Defining a core genome multilocus sequence typing scheme for the global epidemiology of *Vibrio parahaemolyticus*

Narjol Gonzalez-Escalona[1,*], Keith A. Jolley[2], Elizabeth Reed[1], and Jaime Martinez-Urtaza[3]

[1]Center for Food Safety and Applied Nutrition, Food and Drug Administration, College Park, MD, USA, [2]Department of Zoology, University of Oxford, UK , and [3] The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, Somerset, United Kingdom.

Running title: Global epidemiology of *Vibrio parahaemolyticus* by cgMLST

Keywords: whole genome sequencing (WGS), core genome multilocus sequence typing, cgMLST, *Vibrio parahaemolyticus*, clinical, phylogenetic analysis, phylogeny, Single nucleotide polymorphism (SNP)

*Corresponding author. narjol.gonzalez-escalona@fda.hhs.gov. Mailing address, Center for Food Safety and Applied Nutrition, Food and Drug Administration, 5001 Campus Drive, College Park, MD 20740, USA

ABSTRACT

Vibrio parahaemolyticus is an important human foodborne pathogen whose transmission is associated with the consumption of contaminated seafood with a growing number of infections reported over recent years worldwide. A multilocus sequence typing (MLST) database for *V. parahaemolyticus* was created in 2008 and a large number of clones have been identified causing severe outbreaks worldwide (ST3), recurrent outbreaks in certain regions (e.g., ST36) or spreading to other regions where they are non-endemic (e.g., ST88 or ST189). The current MLST scheme uses sequences of 7 genes to generate a sequence type (ST) which results in a powerful tool for inferring the population structure of this pathogen, although with limited resolution, especially compared to pulse field gel electrophoresis (PFGE). Application of whole genome sequencing (WGS) has become routine for traceback investigations with core genome MLST (cgMLST) analysis as one of the most straightforward ways to explore complex genomic data in an epidemiological context. Therefore, there is a need to generate a new, portable, standardized, and more advanced system that provides higher resolution and discriminatory power among *V. parahaemolyticus* strains using WGS data. We sequenced 92 *V. parahaemolyticus* genomes and used the genome of strain RIMD 2210633 as reference (with a total of 4832 genes) to determine which genes were suitable for establishing a *V. parahaemolyticus* cgMLST scheme. This analysis resulted in the identification of 2254 suitable core genes for use in the cgMLST scheme. To evaluate the performance of this scheme, we performed a cgMLST analysis of 92 newly sequenced genomes plus an additional 142 strains with genomes available at NCBI. cgMLST analysis was able to distinguish related and unrelated strains including those with the same ST, clearly showing its enhanced resolution over conventional MLST analysis. It also distinguished outbreak-related from unrelated strains within

the same ST. The sequences obtained from this work were deposited and are available in the public database (http://pubmlst.org/vparahaemolyticus). Application of this cgMLST scheme to the characterization of *V. parahaemolyticus* strains provided by different laboratories from around the world will reveal the global picture of the epidemiology, spread, and evolution of this pathogen and will become a powerful tool for outbreak investigations allowing for the unambiguous comparison of strains with global coverage.

IMPORTANCE

Vibrio parahaemolyticus is an important human foodborne pathogen whose transmission is associated with the consumption of contaminated seafood. Classic typing methods for trace back or outbreak investigations have insufficient discriminatory power (MLST) or are unable to establish dispersion routes and/or evolutionary trends (PFGE). However with the establishment of a database to store the new WGS data and the new cgMLST scheme described here, the aforementioned drawbacks can be eliminated. Therefore, application of this cgMLST scheme to more *V. parahaemolyticus* strains around the world by different laboratories will facilitate a global picture of the epidemiology, spreading, and evolution of this pathogen. Finally, this cgMLST scheme will help in outbreak investigations since this database can be used for unambiguous comparison of data generated from laboratories around the world.

INTRODUCTION

Vibrio parahaemolyticus is an important human foodborne pathogen whose transmission is associated with the consumption of contaminated seafood (1). Most *V. parahaemolyticus* strains that are considered pathogenic carry genes encoding for thermostable direct hemolysin (*tdh*) and/or thermostable direct hemolysin-related hemolysin (*trh*) (2). Usually these potential pathogenic strains represent a small fraction of all environmental strains (3). In addition to these two virulence genes pathogenic *V. parahaemolyticus* carry other virulence related genes, usually located in pathogenicity islands (4-7).

V. parahaemolyticus "pandemic clonal complex" has been the dominant clone causing diseases around the world (3,8-14). The emergence and cross border spreading of strains, mostly belonging to sequence type 3- ST3, raised public health concerns regarding the possibility of a pandemic spread, an uncharacteristic trait for *V. parahaemolyticus*. It was believed that this pandemic strain was the only strain that was spreading among distant regions. However, recent findings have shown that this was not the case and other *V. parahaemolyticus* strains, belonging to diverse clonal complexes, have been spreading between Asia and other parts of the world (15-18). The dispersal routes of these strains remains uncertain at the moment but at least three different mechanisms have been identified associated with the introduction of pathogenic *V. parahaemolyticus*: ballast water, ocean currents and transport of oysters or other mollusks between regions (11,15,16).

A first glance into the population structure and diversity of *V. parahaemolyticus* populations was accomplished by the establishment of the multilocus sequence typing (MLST) scheme for *V. parahaemolyticus* (19) and a centralized database (http://pubmlst.org/vparahaemolyticus) in 2008. This MLST database has enabled researchers from around the world to compare isolates.

Currently more than 2477 strains from diverse regions of the world, belonging to 1681 STs, are available for analyses. Genetic variants identified as prevalent in the different regions of the world can be mapped to identify potential connections between populations from diverse geographical areas and delineate potential routes of dispersion. Although useful, MLST is based on sequence analysis of 7 chosen housekeeping genes and therefore lacks enough resolution when used in outbreak scenarios to discriminate related and unrelated strains at the ST level (19).

The prices for performing whole genome sequencing (WGS) have decreased dramatically during the last 5 years with genomes costing around 50 to 100 USD. Scientists have been using WGS to re-analyze historical collections of pathogens and outbreak strains, resulting in a new way of performing outbreak investigations. WGS analyses, single nucleotide polymorphism (WGS-SNP) (20-26) and core genome MLST analyses (15,16,27-33), have been used extensively for epidemiological trace back investigations of outbreaks. WGS data analyses allow us to better understand both population dynamics and the mechanisms which contribute to increased virulence among foodborne bacterial pathogens.

cgMLST schemes have already been successfully used for the analysis of different epidemiological investigations such as the two recent *V. parahaemolyticus* outbreaks in Maryland (pandemic ST3 strains in MD 2014 and a retrospective analysis of ST8 strains in MD 2010) (15,16), the identification of a novel clone of *V. parahaemolyticus* causing infections in Peru (16), or the description of an emergent *V. parahaemolyticus* pathogenic strain (ST631) causing illnesses in the North Atlantic coast of USA (34). All of the cgMLST schemes used in these analyses were custom made for each strain type and according to a specific epidemiological context where strains were very similar and shared most of the genes from the reference strain (>83%) (12,15,16,34). Therefore there is a need to generate a portable,

112 standardized, and more advanced system for the analysis of *V. parahaemolyticus* strains. Using
113 WGS data will introduce a higher level of resolution and discrimination into the study of
114 populations collected from all around the world which can be analysis using a universal cgMLST
115 scheme for *V. parahaemolyticus*.

116 To establish this universal *V. parahaemolyticus* cgMLST scheme, we sequenced 92 *V.*
117 *parahaemolyticus* genome representatives from the STs prevailing in different areas of the
118 world. We used the genome of strain RIMD 2210633 that contained 4832 total genes, as
119 reference, of which 2254 were selected to create the new *V. parahaemolyticus* cgMLST scheme
120 after analyzing those 92 genomes. Additionally, another 142 genomes, available at NCBI, were
121 included in the study to evaluate the performance of the new cgMLST scheme. The cgMLST
122 analysis was able to distinguish related and unrelated strains including those with the same ST,
123 clearly showing its enhanced resolution over the conventional MLST analysis. The sequences
124 obtained from this work were deposited and are available online in a public cgMLST *V.*
125 *parahaemolyticus* database (http://pubmlst.org/vparahaemolyticus).
126
127
128

RESULTS

Sequencing of representatives strains of *V. parahaemolyticus* for setting up the cgMLST scheme. Ninety two *V. parahaemolyticus* strains, previously used for setting up the MLST scheme for this bacterium (19), were sequenced to reach a > 25X average coverage using MiSeq (Illumina) (Table 1). Genome sequences with low coverage (< 25X) usually result in low sequencing qualities and incorrect assemblies. Forty eight additional strains previously sequenced by Ion Torrent (5), were re-sequenced by MiSeq in order to generate better quality genomes (Table 2) and to be used to validate the cgMLST scheme. *in silico* multilocus sequence typing (MLST- http://pubmlst.org/vparahaemolyticus) analysis of the *de novo* assembled contigs confirmed the identity of every *V. parahaemolyticus* strain (Table 1 and 2).

Development of a cgMLST for *V. parahaemolyticus*. The initial set up of the cgMLST for *V. parahaemolyticus* using the genome of RIMD 2210633 strain as the reference genome (4832 genes total) generated 3709 potential core gene targets for use in the cgMLST scheme after eliminating duplicated genes, truncated and accessory genes. RIMD 2210633 is a prototypic ST3 pandemic strain and was fully sequenced in 2003 using Sanger sequencing technology (35). Only core genes were used for constructing the cgMLST scheme. Of the 3709 potential core genes identified in the comparison of RIMD 2210633 strain with seven other *V. parahaemolyticus* strains (BB22OP, CDC_K4557, FDA_R31, UCM-V493, FORC_008, FORC_006, and FORC_004); only 2254 genes were present in every genome of the additional 92 *V. parahaemolyticus* strains used to define the final cgMLST scheme (Table S1). These 92 strains represented a diverse set of strains isolated from environmental and clinical sources as well as from different locations (Table 1).

Implementation of the *V. parahaemolyticus* cgMLST website. The cgMLST scheme was implemented into the BIGSdb database hosting the original MLST scheme for *V. parahaemolyticus* (http://pubmlst.org/vparahaemolyticus). This database allows for testing contigs of new *V. parahaemolyticus* genomes for the presence and typing of 2254 genes. Briefly the BIGSdb genome comparator tool performs a cgMLST analysis, which produces a color coded cgMLST output (example Table S2) facilitating comparison among isolates (see Material and Methods for specific details).

