4-1-2015

Use of Whole Genome Phylogeny and Comparisons in the Development of a Multiplex-PCR Assay to Identify Sequence Type 36 Vibrio parahaemolyticus

Cheryl A. Whistler
University of New Hampshire, Durham, cheryl.whistler@unh.edu

Jeffrey A. Hall
University of New Hampshire, Durham, Jeffrey.Hall@unh.edu

Feng Xu
University of New Hampshire, Durham

Saba Ilyas
University of New Hampshire, Durham

Puskar Siwakoti
University of New Hampshire, Durham

See next page for additional authors

Recommended Citation

Cheryl A. Whistler, Jeffrey A. Hall, Feng Xu, Saba Ilyas, Puskar Siwakoti, Vaughn S. Cooper, and Stephen H. Jones. Use of Whole Genome Phylogeny and Comparisons in the Development of a Multiplex-PCR Assay to Identify Sequence Type 36 Vibrio parahaemolyticus. J. Clin. Microbiol. Accepted manuscript posted online 1 April 2015, https://dx.doi.org/0.1128/JCM.00034-15

J. Clin. Microbiol. doi:10.1128/JCM.00034-15

Use of Whole Genome Phylogeny and Comparisons in the Development of a Multiplex-PCR Assay to Identify Sequence Type 36 *Vibrio parahaemolyticus*

Cheryl A. Whistler[1, 2]*, Jeffrey A. Hall[1, 2], Feng Xu[1, 2, 3], Saba Ilyas[1], Puskar Siwakoti[1], Vaughn S. Cooper[1, 2], and Stephen H. Jones[2, 4]

[1] Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, New Hampshire

[2] Northeast Center for Vibrio Disease and Ecology, University of New Hampshire, Durham, New Hampshire

[3] Genetics Graduate Program, University of New Hampshire, Durham, New Hampshire

[4] Department of Natural Resources and the Environment, University of New Hampshire, Durham, New Hampshire

*Corresponding author

E-mail: cheryl.whistler@unh.edu

Running title: Detection of ST36 *Vibrio parahaemolyticus*

ABSTRACT

Vibrio parahaemolyticus sequence type (ST) 36 strains that are native to the Pacific Ocean have recently caused multi-state outbreaks of gastroenteritis linked to shellfish harvested from the Atlantic Ocean. Whole genome comparisons of 295 genomes of *V. parahaemolyticus*, including several traced to northeastern US sources, were used to identify diagnostic loci: one putatively encoding an endonuclease (*prp*), and two others potentially conferring O-antigenic properties (*cps* and *flp*). The combination of all three loci was present only in one clade of closely-related strains, of ST36, ST59 and one additional unknown sequence type. However, each locus was also identified outside this clade, with *prp* and *flp* occurring in only two non-clade isolates, and *cps* in four. Based on the distribution of these loci in sequenced genomes, *prp* could identify clade strains with >99% accuracy, but the addition of one more locus would increase accuracy to 100%. Oligonucleotide primers targeting *prp* and *cps* were combined in a multiplex PCR method that defines species using the *tlh* locus, and determines presence of both the *tdh* and *trh* hemolysin-encoding genes which are also present in ST36. Application of the method *in vitro* to a collection of 94 clinical isolates collected over a four year period in three Northeastern US, and 87 environmental isolates, revealed the *prp* and *cps* amplicons were only detected in clinical isolates identified as belonging to the ST36-clade, and in no environmental isolates from the region. The assay should improve detection and surveillance, thereby reducing infections.

INTRODUCTION

Vibrio parahaemolyticus is typically harmless, but pathogenic strains can cause severe inflammatory gastroenteritis infections that rarely progresses to lethal sepsis (1). It is the leading cause of bacterial seafood-borne illness worldwide, with raw or improperly handled seafood as a major vector. In the United States, it has been of greatest concern for shellfish harvested in the Gulf of Mexico and the Pacific Northwest (2-5). Infections linked to shellfish from the northeastern US have been rare, but a steep rise in infections occurred in 2012-2013 that was concurrent with the probable ecological invasion of a serotype O4:K12 sequence type (ST) 36 strain. This strain type has been linked to recurrent infections in the Pacific Northwest for more than a decade, suggesting that it may have expanded its geographic range (6, 7). Furthermore, unlike native strains present in the northern Atlantic that cause infrequent infections, the ST36 strain is responsible for an ongoing multi-state outbreak (7) (Table 1). Rapid identification of this strain complex in clinical samples could aid in the prevention of more widespread infections. Additionally, accurate quantification of this strain in shellfish growing areas could inform harvest strategies that maintain a safe product.

Identifying rare pathogenic strains from among mostly nonpathogenic populations of *V. parahaemolyticus* has been a longstanding challenge. A few strains, such as those in the pandemic clonal complex, serotype O3:K6, can be identified as such through diagnostic attributes (e.g. the presence of locus ORF8), but most infections in the Americas are caused by other strains (5, 8, 9). Extensive analysis has yet to reveal a common diagnostic attribute for pathogenic *V. parahaemolyticus*. Only a

Journal of Clinical Microbiology

few virulence markers are known and routinely applied for pathogen identification, including the *tdh* and *trh* hemolysin genes, but these do not detect all pathogens. Indeed, more than 10% of infections in North America are apparently caused by strains that lack these genes, whose prevalence among non-pathogens and other *Vibrio* species is not known (10-15). Detection of a combination of traits or markers that can identify pathogen lineages of most concern along with virulence traits quickly and affordably could improve the reliability of pathogen discrimination, identification, and surveillance.

The goal of this study was to identify genomic loci diagnostic for ST36 clonal complex related strains, and to develop and apply a specific detection assay in strain identification. This assay will facilitate rapid detection of the ST36 strain complex from clinical samples and allow more targeted monitoring in natural environments.

MATERIALS AND METHODS

Bacteria strains and culture conditions. Ninety-four clinical isolates of *Vibrio parahaemolyticus* from 2010-2013 were provided by cooperating public health laboratories in Massachusetts (MA), New Hampshire (NH) and Maine (ME), only 35 of which were definitively, or deemed likely to be from northeastern US sources (Connecticut, MA, and ME), whereas the remaining 59 were either traced to other geographic locations (Canada and Virginia), from multi-source exposures with some regions outside the Northeast, or from unknown sources. Four environmental isolates from the Great Bay Estuary of NH (G61, G363, G1350, G3654) (16, 17) and ST36 strain F11-3A , a clam isolate from Washington State during an outbreak in 1997 (18) were

included in some analysis for comparison. Strains were grown in heart infusion (HI) medium (Fluka, Buchs, Switzerland) with added NaCl (3%) at 28°C (for environmental strains) or 37°C (for clinical strains) for routine culturing.

