Genomic resolution of an aggressive, widespread,

diverse and expanding meningococcal serogroup B, C and W lineage. J. Infect. **71**(5): 544–552. doi:10.1016/j.jinf.2015.07.007.

Luo, H., and Moran, M.A. 2014. Evolutionary ecology of the marine Roseobacter clade. Microbiol. Mol. Biol. Rev. **78**(4): 573–587. doi:10.1128/MMBR.00020-14.

Luo, H., and Moran, M.A. 2015. How do divergent ecological strategies emerge among marine bacterioplankton lineages? Trends Microbiol. **23**(9): 577–584. Elsevier Ltd. doi:10.1016/j.tim.2015.05.004.

Luo, Y., Ye, J., Jin, D., Ding, G., Zhang, Z., Mei, L., Octavia, S., and Lan, R. 2013. Molecular analysis of non-O1/non-O139 *Vibrio cholerae* isolated from hospitalised patients in China. BMC Microbiol. **13**(1): 52. doi:10.1186/1471-2180-13-52.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., and Spratt, B.G. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA **95**(March): 3140–3145. Available from www.pnas.org.

Maiden, M.C.J., Van Rensburg, M.J.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A., and McCarthy, N.D. 2013. MLST revisited: The gene-by-gene approach to bacterial genomics. Nat. Rev. Microbiol. **11**(10): 728–736. Nature Publishing Group. doi:10.1038/nrmicro3093.

Mandal, S., Mandal, M.D., and Pal, N.K. 2011. Cholera: A great global concern. Asian Pac. J. Trop. Med. **4**(7): 573–580. Hainan Medical College. doi:10.1016/S1995-7645(11)60149-1.

Margos, G., Gatewood, A.G., Aanensen, D.M., Hanincova, K., Terekhova, D., Vollmer, S.A., Cornet, M., Piesman, J., Donaghy, M., Bormane, A., Hurn, M.A., Feil, E.J., Fish, D., Casjens, S., Wormser, G.P., Schwartz, I., and Kurtenbach, K. 2008. MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. **105**(25): 8730–8735. doi:10.1073/pnas.0800323105.

Marsh, J.W., O'Leary, M.M., Shutt, K.A., Sambol, S.P., Johnson, S., Gerding, D.N., and Harrison, L.H. 2010. Multilocus variable-number tandem-repeat analysis and multilocus sequence typing reveal genetic relationships among *Clostridium difficile* isolates genotyped by restriction endonuclease analysis. J. Clin. Microbiol. **48**(2): 412–418. doi:10.1128/JCM.01315-09.

Martín, B., Bover-Cid, S., and Aymerich, T. 2018. MLVA subtyping of *Listeria monocytogenes* isolates from meat products and meat processing plants. Food Res. Int. **106**(October 2017): 225–232. Elsevier. doi:10.1016/j.foodres.2017.12.052.

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. Available from papers2://publication/doi/10.1126/science.1120096.

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. doi:10.1186/1471-2105-14-60.

Mellmann, A., Bletz, S., Böking, T., Kipp, F., and Becker, K. 2016. Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional

Setting. Am. Soc. Microbiol. **54**(August): 2874–2881. doi:10.1128/JCM.00790-16.

Mignard, S., and Flandrois, J.P. 2006. 16S rRNA sequencing in routine bacterial identification: A 30-month experiment. J. Microbiol. Methods **67**(3): 574–581. doi:10.1016/j.mimet.2006.05.009.

Mintz, E., Omolo, J., Ope, M., Gathigi, L., Thuranira, M., Loharikar, A., Abade, A., Ayers, T., De Cock, K.M., Makayotto, L., Oundo, J., Ismail, A.M., Breiman, R.F., Langat, D., Briere, E., Amwayi, S., O'Reilly, C.E., and Njeru, I. 2013. A National Cholera Epidemic With High Case Fatality Rates-Kenya 2009. J. Infect. Dis. **208**(suppl 1): S69–S77. doi:10.1093/infdis/jit220.

Momba, M., and Azab El-Liethy, M. 2017. *Vibrio cholerae* and Cholera biotypes. *In* Global Water Pathogen Project. *Edited by* A. Pruden, N. Ashbolt, and J. Miller. Michigan State University, Michigan. p. online. doi:10.14321/waterpathogens.28.

Moran, M.A., Belas, R., Schell, M.A., Gonzalez, J.M., Sun, F., Sun, S., Binder, B.J., Edmonds, J., Ye, W., Orcutt, B., Howard, E.C., Meile, C., Palefsky, W., Goesmann, A., Ren, Q., Paulsen, I., Ulrich, L.E., Thompson, L.S., Saunders, E., and Buchan, A. 2007. Ecological genomics of marine roseobacters. Appl. Environ. Microbiol. **73**(14): 4559–4569. doi:10.1128/AEM.02580-06.