Evaluation of the cgMLST target gene set. All *V. parahaemolyticus* genomes generated in this study as well as a collection of 142 additional *V. parahaemolyticus* genomes available at NCBI (Table 2) were used to validate this cgMLST scheme (Fig 1). The average percentage of cgMLST targets called was 99.21%. Only five assembled genomes contained incomplete loci: 97-10290 (incomplete loci 2), Guillen_151_Peru (6), P310 (2), C148 (1), and HS-06-05 (7). The output of this general analysis produced an informative Excel file (Table S2) divided into different sheets with each one containing different results as explained in Materials and Methods. cgMLST analysis for the 234 genomes available at the MLST database allowed a fast phylogenetic exploration of *V. parahaemolyticus* genomes (Fig. 1), clearly differentiating strains belonging to different STs, clustering strains with same STs, and also allowed for further discrimination among strains within a specific ST.

Evaluation of the cgMLST scheme using genomes of strains belonging to four known STs (outbreak related and non-outbreak related strains). The performance of this cgMLST scheme was tested using six different sets of informative *V. parahaemolyticus* strains whose genomes were available and that clustered together in the global dataset (Fig. 1). In addition to the unique pandemic clone of *V. parahaemolyticus* identified to date (CC3), other

major groups with a relevance at local or transnational scale were also analyzed: 1) strains belonging to ST36 (CC36) (outbreak and non-outbreak related) (5,19,36) (Fig. 2B), 2) strains belonging to ST8 (CC8) that were outbreak related, MD 2010 (15) (Fig. 2C), 3) strains belonging to ST120 (CC120) from the same outbreak (Peru 2009) and recently characterized (16) (Fig. 2D), and 4) strains belonging to ST631, a new emergent clone in the east coast of USA (5,34,36) (Fig. 2E).

CC3. The first test of the new cgMLST was performed using strains belonging to the pandemic clone (CC3) using a panel of 30 strains (all ST3) epidemiologically unrelated along with some additional strains collected in the course of a single epidemiological event (typically the same outbreak), including the recently reported strains MDVP16, MDVP7, and MDVP18 that caused a small outbreak in MD in 2014 (12) (Fig. 2A).

The cgMLST analysis of the genomes identified as CC3 by MLST consistently grouped the strains according to their serotype. Strains of serotypes O3:K6, O1:Kunk, O1:K25, and O4:K68 strains were efficiently discriminated and included in independent clusters. A high level of diversity was found within each cluster, even though these strains were highly related by PFGE profiling and random amplified of polymorphic DNA (RADP). The cgMLST was highly effective in separating strains that were less related to each other (e.g., see O3:K6 group). Noteworthy, cgMLST analysis showed that the first reported outbreaks of pandemic *V. parahaemolyticus* in USA in 1998 (NY and TX), were caused by two different strains and differing by at least 14 loci from each other (detailed analysis can be found in Table S3). The strains causing the outbreak in MD in 2014 were grouped together and divergent from the original O3:K6 strains (old ST3 strains) by > 30 loci. Strains MDVP17 and 18 were

undistinguishable and differed by 1 loci from MDVP16, confirming that this outbreak in MD in 2014 was caused by a single strain.

CC36. CC36 includes strains typically causing infections in the Pacific Northwest of the USA and Canada (2,19,37). Figure 2B shows the analysis of strains belonging to CC36 from USA and Canada isolated over the last 20 years from clinical and environmental sources. The cgMLST analysis clearly separated them into two distinct groups, strains isolated before 2000 (old or classic clone ST36) and after 2000 (new clone ST36). The results of the cgMLST analysis can be found in Table S4. For illustration purposes here in this analysis, we focused on the known outbreak strains isolated in MD during the period 2012-2013. In 2012 there was an outbreak on the East Coast of USA caused by a unique ST36 clone (5,38). This clone is represented by strain MDVP12 (grouped as strain 2 in the tree). However, as can be observed during the 2013 season, the remaining MDVP strains, there were at least 3 different strains causing clinical cases during that year.

CC8. Strains belonging to this CC8 have been described primarily causing illnesses in Asia (15), however strains belonging to CC8 caused a small outbreak in MD in 2010 (15). Haendiges *et al*. (20) showed that these clinical ST8 strains were almost indistinguishable from strains isolated from oysters in MD and that they were different from other ST8 strains that were available at NCBI. Therefore, we chose these strains to test the performance of the newly developed *V. parahaemolyticus* cgMLST. Figure 2C shows the cgMLST analysis of these ST8 strains from an outbreak in MD in 2010 and their relation to two other strains isolated in Canada. The qualities of the ST8 sequences available from NCBI were under par and were not included in this analysis, because they were sequenced at low coverage and too many contigs were generated in their assembly (>300) indicative of the low quality of the sequences. The cgMLST

analysis results (Table S5) clearly indicate that all ST8 strains from the MD outbreak in 2010 clustered together (differing up to 2 loci) revealing that the outbreak was caused by the same strain and differed > 500 loci from the ST8 strains isolated in Canada in 2006 and 2007.

CC120. Strains belonging to this CC120 and that were ST120 suddenly emerged in Peru during the course of a cross-country epidemic event in 2009 causing infections in different cities throughout the country (16). Figure 2D shows the cgMLST analysis with a set of 20 strains belonging to ST120 previously characterized by another cgMLST (custom reference based) causing an outbreak of gastroenteritis in Peru in 2009 (16). Results from the cgMLST analysis (Table S6) identified 11 of the 20 strains were undistinguishable and the remaining 9 differed by 1 to 3 loci, indicating the high clonality of these strains and that they were indeed part of the same outbreak.

ST631. Strains belonging to ST631, also previously characterized by another custom made cgMLST (34), were tested with the *V. parahaemolyticus* cgMLST. These strains belong to a new emergent *V. parahaemolyticus* clone causing the second highest number of *V. parahaemolyticus* illnesses in the East Coast of USA. The cgMLST analysis results (Table S7) identified a highly clonal structure within this group (with two strains being undistinguishable – MDVP8 and 9) differing between 1-10 loci, which contrasted with the differences found when compared to ST631 strains isolated in Canada (> 22 loci) (see Table S7).

DISCUSSION

This study describes the implementation and evaluation of a cgMLST scheme for *V. parahaemolyticus* using a geographically diverse panel of *V. parahaemolyticus* strains with global coverage. A database from this study was created and is freely available online (http://pubmlst.org/vparahaemolyticus). The cgMLST scheme, consisted of 2254 target genes, was validated using 142 additional *V. parahaemolyticus* strains form diverse sources and geographical locations. The new database is a valuable and reliable tool for the unambiguous comparison of data generated from laboratories around the world.

The re-sequenced 140 genomes provided by this study to the NCBI database, encompass a diverse repertoire of strains of historical importance. These genomes were instrumental in the creation of this universal cgMLST scheme for *V. parahaemolyticus* and represent a diverse set that can be used for other research endeavors such as virulence typing, PCR detection of specific lineages, evolution, and spreading of different *V. parahaemolyticus* strains around the world (39,40). This database will allow for testing contigs of new *V. parahaemolyticus* genomes for the presence and typing of 2254 genes. The steady incorporation of new genomes into this database will improve surveillance of this important foodborne pathogen worldwide and provide early detection of new variants being introduced into new locations where they are not usually found as was shown for ST189 (17), ST3 in MD 2014(12), ST8 in MD 2010 (15), ST120 in Peru 2009 (16), among others.

The suggested analysis starts with running a default cgMLST analysis with all of the *V. parahaemolyticus* genomes available in the database and the new *V. parahaemolyticus* genomes being tested can be localized in the NeighborNet tree (Fig 1). This type of analysis allows for a fast phylogenetic examination of *V. parahaemolyticus* genomes. Then a more detailed analysis

can be produced including only the relevant strains contained in the initial tree that clustered with the *V. parahaemolyticus* genomes tested (Fig. 2). Also, two types of output of the analysis can be performed: a fast analysis output, in which only the allelic information is used and a more detailed (although slower) output , where not only are the alleles differences provided but also an alignment containing the sequences for all the variable genes (loci). The more detailed output (which is generated in order to be able to generate phylogenetic trees outside of the web site), can be used to perform additional tests such as SNP-based phylogeny reconstruction using sequence based algorithms such as Maximum likelihood (41), time of evolution (42), or to find a specific sequence signature for an specific lineage or clone.

The evaluation of this universal *V. parahaemolyticus* cgMLST was performed using five sets of strains known to be part of the same outbreak or unrelated but having the same ST. As expected, cgMLST was extremely efficient in partitioning even among the highly clonal ST3 (pandemic strains), dividing the strains causing an outbreak in USA in 1998 in two different locations (NY and TX) into two different groups (Fig. 2A). This result is in line with findings from other ongoing studies also identifying these two strains (TX and NY, 1998) having a different origin (unpublished data, personal communication JMU). Furthermore, it partitioned the pool of strains in concordance with their serotypes, with all the O3:K6 strains clustering loosely together while strains from each other serotype were grouped consistently according to the serotype. This analysis also showed that the ST3 from the outbreak in MD in 2014 (12) were almost identical strains (only 1 SNP difference in one strain among the 2254 genes analyzed) and very different from the other ST3 strains analyzed. This conclusion was not possible to arrive at previously due to the inherent problems with the sequence quality and analysis performed in the earlier publication (12).

A similar result was achieved with the other sets of strains employed for each individual analysis. Strains belonging to CC36 from USA and Canada were separated by the *V. parahaemolyticus* cgMLST analysis into two distinct groups as observed preliminarily elsewhere (5,36) with strains isolated before 2000 (classic ST36 clone) and after 2000 (new ST36 clone). It also showed that ST36 strains causing outbreak in 2013 in MD belonged at least to 3 different lineages. This example clearly shows the performance of the cgMLST for fast clustering and differentiation of strains during an outbreak. Overall MLST discriminatory power expressed by the formula of Simpson's index of diversity (D) for the genomes analyzed was D = 0.947, which shows that MLST is quite discriminatory but that is not enough to discriminate within strains of the same ST. While overall the D of cgMLST was 0.9921, showing a significantly higher discriminatory power than MLST.

This cgMLST analysis has several advantages compared to SNP based methodologies: it is rapid, reproducible, there is no need for high-performance computers or bioinformatic skills, easy visualization and location on the genome of the loci that differ between or among strains analyzed, the results can be easily transferred among different laboratories, and the information for each genome from all around the world will be stored in the database for future use. By contrast, a limitation of the cgMLST approach is that the analysis is reduced to only coding regions. Of the 4832 open reading frames (ORFs) used as reference (present in RIMD strain) only 50% is shared by the highly diverse *V. parahaemolyticus* strains used in this study, representing only a fraction of the genome. Therefore, if more detailed or enhanced resolution is needed, whole genome MLST (wgMLST) using an uploaded annotated reference of a related strain (supported within the website), or a genome-wide SNP analysis is recommended.

