JCM
Journals.ASM.org

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

Cheryl A. Whistler,[a,b] Jeffrey A. Hall,[a,b] Feng Xu,[a,b,c] Saba Ilyas,[a] Puskar Siwakoti,[a] Vaughn S. Cooper,[a,b] Stephen H. Jones[b,d]

Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, New Hampshire, USA[a]; Northeast Center for Vibrio Disease and Ecology, University of New Hampshire, Durham, New Hampshire, USA[b]; Genetics Graduate Program, University of New Hampshire, Durham, New Hampshire, USA[c]; Department of Natural Resources and the Environment, University of New Hampshire, Durham, New Hampshire, USA[d]

***Vibrio parahaemolyticus* sequence type 36 (ST36) strains that are native to the Pacific Ocean have recently caused multistate 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 U.S. 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 in only 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 nonclade isolates and *cps* in four. Based on the distribution of these loci in sequenced genomes, *prp* identified clade strains with >99% accuracy, but the addition of one more locus increased 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 the 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 4-year period in three northeastern U.S. states and 87 environmental isolates revealed that the *prp* and *cps* amplicons were detected only 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.**

Vibrio parahaemolyticus is typically harmless, but pathogenic strains can cause severe inflammatory gastroenteritis infections that rarely progress 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 United States have been rare, but a steep rise in infections occurred in 2012 to 2013 that was concurrent with the probable ecological invasion of a serotype O4:K12 sequence type 36 (ST36) 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 multistate outbreak (7) (Fig. 1). Rapid identification of this strain complex in clinical samples might 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 among mostly nonpathogenic populations of *V. parahaemolyticus* has been a long-standing challenge. A few strains, such as those in the pandemic clonal complex serotype O3:K6 can be identified as such through certain diagnostic attributes (e.g., the presence of locus open reading frame 8 [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 few virulence markers are known and routinely applied to pathogen identification, including the *tdh* and *trh* hemolysin genes, but these do not detect all pathogens. Indeed, >10% of infections in North America are apparently caused by strains that lack these genes, whose prevalence among nonpathogens and other *Vibrio* species is not known (10–15). The detection of a combination of traits or markers that can identify pathogen lineages of most concern, along with quickly and affordably identifying virulence traits, could improve the reliability of pathogen discrimination, identification, and surveillance.

The goal of this study was to identify the genomic loci diagnostic for ST36 clonal complex-related strains and to develop and apply a specific detection assay for use 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

Bacterial strains and culture conditions. Ninety-four clinical isolates of *V. parahaemolyticus* collected from 2010 to 2013 were provided by cooperating public health laboratories in Massachusetts, New Hampshire, and

Received 8 January 2015 Returned for modification 1 February 2015
Accepted 16 March 2015
Accepted manuscript posted online 1 April 2015
Citation Whistler CA, Hall JA, Xu F, Ilyas S, Siwakoti P, Cooper VS, Jones SH. 2015. Use of whole-genome phylogeny and comparisons for development of a multiplex PCR assay to identify sequence type 36 *Vibrio parahaemolyticus*. J Clin Microbiol 53:1864–1872. doi:10.1128/JCM.00034-15.
Editor: C.-A. D. Burnham
Address correspondence to Cheryl A. Whistler, cheryl.whistler@unh.edu.