Moran, M.A., González, J.M., and Kiene, R.P. 2003. Linking a bacterial taxon to sulfur cycling in the sea: studies of the marine Roseobacter group. Geomicrobiol. J. **20**(4): 375–388. Taylor & Francis.

Moran, M.A., Reisch, C.R., Kiene, R.P., and Whitman, W.B. 2012. Genomic Insights into

Bacterial DMSP Transformations. Ann. Rev. Mar. Sci. **4**(1): 523–542. doi:10.1146/annurev-marine-120710-100827.

Moura, A., Criscuolo, A., Pouseele, H., Maury, M.M., Leclercq, A., Tarr, C., Björkman, J.T., Dallman, T., Reimer, A., Enouf, V., Larsonneur, E., Carleton, H., Bracq-Dieye, H., Katz, L.S., Jones, L., Touchon, M., Tourdjman, M., Walker, M., Stroika, S., Cantinelli, T., Chenal-Francisque, V., Kucerova, Z., Rocha, E.P.C., Nadon, C., Grant, K., Nielsen, E.M., Pot, B., Gerner-Smidt, P., Lecuit, M., and Brisse, S. 2016. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nat. Microbiol. **2**(2): 1–10. Nature Publishing Group. doi:10.1038/nmicrobiol.2016.185.

Murray, R.G.E., Brenner, D.J., Colwell, R.R., De Vos, P., Goodfellow, M., Grimont, P.A.D., Pfennig, N., Stackebrandt, E., and Zavarzin, G.A. 1990. Report of the ad hoc committee on approaches to taxonomy within the Proteobacteria. Int. J. Syst. Bacteriol. **40**(2): 213–215. doi:10.1099/00207713-40-2-213.

Murray, R.G.E., and Holt, J.G. 1989. The History of Bergey's Manual. Archives **19**(1): 1–13. doi:10.1016/j.ghir.2009.04.009.

Mutreja, A., Kim, D.W., Thomson, N.R., Connor, T.R., Lee, J.H., Kariuki, S., Croucher, N.J., Choi, S.Y., Harris, S.R., Lebens, M., Niyogi, S.K., Kim, E.J., Ramamurthy, T., Chun, J., Wood, J.L.N., Clemens, J.D., Czerkinsky, C., Nair, G.B., Holmgren, J., Parkhill, J., and Dougan, G. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature **477**(7365): 462–465. Nature Publishing Group. doi:10.1038/nature10392.

Na, S.-I., Kim, Y.O., Yoon, S.-H., Ha, S., Baek, I., and Chun, J. 2018. UBCG: Up-to-date

bacterial core gene set and pipeline for phylogenomic tree reconstruction. J. Microbiol. **56**(4): 280–285. doi:10.1007/s12275-018-8014-6.

Neumann, B., Prior, K., Bender, J.K., Harmsen, D., Klare, I., Fuchs, S., Bethe, A., Zühlke, D., Göhler, A., Schwarz, S., Schaffer, K., Riedel, K., Wieler, L.H., and Werner, G. 2019. A Core Genome Multilocus Sequence Typing Scheme for *Enterococcus faecalis*. J. Clin. Microbiol. **57**(3): e01686-18. doi:10.1128/JCM.01686-18.

Octavia, S., Salim, A., Kurniawan, J., Lam, C., Leung, Q., Ahsan, S., Reeves, P.R., Nair, G.B., and Lan, R. 2013. Population Structure and Evolution of Non-O1/Non-O139 *Vibrio cholerae* by Multilocus Sequence Typing. PLoS One **8**(6): e65342. doi:10.1371/journal.pone.0065342.

Orata, F.D., Keim, P.S., and Boucher, Y. 2014. The 2010 Cholera Outbreak in Haiti: How Science Solved a Controversy. PLoS Pathog. **10**(4): e1003967. doi:10.1371/journal.ppat.1003967.

Orata, F.D., Meier-Kolthoff, J.P., Sauvageau, D., and Stein, L.Y. 2018. Phylogenomic Analysis of the Gammaproteobacterial Methanotrophs (Order Methylococcales) Calls for the Reclassification of Members at the Genus and Species Levels. Front. Microbiol. **9**: 3162. doi:10.3389/fmicb.2018.03162.

Orata, F.D., Xu, Y., Gladney, L.M., Rishishwar, L., Case, R.J., Boucher, Y., Jordan, I.K., and Tarr, C.L. 2016. Characterization of clinical and environmental isolates of *Vibrio cidicii* sp. nov., a close relative of *Vibrio navarrensis*. Int. J. Syst. Evol. Microbiol. **66**(10): 4148–4155. doi:10.1099/ijsem.0.001327.

P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. **75**(23): 7537–7541. doi:10.1128/AEM.01541-09.

Parker, C.T., Tindall, B.J., and Garrity, G.M. 2019. International Code of Nomenclature of Prokaryotes. Int. J. Syst. Evol. Microbiol. **69**(1A): S1–S111. doi:10.1099/ijsem.0.000778.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. **36**(10): 996–1004. Nature Publishing Group. doi:10.1038/nbt.4229.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. 2015. CheckM : assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. **25**: 1043–1055.