Multi-locus sequence analysis. Template genomic DNA was isolated using the Wizard Genomic DNA Purification Kit (Promega, WI, USA), using columns and manufacturer-provided recipes (Epochlifescience Inc., TX, US), or using cetyltrimethylammonium bromide protein precipitation and organic extraction (19). The *dnaE, dtdS, pntA* and *tnaA* amplicons were generated using published primers and cycling parameters (18), using Master Taq polymerase (5 PRIME, MD US), and sequenced by the Sanger method at the UNH Hubbard Center for Genome Studies or by Functional Biosciences (WI, US). Four strains from the collection of 94 clinical isolates were not included in the analysis due to failure of one or more locus to yield an amplicon, or failure of the sequencing reaction to yield sequence with homology to the expected locus. Each raw forward and reverse sequence was assembled, aligned, and trimmed to match the corresponding amplicon sequence from the public database. The allele designation for each locus and probable sequence type was identified from www.pubmlst.org. The allele combination at these four loci for ST36 was determined (dnaE: 21, dtdS: 23, pntA: 23, and tnaA: 16) and other sequence types with this combination of alleles identified as ST37 and ST39. The genome of the single ST37 isolate (10290, GCA_000454205.1) was re-analyzed using the REALPHY-generated FASTQ file (20) as an input to the short read sequence typing (SRST2) pipeline (21) to determine the sequence type. The four concatenated loci (1868 bp) for 90 clinical and 192 environmental isolates from the Northeast were aligned by CLUSTALW and used to

Journal of Clinical Microbiology

construct a Neighbor-joining phylogeny using Jukes-Cantor model using the Mega 6.0 software (22). Only three environmental isolates were deemed related sufficiently to the ST36 clade to warrant their inclusion in the *in vitro* analysis. Neighbor-joining trees were again constructed on the 90 clinical isolates and three sequenced reference strains, with statistical support assessed by 1,000 bootstrap re-assemblies.

Genome sequencing, assembly, annotation and typing. Four ST36 isolates were chosen for whole genome sequencing using an Illumina- HiSeq2500 device at the Hubbard Center for Genome Studies at the University of New Hampshire: isolate MAVP-26 is a 2013 isolate traced to oysters harvested from MA north of Cape Cod; isolate MAVP-36 is a 2013 isolate traced to oysters harvested from MA south of Cape Cod; isolate MAVP-45 is a 2013 multisource isolate traced to oysters harvested at three MA sites, two sites matching those for isolates MAVP-26 and MAVP-36, and one site on Cape Cod; and MAVP-V, is a 2011 isolate from an unknown source. Genomic DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega, WI, USA) or by a cetyltrimethylammonium bromide and organic extraction method (19). The DNA quality was assessed visually by electrophoresis. Sequencing libraries were generated from 1µg of genomic DNA as determined using the Qubit 2.0 fluorimeter (LifeTech, CA, US). DNA was sheared on the Covaris M220 Ultrasonicator to a mean size of 500 bp. Libraries were generated using the TruSeq Kit and targeted size selection of 500 bp was completed using the optional gel-extraction method in the TruSeq protocol (Illumina). Genomes were sequenced using a rapid output mode run producing 150 bp paired-ends with 249x for MAVP-26 (SAMN03107383), 238x for MAVP-36 (SAMN03107385), 355x for MAVP-45 (SAMN03177810), and 847x for MAVP-V

Journal of Clinical Microbiology

(SAMN03177809). Raw sequences were processed and *de novo* assembled using the A5 pipeline (23). The sequence types were subsequently determined using the SRST2 pipeline (21).

Whole genome comparisons and bioinformatics analysis of unique content. Because the ST36 10329 draft genome (NZ_AFBW01000001.1 - 33.1) is not closed, and is currently assembled as 33 contigs, the regions of shared and unique genome content in comparison to RIMD 2210633 (NC_004605.1, NC_004603.1) and BB22OP (NC_019955.1, NC_019971.1) for each individual contig were visualized using the BRIG program (24). Contigs harboring substantially unique content were then individually aligned with RIMD 2210633 and BB22OP using Mauve (25) to identify the coordinates of the unique regions. The coding sequences in these unique regions were subsequently annotated and ORFs identified using Prokka 1.8 using a *Vibrio*-specific database in NCBI for these annotations (26).

The distribution of identified ORFs was determined by query against all draft genomes of *V. parahaemolyticus* available at the time of this analysis (n=289) in NCBI genomes (http://www.ncbi.nlm.nih.gov/genome/691, accessed 10/12/2014). Loci identified as ORFs of sufficient size (≥1 kb) and variation to facilitate primer design that were harbored in nearly every strain in the10329 NCBI genome group, of which ST36, ST59, and ST678 are members, but virtually absent outside this genome group were selected as potential PCR targets, and were *prp* (EGF42613), *cps* (EGF42671), and *flp* (EGF42675). Each locus was further analyzed using the BLAST algorithm by query against the nucleotide collection and the non-redundant protein sequences using default

settings, to evaluate their broader distribution and potential function. The distribution of each locus was also evaluated in NCBI genomes by query against in the genus *Vibrio* (taxid: 662), excluding *Vibrio parahaemolyticus* (taxid: 691) using the default settings for blastn.

Reconstruction of whole genome phylogenies. The assembled genomes from every strain harboring one or more of the identified diagnostic loci (*prp*, *cps*, *flp*) were acquired from NCBI genomes phylogeny (http://www.ncbi.nlm.nih.gov/genome/?term=vibrio%20parahaemolyticus), which, along with MAVP-26, MAVP-36, MAVP-45, MAVP-V, were analyzed using REALPHY v. 1.09 (20). To produce the most accurate phylogeny, the analysis was then limited to the highest quality genomes (based on NCBI genomes statistics including level, number of contigs, and N50) from the 10329 genome group, along with strains from other genome groups harboring one or more loci (i.e. NIHCB0757, S159, S048), and representative strains from genome groups that were phylogenetically adjacent to group 10329 and lack any of the three loci (i.e. S120 and S100). Sequences were analyzed using 10290 (GCA_000454205.1) 10329 (NZ_AFBW01000001.1 - 33.1), and 10296 (GCA_000500105.1) as three reference ST36 strains where the alignment positions were extracted and then merged into a single alignment. Neighbor-joining phylogenies were reconstructed using the maximum likelihood method in PhyML, with a GTR substitution matrix and a gamma-distributed rate heterogeneity model (27). Phylogenies were visualized as trees using FigTree 1.4.2 (28). The branch length reflects nucleotide changes per total number of nucleotides in the sequence.