doi:10.1128/JCM.00034-15

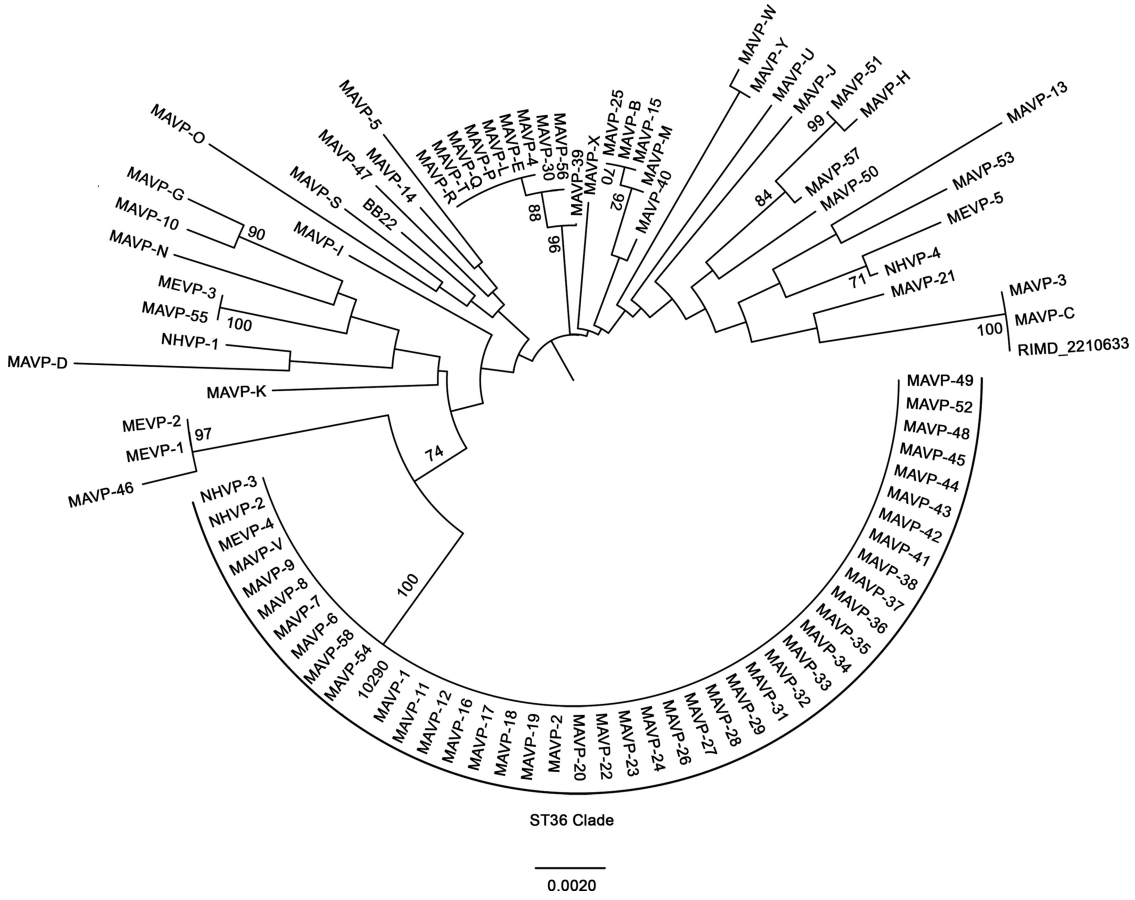

FIG 1 Identification of ST36 clade strains from among northern New England clinical isolates of *V. parahaemolyticus*. The relationships of 90 clinical isolates reported in the Northeast between 2010 and 2013, each with a unique assigned identifier including the reporting state (MA, NH, and ME), VP, 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 (*dnaE*, *dtdS*, *pntA*, and *tnaA* sequences [1,868 bp]) by using a Jukes-Cantor model, with statistical support assessed by 1,000 bootstrap reassemblies. 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 <70% bootstrap support are unlabeled.

Maine, only 35 of which were definitively or deemed likely to be from northeastern U.S. sources (CT, MA, and ME), whereas the remaining 59 were traced to either other geographic locations (Canada and VA), multisource exposures with some regions outside the Northeast, or unknown sources. Four environmental isolates from the Great Bay Estuary of NH (*V. parahaemolyticus* G61, G363, G1350, and G3654) (16, 17) and *V. parahaemolyticus* ST36 strain F11-3A, a clam isolate from Washington state during an outbreak in 1997 (18), were included in some analyses for comparison. The 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.

Multilocus sequence analysis. Template genomic DNA was isolated using the Wizard genomic DNA purification kit (Promega, WI, USA), using columns and manufacturer-provided recipes (Epoch Life Science, Inc., TX, USA), 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, USA), and sequenced by the Sanger method at the University of New Hampshire (UNH) Hubbard Center for Genome Studies (Durham, NH) or by Functional Biosciences (WI, USA). Four strains from the collection of 94 clinical isolates were not included in the analysis due to failure of one or more loci to yield an amplicon or the failure of the sequencing reaction to yield a 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 were identified from the PubMLST website (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 were identified as ST37 and ST39. The genome of the single ST37 isolate (*V. parahaemolyticus* 10290, genome assembly GCA_000454205.1) was reanalyzed 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 (1,868 bp) for 90 clinical and 192 environmental isolates from the Northeast were aligned by Clustal W and used to construct a neighbor-joining phylogeny using a Jukes-Cantor model with the Mega 6.0 software (22). Only three environmental isolates were deemed to be sufficiently related to the ST36 clade to warrant their inclusion in the *in vitro* analysis. Neighbor-joining trees were again constructed with the 90 clinical isolates and three sequenced reference strains, with statistical support assessed by 1,000 bootstrap reassemblies.