V. parahaemolyticus is a natural inhabitant of a wide range of marine habitats with a life cycle encompassing different stages as free living organism in seawater, as a component of the microbiota of a vast range of marine organisms, but also as a pathogen in the human gut (43). As a result of this complex style of life, this organism is extremely diverse in terms of genomic variation with a large genomic repertory which enables it to adapt and survive in different habitats under the constant variations of the environmental conditions typical of coastal areas. In addition to mutation, homologous recombination and horizontal gene transfer have been found major contributions to genomic variation in *V. parahaemolyticus* populations in the need for a rapid adaptation to new habitats under changing environmental conditions (17,19,44,45). These particular features make the phylogenetic analysis of *V. parahaemolyticus* especially challenging where the identification of the different sources contributing to genetic variation of genomes is needed. For all these reasons, the cgMLST scheme described here represents a notable advance for the genomic analysis of complex organisms such as *V. parahaemolyticus*, providing a permanent platform to store the available genomes and streamlining the analytical process with the selection of the core genes shared by all the genome and a rapid identification of the variation within each gene, without the need of dealing with complex and time-consuming bioinformatics tools and enabling a urgent response within a context of epidemiological investigation.

In conclusion, we have created a standardized cgMLST scheme that allows for fast typing of *V. parahaemolyticus* from WGS data in a publicly available database. This cgMLST scheme was tested with a diverse set of strains belonging to the same or unrelated outbreaks and was able to differentiate them accordingly, therefore showing a great potential for use in outbreak investigations. Application of this cgMLST scheme to *V. parahaemolyticus* strains collected by different laboratories around the world will help define the global picture of the

epidemiology, spreading, and evolution of this pathogen. All of this information will be critical in its application for outbreak investigations providing a unique repository of genomes that can be used for unambiguous comparisons of data generated worldwide. Finally, since *V. parahaemolyticus* is a bacteria highly intertwined with environmental changes, it is our goal to develop a tool that would be able to integrate the results obtained from the cgMLST scheme analysis of the entire database, as it continues to grow, into a geographical visualization that together with environmental variables (e.g. salinity, temperature) would help to determine worldwide dispersal rates of this pathogen, and help in modifying risk assessments for this bacteria in different regions.

MATERIALS AND METHODS

Bacterial strains and media. The *V. parahaemolyticus* strains sequenced in this study are listed, along with their assigned CFSAN numbers, in Table 1. Strains were selected based on their origin, ST and date of isolation with representatives of all the major clinical clones of *V. parahaemolyticus* prevailing in the different regions of the world. All isolates were retrieved from storage (-80ºC freezer), transferred to Luria-Bertani (LB) medium with 3% NaCl and incubated at 37 ºC with shaking at 250 rpm. Strains were confirmed in the original studies as *V. parahaemolyticus* and subsequently confirmed in this study by *in silico* MLST and *in silico* presence of a *V. parahaemolyticus* -specific gene (Vp-toxR- AB029907) in the genome.

DNA extraction and quantification. Genomic DNA from each strain was isolated from overnight cultures using the DNeasy Blood and Tissue Kit (QIAGEN, Valencia, CA). The concentration was determined using a Qubit double-stranded DNA HS assay kit and a Qubit 2.0 fluorometer (Thermo Scientific, Waltham, MA), according to each manufacturer's instructions.

Whole genome sequencing, contigs assembly and annotation. Strains were sequenced (Table 1 and some in Table 2) using an Illumina MiSeq sequencer (Illumina, CA) with 2x250 bp pair-end chemistry, according to manufacturer's instructions, > 25X average coverage. The genome libraries were constructed using Nextera XT DNA sample prep kit (Illumina). Genomic sequence contigs were *de novo* assembled using default settings within CLC Genomics Workbench v8.5.1 (QIAGEN) with a minimum contig size threshold of 500 bp in length. The draft genomes were annotated using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP, http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html) (46).

368 ***in silico* MLST phylogenetic analysis.** The initial analysis and identification of the strains
369 was performed using an *in silico V. parahaemolyticus* MLST, based on information available at
370 the *V. parahaemolyticus* MLST website (http://pubmlst.org/vparahaemolyticus/) and using
371 Ridom SeqSphere+ software v3.1.0 (Ridom; Münster, Germany)
372 (http://www.ridom.com/seqsphere). Seven loci (*dnaE, gyrB, recA, dtdS, pntA, pyrC,* and *tnaA*),
373 previously described for *V. parahaemolyticus* (19), were used for MLST analysis. The same *V.*
374 *parahaemolyticus* MLST database was also used to assign numbers for alleles and sequence
375 types (STs).
376 **cgMLST target gene definition.** The cgMLST scheme for *V. parahaemolyticus* was created
377 using Ridom Seqsphere software v3.1.0 with the genome of strain RIMD 2210633 as reference
378 (Ridom; Münster, Germany). The cgMLST scheme was composed using the cgMLST target
379 definer tool with default settings within the software. The reference genome contains 4832 genes
380 in total (35). The only seven closed *V. parahaemolyticus* genomes available at NCBI were used
381 to establish a list of core and accessory genome genes (strains: BB22OP, CDC_K4557,
382 FDA_R31, UCM-V493, FORC_008, FORC_006, and FORC_004). Core genes, genes shared by
383 all the strains queried, and accessory genes that were only present in some of the queried
384 genomes but not in all, were identified. Genes that are present in more than one copy in any of
385 the eight genomes were removed from the analysis. A genome wide gene-by-gene cgMLST
386 comparison was performed with every genome queried against the reference.
387 **Establishment of the cgMLST for *V. parahaemolyticus* website.** The *V. parahaemolyticus*
388 MLST website (http://pubmlst.org/vparahaemolyticus/) is run using the BIGSdb platform (47)
389 designed for gene-by-gene analysis of whole genome assemblies. Establishing the cgMLST
390 scheme was a matter of defining the core gene loci within the database and grouping these in to a

scheme. The first allele (allele 1) for each locus was defined from the RIMD 2210633 strain and added to the database in order to seed it. New variants of each locus were found using the BIGSdb manual web-based scan tools and automated offline allele definer. This identified new variants by performing a BLAST query of the genome assembly against a database of known alleles. New alleles were assigned automatically if they had an identity of >= 98% with an existing allele over an alignment length of >= 98% of the allele and contained an initial start codon, a final stop codon and were in frame with no internal stop codons. New alleles that did not match the above description were manually curated. Allele designations and positions for each locus in each genome assembly were recorded within the database.

Genealogical reconstructions using the cgMLST scheme. Gene-by-gene analysis was performed using the BIGSdb Genome Comparator tool (47). This analysis produced an output showing allelic variation at each locus, further categorized into loci that are: 1) variable among all strains, 2) same among all strains, 3) incomplete in some isolates; also included in the output are: 4) unique strains, 5) a distance matrix, and finally 6) the parameters used for comparison. The distance matrix generated by the analysis is based on allelic differences across the cgMLST loci, with every locus with a different allele counted as a single difference in pairwise comparisons of isolates. The genealogies were reconstructed from this distance matrix using the NeighborNet algorithm (48) implemented in SplitsTree4 (49), and were either integrated into the PubMLST website or using the desktop package.

Evaluation of the cgMLST target gene set. A collection of 142 additional *V. parahaemolyticus* genomes available at NCBI (Table 2) was used to validate the cgMLST scheme. Some of these genomes were sequenced *de novo*, because cgMLST performed best with high quality sequences: >25X coverage and without *indels* due to homopolymers or sequencing errors, that

might arise from some sequencing techniques such as 454 and Ion Torrent (Table 2). These strains have been isolated from varied sources (environmental and clinical) around the world, and constitute a diverse set of *V. parahaemolyticus* strains. Some of them belonged to the same outbreak and others belonged to the same ST but were not epidemiological related. All isolates have been previously evaluated by MLST typing (http://pubmlst.org/vparahaemolyticus). The index of discrimination or discriminatory power (D) of cgMLST and MLST was calculate using the Simpson's index of diversity as described (50).

Nucleotide sequence accession numbers. The draft genome sequences for all 129 *V. parahaemolyticus* strains used in our analyses are available in GenBank under the accession numbers listed in Table 1(92 strains) and 2 (37 strains).

ACKNOWLEDGMENTS

This project was supported by the FDA Foods Program Intramural Funds. Development of the PubMLST site is supported by the Wellcome Trust. JMU was funded through a NERC project (NE/P004121/1).

FIGURES

Figure 1. cgMLST analysis of the 234 *V. parahaemolyticus* genomes available at the *V. parahaemolyticus* MLST database using the genome comparator tool implemented within the MLST database (NeighborNet phylogenetic network). Visualization of the nexus file exported from the cgMLST analysis report in Splits Tree software (48). The names at the nodes were removed for easy visualization. The original tree with the nodes names is available as Supplementary Figure 1.