Development and application of a multiplex-PCR amplicon assay. The similar size of the *tlh* (~450 bp) and *trh* (500 bp) amplicons produced by an existing multiplex PCR assay makes their resolution challenging, especially since the length of the *tlh* gene is somewhat variable. Therefore, we sought to redesign the *tlh* PCR to improve the existing multiplex assay (9). The 44 longest published *tlh* sequences derived from *V. parahaemolyticus* were identified from NCBI. These were aligned using the MEGA 6.0 software suite (22) and used to identify regions suitable for a new forward primer with 100% sequence identity across all aligned sequences. Primer design was optimized to minimize secondary structure, to have compatible annealing temperature, and to promote minimal cross-dimerization with the other multiplex primers in the existing assay using the NetPrimer program as a tool (PREMIER Biosoft, CA, US). When used with the published reverse primer, R-TLH, the new F2-TLH primer produces an amplicon of ~401 bp (Table 1) which cannot be accurately resolved along with the ORF8 amplicon (369 bp) specific for the pandemic ST3, O3:K6 strain; however, analysis of regional isolates (Fig. 1) indicates that the pandemic strain is not prevalent among clinical isolates from the northeastern region of the US, and we reasoned that the inclusion of the ORF8 primers for routine analysis is not critical and could be applied secondarily. The F2-TLH primer was evaluated in multiplex with the R-TLH primer, and published *trh* and *tdh* primer pairs in triplicate in a three-amplicon multiplex assay on ~5µg genomic DNA as a template using AccuStart PCR Supermix (Quanta, MD, US) in 10 µl volume with an initial denaturation at 94° C for 3 minutes, followed by 30 cycles with a denaturation at 94° C for 1 minute, primer annealing at 55° C for 1 minute, and extension at 72° C for 1 minute, and with a final extension at 72° C at the completion of

Journal of Clinical Microbiology

199 the cycling for 5 minutes (9). Amplicons were evaluated by electrophoresis of 1.5 µl of
200 sample on 1.2% SeaKem LE agarose (Lonza, Rockland, ME USA) gel, with 1x Gelred
201 (Phenix Research Products, Candler, NC USA) in TAE buffer, compared with 1 Kb Plus
202 DNA Ladder (Invitrogen, Grand Island, NY USA).

203 To develop a PCR-based assay to identify ST36 and related strains, a total of 25
204 individual *prp* sequences were obtained from NCBI genomes, and aligned using the
205 MEGA 6.0 software suite (22) along with the *prp* sequences from MAVP-26, MAVP-36,
206 MAVP-V and MAVP-45 to identify highly conserved regions. Oligonucleotide primers
207 were designed to these regions with optimal amplicon size separation by
208 electrophoresis and minimal primer cross-dimerization with the existing multiplex PCR
209 primers, including the newly designed F2-TLH primer (above) and minimal secondary
210 structure which was determined as previously described. A similar strategy was used to
211 design the *cps* amplicon assay. Amplification of the *prp* and *cps* loci were evaluated in
212 individual and multiplex assays using genomic DNA from positive (F11-3A, a 1997
213 isolate from the Pacific Northwest) (18), and negative (G61, an environmental isolate
214 from NH) (16, 17) control strains using the published cycling parameters (9), and the
215 amplicons visualized as previously described.

216 **Validation of PCR amplicon assays**. To evaluate the performance of the individual
217 amplicon and multiplex PCR assays, PCR amplifications were completed with reagents,
218 cycling, and electrophoretic analysis as described previously on either ~5 µg purified
219 genomic DNA, which is used routinely for clinical and archived isolates, or on 1 µl of
220 crude lysate, which is used routinely for analysis of putative *V. parahaemolyticus*

isolates from environmental sources during high throughput isolate screening. Purified genomic DNA was obtained by using cetyltrimethylammonium bromide protein precipitation and organic extraction (19), and used as a template. Crude lysates were generated by a boiling lysis protocol (29). Briefly, cultures inoculated with a single isolated colony were grown for a minimum of 6hr or up to 24 hr in HI broth with 3% NaCl, and the cells from 1 mL pelleted by centrifugation, re-suspended in 1 mL diH_20, and lysed by boiling for 10 min. The cell debris was pelleted, and the cleared supernatant used as a template. For assay validation, we used the 94 Northeast regional clinical strains (43 of which were identified as the ST36 clade by four locus MLST, and four confirmed as ST36 based on all seven loci, see Results) and three related environmental strains (referred to hereafter as the reference set), with G61 and F11-3A as standards in each assay. Additionally, 50 environmental isolates from oysters harvested in NH (hereafter referred to as the unknown NH environmental set) or 84 environmental isolates from MA (hereafter referred to as the unknown MA environmental set) recovered on CHROMAgar Vibrio as purple colonies and cultured on T-soy agar as previously described (29), were used to further quantify the rate of false positives, and assay precision (number of replicate assays producing the same results).

The proportion of known positives that by the assay test positive, and match the result of the control template (the assay accuracy and sensitivity) of the newly designed F2-TLH primer compared to the published forward primer (F-TLH: 5′-AAAGCGGATTATGCAGAAGCACTG-3′) (9) was evaluated on crude lysates of the reference set using the published R-TLH primer in a three gene multiplex assay also using published primers for *tdh* and *trh* (Table 1), with precision (reproducibility)

Journal of Clinical Microbiology

244 determined from duplicate assays on the same sample. Both the F-TLH primer and the
245 F2-TLH primer yielded a band of the correct size from each sample (matching that from
246 standards F11-3A and G61). The rate that negatives were identified as negative
247 (specificity) of the F2-TLH primer was assessed similarly on crude lysates from the
248 unknown NH environmental set (not all of which were *V. parahaemolyticus*) where the
249 F2-TLH primer only yielded an amplicon from samples that were also amplified by the
250 F-TLH primer.