Parks, D.H., Porter, M., Churcher, S., Wang, S., Blouin, C., Whalley, J., Brooks, S., and Beiko, R.G. 2009. GenGIS: A geospatial information system for genomic data. Genome Res. **19**(10): 1896–1904. doi:10.1101/gr.095612.109.

Pearce, M.E., Alikhan, N.-F., Dallman, T.J., Zhou, Z., Grant, K., and Maiden, M.C.J. 2018. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar Enteritidis outbreak. Int. J. Food Microbiol. **274**(October 2017): 1–11. Elsevier.

doi:10.1016/j.ijfoodmicro.2018.02.023.

Philippot, L., Andersson, S.G.E., Battin, T.J., Prosser, J.I., Schimel, J.P., Whitman, W.B., and Hallin, S. 2010. The ecological coherence of high bacterial taxonomic ranks. Nat. Rev. Microbiol. **8**(7): 523–529. Nature Publishing Group. doi:10.1038/nrmicro2367.

Pightling, A.W., Petronella, N., and Pagotto, F. 2014. Choice of Reference Sequence and Assembler for Alignment of *Listeria monocytogenes* Short-Read Sequence Data Greatly Influences Rates of Error in SNP Analyses. PLoS One **9**(8): e104579. doi:10.1371/journal.pone.0104579.

Poretsky, R., Rodriguez-R, L.M., Luo, C., Tsementzi, D., and Konstantinidis, K.T. 2014. Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. PLoS One **9**(4): e93827. doi:10.1371/journal.pone.0093827.

Porwollik, S., Boyd, E.F., Choy, C., Cheng, P., Florea, L., Proctor, E., and McClelland, M. 2004. Characterization of *Salmonella enterica* Subspecies I Genovars by Use of Microarrays. J. Bacteriol. **186**(17): 5883–5898. doi:10.1128/JB.186.17.5883-5898.2004.

Pujalte, M.J., Lucena, T., Ruvira, M.A., Arahal, D.R., and Macian, C.M. 2007. The Family *Rhodobacteraceae*. *In* The Prokaryotes - Alphaproteobacteria and Betaproteobacteria. *Edited by* E. Rosenberg, E.F. DeLong, F. Thompson, S. Lory, and E. Stackebrandt. Elsevier. pp. 440–498. doi:10.1016/B978-0-08-047514-1.50009-3.

Qin, Q.L., Xie, B.B., Zhang, X.Y., Chen, X.L., Zhou, B.C., Zhou, J., Oren, A., and Zhang, Y.Z. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. J.

Bacteriol. **196**(12): 2210–2215. doi:10.1128/JB.01688-14.

Qin, T., Zhang, W., Liu, W., Zhou, H., Ren, H., Shao, Z., Lan, R., and Xu, J. 2016. Population structure and minimum core genome typing of Legionella pneumophila. Sci. Rep. **6**: 21356. Nature Publishing Group. doi:10.1038/srep21356.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna, Austria. Available from https://www.r-project.org/.

Rabus, R., Wöhlbrand, L., Thies, D., Meyer, M., Reinhold-Hurek, B., and Kämpfer, P. 2019. Aromatoleum gen. nov., a novel genus accommodating the phylogenetic lineage including Azoarcus evansii and related species, and proposal of Aromatoleum aromaticum sp. nov., Aromatoleum petrolei sp. nov., *Aromatoleum bremense*. Int. J. Syst. Evol. Microbiol. **69**(4): 982–997. doi:10.1099/ijsem.0.003244.

Ramette, A., and Tiedje, J.M. 2007. Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. Microb. Ecol. **53**(2): 197–207. doi:10.1007/s00248-005-5010-2.

Ranjan, R., Rani, A., Metwally, A., McGee, H.S., and Perkins, D.L. 2016. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. **469**(4): 967–977. doi:10.1016/j.bbrc.2015.12.083.

Reimer, A., Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., Talkington, D., Rowe, L., Olsen-Rasmussen, M., Frace, M., Sammons, S., Dahourou, G., Boncy, J., Smith, A., Mabon, P., Petkau, A., Graham, M., Gilmour, M., and Gerner-Smidt, P. 2011. Comparative Genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. Emerg. Infect. Dis. **17**(11):

2113. doi:10.3201/eid1711.110794.

Reisch, C.R., Moran, M.A., and Whitman, W.B. 2011. Bacterial catabolism of dimethylsulfoniopropionate (DMSP). Front. Microbiol. **2**: 172. doi:10.3389/fmicb.2011.00172.

Ribeiro-Gonçalves, B., Francisco, A.P., Vaz, C., Ramirez, M., and Carriço, J.A. 2016. PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Res. **44**(W1): W246–W251. doi:10.1093/nar/gkw359.