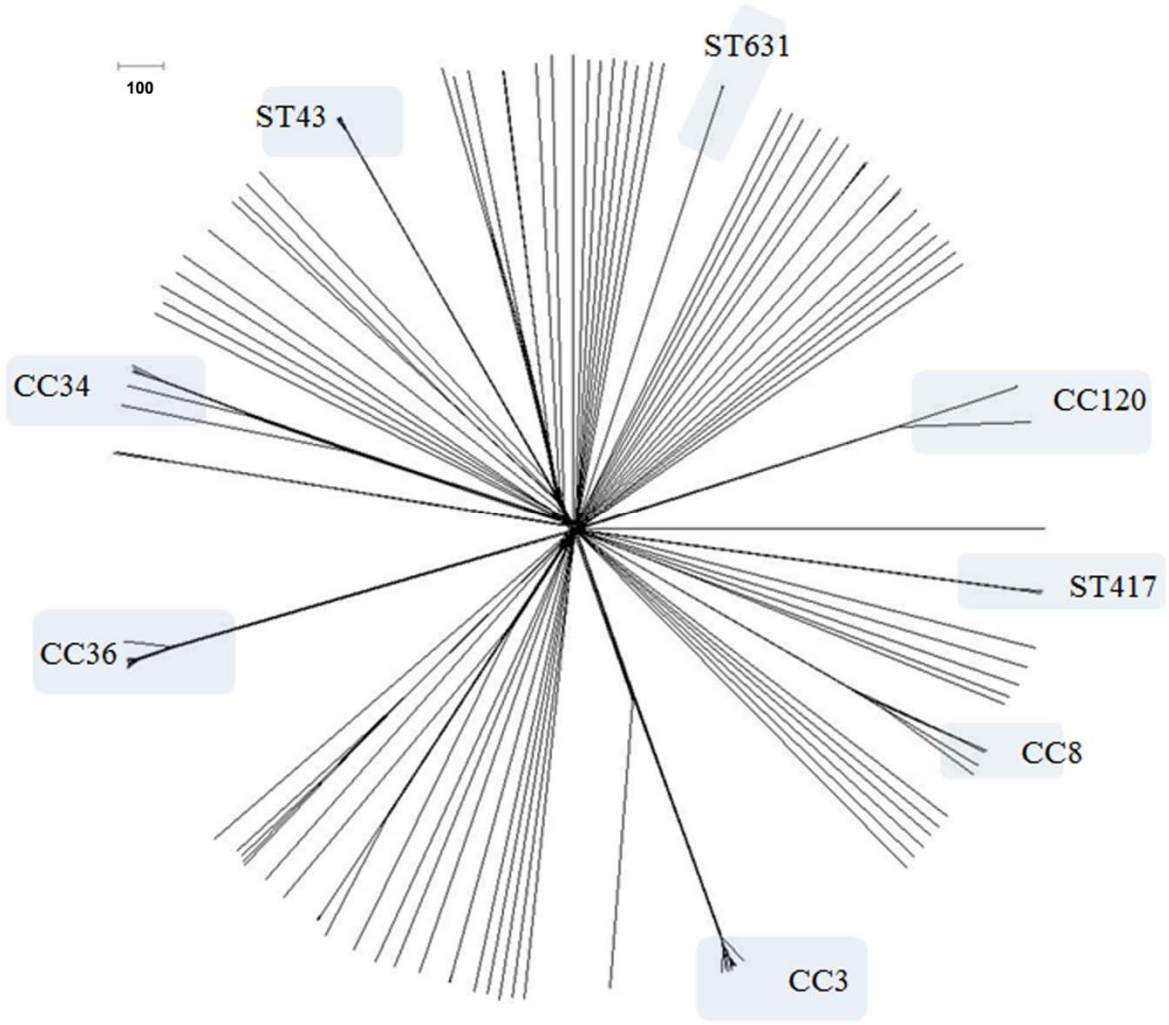
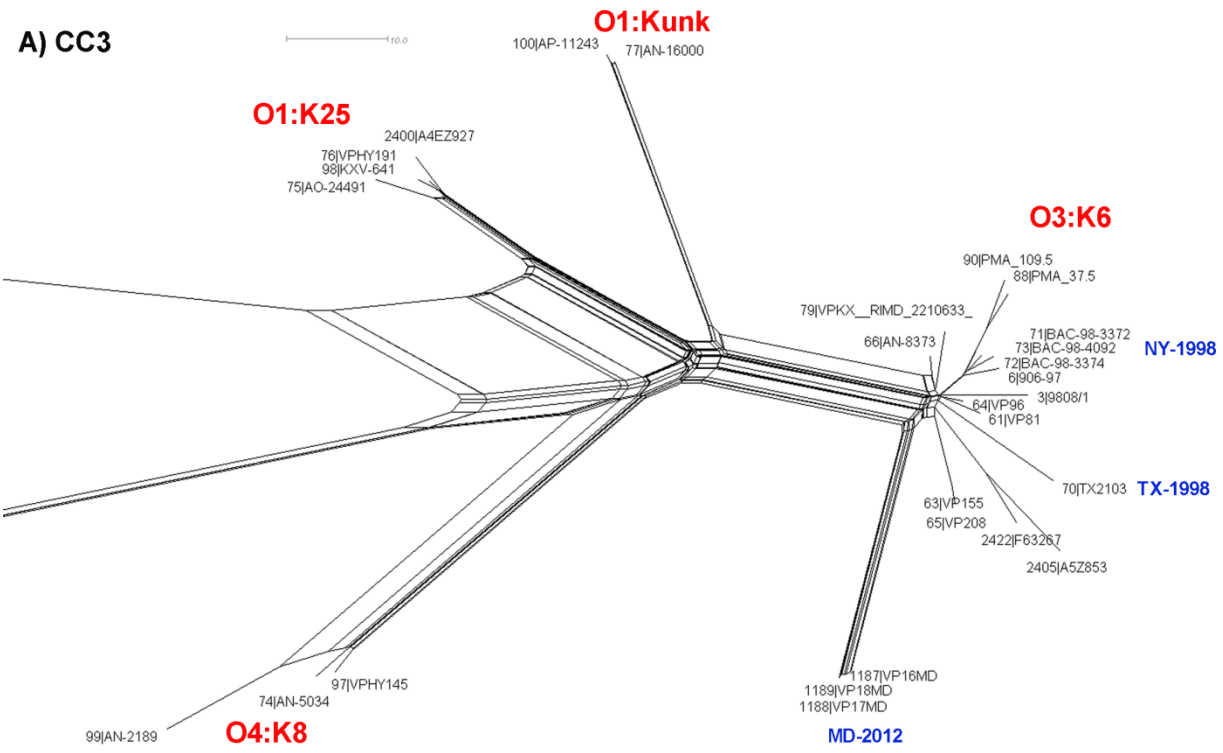


Figure 2. cgMLST analysis of representative *V. parahaemolyticus* strains from same outbreaks and/or unrelated displaying the same ST identified in Figure 1. A) CC3 outbreak (12) and no outbreak related (19), B) CC36 - ST36 strains outbreak and no outbreak related (5,19,36), C) CC8 – ST8 outbreak related and unrelated (15), D) ST417, E) CC120 outbreak strains Peru 2009 (16) –strains: 281-09, 241-09, 379-09, CO1409, CO1609, P310, Guillen_151_Peru, C226-09, C224-09, C235, PIURA_17, C237, and 239-09, were identical by cgMLST (represented by letter a), and F) ST631strains (5,34,36). The scale represents the number of allele differences.

449



450

451

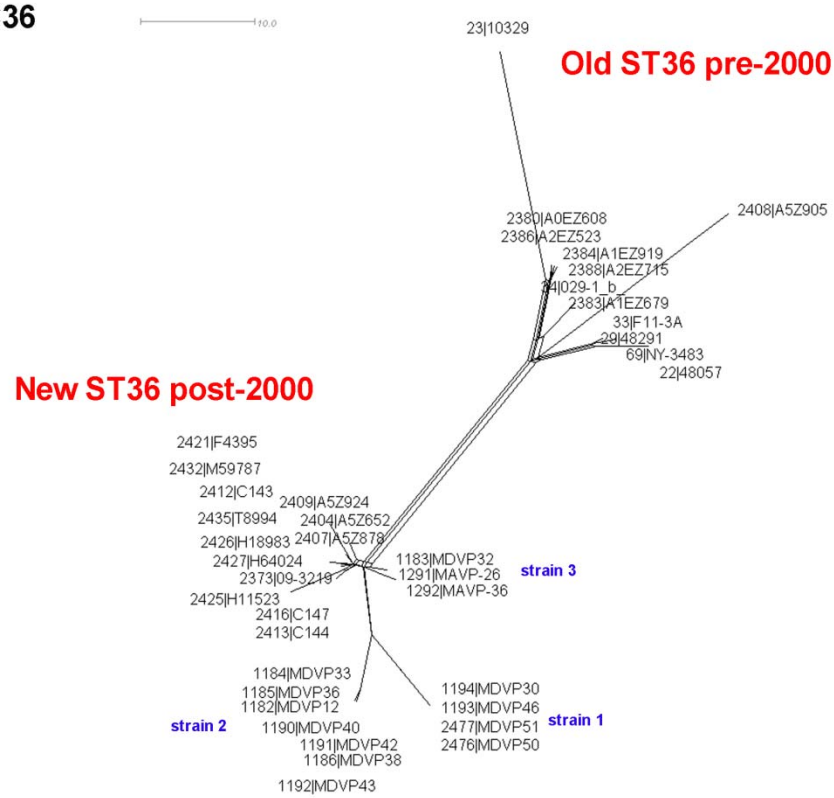
452

453

454

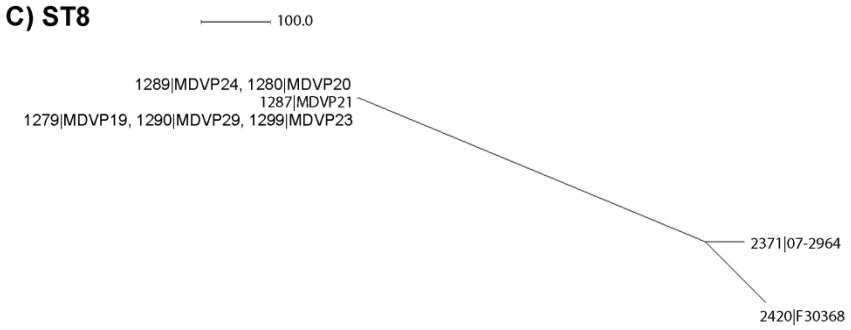
455

B) CC36


10.0
23|10329
Old ST36 pre-2000
2380|A0EZ608
2386|A2EZ523
2384|A1EZ919
2388|A2EZ715
34|029-1_b
2383|A1EZ679
33|F11-3A
29|48291
69|NY-3483
22|48057
2408|A5Z905
New ST36 post-2000
2421|F4395
2432|M59787
2412|C143
2409|A5Z924
2435|T8994
2404|A5Z652
2426|H18983
2407|A5Z878
2427|H64024
2373|09-3219
2425|H11523
2416|C147
2413|C144
1183|MDVP32
1291|MAVP-26
1292|MAVP-36
strain 3
1184|MDVP33
1185|MDVP36
1182|MDVP12
strain 2
1190|MDVP40
1191|MDVP42
1186|MDVP38
1192|MDVP43
1194|MDVP30
1193|MDVP46
2477|MDVP51
2476|MDVP50
strain 1