251 The accuracy, sensitivity, and specificity of three *prp* primer pairs were first
252 evaluated as a single amplicon assay on controls (MAVP-26, F11-3A, and G61) and
253 then in a four-gene multiplex (with *tlh*, *tdh* and *trh* primer pairs) (Table 1) on purified
254 DNA of a subset of the reference set (12 ST36-clade strains, and 7 non-clade strains)
255 and replicated in three separate trials to identify which primers had the best precision
256 and overall performance. The F2/R2-ST36*prp* and R3/R3-ST36*prp* primer pairs were
257 selected and tested on crude lysates of the complete reference set and the unknown
258 MA environmental set. The accuracy and sensitivity of the *cps* amplification was
259 assessed first on purified DNA from the subset of the reference set used for analysis of
260 *prp* (19 isolates), and then on all 43 isolates identified as ST36 using crude lysates in
261 the 5-gene multiplex assay using F3/R3-ST36*prp* primer pair, with precision determined
262 by replication (see Results). The range of detection (analytical sensitivity) of the F2-TLH,
263 F3/R3-ST36*prp* and F/R-ST36*cps* primer pairs was examined in a five-amplicon
264 multiplex assay (with *trh*, and *tdh*) on purified and serially diluted DNA from F11-3A as a
265 template. Visualization of all five amplicons from 1.5 µl PCR product was optimal when

266 between 50 µg and 5 ng of genomic DNA was used as a starting template, with
267 decreased but visible detection of all five amplicons as low as 50 pg.

268

269 RESULTS

270 **Identification of ST36-clade and related strains from among Northeast US clinical**
271 **and environmental isolates**

272 Although multi-locus sequence typing based on seven loci is a widely-used
273 method for *V. parahaemolyticus* strain identification (4, 17, 18, 21, 24, 30), it can be
274 cost-prohibitive, and it is not done routinely with more than a few unique isolates, or to
275 all strains of similar type associated with an outbreak (6, 7). However, a subset of only
276 four of the loci is less costly, and can sufficiently inform whether strains are related.
277 Additionally, for the loci chosen in this study, the combination of alleles in ST36 is only
278 shared with ST37 and ST39. Only two strains, one of each of these other sequence
279 types, have been reported (18, 30). Analysis of the 7 loci extracted from the draft
280 genome of the single reported ST37 isolate (10290) indicated this isolate is ST36. Thus,
281 the combination of these four alleles is only in ST36 isolates with the exception of a
282 single ST39 reported isolate suggesting most isolates with this combination of alleles
283 are ST36. We applied this four-locus multi-locus sequence analysis (MLSA) approach to
284 examine the relationships of clinical isolates from infections reported in MA, NH, and
285 ME between 2010-2013, during which time infections from the ST36 strain were first
286 reported from Atlantic sources (6). A total of 43 isolates were identical to ST36 at these
287 four loci and as such are identified as the ST36-clade (18). The relationships of these

ST36-clade isolates to 47 clinical and 192 environmental isolates from the region was determined. Three additional clinical isolates that are MAVP-46, MEVP-1, MEVP-2, and only three environmental isolates from New Hampshire, one of a previously unreported sequence type (G3654) and two ST34 isolates (G1350, and G363), were related to yet still distinct from the ST36 clade (Fig. 1) (31).

Four strains from among the Northeast clinical ST36-clade were selected for whole genome sequencing as representatives of the population. MAVP-V was isolated in 2011, predating reported infections from ST36 in the Atlantic, was not traced to a regional source, and was not part of the 2013 regional outbreak. MAVP-26, MAVP-36, MAVP-45 were isolated in 2013, were from the regional outbreak and were traced to at least two, and potentially three, different shellfish harvest sites in MA. The analysis of all seven housekeeping loci confirmed that all four of these isolates are ST36.

Comparative genomics and identification of loci of potential diagnostic utility

To identify genetic differences that are potentially useful for development of an assay to identify ST36, we performed whole genome comparisons between the published draft genome for serotype O4:K12 ST36 strain 10329 (32) and the genomes of two other pathogenic strains, which are the pandemic strain RIMD 2210633 (33) and pre-pandemic strain BB22OP (34). Few (six) coding regions in three different genome contigs appeared unique to strain 10329. We then systematically examined whether any of these regions were potentially diagnostic of ST36 based on comparisons with all draft genomes of *V. parahaemolyticus* available at the time of this analysis (289 total) in NCBI genomes (http://www.ncbi.nlm.nih.gov/genome/691, accessed 10/12/2014).

Notably, NCBI genomes places ST36 strains within genome group 10329 which harbors several sequence types, all of which share at least 92% identity. Loci were considered potentially diagnostic if they: 1) were present in virtually every sequenced isolate in the 10329 genome group; 2) were not frequently present in other, distant genome groups; and 3) were also present in all four sequenced Northeastern ST36-clade strains.

This process of elimination focused attention on two different regions of contig 10329_28. These regions likely reside on chromosome I based on homology of their flanking DNA with the reference genomes RIMD 2210633, and BB22OP. An ORF identified as a pathogenesis-related protein (locus *prp*) based on its similarity to a single annotated ORF in *Vibrio* sp. Ex25 (YP_003285914.1) was selected as a potential assay locus (Fig. 2). This locus is particularly unique in that nucleotide sequence similarity searches querying the non-redundant database in NCBI revealed no matches. A similar analysis queried against all *Vibrio* sp. draft genomes only returned similar sequences in select *V. parahaemolyticus* strains, three *Vibrio cholerae* genomes (90-97% identity), and 2 *Vibrio albensis* genomes (90% identity). Sequence similarity searches of the non-redundant database using the translated *prp* locus revealed the gene more likely encodes an endonuclease, or a DNA helicase. We propose the designation of *prp* for this locus until its function is better defined allowing accurate gene annotation. Two additional ORFs, one encoding a capsular polysaccharide (locus *cps*) and another encoding O-antigen flippase (locus *flp*) in a second region of the same contig, were chosen as assay targets due to their potential role in conferring the O4 antigenic property of the strain, which is a diagnostic trait used by some clinical laboratories. Searches using *cps* as a query returned matching sequences of similar length only in

select *V. parahaemolyticus* strains, and *Vibrio* sp. AND4 (66% identity). The *flp* locus was only in select *V. parahaemolyticus* strains, and a single *Vibrio cyclitrophicus* genome (69% identity). These three loci are conserved in ST36 strains, and have limited distribution outside the NCBI designated 10329 genome group (Table 2).