Richter, M., and Rosselló-Móra, R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. **106**(45): 19126–19131. doi:10.1073/pnas.0906412106.

Romanenko, L.A., Tanaka, N., Svetashev, V.I., and Mikhailov, V. V. 2011. *Primorskyibacter sedentarius* gen. nov., sp. nov., a novel member of the class *Alphaproteobacteria* from shallow marine sediments. Int. J. Syst. Evol. Microbiol. **61**(7): 1572–1578. doi:10.1099/ijs.0.025551-0.

Rosselló-Móra, R., and Amann, R. 2015. Past and future species definitions for Bacteria and Archaea. Syst. Appl. Microbiol. **38**(4): 209–216. doi:10.1016/j.syapm.2015.02.001.

Royer, G., Fourreau, F., Boulanger, B., Mercier-Darty, M., Ducellier, D., Cizeau, F., Potron, A., Podglajen, I., Mongardon, N., and Decousser, J.-W. 2019. Local outbreak of extended-spectrum β-lactamase SHV2a-producing *Pseudomonas aeruginosa* reveals the emergence of a new specific sub-lineage of the international ST235 high-risk clone. J. Hosp. Infect.

doi:10.1016/j.jhin.2019.07.014.

Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H.L., Allerberger, F., Harmsen, D., and Mellmann, A. 2015. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *listeria monocytogenes*. J. Clin. Microbiol. **53**(9): 2869–2876. doi:10.1128/JCM.01193-15.

Sabat, A.J., Budimir, A., Nashev, D., Sá-Leão, R., van Dijl, J.M., Laurent, F., Grundmann, H., Friedrich, A.W., and on behalf of the ESCMID Study Group. 2013. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Eurosurveillance **18**(4): 20380. European Centre for Disease Prevention and Control. doi:10.2807/ese.18.04.20380-en.

Sack, D.A., Sack, R.B., and Chaignat, C.-L. 2006. Getting Serious about Cholera. N. Engl. J. Med. **355**(7): 649–651. doi:10.1056/nejmp068144.

Safa, A., Nair, G.B., and Kong, R.Y.C. 2010. Evolution of new variants of *Vibrio cholerae* O1. Trends Microbiol. **18**(1): 46–54. doi:10.1016/j.tim.2009.10.003.

Sails, A.D., Swaminathan, B., and Fields, P.I. 2003. Clonal complexes of *Campylobacter jejuni* identified by multilocus sequence typing correlate with strain associations identified by multilocus enzyme electrophoresis. J. Clin. Microbiol. **41**(9): 4058–4067. doi:10.1128/JCM.41.9.4058-4067.2003.

Salim, A., Lan, R., and Reeves, P.R. 2005. *Vibrio cholerae* pathogenic clones. Emerg. Infect. Dis. **11**(11): 1758–1760. doi:10.3201/eid1111.041170.

Sardar, T., Mukhopadhyay, S., Bhowmick, A.R., and Chattopadhyay, J. 2013. An optimal cost effectiveness study on Zimbabwe cholera seasonal data from 2008-2011. PLoS One **8**(12): e81231. doi:10.1371/journal.pone.0081231.

Sasi Jyothsna, T.S., Tushar, L., Sasikala, C., and Ramana, C. V. 2016. *Paraclostridium benzoelyticum* gen. nov., sp. nov., isolated from marine sediment and reclassification of *Clostridium bifermentans* as *Paraclostridium bifermentans* comb. nov. Proposal of a new genus *Paeniclostridium* gen. nov. to . Int. J. Syst. Evol. Microbiol. **66**(3): 1268–1274. doi:10.1099/ijsem.0.000874.

Schürch, A.C., Arredondo-Alonso, S., Willems, R.J.L., and Goering, R.V. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clin. Microbiol. Infect. **24**(4): 350–354. doi:10.1016/j.cmi.2017.12.016.

Sen, A., Daubin, V., Abrouk, D., Gifford, I., Berry, A.M., and Normand, P. 2014. Phylogeny of the class *Actinobacteria* revisited in the light of complete genomes. The orders "*Frankiales*" and *Micrococcales* should be split into coherent entities: proposal of *Frankiales* ord. nov., <i≥Geodermatophilales</i> ord. Int. J. Syst. Evol. Microbiol. **64**: 3821–3832. doi:10.1099/ijs.0.063966-0.

Seymour, J.R., Simo, R., Ahmed, T., and Stocker, R. 2010. Chemoattraction to Dimethylsulfoniopropionate Throughout the Marine Microbial Food Web. Science **329**(5989): 342–345. doi:10.1126/science.1188418.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N.,

Schwikowski, B., and Ideker, T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. **13**(11): 2498–2504. doi:10.1101/gr.1239303.

Shiba, T. 1991. Roseobacter litoralis gen. nov., sp. nov., and Roseobacter denitrificans sp. nov., aerobic pink-pigmented bacteria which contain bacteriochlorophyll a. Syst. Appl. Microbiol. **14**(2): 140–145.