456

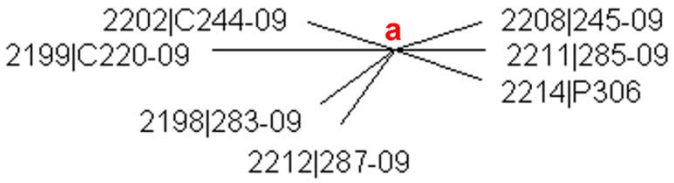
C) ST8


100.0
1289|MDVP24, 1280|MDVP20
1287|MDVP21
1279|MDVP19, 1290|MDVP29, 1299|MDVP23
2371|07-2964
2420|F30368

457

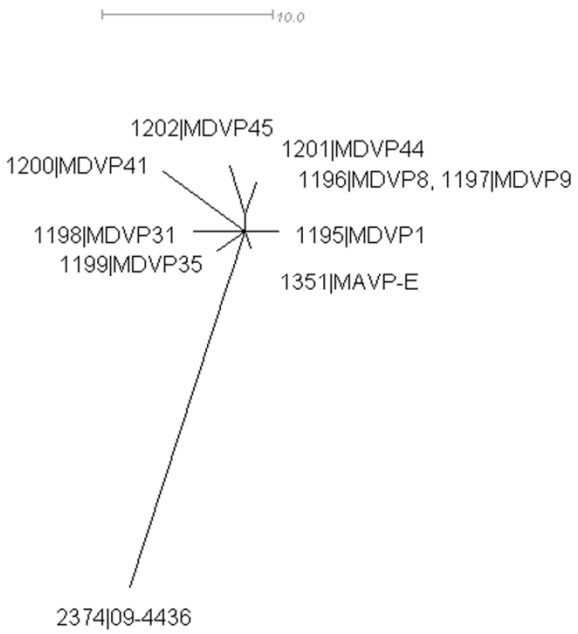
D) CC120


1.0


2202|C244-09
2199|C220-09
a
2208|245-09
2211|285-09
2214|P306
2198|283-09
2212|287-09

458

E) ST631



459

460

461 **TABLES**

462 **Table 1**. List of *V. parahaemolyticus* strains sequenced in this study.

isolate	CFSAN	year	country	source	ST	serotype	Accesion No. or SRA No. NCBI	coverage
428/00	CFSAN018752	1998	Spain	C	17	O4:K11	LHAU00000000	145
30824	CFSAN018753	1999	Spain	C	17	O4:K11	LHAV00000000	88
9808/1	CFSAN018754	2004	Spain	C	3	O3:K6	LHAW00000000	131
UCM-V441	CFSAN018755	2002	Spain	E	52	O4:Kunk	LHAX00000000	107
UCM-V586	CFSAN018756	2003	Spain	E	45	O8:k22	LHAY00000000	114
906-97	CFSAN018757	1997	Peru	C	3	O3:K6	LHAZ00000000	127
357-99	CFSAN018758	1999	Peru	C	19	O3:Kunk	LHBA00000000	148
K0976	CFSAN001174	2004	USA	E	50	O6:K18	LHBB00000000	73
K1068	CFSAN018760	2004	USA	E	61	O5:Kunk	LHBC00000000	83
K1297	CFSAN018761	2004	USA	E	12	O5:K17	LHBD00000000	102
K1314	CFSAN018762	2004	USA	E	12	O4:K63	LHBE00000000	34
K1202	CFSAN018763	2004	USA	E	43	O4:K63	LHBF00000000	115
K1322	CFSAN018764	2004	USA	E	58	O3:K56	LHBG00000000	108
K1186	CFSAN018765	2004	USA	E	58	O3:K20	LHBH00000000	72
K1296	CFSAN018766	2004	USA	E	9	O10:K68	LHBI00000000	77
K1303	CFSAN018767	2004	USA	E	20	O1:Kunk	LHBJ00000000	131
NY3547	CFSAN001172	1998	USA	E	98	O4:K55	LHQW00000000	53
ATCC 17802	CFSAN022339	1951	Japan	C	1	O1:K1	MQUE00000000	92
K1193	CFSAN022890	2004	USA	E	15	O1:K9	SRR5070562	77
K1317	CFSAN022891	2004	USA	E	23	O1:K54	SRR5070560	129
K1302	CFSAN022892	2004	USA	E	50	O1:K25	SRR5070559	50
48262	CFSAN022893	1990	USA	C	43	O1:K56	SRR5070561	93
HC-01-22	CFSAN022894	2001	USA	C	43	O4:K63	SRR5070563	78

049-2A3	CFSAN022895	1997	USA	E	57	O4:K29	SRR5070568	65
HC-01-20	CFSAN022896	2001	USA	E	199	O1:Kunk	SRR5070567	96
M25-0B	CFSAN022897	1993	USA	E	22	O4:Kunk	SRR5070565	84
HC-01-06	CFSAN022898	2001	USA	E	199	O1:Kunk	SRR5070566	37
9546257	CFSAN022899	1995	USA	C	32	O4:K8	SRR5070569	144
98-506-B102	CFSAN022900	1998	USA	E	30	O5:K11	SRR5070574	91
98-506-B103	CFSAN022901	1998	USA	E	30	O5:K11	SRR5070571	112
98-513-F51	CFSAN022902	1998	USA	E	34	O4:K9	SRR5070570	95
98-548-D11	CFSAN023517	1998	USA	E	34	O4:K9	SRR5070572	110
98-605-A9	CFSAN023518	1998	USA	E	30	O5:K17	SRR5070573	43
98-605-A10	CFSAN023519	1998	USA	E	30	O5:K17	SRR5070586	99
99-524-A9	CFSAN023520	1999	USA	E	53	O3:K34	SRR5070584	98
99-780-C12	CFSAN023521	1999	USA	E	29	O11:Kuk	SRR5070588	148
DI-B11	CFSAN023522	1999	USA	E	54	O1:K22	SRR5070587	110
DI-A8	CFSAN023523	2000	USA	E	46	O1:K30	SRR5070585	136
DI-B-6-4	CFSAN023524	2000	USA	E	47	O1:K30	SRR5070601	102
CP-B-5	CFSAN023525	2000	USA	E	23	O1:K54	SRR5070598	132
DI-B-1	CFSAN023526	2000	USA	E	23	O1:K54	SRR5070600	82
DI-A-6-1	CFSAN023527	2000	USA	E	24	O1:K55	SRR5070597	142
DI-E5	CFSAN023528	2000	USA	E	60	O1:K55	SRR5070599	79
DI-B9	CFSAN023529	1999	USA	E	25	O1:K56	SRR5070649	103
DI-H8	CFSAN023530	1999	USA	E	26	O1:K56	SRR5070650	89
DI-C2	CFSAN023531	1999	USA	E	35	O4:K9	SRR5070648	70
DI-C5	CFSAN023532	1999	USA	E	35	O4:K9	SRR5070651	65
U5474	CFSAN023549	1980	Bangladesh	C	87	O3:K6	SRR5071102	93
PMA 1.5	CFSAN023550	2005	Chile	E	28	O3:K6	SRR5071104	24
PMA 2.5	CFSAN023551	2005	Chile	E	10	O4:Kunk	SRR5071129	30
PMA 3.5	CFSAN023552	2005	Chile	E	16	O4:Kunk	SRR5071130	71

PMA 16.5	CFSAN023553	2005	Chile	E	48	O4:K12	SRR5071131	95
PMA 45.5	CFSAN023555	2005	Chile	E	49	O3:K6	SRR5071133	122
PMA 79	CFSAN023557	2004	Chile	E	56	O2:Kunk	SRR5071135	43
PMA 112	CFSAN023558	2004	Chile	E	6	O3:K6	SRR5071136	45
PMA 189	CFSAN023559	2004	Chile	E	7	O3:K6	SRR5071134	136
PMA 337	CFSAN023560	2004	Chile	E	11	O7:unk	SRR5071137	59
PMA 339	CFSAN023561	2004	Chile	E	55	O4:Kunk	SRR5071139	36
PMA 3316	CFSAN023562	2004	Chile	E	13	O3:K6	SRR5071141	73
VpHY145	CFSAN023563	1999	Thailand	C	3	O4:K68	SRR5071143	83
KXV-641	CFSAN023564	1998	Japan	C	3	O1:K25	SRR5071140	52
AN-2189	CFSAN023565	1998	Bangladesh	C	3	O4:K68	SRR5071142	80
AP-11243	CFSAN023566	2000	Bangladesh	C	3	O1:Kunk	SRR5071144	59
PMA 109.5	CFSAN023556	2005	Chile	E	3	O3:K6	SRR5071138	33
PMA 37.5	CFSAN023554	2005	Chile	E	3	O3:K6	SRR5071132	37
TX2103	CFSAN023541	1998	USA	C	3	O3:K6	SRR5071094	103
BAC-98-3372	CFSAN023542	1998	USA	C	3	O3:K6	SRR5071092	104
BAC-98-3374	CFSAN023543	1998	USA	C	42	O3:K6	SRR5071095	118
BAC-98-4092	CFSAN023544	1998	USA	C	3	O3:K6	SRR5071096	126
AN-5034	CFSAN023545	1998	Bangladesh	C	3	O4:K68	SRR5071093	85
AO-24491	CFSAN023546	1999	Bangladesh	C	3	O1:K25	SRR5071106	94
VpHY191	CFSAN023547	1999	Thailand	C	3	O1:K25	SRR5071105	108
AN-16000	CFSAN023548	1998	Bangladesh	C	3	O1:Kunk	SRR5071103	90
Vp81	CFSAN023533	1996	India	C	3	O3:K6	SRR5070652	96
Vp155	CFSAN023535	1996	India	C	3	O3:K6	SRR5071101	132
Vp96	CFSAN023536	1996	India	C	3	O3:K6	SRR5071097	92
Vp208	CFSAN023537	1997	India	C	3	O3:K6	SRR5071099	123
AN-8373	CFSAN023538	1998	Bangladesh	C	3	O3:K6	SRR5071098	100
Vp2	CFSAN023540	1998	Korea	C	3	O3:K6	SRR5071100	95

029-1(b)	CFSAN001611	1997	USA	E	36	O4:K12	JNTW02000000	104
48057	CFSAN001612	1990	USA	C	36	O4:K12	JNTX02000000	118
K1198	CFSAN001614	2004	USA	E	59	O4:K12	JNTY02000000	150
10292	CFSAN001617	1997	USA	C	50	O6:K18	JNTZ02000000	85
48291	CFSAN001618	1990	USA	C	36	O12:K12	JNUA02000000	99
F11-3A	CFSAN001619	1988	USA	E	36	O4:K12	JNUB02000000	113
NY-3483	CFSAN001620	1998	USA	C	36	O4:K12	JNUC02000000	72
K1203	CFSAN001173	2004	USA	E	59	O4:K12	JNUD02000000	47
98-513-F52	CFSAN001160	1998	USA	E	34	O4:K9	JNUE02000000	39
10290	CFSAN001613	1997	USA	C	37	O4:K12	JNUF02000000	51
JJ21-1C	CFSAN001615	1990	USA	E	38	O4:KUK	LHPD00000000	64
W9OA	CFSAN001616	1982	USA	E	59	O4:K12	LHPE00000000	39
VP43-1A	CFSAN001621	1992	USA	E	36	O4:KUK	LHQV00000000	92

C- clinical. E- environmental.