To determine the extent that one or a combination of these loci are phylogenetically informative, we examined the association of the three loci with relatedness of strains determined from whole genome phylogenies. These phylogenies were constructed with a subset of high quality genomes (see MATERIALS AND METHODS) from each NCBI genome group lineage that harbored at least one of the three loci under evaluation. The phylogeny also included a few strains that are phylogenetically closely related to (i.e. on adjacent branches with) the 10329 genome group but that lacked the loci, thereby aiding in visualization of the close relative that lacked the loci as part of this tree. Because this phylogeny is limited to fewer, high quality, and complete genomes, it utilized a higher proportion of informative sites than the BLAST phylogeny which includes a substantial number of incomplete genomes and thus excludes many informative sites (http://www.ncbi.nlm.nih.gov/genome/?term=vibrio%20parahaemolyticus) (Fig. 2). The *prp*, *cps* and *flp* loci only co-occur in a single clade of closely related- strains that are ST36, ST59 and one other unknown ST for which there was only one draft genome (vpV223/04) (Fig. 2, Table 2). A single non-ST36 high quality genome (MDVP13, ST678) in the 10329 genome group apparently lacks *prp*, and this genome harbors both *cps* and *flp*, but based on whole genome phylogeny, this strain does not group within the same clade as ST36 and ST59 (Fig. 2, Table 2, Fig. 3). Five other genomes outside

the 10329 genome group harbored one or two of the three loci but not every strain in these genome groups harbored these genes (See Table 2).

A total of 295 *V. parahaemolyticus* draft and complete genomes from isolates of a broad geographic and phylogenetic distribution were used to predict the sensitivity and specificity of these loci in strain identification. This analysis suggested that the *prp* locus, which, along with *flp* has the most limited distribution, would accurately identify *Vibrio parahaemolyticus* isolates as members of the 10329 genome group, with only a 0.3% false negative rate (only MDVP13, ST678)), and a 1% false positive rate (3 strains, including vpV223/04). Inclusion of just one additional locus (e.g. *cps*) for positive identification reduced the rate of false positives from 1% to only 0.3%; notably, the one false positive strain (vpV223/04) may fall within the 10329 genome group once analyzed in NCBI genomes as this strain has not been included in the NCBI BLAST phylogeny, but it is still closely related to ST36 and ST59 and is within the same clade (Fig. 2). These data indicate an assay utilizing *prp* may be sufficiently accurate for routine screening, but addition of a second amplicon (*cps*) and requirement of both amplicons would increase accuracy of identification of ST36-related strains to 100% if including all three sequence types within the clade harboring ST36.

Analysis of the distribution of *prp* and *cps* in clinical and environmental isolates from the Northeastern US using multiplex PCR.

To examine the utility of loci identified by whole genome comparisons for strain identification not only *in silico*, but *in vitro*, we developed a multiplex PCR-detection assay. Oligonucleotide primers that produce *prp*- and *cps*-specific amplicons were

developed for simultaneous detection with both hemolysin-encoding genes (*tdh* and *trh*) and the species specific locus (*tlh*) to improve an existing multiplex PCR assay (9) (Fig. 3, Table 1, See Methods). The single *cps* primer pair, and each of three *prp* primer pairs amplified the predicted size bands from positive control ST36 strain F11-3A but produced no bands with a reference environmental ST1125 strain G61 (data not shown). When used in a four-locus multiplex with primers that also amplify *tlh*, *trh* and *tdh*, either the F2/R2-ST36*prp* or the F3/R3-ST36*prp* primer pairs yielded bands of predicted size for all amplicons in F11-3A (data not shown). When the single *cps* primer pair was combined with the F3/R3-ST36*prp* primer pair giving optimal separation in a five-gene multiplex assay, all five amplicons were detected from F11-3A (Fig. 3). The intensity of *cps* was relatively lower, perhaps as a result of decreased efficiency for this amplicon relative to the other smaller amplicons (Fig. 3).

The above assays were applied to the reference set of 94 clinical isolates and the three most closely related environmental isolates (i.e. G1350, G363, and G3654) from the Northeast, and an unknown MA environmental set of 84 isolates for further assessment of specificity. Based on our bioinformatics analysis, we predicted the combination of *prp* and *cps* will only be associated with 10329 genome group strains, which for the reference set, are the 43 ST36-clade isolates (potentially ST36, ST37 and ST39 based on four-locus MLSA), and not in any other strain lineages from the region (Fig. 1). A four-gene multiplex assay including either the F2/R2-ST36*prp* or F3/R3-ST36*prp* primer pairs produced amplicons from all four diagnostic loci, including *prp*, in all 43 isolates that grouped within the ST36 clade, and did so reproducibly in duplicated assays (data not shown). Furthermore, the *prp* amplicon was not detected in any other

Journal of Clinical Microbiology

clinical or environmental isolate from the Northeast, including the six isolates identified as most closely related to the ST36 clade (Fig. 3, Fig. 1)). Thus, the four-amplicon assay using either the F2/R2-ST36*prp* or F3/R3-ST36*prp* primer pairs was 100% specific, accurate and precise on both purified DNA and on freshly prepared crude lysates (see MATERIALS AND METHODS). When the *cps* locus primers were included in a five-locus multiplex assay in replicated assays on either purified DNA or crude lysate including the F3/R3-ST36*prp* primer pairs, the *cps* amplicon also was detected in all 43 isolates that grouped within the ST36 clade, and in no other isolate indicating 100% accuracy and precision of the assay (Fig. 3, and data not shown).

DISCUSSION

The Northeast gastroenteritis outbreak attributed to a non-native ST36 strain of *Vibrio parahaemolyticus* in 2012 (6), with widespread infections in 2013 over multiple states, indicates the ST36 strain has established residency and continues to be a significant public health concern (7). This spurred the development of a rapid PCR-based strain identification assay informed by the extensive genome data that is now publicly available. Serotype associated genetic markers have proven useful for PCR-based identification of the pandemic *V. parahaemolyticus* ST3, serotype O3:K6; although a few O3:K6 isolates were later identified as lacking the ORF8 phage-associated gene used for typing (9, 35-37). The development of ORF8 marker-based detection strategies predates the current time when a large number of genomes are publicly available that would better inform assays, improving their specificity, or at a minimum aiding in the interpretation of results within the context of evolving pathogen

lineages. With the caveat that the quality and completeness of draft genomes vary and must guide the interpretation of results, our *in silico* comparative analysis and whole genome phylogeny indicates the *prp* locus has a very narrow distribution, is conserved in ST36 and therefore may be used for strain typing (Fig. 3). Furthermore, co-occurrence of *prp* with *cps* (or *flp*) was, without exception, restricted to and conserved in a clade of closely-related strains containing ST36, ST59, and just one other unknown sequence type for which there is only a single draft genome (Table 2, Fig. 2) suggesting the combined presence of two loci could accurately identify ST36-clade strains.