Shin, Y.H., Kim, J., Suckhoom, A., Kantachote, D., and Kim, W. 2017. *Limibaculum halophilum* gen . nov ., sp . nov ., a new member of the family *Rhodobacteraceae*. Int. J. Syst. Evol. Microbiol. **67**: 3812–3818. doi:10.1099/ijsem.0.002200.

Simon, M., Scheuner, C., Meier-Kolthoff, J.P., Brinkhoff, T., Wagner-Döbler, I., Ulbrich, M., Klenk, H.P., Schomburg, D., Petersen, J., and Göker, M. 2017. Phylogenomics of *Rhodobacteraceae* reveals evolutionary adaptation to marine and non-marine habitats. ISME J. **11**(6): 1483–1499. doi:10.1038/ismej.2016.198.

Sneath, P.H.A. 2005. Numerical taxonomy. *In* Bergey's Manual of Systematic Bacteriology. Springer. pp. 39–42.

Souvorov, A., Agarwala, R., and Lipman, D.J. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. **19**(1): 153. Genome Biology. doi:10.1186/s13059-018-1540-z.

Srinivas, A., Kumar, B. V., Sree, B.D., Tushar, L., Sasikala, C., and Ramana, C. V. 2014. *Rhodovulum salis* sp. nov. and *Rhodovulum viride* sp. nov., phototrophic Alphaproteobacteria isolated from marine habitats. Int. J. Syst. Evol. Microbiol. **64**(Pt 3):

957–962. doi:10.1099/ijs.0.058974-0.

Stackebrandt, E., and Ebers, J. 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today **33**: 152–155.

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. Int. J. Syst. Bacteriol. **44**(4): 846–849. doi:10.1099/00207713-44-4-846.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9): 1312–1313. doi:10.1093/bioinformatics/btu033.

Stanier, R.Y., and Van Niel, C.B. 1941. The main outlines of bacterial classification. J. Bacteriol. **42**(4): 437. American Society for Microbiology (ASM).

Strom, S., Wolfe, G., Slajer, A., Lambert, S., and Clough, J. 2003. Chemical defense in the microplankton II: Inhibition of protist feeding by β-dimethylsulfoniopropionate (DMSP). Limnol. Oceanogr. **48**(1): 230–237. doi:10.4319/lo.2003.48.1.0230.

Struelens, M.J., and Brisse, S. 2013. From molecular to genomic epidemiology: Transforming surveillance and control of infectious diseases. Eurosurveillance **18**(4): 20386. doi:10.2807/ese.18.04.20386-en.

Sunda, W.G., Kieber, D., and Kiene, R.P. 2002. An antioxidant function of DMSP and DMS in marine algae Oceanic dimethylsulfide (DMS) photolysis View project. : 317–320. doi:10.1038/nature00851.

Suresh, G., Lodha, T.D., Indu, B., Sasikala, C., and Ramana, C. V. 2019. Taxogenomics

Resolves Conflict in the Genus *Rhodobacter*: A Two and Half Decades Pending Thought to Reclassify the Genus *Rhodobacter*. Front. Microbiol. **10**: 2480. doi:10.3389/fmicb.2019.02480.

Thomas, J.C.C., Vargas, M.R.R., Miragaia, M., Peacock, S.J.J., Archer, G.L.L., and Enright, M.C.C. 2007. Improved Multilocus Sequence Typing Scheme for *Staphylococcus epidermidis*. J. Clin. Microbiol. **45**(2): 616–619. doi:10.1128/JCM.01934-06.

Thompson, C.C., Amaral, G.R., Campeão, M., Edwards, R.A., Polz, M.F., Dutilh, B.E., Ussery, D.W., Sawabe, T., Swings, J., and Thompson, F.L. 2015. Microbial taxonomy in the post-genomic era: Rebuilding from scratch? Arch. Microbiol. **197**(3): 359–370. doi:10.1007/s00203-014-1071-2.

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. Int. J. Syst. Evol. Microbiol. **60**(1): 249–266. doi:10.1099/ijs.0.016949-0.

Torpdahl, M., Sørensen, G., Lindstedt, B.-A., and Nielsen, E.M. 2007. Tandem repeat analysis for surveillance of human *Salmonella Typhimurium* infections. Emerg. Infect. Dis. **13**(3): 388–95. doi:10.3201/eid1303.060460.

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. Int. J. Syst. Evol. Microbiol. **54**(4): 1263–1269. doi:10.1099/ijs.0.03006-0.

Treangen, T.J., Ondov, B.D., Koren, S., and Phillippy, A.M. 2014. The harvest suite for rapid

core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. **15**(11): 524. doi:10.1186/s13059-014-0524-x.