466 **Table 2**. List of *V. parahaemolyticus* genomes from NCBI used for further testing of the newly

467 created cgMLST.

isolate	CFSAN	year	country	source	ST	serotype	accession No. NCBI	reference
From our lab								
MDVP1[a]	CFSAN007429	2012	USA	C	631	unk	JNSM02000000	this study
MDVP8[a]	CFSAN007430	2012	USA	C	631	unk	JNSN02000000	this study
MDVP9[a]	CFSAN007431	2012	USA	C	631	unk	JNSO02000000	this study
MDVP31[a]	CFSAN007432	2013	USA	C	631	unk	JNSP02000000	this study
MDVP35[a]	CFSAN007433	2013	USA	C	631	unk	JNSQ02000000	this study
MDVP41[a]	CFSAN007434	2013	USA	C	631	unk	JNSR02000000	this study
MDVP44[a]	CFSAN007435	2013	USA	C	631	unk	JNSS02000000	this study
MDVP45[a]	CFSAN007436	2013	USA	C	631	unk	JNST02000000	this study
MDVP2[a]	CFSAN007437	2012	USA	C	651	unk	JNSU02000000	this study
MDVP3[a]	CFSAN007438	2012	USA	C	652	unk	JNSV02000000	this study
MDVP4[a]	CFSAN007439	2012	USA	C	653	unk	JNSW02000000	this study
MDVP34[a]	CFSAN007440	2013	USA	C	653	unk	JNSX02000000	this study
MDVP5[a]	CFSAN007441	2012	USA	C	113	unk	JNSY02000000	this study
MDVP7[a]	CFSAN007442	2012	USA	C	34	unk	JNSZ02000000	this study
MDVP11[a]	CFSAN007443	2012	USA	C	1116	unk	JNTA02000000	this study
MDVP6[a]	CFSAN007444	2012	USA	C	677	unk	JNTB02000000	this study
MDVP10[a]	CFSAN007445	2012	USA	C	43	unk	JNTC02000000	this study
MDVP13[a]	CFSAN007446	2012	USA	C	678	unk	JNTD02000000	this study
MDVP14[a]	CFSAN007447	2012	USA	C	162	unk	JNTE02000000	this study
MDVP15[a]	CFSAN007448	2012	USA	C	679	unk	JNTF02000000	this study
MDVP39[a]	CFSAN007455	2013	USA	E	896	unk	JNTL02000000	this study

090-96-70 [a]	CFSAN001595	1996	Peru	C	189a	O4:K8	JFFP02000000	this study
VP16MD [a]	CFSAN007449	2012	USA	C	3	unk	JNTG02000000	this study
VP17MD [a]	CFSAN007450	2012	USA	C	3	unk	JNTH02000000	this study
VP18MD [a]	CFSAN007451	2012	USA	C	3	unk	JNTI02000000	this study
MDVP19 [a]	CFSAN007452	2010	USA	C	8	unk	JNTJ02000000	(15)
MDVP20 [a]	CFSAN007453	2010	USA	C	8	unk	JNTK02000000	(15)
MDVP22 [a]	CFSAN007454	2010	USA	E	676	unk	JNUO02000000	(15)
MDVP25 [a]	CFSAN007456	2010	USA	E	810	unk	JNUK02000000	(15)
MDVP26 [a]	CFSAN007457	2010	USA	E	811	unk	JNUL02000000	(15)
MDVP27 [a]	CFSAN007458	2010	USA	E	34	unk	JNUM02000000	(15)
MDVP28 [a]	CFSAN007459	2010	USA	E	768	unk	JNUN02000000	(15)
MDVP21 [a]	CFSAN012491	2010	USA	E	8	unk	JNUG02000000	(15)
MDVP23 [a]	CFSAN012492	2010	USA	E	8	unk	JNUH02000000	(15)
MDVP24 [a]	CFSAN012493	2010	USA	E	8	unk	JNUI02000000	(15)
MDVP29 [a]	CFSAN012494	2010	USA	E	8	unk	JNUJ02000000	(15)
281-09 [a]	CFSAN025052	2009	Peru	C	120	O3:K59	LKQB00000000	(16)
283-09 [a]	CFSAN025053	2009	Peru	C	120	O3:K59	LKQA00000000	(16)
C220-09 [a]	CFSAN025054	2009	Peru	C	120	O3:KUT	LKQC00000000	(16)
C224-09 [a]	CFSAN025055	2009	Peru	C	120	O3:K59	LKQD00000000	(16)
C226-09 [a]	CFSAN025056	2009	Peru	C	120	O3:K59	LKQE00000000	(16)
C244-09 [a]	CFSAN025057	2009	Peru	C	120	O3:K59	LKQF00000000	(16)
C235 [a]	CFSAN025058	2009	Peru	C	120	O3:K59	LKQG00000000	(16)
PIURA 17 [a]	CFSAN025059	2009	Peru	C	120	O3:K59	LKQH00000000	(16)
C237 [a]	CFSAN025060	2009	Peru	C	120	O3:K59	LKQI00000000	(16)
239-09 [a]	CFSAN025061	2009	Peru	C	120	O3:K59	LKQJ00000000	(16)
241-09 [a]	CFSAN025062	2009	Peru	C	120	O3:K59	LKQK00000000	(16)
245-09 [a]	CFSAN025063	2009	Peru	C	120	O3:K59	LKQL00000000	(16)

CO1409 [a]	CFSAN025064	2009	Peru	C	120	O3:K59	LKQM00000000	(16)
CO1609 [a]	CFSAN025065	2009	Peru	C	120	O3:K59	LKQN00000000	(16)
285-09 [a]	CFSAN025066	2009	Peru	C	120	O3:K59	LKQO00000000	(16)
287-09 [a]	CFSAN025067	2009	Peru	C	120	O3:K59	LKQP00000000	(16)
379-09 [a]	CFSAN025068	2009	Peru	C	120	O3:K59	LKQQ00000000	(16)
P306 [a]	CFSAN029653	2009	Peru	E	120	O3:K59	LKQR00000000	(16)
Guillen 151 Peru [a]	CFSAN029654	2009	Peru	E	120	O3:K59	LKQS00000000	(16)
P310 [a]	CFSAN029656	2009	Peru	E	120	O3:K59	LKQT00000000	(16)
From other labs								
10-4287 [a]	NA	2003	Canada	C	50	O6:K18	JYJU00000000	unpublished
BB22OP[d]	NA	1995	Bangladesh	E	88	O4:K8	NC_019955.1, NC_019971.1	(51)
CDC_K4557[b]	NA	2006	USA	C	799	O1:K53	NC_021822.1, NC_021848.1	(52)
FDA_R31[b]	NA	2007	USA	E	23	O1:Kunk	NC_021847.1, NC_021821.1	(52)
RIMD 2210633[e]	NA	2003	Japan	C	3	O3:K6	NC_004605.1, NC_004603.1	(35)
FORC_008 [a,b,d]	NA	2004	South Korea	E	984	unk	NZ_CP009982.1, NZ_CP009983.1	Unpublished
UCM-V493 [a,b]	NA	2002	Spain	E	471	O2:K28	CP007004, CP007005	(53)
CHN25[d]	NA	2011	China	E	395	unk	NZ_CP010884.1, NZ_CP010883.1	Unpublished
FORC_004 [b]	NA	2014	South Korea	E	1628	unk	NZ_CP009848.1, NZ_CP009847.1	Unpublished
FORC_006 [a,b]	NA	2014	South Korea	E	1630	unk	NZ_CP009765.1,	Unpublished

FORC_014 [b]	NA	2015	South Korea	E	1629	unk	NZ_CP009766.1 NZ_CP011407.1, NZ_CP011406.1	Unpublished
KVp10 [a]	NA	2007	Sweden	E	1579	unk	MBTR01	Unpublished
R10B2_71 [a]	NA	1997	USA	E	1556	unk	MCFR01	Unpublished
04-2192 [a]	NA	2004	Canada	C	629	unk	LQCB01	Unpublished
04-2550 [a]	NA	2004	Canada	C	630	unk	LRAH01	Unpublished
05-3133 [a]	NA	2005	Canada	C	43	unk	LRAI01	Unpublished
05-4792 [a]	NA	2005	Canada	C	199	unk	LPUZ01	Unpublished
07-2964 [a]	NA	2007	Canada	C	8	unk	LRSV01	Unpublished
09-1772 [a]	NA	2009	Canada	C	417	unk	LRSX01	Unpublished
09-3219 [a]	NA	2009	Canada	C	36	unk	LRSW01	Unpublished
09-4436 [a]	NA	2009	Canada	C	631	unk	LRAJ01	Unpublished
09-4661 [a]	NA	2009	Canada	C	417	unk	LNTR01	Unpublished
09-4662 [a]	NA	2009	Canada	C	417	unk	LRTH01	Unpublished
09-4665 [a]	NA	2009	Canada	C	417	unk	LRFL01	Unpublished
09-4666 [a]	NA	2009	Canada	C	417	unk	LQCC01	Unpublished
A0EZ383 [a]	NA	2000	Canada	C	638	unk	LRSY01	Unpublished
A0EZ608 [a]	NA	2000	Canada	C	36	unk	LRFM01	Unpublished
A0EZ664 [a]	NA	2000	Canada	C	50	unk	LRFN01	Unpublished
A0EZ713 [a]	NA	2000	Canada	C	50	unk	LRFO01	Unpublished
A1EZ679 [a]	NA	2001	Canada	C	36	unk	LRSZ01	Unpublished
A1EZ919 [a]	NA	2001	Canada	C	36	unk	LNTX01	Unpublished
A1EZ952 [a]	NA	2001	Canada	C	43	unk	LRTI01	Unpublished
A2EZ523 [a]	NA	2002	Canada	C	36	unk	LRTA01	Unpublished
A2EZ614 [a]	NA	2002	Canada	C	43	unk	LRFP01	Unpublished
A2EZ715 [a]	NA	2002	Canada	C	36	unk	LRFQ01	Unpublished