The distribution of *prp* and *cps* in an ecologically and epidemiologically relevant collection of clinical and a limited number of environmental isolates from the Northeast (see MATERIALS AND METHODS) where the ST36 strain has become prevalent among clinical samples (Fig. 1) indicates *prp* is exclusively and always detected in ST36-clade strains (see RESULTS and Fig. 3). This suggests the locus could accurately distinguish these strains from close relatives. Although not surveyed as broadly, *cps* was detected in each of the ST36-clade strains, and in none of the environmental set (See RESULTS and Fig. 3). Importantly, the accuracy, sensitivity, specificity and precision of the F2/R-THL, F3/R3-ST36*prp* and F/R-ST36*cps* primer pairs in multiplex reactions were all 100% using crude lysates of the reference and unknown set. However, high performance of any PCR-based assay requires quality samples with optimal concentration of template DNA or freshly prepared crude lysates, and skill in performing the assay to prevent cross-contamination, which can be assessed through use of proper controls and replication. Because primers were designed from alignments of these genes with regions having 100% sequence identity

(see MATERIALS AND METHODS), we anticipate that they will have high accuracy and sensitivity when applied more broadly, although some level of non-detection and false detection is still possible. Confirmation could be done by additional genotyping, such as for the *flp* locus (Fig. 2, Table 1) (16), by application of one of the other primer sets for *prp* (Table 1), through other typing methods including PFGE and serotyping (15) by four or seven-gene MLST (17, 18, 21, 30) (Fig. 1), or when resources are available by whole genome sequencing and phylogeny (Fig. 2) (20, 27). For isolates identified by this method as ST36 that are traced to regions that currently are not know to contain these as residents, some additional analysis would be warranted. Since *V. parahaemolyticus* is known to undergo recombination (4, 17) that could result in mobilization these elements to non-ST36 isolates, any isolate harboring these loci would be of considerable interest for understanding pathogen evolution.

Even though this study describes the application of this method only to a regional collection, the threat by the Pacific-native ST36 strain is not limited to the Northeast and outbreaks have also occurred in the mid-Atlantic US coast and Spain (6) suggesting this clonal complex of strains may be spreading more broadly. We anticipate the method will help determine the extent of this strain's geographic expansion beyond the Northeast, establishment of stable local populations, and the seasonal dynamics of these strains, thereby aiding in management of shellfish harvesting and reducing public health risk. The method may also be readily applied in clinical analyses to enable a more rapid response to outbreaks to prevent additional infections, and to potentially inform a laboratory diagnostic test for accurate strain identification.

ACKNOWLEDGMENTS

We are grateful for clinical strains and wish to thank specifically: Associate Commissioner S. Condon and K. Foley of the Massachusetts Department of Public Health, and M. Hickey and C. Schillaci from the Massachusetts Department of Marine Fisheries; JC Mahoney; J.K. Kanwit of the Maine Department of Marine Resources and A. Robbins of the Maine Center for Disease Control and Prevention; and K. DeRosia-Banick, Connecticut Department of Agriculture. Access to additional environmental strains was provided by M. Taylor, and assistance with genome sequencing provided by W. K. Thomas. Partial funding for this work was provided by the USDA National Institute of Food and Agriculture Hatch NH00574, NH00609 (accession 233555) and NH00625 (accession 1004199). Additional funding provided by the National Oceanic and Atmospheric Administration College Sea Grant program and grants R/CE-137, R/SSS-2, R/HCE-3. Support also provided through the National Institutes of Health 1R03AI081102-01, National Science Foundation EPSCoR IIA-1330641, and National Science Foundation DBI 1229361 NSF MRI. This is Scientific Contribution Number 2582 for the New Hampshire Agricultural Experiment Station.

LITERATURE CITED

1. **Daniels NA, MacKinnon L, Bishop R, Altekruse S, Ray B, Hammond RM, Thompson S, Wilson S, Bean NH, Griffin PM.** 2000. *Vibrio parahaemolyticus* infections in the United States, 1973–1998. J Infect Dis **181:**1661-1666.

2. **Altekruse S, Bishop R, Baldy L, Thompson S, Wilson S, Ray B, Griffin P.** 2000. *Vibrio* gastroenteritis in the US Gulf of Mexico region: the role of raw oysters. Epidemiol Infect **124:**489-495.

3. **Johnson CN, Bowers JC, Griffitt KJ, Molina V, Clostio RW, Pei S, Laws E, Paranjpye RN, Strom MS, Chen A.** 2012. Ecology of *Vibrio parahaemolyticus* and *Vibrio vulnificus* in the coastal and estuarine waters of Louisiana, Maryland, Mississippi, and Washington (United States). Appl Environ Microbiol **78:**7249-7257.

4. **Turner JW, Paranjpye RN, Landis ED, Biryukov SV, González-Escalona N, Nilsson WB, Strom MS.** 2013. Population structure of clinical and environmental *Vibrio parahaemolyticus* from the Pacific Northwest coast of the United States. PLoS ONE **8(2):**e55726 DOI:55710.51371/journal.pone.0055726.

5. **Johnson C, Flowers A, Noriea N, Zimmerman A, Bowers J, DePaola A, Grimes D.** 2010. Relationships between environmental factors and pathogenic vibrios in the northern Gulf of Mexico. Appl Environ Microbiol **76:**7076-7084.

6. **Martinez-Urtaza J, Baker-Austin C, Jones JL, Newton AE, Gonzalez-Aviles GD, DePaola A.** 2013. Spread of Pacific Northwest *Vibrio parahaemolyticus* strain. N Engl J Med **369:**1573-1574.

7. **Newton AE, Garrett N, Stroika SG, Halpin JL, Turnsek M, Mody RK, Division of Foodborne W, Environmental D.** 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.

8. **Nair GB, Ramamurthy T, Bhattacharya SK, Dutta B, Takeda Y, Sack DA.** 2007. Global dissemination of *Vibrio parahaemolyticus* serotype O3: K6 and its serovariants. Clin Microbiol Rev **20:**39-48.

9. **Panicker G, Call DR, Krug MJ, Bej AK.** 2004. Detection of pathogenic *Vibrio spp.* in shellfish by using multiplex PCR and DNA microarrays. Appl Environ Microbiol **70:**7436-7444.