Trotsenko, Y.A., Doronina, N. V, and Tourova, T.P. 2000. *Methylarcula marina* gen. nov., sp. nov. and *Methylarcula terricola* sp. nov.: novel aerobic, moderately halophilic, facultatively methylotrophic bacteria from coastal saline environments. Int. J. Syst. Evol. Microbiol. **50**(5): 1849–1859. doi:10.1099/00207713-50-5-1849.

Vallina, S.M., and Simó, R. 2007. Strong relationship between DMS and the solar radiation dose over the global surface ocean. Science **315**(5811): 506–508. American Association for the Advancement of Science.

Vandamme, P., Pot, B., Gillis, M., De Vos, P., Kersters, K., and Swings, J. 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol. Rev. **60**(2): 407–438. Available from http://www.ncbi.nlm.nih.gov/pubmed/8801440.

Vergnaud, G., and Pourcel, C. 2009. Multiple locus variable number of tandem repeats analysis. *In* Molecular Epidemiology of Microorganisms. Springer. pp. 141–158.

Visscher, P.T., Diaz, M.R., and Taylor, B.F. 1992. Enumeration of bacteria which cleave or demethylate dimethylsulfoniopropionate in the Caribbean Sea. Mar. Ecol. Prog. Ser. Oldend. **89**(2): 293–296.

Vogler, A.J., Birdsell, D.N., Lee, J., Vaissaire, J., Doujet, C.L., Lapalus, M., Wagner, D.M., and Keim, P. 2011. Phylogeography of *Francisella tularensis* ssp. holarctica in France. Lett. Appl. Microbiol. **52**(2): 177–180. doi:10.1111/j.1472-765X.2010.02977.x.

Volpe Sperry, E.K., Kathariou, S., Edwards, J.S., and Wolf, L.A. 2008. Multiple-locus variable-number tandem-repeat analysis as a tool for Subtyping *Listeria monocytogenes* strains. J. Clin. Microbiol. **46**(4): 1435–1450. doi:10.1128/JCM.02207-07.

Waite, D.W., Vanwonterghem, I., Rinke, C., Parks, D.H., Zhang, Y., Takai, K., Sievert, S.M., Simon, J., Campbell, B.J., Hanson, T.E., Woyke, T., Klotz, M.G., and Hugenholtz, P. 2017. Comparative Genomic Analysis of the Class *Epsilonproteobacteria* and Proposed Reclassification to *Epsilonbacteraeota* (phyl. nov.). Front. Microbiol. **8**: 682. doi:10.3389/fmicb.2017.00682.

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*. Int. J. Syst. Evol. Microbiol. **64**(2): 551–558. doi:10.1099/ijs.0.050351-0.

Wang, R., Yu, D., Yue, J., and Kan, B. 2016. Variations in SXT elements in epidemic *Vibrio cholerae* O1 El Tor strains in China. Sci. Rep. **6**(1): 22733. Nature Publishing Group. doi:10.1038/srep22733.

Wang, X., Jordan, I.K., and Mayer, L.W. 2015. A Phylogenetic Perspective on Molecular Epidemiology. *In* Molecular Medical Microbiology. Elsevier. pp. 517–536. doi:10.1016/B978-0-12-397169-2.00029-9.

Wang, X.M., Noble, L., Kreiswirth, B.N., Eisner, W., McClements, W., Jansen, K.U., and Anderson, A.S. 2003. Evaluation of a multilocus sequence typing system for *Staphylococcus epidermidis*. J. Med. Microbiol. **52**(11): 989–998.

doi:10.1099/jmm.0.05360-0.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., Starr, M.P., and Trüper, H.G. 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int. J. Syst. Bacteriol. **37**(4): 463–464.

Weill, F.-X., Domman, D., Njamkepo, E., Almesbahi, A.A., Naji, M., Nasher, S.S., Rakesh, A., Assiri, A.M., Sharma, N.C., Kariuki, S., Pourshafie, M.R., Rauzier, J., Abubakar, A., Carter, J.Y., Wamala, J.F., Seguin, C., Bouchier, C., Malliavin, T., Bakhshi, B., Abulmaali, H.H.., Kumar, D., Njoroge, S.M., Malik, M.R., Kiiru, J., Luquero, F.J., Azman, A.S., Ramamurthy, T., Thomson, N.R., and Quilici, M.-L. 2018. Genomic insights into the 2016–2017 cholera epidemic in Yemen. Nature **565**(7738): 230–233. Springer US. doi:10.1038/s41586-018-0818-3.

Weill, F.-X., Domman, D., Njamkepo, E., Tarr, C., Rauzier, J., Fawal, N., Keddy, K.H., Salje, H., Moore, S., Mukhopadhyay, A.K., Bercion, R., Luquero, F.J., Ngandjio, A., Dosso, M., Monakhova, E., Garin, B., Bouchier, C., Pazzani, C., Mutreja, A., Grunow, R., Sidikou, F., Bonte, L., Breurec, S., Damian, M., Njanpop-Lafourcade, B.-M., Sapriel, G., Page, A.-L., Hamze, M., Henkens, M., Chowdhury, G., Mengel, M., Koeck, J.-L., Fournier, J.-M., Dougan, G., Grimont, P.A.D., Parkhill, J., Holt, K.E., Piarroux, R., Ramamurthy, T., Quilici, M.-L., and Thomson, N.R. 2017. Genomic history of the seventh pandemic of cholera in Africa. Science **358**(6364): 785–789. doi:10.1126/science.aad5901.