A2EZ743 [a]	NA	2002	Canada	C	324	unk	LRFR01	Unpublished
A3EZ136 [a]	NA	2003	Canada	C	3	unk	LRFS01	Unpublished
A3EZ634 [a]	NA	2003	Canada	C	50	unk	LRTB01	Unpublished
A3EZ710 [a]	NA	2003	Canada	C	43	unk	LRTC01	Unpublished
A3EZ711 [a]	NA	2003	Canada	C	43	unk	LRTD01	Unpublished
A3EZ770 [a]	NA	2003	Canada	C	50	unk	LRTE01	Unpublished
A3EZ799 [a]	NA	2003	Canada	C	43	unk	LRTF01	Unpublished
A3EZ936 [a]	NA	2003	Canada	C	1060	unk	LRTG01	Unpublished
A4EZ700 [a]	NA	2004	Canada	C	43	unk	LOBT01	Unpublished
A4EZ703 [a]	NA	2004	Canada	C	141	unk	LODO01	Unpublished
A4EZ724 [a]	NA	2004	Canada	C	43	unk	LOHO01	Unpublished
A4EZ927 [a]	NA	2004	Canada	C	3	unk	LOHN01	Unpublished
A4EZ964 [a]	NA	2004	Canada	C	636	unk	LQGX01	Unpublished
A5Z1022 [a]	NA	2005	Canada	C	15	unk	LRFT01	Unpublished
A5Z273 [a]	NA	2005	Canada	C	?	unk	LQCD01	Unpublished
A5Z652 [a]	NA	2005	Canada	C	36	unk	LQCE01	Unpublished
A5Z853 [a]	NA	2005	Canada	C	3	unk	LQCF01	Unpublished
A5Z860 [a]	NA	2005	Canada	C	43	unk	LQCS01	Unpublished
A5Z878 [a]	NA	2005	Canada	C	36	unk	LQCT01	Unpublished
A5Z905 [a]	NA	2005	Canada	C	36	unk	LQCU01	Unpublished
A5Z924 [a]	NA	2005	Canada	C	36	unk	LQCV01	Unpublished
C140 [a]	NA	2008	Canada	C	332	unk	LQCW01	Unpublished
C142 [a]	NA	2008	Canada	C	417	unk	LPVA01	Unpublished
C143 [a]	NA	2008	Canada	C	36	unk	LPVB01	Unpublished
C144 [a]	NA	2008	Canada	C	36	unk	LPVC01	Unpublished
C145 [a]	NA	2008	Canada	C	417	unk	LPVK01	Unpublished
C146 [a]	NA	2008	Canada	C	1060	unk	LPVL01	Unpublished

C147 [a]	NA	2008	Canada	C	36	unk	LPVM01	Unpublished
C148 [a]	NA	2008	Canada	C	43	unk	LPVN01	Unpublished
C150 [a]	NA	2008	Canada	C	417	unk	LPVU01	Unpublished
F1419 [a]	NA	2006	Canada	C	43	unk	LRSU01	Unpublished
F30368 [a]	NA	2006	Canada	C	8	unk	LRFV01	Unpublished
F4395 [a]	NA	2006	Canada	C	36	unk	LRFU01	Unpublished
F63267 [a]	NA	2006	Canada	C	3	unk	LRFW01	Unpublished
H11523 [a]	NA	2006	Canada	C	36	unk	LRFY01	Unpublished
H18983 [a]	NA	2006	Canada	C	36	unk	LRST01	Unpublished
H64024 [a]	NA	2006	Canada	C	36	unk	LRFZ01	Unpublished
M59787 [a]	NA	2006	Canada	C	36	unk	LRJZ01	Unpublished
T8994 [a]	NA	2006	Canada	C	36	unk	LRGA01	Unpublished
W501 [a]	NA	2006	Canada	C	635	unk	LRFX01	Unpublished
HS-06-05 [a]	NA	2014	Canada	E	614	unk	LIRS01	Unpublished
ISF-29-3 [a]	NA	2011	Canada	E	1518	unk	LFYM01	Unpublished
ISF-54-12 [a]	NA	2011	Canada	E	1631	unk	LIRR01	Unpublished
S357-21 [a]	NA	2010	Canada	E	102	unk	LFYN01	Unpublished
S372-5 [a]	NA	2011	Canada	E	324	unk	LIRQ01	Unpublished
ISF-94-1 [a]	NA	2011	Canada	E	1632	unk	LIRT01	Unpublished
RM-14-5 [a]	NA	2014	Canada	E	1663	unk	LFXK01	Unpublished
Gxw_7004 [c]	NA	2007	China	C	3	unk	LPZS01	Unpublished
Gxw_9143 [c]	NA	2009	China	C	265	unk	LPZT01	Unpublished
K23 [a]	NA	2013	India	E	1052	unk	LQGU01	(54)

C- clinical. E- environmental. unk – unknown. Sequencing platform -[a]Miseq, [b]PacBio, [c]HiSeg, [d]454, and [e]Sanger.

REFERENCES

1. **Mead, P. S., L. Slutsker, V. Dietz, L. F. McCaig, J. S. Bresee, C. Shapiro, P. M. Griffin, and R. V. Tauxe**. 1999. Food-related illness and death in the United States. Emerg. Infect Dis **5**:607-625.
2. **Turner, J. W., R. N. Paranjpye, E. D. Landis, S. V. Biryukov, N. Gonzalez-Escalona, W. B. Nilsson, and M. S. Strom**. 2013. Population structure of clinical and environmental *Vibrio parahaemolyticus* from the Pacific Northwest coast of the United States. PLoS. One. **8**:e55726.
3. **DePaola, A., C. A. Kaysner, J. Bowers, and D. W. Cook**. 2000. Environmental investigations of *Vibrio parahaemolyticus* in oysters after outbreaks in Washington, Texas, and New York (1997 and 1998). Appl. Environ. Microbiol. **66**:4649-4654.
4. **Noriea, N. F., III, C. N. Johnson, K. J. Griffitt, and D. J. Grimes**. 2010. Distribution of type III secretion systems in *Vibrio parahaemolyticus* from the northern Gulf of Mexico. J. Appl. Microbiol. **109**:953-962.
5. **Haendiges, J., R. Timme, M. W. Allard, R. A. Myers, E. W. Brown, and N. Gonzalez-Escalona**. 2015. Characterization of *Vibrio parahaemolyticus* clinical strains from Maryland (2012-2013) and comparisons to a locally and globally diverse *V. parahaemolyticus* strains by whole-genome sequence analysis. Front Microbiol. **6**:125.
6. **Park, K. S., T. Iida, Y. Yamaichi, T. Oyagi, K. Yamamoto, and T. Honda**. 2000. Genetic characterization of DNA region containing the trh and ure genes of *Vibrio parahaemolyticus*. Infect. Immun. **68**:5742-5748.

7. **Boyd, E. F., A. L. Cohen, L. M. Naughton, D. W. Ussery, T. T. Binnewies, O. C. Stine, and M. A. Parent**. 2008. Molecular analysis of the emergence of pandemic *Vibrio parahaemolyticus*. BMC. Microbiol. **8**:110.

8. **Matsumoto, C., J. Okuda, M. Ishibashi, M. Iwanaga, P. Garg, T. Rammamurthy, H. C. Wong, A. DePaola, Y. B. Kim, M. J. Albert, and M. Nishibuchi**. 2000. Pandemic spread of an O3:K6 clone of *Vibrio parahaemolyticus* and emergence of related strains evidenced by arbitrarily primed PCR and *toxRS* sequence analyses. J Clin Microbiol **38**:578-585.

9. **Okuda, J., M. Ishibashi, E. Hayakawa, T. Nishino, Y. Takeda, A. K. Mukhopadhyay, S. Garg, S. K. Bhattacharya, G. B. Nair, and M. Nishibuchi**. 1997. Emergence of a unique O3:K6 clone of Vibrio parahaemolyticus in Calcutta, India, and isolation of strains from the same clonal group from Southeast Asian travelers arriving in Japan. J. Clin. Microbiol. **35**:3150-3155.

10. **Chowdhury, N. R., S. Chakraborty, T. Ramamurthy, M. Nishibuchi, S. Yamasaki, Y. Takeda, and G. B. Nair**. 2000. Molecular evidence of clonal *Vibrio parahaemolyticus* pandemic strains. Emerg. Infect. Dis **6**:631-636.

11. **Gonzalez-Escalona, N., V. Cachicas, C. Acevedo, M. L. Rioseco, J. A. Vergara, F. Cabello, J. Romero, and R. T. Espejo**. 2005. *Vibrio parahaemolyticus* diarrhea, Chile, 1998 and 2004. Emerg. Infect. Dis. **11**:129-131.

12. **Haendiges, J., M. Rock, R. A. Myers, E. W. Brown, P. Evans, and N. Gonzalez-Escalona**. 2014. Pandemic *Vibrio parahaemolyticus*, Maryland, USA, 2012. Emerg. Infect. Dis. **20**:718-720.

13. **Martinez-Urtaza, J., L. Simental, D. Velasco, A. DePaola, M. Ishibashi, Y. Nakaguchi, M. Nishibuchi, D. Carrera-Flores, C. Rey-Alvarez, and A. Pousa**. 2005. Pandemic *Vibrio parahalemolyticus* O3 : K6, Europe. Emerg. Infect. Dis **11**:1319-1320.
14. **Ansaruzzaman, M., M. Lucas, J. L. Deen, N. A. Bhuiyan, X. Y. Wang, A. Safa, M. Sultana, A. Chowdhury, G. B. Nair, D. A. Sack, L. von Seidlein, M. K. Puri, M. Ali, C. L. Chaignat, J. D. Clemens, and A. Barreto**. 2005. Pandemic serovars (O3:K6 and O4:K68) of *Vibrio parahaemolyticus* associated with diarrhea in Mozambique: spread of the pandemic into the African continent. J Clin Microbiol **43**:2559-2562.
15. **Haendiges, J., J. Jones, R. A. Myers, C. S. Mitchell, E. Butler, M. Toro, and N. Gonzalez-Escalona**. 2016. A Nonautochthonous U.S. Strain of *Vibrio parahaemolyticus* Isolated from Chesapeake Bay Oysters Caused the Outbreak in Maryland in 2010. Appl. Environ. Microbiol. **82**:3208-3216.
16. **Gonzalez-Escalona, N., R. G. Gavilan, M. Toro, M. L. Zamudio, and J. Martinez-Urtaza**. 2016. Outbreak of Vibrio parahaemolyticus Sequence Type 120, Peru, 2009. Emerg. Infect Dis. **22**:1235-1237.
17. **Gonzalez-Escalona, N., R. G. Gavilan, E. W. Brown, and J. Martinez-Urtaza**. 2015. Transoceanic Spreading of Pathogenic Strains of *Vibrio parahaemolyticus* with Distinctive Genetic Signatures in the *rec*A Gene. PLoS. One. **10**:e0117485.
18. **Martinez-Urtaza, J., C. Baker-Austin, J. L. Jones, A. E. Newton, G. D. Gonzalez-Aviles, and A. DePaola**. 2013. Spread of Pacific Northwest *Vibrio parahaemolyticus* strain. N. Engl. J. Med. **369**:1573-1574.

19. **Gonzalez-Escalona N, Martinez-Urtaza J, Romero J, Espejo RT, Jaykus LA, and Depaola A**. 2008. Determination of molecular phylogenetics of *Vibrio parahaemolyticus* strains by multilocus sequence typing. J Bacteriol **190**:2831-2840.

20. **Allard, M. W., Y. Luo, E. Strain, C. Li, C. E. Keys, I. Son, R. Stones, S. M. Musser, and E. W. Brown**. 2012. High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics **13**:32.