10. **Klein SL, West CKG, Mejia DM, Lovell CR.** 2014. Genes similar to the *Vibrio parahaemolyticus* virulence-related genes *tdh, tlh*, and *vscC2* occur in other Vibrionaceae species isolated from a pristine estuary. Appl Environ Microbiol **80:**595-602.

11. **West CKG, Klein SL, Lovell CR.** 2013. High frequency of virulence factor genes tdh, trh, and tlh in *Vibrio parahaemolyticus* strains isolated from a pristine estuary. Appl Environ Microbiol **79:**2247-2252.

12. **Honda T, Iida T.** 1993. The pathogenicity of *Vibrio parahaemolyticus* and the role of the thermostable direct haemolysin and related haemolysins. Rev Med Microbiol **4:**106-113.

13. **Hiyoshi H, Kodama T, Iida T, Honda T.** 2010. Contribution of *Vibrio parahaemolyticus* virulence factors to cytotoxicity, enterotoxicity, and lethality in mice. Infect Immun **78:**1772-1780.

14. **Jones JL, Lüdeke CH, Bowers JC, Garrett N, Fischer M, Parsons MB, Bopp CA, DePaola A.** 2012. Biochemical, serological, and virulence characterization of clinical and oyster *Vibrio parahaemolyticus* isolates. J Clin Microbiol**:**JCM.00196-00112.

15. **Banerjee SK, Kearney AK, Nadon CA, Peterson C-L, Tyler K, Bakouche L, Clark CG, Hoang L, Gilmour MW, Farber JM.** 2014. Phenotypic and Genotypic Characterization of Canadian Clinical Isolates of *Vibrio parahaemolyticus* Collected from 2000 to 2009. J Clin Microbiol **52:**1081-1088.

16. **Mahoney JC, Gerding MJ, Jones SH, Whistler CA.** 2010. Comparison of the pathogenic potentials of environmental and clinical *Vibrio parahaemolyticus* strains indicates a role for temperature regulation in virulence. Appl Environ Microbiol **76:**7459-7465.

17. **Ellis CN, Schuster BM, Striplin MJ, Jones SH, Whistler CA, Cooper VS.** 2012. Influence of seasonality on the genetic diversity of *Vibrio parahaemolyticus* in New Hampshire shellfish waters as determined by multilocus sequence analysis. Appl Environ Microbiol **78:**3778-3782.

18. **González-Escalona N, Martinez-Urtaza J, Romero J, Espejo RT, Jaykus L-A, DePaola A.** 2008. Determination of molecular phylogenetics of *Vibrio parahaemolyticus* strains by multilocus sequence typing. J Bacteriol **190:**2831-2840.

19. **Ausubel F, R. Brent, R.E. Kingston, D/ D/ Moore, J.G. Seidman, J.A. Smith, and K. Struhl.** 1990 Current protocols in molecular biology. , p 4648. Wiley and Sons, Inc., New York, NY.

20. **Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E.** 2014. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol Biolo Evol **31:**1077-1088.

21. **Inouye M, Conway TC, Zobel J, Holt KE.** 2012. Short read sequence typing (SRST): multi-locus sequence types from short reads. BMC Genomics **13:**338.

22. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol **30:**2725-2729.

23. **Tritt A, Eisen JA, Facciotti MT, Darling AE.** 2012. An integrated pipeline for *de novo* assembly of microbial genomes. PLoS ONE **7:**e42304 DOI: 42310.41371/journal.pone.0042304.

24. **Alikhan N-F, Petty NK, Zakour NLB, Beatson SA.** 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics **12:**402 DOI:410.1186/1471-2164-1112-1402.

25. **Darling AC, Mau B, Blattner FR, Perna NT.** 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res **14:**1394-1403.

26. **Seemann T.** 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics**:**DOI:10.1093/bioinformatics/btu1153.

27. **Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59:**307-321.

28. **Rambaut A.** 2012. FigTree v1. 4. Accessed Dec.5th.

29. **Schuster BM, Tyzik AL, Donner RA, Striplin MJ, Almagro-Moreno S, Jones SH, Cooper VS, Whistler CA.** 2011. Ecology and genetic structure of a northern temperate *Vibrio cholera*e population related to toxigenic isolates. Appl Environ Microbiol **77:**7568-7575.

30. **Jolley KA, Chan M-S, Maiden MC.** 2004. mlstdbNet–distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics **5:**86.

31. **Xu F, Ilyas S, Hall JA, Jones SH, Cooper VS, Whistler CA.** 2015. Genetic characterization of clinical and environmental *Vibrio parahaemolyticus* from the Northeast USA reveals emerging resident and non-indigenous pathogen lineages. Front Microbiol **6:**272. doi: 10.3389/fmicb.2015.00272

32. **Gonzalez-Escalona N, Strain E, De Jesús A, Jones J, DePaola A.** 2011. Genome sequence of the clinical O4: K12 serotype *Vibrio parahaemolyticus* strain 10329. J Bacteriol **193:**3405-3406.

33. **Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, Iijima Y, Najima M, Nakano M, Yamashita A.** 2003. Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of V cholerae. The Lancet **361:**743-749.

34. **Jensen RV, DePasquale SM, Harbolick EA, Hong T, Kernell AL, Kruchko DH, Modise T, Smith CE, McCarter LL, Stevens AM.** 2013. Complete genome sequence of prepandemic *Vibrio parahaemolyticus* BB22OP. Genome announcements **1:**e00002-00012.

35. **Nasu H, Iida T, Sugahara T, Yamaichi Y, Park K-S, Yokoyama K, Makino K, Shinagawa H, Honda T.** 2000. A Filamentous Phage Associated with Recent PandemicVibrio parahaemolyticus O3: K6 Strains. J Clin Microbiol **38:**2156-2161.

36. **Okura M, Osawa R, Iguchi A, Arakawa E, Terajima J, Watanabe H.** 2003. Genotypic analyses of Vibrio parahaemolyticus and development of a pandemic group-specific multiplex PCR assay. J Clin Microbiol **41:**4676-4682.

37. **Myers ML, Panicker G, Bej AK.** 2003. PCR detection of a newly emerged pandemic Vibrio parahaemolyticus O3: K6 pathogen in pure cultures and seeded waters from the Gulf of Mexico. Appl Environ Microbiol **69:**2194-2200.