Wells, J.M., Raju, B.C., Hung, H., Weisburg, W.G., Mandelco-paul, L., and Brenner, D.O.N.J. 1987. *Xylella fastidiosa* gen. nov., sp. nov: Gram-Negative, Xylem-limited, tastidious plant

bacteria related to *Xanthomonas* spp. Int. J. Syst. Bacteriol. **37**(2): 136–143.

Wickham, H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Available from http://ggplot2.org.

Wirth, J.S., and Whitman, W.B. 2018. Phylogenomic analyses of a clade within the roseobacter group suggest taxonomic reassignments of species of the genera Aestuariivita, Citreicella, Loktanella, Nautella, Pelagibaca, Ruegeria, Thalassobius,. Int. J. Syst. Evol. Microbiol. **68**(7): 2393–2411. doi:10.1099/ijsem.0.002833.

Wisplinghoff, H., Rosato, A.E., Enright, M.C., Noto, M., Craig, W., and Archer, G.L. 2003. Related Clones Containing SCC mec Type IV Predominate among Clinically Significant *Staphylococcus epidermidis* Isolates Related Clones Containing SCCmec Type IV Predominate among Clinically Significant *Staphylococcus epidermidis* Isolates. Antimicrob. Agents Chemother. **47**(11): 3574–3579. doi:10.1128/AAC.47.11.3574.

Woese, C.R. 1987. Bacterial evolution. Microbiol. Rev. **51**(2): 221–271. Available from https://www.ncbi.nlm.nih.gov/pubmed/2439888.

Wong, V.K., Baker, S., Connor, T.R., Pickard, D., Page, A.J., Dave, J., Murphy, N., Holliman, R., Sefton, A., Millar, M., Dyson, Z.A., Dougan, G., and Holt, K.E. 2016. An extended genotyping framework for *Salmonella enterica* serovar Typhi, the cause of human typhoid. Nat. Commun. **7**(1): 12827. doi:10.1038/ncomms12827.

World Health Organization. 2017. Weekly epidemiological record. Available from http://www.who.int/wer.

Wurtzel, O., Sesto, N., Mellin, J.R., Karunker, I., Edelheit, S., Bécavin, C., Archambaud, C., Cossart, P., and Sorek, R. 2012. Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. Mol. Syst. Biol. **8**(1): 583. doi:10.1038/msb.2012.11.

Yamamoto, S., and Harayama, S. 1996. Phylogenetic analysis of *Acinetobacter* strains based on the nucleotide sequences of gyrB genes and on the amino acid sequences of their products. Int. J. Syst. Bacteriol. **46**(2): 506–511. doi:10.1099/00207713-46-2-506.

Yang, G., Chen, M., Zhou, S., Liu, Z., and Yuan, Y. 2013. *Sinorhodobacter ferrireducens* gen. nov., sp. nov., a non-phototrophic iron-reducing bacterium closely related to phototrophic *Rhodobacter* species. Antonie Van Leeuwenhoek **104**(5): 715–724. doi:10.1007/s10482-013-9979-0.

Zan, J., Liu, Y., Fuqua, C., and Hill, R.T. 2014. Acyl-homoserine lactone quorum sensing in the Roseobacter clade. Int. J. Mol. Sci. **15**(1): 654–669. doi:10.3390/ijms15010654.

Zengler, K., Toledo, G., Rappe, M., Elkins, J., Mathur, E.J., Short, J.M., and Keller, M. 2002. Cultivating the uncultured. Proc. Natl. Acad. Sci. **99**(24): 15681–15686. doi:10.1073/pnas.252630999.

Zhou, Z., Alikhan, N., Sergeant, M.J., Luhmann, N., Vaz, C., Francisco, A.P., Carriço, J.A., and Achtman, M. 2018. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res. **28**(9): 1395–1404. doi:10.1101/gr.232397.117.

Appendices

Appendix A: Supplementary Data for Chapter 2

*All tables in **tsv** format can be accessed online at: https://doi.org/10.7939/DVN/K8C7RG

*All tables in **excel** format can be accessed online at: https://doi.org/10.7939/DVN/WK4IB9

Table A1: Meta-information of all isolates used in this study

Table A2: List of mono-, para- and polyphyletic genera.

Table A3: dDDH results for all within genus comparisons and ANI results for all genus were species level misclassifications exist.

Table A4: 6S rRNA gene sequence similarity, AAI and 1st, 2nd, and 3rd codon position similarities for all within and between all genera.