21. **Bakker, H. C., A. I. Switt, C. A. Cummings, K. Hoelzer, L. Degoricija, L. D. Rodriguez-Rivera, E. M. Wright, R. Fang, M. Davis, T. Root, D. Schoonmaker-Bopp, K. A. Musser, E. Villamil, H. Waechter, L. Kornstein, M. R. Furtado, and M. Wiedmann**. 2011. A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to a common *Salmonella enterica* subsp. enterica serovar Montevideo pulsed-field gel electrophoresis type. Appl. Environ Microbiol **77**:8648-8655.

22. **Chin, C.-S., J. Sorenson, J. B. Harris, W. P. Robins, R. C. Charles, R. R. Jean-Charles, J. Bullard, D. R. Webster, A. Kasarskis, P. Peluso, E. E. Paxinos, Y. Yamaichi, S. B. Calderwood, J. J. Mekalanos, E. E. Schadt, and M. K. Waldor**. 2011. The Origin of the Haitian Cholera Outbreak Strain. N Engl J Med **364**:33-42.

23. **Rasko, D. A., D. R. Webster, J. W. Sahl, A. Bashir, N. Boisen, F. Scheutz, E. E. Paxinos, R. Sebra, C. S. Chin, D. Iliopoulos, A. Klammer, P. Peluso, L. Lee, A. O. Kislyuk, J. Bullard, A. Kasarskis, S. Wang, J. Eid, D. Rank, J. C. Redman, S. R. Steyert, J. Frimodt-Moller, C. Struve, A. M. Petersen, K. A. Krogfelt, J. P. Nataro, E. E. Schadt, and M. K. Waldor**. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med **365**:709-717.

24. **Allard, M. W., Y. Luo, E. Strain, J. Pettengill, R. Timme, C. Wang, C. Li, C. E. Keys, J. Zheng, R. Stones, M. R. Wilson, S. M. Musser, and E. W. Brown**. 2013. On the evolutionary history, population genetics and diversity among isolates of *Salmonella* Enteritidis PFGE pattern JEGX01.0004. PLoS. One. **8**:e55254.

25. **Gonzalez-Escalona, N., R. Timme, B. H. Raphael, D. Zink, and S. K. Sharma**. 2014. Whole-Genome Single-Nucleotide-Polymorphism Analysis for Discrimination of *Clostridium botulinum* Group I Strains. Appl. Environ. Microbiol. **80**:2125-2132.

26. **Hoffmann, M., Y. Luo, S. R. Monday, N. Gonzalez-Escalona, A. R. Ottesen, T. Muruvanda, C. Wang, G. Kastanis, C. Keys, D. Janies, I. F. Senturk, U. V. Catalyurek, H. Wang, T. S. Hammack, W. J. Wolfgang, D. Schoonmaker-Bopp, A. Chu, R. Myers, J. Haendiges, P. S. Evans, J. Meng, E. A. Strain, M. W. Allard, and E. W. Brown**. 2016. Tracing Origins of the *Salmonella* Bareilly Strain Causing a Food-borne Outbreak in the United States. J. Infect Dis. **213**:502-508.

27. **Mellmann, A., D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S. F. McLaughlin, J. K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P. M. Brzoska, R. L. Moore, S. Guenther, J. M. Rothberg, and H. Karch**. 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS One **6**:e22751.

28. **Kovanen, S. M., R. I. Kivisto, M. Rossi, T. Schott, U. M. Karkkainen, T. Tuuminen, J. Uksila, H. Rautelin, and M. L. Hanninen**. 2014. Multilocus sequence typing (MLST) and whole-genome MLST of *Campylobacter jejuni* isolates from human

582 infections in three districts during a seasonal peak in Finland. J. Clin. Microbiol.
583 **52**:4147-4154.
584 29. **Jolley, K. A. and M. C. Maiden**. 2014. Using MLST to study bacterial variation:
585 prospects in the genomic era. Future. Microbiol. **9**:623-630.
586 30. **Schmid, D., F. Allerberger, S. Huhulescu, A. Pietzka, C. Amar, S. Kleta, R. Prager,**
587 **K. Preussel, E. Aichinger, and A. Mellmann**. 2014. Whole genome sequencing as a
588 tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011-
589 2013. Clin. Microbiol. Infect **20**:431-436.
590 31. **Kohl, T. A., R. Diel, D. Harmsen, J. Rothganger, K. M. Walter, M. Merker, T.**
591 **Weniger, and S. Niemann**. 2014. Whole-genome-based *Mycobacterium tuberculosis*
592 surveillance: a standardized, portable, and expandable approach. J. Clin. Microbiol.
593 **52**:2479-2486.
594 32. **Gonzalez-Escalona, N., M. Toro, L. V. Rump, G. Cao, T. G. Nagaraja, and J. Meng**.
595 2016. Virulence Gene Profiles and Clonal Relationships of Escherichia coli O26:H11
596 Isolates from Feedlot Cattle as Determined by Whole-Genome Sequencing. Appl.
597 Environ. Microbiol. **82**:3900-3912.
598 33. **Chen, Y., N. Gonzalez-Escalona, T. S. Hammack, M. W. Allard, E. A. Strain, and E.**
599 **W. Brown**. 2016. Core Genome Multilocus Sequence Typing for Identification of
600 Globally Distributed Clonal Groups and Differentiation of Outbreak Strains of Listeria
601 monocytogenes. Appl. Environ. Microbiol. **82**:6258-6272.
602 34. **Xu, F., N. Gonzalez-Escalona, J. Haendiges, R. A. Myers, J. Ferguson, T. Stiles, E.**
603 **Hickey, M. Moore, J. M. Hickey, C. Schillaci, L. Mank, K. DeRosia-Banick, N.**
604 **Matluk, A. Robbins, R. P. Sebra, V. S. Cooper, S. H. Jones, and C. A. Whistler**.

2016. Vibrio parahaemolyticus sequence type 631, an emerging foodborne pathogen in North America. J. Clin. Microbiol.

35. **Makino, K., K. Oshima, K. Kurokawa, K. Yokoyama, T. Uda, K. Tagomori, Y. Iijima, M. Najima, M. Nakano, A. Yamashita, Y. Kubota, S. Kimura, T. Yasunaga, T. Honda, H. Shinagawa, M. Hattori, and T. Iida**. 2003. Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of V cholerae. Lancet **361**:743-749.

36. **Xu, F., S. Ilyas, J. A. Hall, S. H. Jones, V. S. Cooper, and C. A. Whistler**. 2015. Genetic characterization of clinical and environmental Vibrio parahaemolyticus from the Northeast USA reveals emerging resident and non-indigenous pathogen lineages. Frontiers in Microbiology **6**:272.

37. **Banerjee, S. K., A. K. Kearney, C. A. Nadon, C. L. Peterson, K. Tyler, L. Bakouche, C. G. Clark, L. Hoang, M. W. Gilmour, and J. M. Farber**. 2014. Phenotypic and genotypic characterization of Canadian clinical isolates of *Vibrio parahaemolyticus* collected from 2000 to 2009. J Clin Microbiol. **52**:1081-1088.

38. **Newton, A. E., N. Garrett, S. G. Stroika, J. L. Halpin, M. Turnsek, and R. K. Mody**. 2014. Notes from the Field: Increase in *Vibrio parahaemolyticus* Infections Associated with Consumption of Atlantic Coast Shellfish - 2013. MMWR Morb. Mortal. Wkly. Rep. **63**:335-336.

39. **Baker-Austin, C., J. Trinanes, N. Gonzalez-Escalona, and J. Martinez-Urtaza**. 2016. Non-Cholera Vibrios: The Microbial Barometer of Climate Change. Trends Microbiol.

40. **Martinez-Urtaza, J., J. Trinanes, N. Gonzalez-Escalona, and C. Baker-Austin**. 2016. Is El Nino a long-distance corridor for waterborne disease? Nat. Microbiol. **1**:16018.

41. **Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar**. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**:2731-2739.

42. **Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut**. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. **29**:1969-1973.

43. **Joseph, S. W., R. R. Colwell, and J. B. Kaper**. 1982. Vibrio parahaemolyticus and related halophilic Vibrios. Crit Rev. Microbiol. **10**:77-124.

44. **Gavilan, R. G., M. L. Zamudio, and J. Martinez-Urtaza**. 2013. Molecular epidemiology and genetic variation of pathogenic *Vibrio parahaemolyticus* in Peru. PLoS. Negl. Trop. Dis. **7**:e2210.

45. **Cui, Y., X. Yang, X. Didelot, C. Guo, D. Li, Y. Yan, Y. Zhang, Y. Yuan, H. Yang, J. Wang, J. Wang, Y. Song, D. Zhou, D. Falush, and R. Yang**. 2015. Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen *Vibrio parahaemolyticus*. Mol. Biol. Evol. **32**:1396-1410.

46. **Klimke, W., R. Agarwala, A. Badretdin, S. Chetvernin, S. Ciufo, B. Fedorov, B. Kiryutin, K. O'Neill, W. Resch, S. Resenchuk, S. Schafer, I. Tolstoy, and T. Tatusova**. 2009. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Res **37**:D216-D223.

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

48. **Bryant, D. and V. Moulton**. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. **21**:255-265.

49. **Huson, D. H. and D. Bryant**. 2006. Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. **23**:254-267.

50. **Hunter, P. R. and M. A. Gaston**. 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. **26**:2465-2466.

51. **Jensen, R. V., S. M. Depasquale, E. A. Harbolick, T. Hong, A. L. Kernell, D. H. Kruchko, T. Modise, C. E. Smith, L. L. McCarter, and A. M. Stevens**. 2013. Complete Genome Sequence of Prepandemic *Vibrio parahaemolyticus* BB22OP. Genome Announc. **1**.

52. **Ludeke, C. H., N. Kong, B. C. Weimer, M. Fischer, and J. L. Jones**. 2015. Complete Genome Sequences of a Clinical Isolate and an Environmental Isolate of *Vibrio parahaemolyticus*. Genome Announc. **3**:pii: e00216-15.

53. **Kalburge, S. S., S. W. Polson, C. K. Boyd, L. Katz, M. Turnsek, C. L. Tarr, J. Martinez-Urtaza, and E. F. Boyd**. 2014. Complete Genome Sequence of *Vibrio parahaemolyticus* Environmental Strain UCM-V493. Genome Announc. **2**:pii: e00159-14.

54. **Prabhakaran, D. M., G. Chowdhury, G. P. Pazhani, T. Ramamurthy, and S. Thomas**. 2016. Draft Genome Sequence of an Environmental trh+ *Vibrio parahaemolyticus* K23 Strain Isolated from Kerala, India. Genome Announc. **4**.