FIGURE LEGENDS**:**

Figure 1. Identification of ST36 clade strains from among northern New England clinical isolates of *V. parahaemolyticus*. The relationships of ninety clinical isolates reported in the Northeast between 2010 and 2013, each with a unique, assigned identifier including reporting state (MA, NH, and ME) VP and a letter (MA isolates prior to 2013) or number (all other isolates) was evaluated by a consensus neighbor-joining tree constructed from four concatenated housekeeping gene loci that are *dnaE, dtdS, pntA,* and *tnaA* sequences (1868 bp) by using a Jukes-Cantor model, with statistical support assessed by 1,000 bootstrap re-assemblies. Three well-characterized strains with complete or draft genomes (RIMD 2210633, BB22OP, and 10290) were included for reference. The bar indicates 0.2% divergences, and branches with less than 70% bootstrap support are unlabeled.

Figure 2: **Distribution of potentially diagnostic loci in ST36 and related draft genomes**. Genome sequence alignment based phylogenies using 10290, 10329, and 12310 as references were reconstructed using REALPHY v1.09 with a representative

sub-set of sequenced isolates where the merged alignment represents 75% coverage of sites of the largest reference genome (10290). The distribution of each of three potentially diagnostic loci based on queries against *V. parahaemolyticus* in NCBI genomes is represented by (+) for gene present, and (-) for gene absent. The distribution of these loci in all available draft genomes is indicated in Table 2. *Isolate VP-2007-007 was identified as ST306 using the SRST2 program (21).

Figure 3: **Improved multiplex PCR assay for identification of ST36 *Vibrio parahaemolyticus***. The presence of virulence-associated *tdh* and *trh* amplicons, strain-associated *prp* (using F3/R3ST36*prp* primers) and *cps* amplicons, and the species specific marker *tlh* on seven Northeast ST36 clade members, and four isolates identified from adjacent, related clades with F11-3A and G61 as controls, using published and newly designed primers (Table 1).

TABLES:

TABLE 1: Oligonucleotide primers used for amplification by PCR

Gene /locus	Primer sequence	Amplicon size (bp)	Source	Use in PCR[1]
tlh	F2: AGAACTTCATCTTGATGACACTGC R: GCTACTTTCTAGCATTTTCTCTGC	401	This Study (9)	M
tdh	F: GTAAAGGTCTCTGACTTTTGGAC R: TGGAATAGAACCTTCATCTTCACC	269	(9)	M
trh	F: CATAACAAACATATGCCCATTTCCG R: TTGGCTTCGATATTTTCAGTATCT	500	(9)	M
ST36*prp*	F: CGGCTTGAGTTTTCGTCATT R: CCACACCTGCTGGTTATTTAGTTC	609	This Study	S
ST36*prp*	F2: TGCGGAATCTGATCTTTATCCTC R2: AACTGTTGGGTCTTCGTCTAACC	1028	This Study	M
ST36*prp*	F3: CCCGAGGCACATCTTCACC R3: TAAACCACTAACATCTTCATCTACC	699	This Study	M
ST36*cps*	F1: TTGAGAATTACTTCCGATTATGTAGA R1: TAAACGCATTAGCGAATAGTGC	889	This Study	M
ST36*flp*	F1: TGGTTGTGTTTAGAGCAGGG R1: TGTTGGTAATACGATAAGAATGAGA	747	This Study	M

[1]Application in PCR is either compatible in multiplex (M) or only useful for single gene amplification (S)

TABLE 2. Distribution of diagnostic loci in all draft genomes of *Vibrio parahaemolyticus*[a]

Strain	NCBI Genome Group[b]	Sequence Type	*Prp*	*cps*	*flp*	Isolation Location[c]	Source[d]	Year[e]
vpV223/04	n/a	Unk	+	+	+	n/a	n/a	n/a
vpS038	10329	59	+	+	+	USA	E	1982
K1203	10329	59	+	+	+	AK	E	2004
K1198	10329	59	+	+	+	AK	E	2004
MDVP12	10329	36	+	+	+	MD	C	2012
MDVP30	10329	36	+	+	+	MD	C	2013
MDVP32	10329	36	+	+	+	MD	C	2013
MDVP33	10329	36	+	+	+	MD	C	2013
MDVP36	10329	36	+	+	+	MD	C	2013
MDVP38	10329	36	+	+	+	MD	C	2013
MDVP40	10329	36	+	+	+	MD	C	2013
MDVP42	10329	36	+	+	+	MD	C	2013
MDVP43	10329	36	+	+	+	MD	C	2013
MAVP-36	10329	36	+	+	+	MA	C	2013
MAVP-26	10329	36	+	+	+	MA	C	2013
MAVP-45	10329	36	+	+	+	MA	C	2013
MAVP-V	10329	36	+	+	+	MA	C	2011
12310	10329	36	+	+	+	WA	C	2006
vp3256	10329	36	+	+	+	USA	C	2007
F11-3A	10329	36	+	+	+	WA	E	1988
48291	10329	36	+	+	+	WA	C	1990
10296	10329	36	+	+	+	WA	C	1997
NY-3483	10329	36	+	+	+	NY	E	1998
029-1(b)	10329	36	+	+	+	OR	E	1997
10290	10329	36	+	+	+	WA	C	1997
48057	10329	36	+	+	+	WA	C	1990
10329	10329	36	+	+	+	WA	C	1998
CFSAN007462	10329	36	+	+	+	MD	C	2013
vpS037	10329	36	+	+	+	USA	C	1994
MDVP13	10329	678	-	+	+	MD	C	2012
vpS058	NIHCB0757	143	-	+	+	Japan	C	1970
Vp970107[f]	S159	43	-	+	-	USA	C	1997
MDVP28	S159	768	-	+	-	USA	E	2010
vpS048	S048	322	+	-	-	USA	E	1997
FIM-S1392	SNUVpS-1	Unk	+	-	-	Mexico	E	2014
10292	S129	50	-	-	-	WA	C	1997
MDVP2	S129	651	-	-	-	MD	C	2012
MDVP39	S129	896	-	-	-	MD	C	2013
VP2007-007	S100	307	-	-	-	USA	E	2007

[a] Presence (+) or absence (-) of each locus was determined for all high quality draft genomes. For high quality genomes which had no sequence type publicly identified, the sequence type was identified using the SRST2 program (21); Unk: sequence type is not known due to new sequence type or incomplete sequences at the 7 loci, n/a: Information was unavailable.

[b]NCBI genome groups are determined from: http://www.ncbi.nlm.nih.gov/genome/691

[c]Location of reported infection or isolation by US state; [d]source identified as clinical (C) or environmental (E); [e]year of isolation; [f] only partial coding sequence for *cps* identified from this genome