Table A5: 16S rRNA gene sequence similarity, AAI and 1st, 2nd, and 3rd codon position similarities for all within and between recognized monophyletic genera comparisons.

Table A6: Genomic metrics (16S rRNA gene sequence similarity, AAI and codon position similarities) for poly- and paraphyletic genera comparisons

Table A7: Genes present and absent matrix for assessing DMSP demethylation and DMSP cleavage pathways.

Table A8: Genes present and absent matrix for assessing AHL-QS pathways.

Table A9: Genome accession numbers and original publications for all 342 type strains obtained from NCBI (331 type strains used in all analyses with the 11 genomes that were subseqnetly removed based on quality filter criteria).

Table A10: Completeness and contamination results for all 342 type trains.

Appendix B: Supplementary Data for Chapter 3

*All tables in **tsv** format can be accessed online at: https://doi.org/10.7939/DVN/LAFNCM

*All tables in **excel** format can be accessed online at: https://doi.org/10.7939/DVN/6E6Z9E

Table B1: V. cholerae isolates from the Yemen cholera outbreak and neighbouring countries, as well as other isolates from different lineages.

Table B2: Meta-information for all 1,262 isolates used in this study.

Table B3: Completeness for own cgMLST scheme (which consists of 2,443 genes). All genomes with less than 90% completeness for our own cgMLST scheme were subsequently removed.

Table B4: Genome completeness information for the final set of 679 genomes. Completeness for own cgMLST scheme is represented as the percentage of the of the 2,443 core genes present in each genome.

Table B5: Allelic profiles for all isolates for the cgMLST, 2013 MLST (Octavia et al. 2013) and 2016 MLST scheme (Kirchberger et al. 2016). All missing genes are indicated as NA. (The most likely cgST are indicated in parenthesis where applicable)

Table B6: All NCBI accession numbers for isolates (where available), PubMLST IDs and link to online storage of genomes.

Appendix C: List of publications

This is the complete list of publications I was involved in. It includes articles published as part of my MSc degree and also those resulting from collaboration with other laboratories.

C1: Published articles as part of my MSc degree

1. Islam MT, **Liang K**, Im MS, Winkjer J, Busby S, Tarr CL, Boucher Y. 2018. Draft Genome Sequences of Nine *Vibrio* sp. Isolates from across the United States Closely Related to Vibrio cholerae. Microbiol Resour Announc 7:e00965-18.
    - MT wrote the manuscript. IMT, KL performed bioinformatic analysis. WJ, BS, and CLT provided samples and sequenced the strains. YB supervised the project

2. **Liang K**, Orata FD, Winkjer NS, Rowe LA, Tarr CL, 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 Announc 5:e00905-17.
    - KL and FDO performed bioinformatic analysis and wrote the manuscript. NSW, LAR, and CLT provided the strains, sequenced and assembled the genome. YB supervised the project.

3. **Liang K**, Islam MT, Hussain N, Winkjer NS, Im MS, Rowe LA, Tarr CL, Boucher Y. 2019. Draft Genome Sequences of Eight Vibrio sp. Clinical Isolates from across the United States That Form a Basal Sister Clade to *Vibrio cholerae*. Microbiol Resour

Announc 8:e01473-18.

- KL and IMT performed bioinformatic analysis and wrote the manuscript with revisions from NH. MSI, LAR, and CLT provided and sequenced the strains. YB supervisded the project.

C2: Published articles from collaboration with other laboratories

1. Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG, Himmelbach A, Mascher M, Bagnaresi P, Faccioli P, Cozzi P, Lauria M, Lazzari B, Stella A, Manconi A, Gnocchi M, Moscatelli M, Avni R, Deek J, Biyiklioglu S, Frascaroli E, Corneti S, Salvi S, Sonnante G, Desiderio F, Marè C, Crosatti C, Mica E, Özkan H, Kilian B, De Vita P, Marone D, Joukhadar R, Mazzucotelli E, Nigro D, Gadaleta A, Chao S, Faris JD, Melo ATO, Pumphrey M, Pecchioni N, Milanesi L, Wiebe K, Ens J, MacLachlan RP, Clarke JM, Sharpe AG, Koh CS, **Liang KYH**, Taylor GJ, Knox R, Budak H, Mastrangelo AM, Xu SS, Stein N, Hale I, Distelfeld A, Hayden MJ, Tuberosa R, Walkowiak S, Mayer KFX, Ceriotti A, Pozniak CJ, Cattivelli L. 2019. Durum wheat genome highlights past domestication signatures and future improvement targets. Nat Genet 51:885–895.

2. Wishart DS, Feunang YD, Marcu A, Guo AC, **Liang K**, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N, Sayeeda Z, Lo E, Assempour N, Berjanskii M, Singhal S,

Arndt D, Liang Y, Badran H, Grant J, Serra-Cayuela A, Liu Y, Mandal R, Neveu V, Pon A, Knox C, Wilson M, Manach C, Scalbert A. 2018. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46:D608–D617.