Genome sequencing, assembly, annotation, and typing. Four ST36 isolates were chosen for whole-genome sequencing using an Illumina HiSeq 2500 device at the Hubbard Center for Genome Studies at the University of New Hampshire: *V. parahaemolyticus* isolate MAVP-26 is a 2013 isolate traced to oysters harvested from MA, north of Cape Cod; *V. parahaemolyticus* isolate MAVP-36 is a 2013 isolate traced to oysters

harvested from MA, south of Cape Cod; *V. parahaemolyticus* isolate MAVP-45 is a 2013 multisource isolate traced to oysters harvested at three MA sites, with 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 1 μg of genomic DNA, as determined using the Qubit 2.0 fluorimeter (LifeTech, CA, USA). DNA was sheared on the Covaris M220 ultrasonicator to a mean size of 500 bp. Libraries were generated using the TruSeq kit, and a targeted size selection of 500 bp was completed using the optional gel extraction method in the TruSeq protocol (Illumina). The genomes were sequenced using a rapid output mode run producing 150-bp paired-ends with 249× coverage for MAVP-26 (SAMN03107383), 238× coverage for MAVP-36 (SAMN03107385), 355× coverage for MAVP-45 (SAMN03177810), and 847× coverage for MAVP-V (SAMN03177809). The 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 (AFBW01000001 to AFBW01000033) is not closed and is currently assembled as 33 contigs, the regions of shared and unique genome content in comparison to those of *V. parahaemolyticus* strains RIMD 2210633 (GenBank accession no. NC_004605.1 and NC_004603.1) and BB22OP (GenBank accession no. NC_019955.1 and 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 the open reading frames (ORFs) identified using Prokka 1.8, using a *Vibrio*-specific database in NCBI for these annotations (26).

The distribution of the identified ORFs was determined by a query against all draft genomes of *V. parahaemolyticus* available at the time of this analysis ($n = 289$) in the NCBI *V. parahaemolyticus* genome list (http://www.ncbi.nlm.nih.gov/genome/691). Loci identified as ORFs of sufficient size (≥1 kb) and variation to facilitate primer design that were harbored in nearly every strain in the 10329 NCBI genome group, of which ST36, ST59, and ST678 are members, but that are virtually absent outside this genome group were selected as potential PCR targets; these were *prp* (GenBank accession no. EGF42613), *cps* (GenBank accession no. EGF42671), and *flp* (GenBank accession no. EGF42675). Each locus was further analyzed using the BLAST algorithm by a query against the nucleotide collection and the nonredundant protein sequences, using default settings, to evaluate their broader distribution and potential function. The distribution of each locus was also evaluated in the NCBI *V. parahaemolyticus* genome list by a query against in the genus *Vibrio* (taxid: 662), excluding *V. 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*, and/or *flp*) were acquired from NCBI genome phylogeny (http://www.ncbi.nlm.nih.gov/genome/691), which, along with MAVP-26, MAVP-36, MAVP-45, MAVP-V, were analyzed using RealPhy version 1.09 (20). To produce the most accurate phylogeny, the analysis was then limited to the highest quality genomes (based on NCBI genome statistics, including level, number of contigs, and N_{50}) from the 10329 genome group, along with strains from other genome groups harboring one or more loci (i.e., NIHCB0757, S159, and/or 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). The sequences were analyzed using 10290 (GCA_000454205.1) 10329 (AFBW01000001 to AFBW01000033), and 10296 (GCA_000500105.1) as three reference ST36 strains, in which the alignment positions were extracted and then merged into a single alignment. Neighbor-joining phylogenies were reconstructed using the maximum likelihood method in PhyML, using a general time-reversible (GTR) substitution matrix and a gamma-distributed rate heterogeneity model (27). The phylogenies were visualized as trees using FigTree 1.4.2 (28). Each branch length reflects the nucleotide changes per total number of nucleotides in the sequence.

Development and application of a multiplex PCR amplicon assay. The similar sizes of the *tlh* (~450 bp) and *trh* (500 bp) amplicons produced by an existing multiplex PCR assay make 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 that were suitable for a new forward primer with 100% sequence identity across all aligned sequences. The primer design was optimized to minimize secondary structure, have compatible annealing temperature, and promote minimal cross-dimerization with the other multiplex primers in the existing assay, using the NetPrimer program as a tool (Premier Biosoft, CA, USA). When used with the published reverse primer R-TLH, the new F2-TLH primer produces an amplicon of ~401 bp (Table 1), which cannot accurately be resolved along with the ORF8 amplicon (369 bp) specific for the pandemic ST3 O3:K6 strain; however, an analysis of regional isolates (Fig. 1) indicates that the pandemic strain is not prevalent among clinical isolates from the northeastern region of the United States, 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; this was performed on ~5 μg of genomic DNA as a template using AccuStart PCR SuperMix (Quanta, MD, USA) in a 10-μl volume, with an initial denaturation at 94°C for 3 min, followed by 30 cycles of denaturation at 94°C for 1 min, primer annealing at 55°C for 1 min, and extension at 72°C for 1 min, and a final extension at 72°C at the completion of the cycling for 5 min (9). The amplicons were evaluated by electrophoresis of 1.5 μl of sample on 1.2% SeaKem LE agarose (Lonza, Rockland, ME, USA) gel with 1× GelRed (Phenix Research Products, Candler, NC, USA) in Tris-acetate-EDTA (TAE) buffer compared against a 1-kb Plus DNA ladder (Invitrogen, Grand Island, NY, USA).

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

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

TABLE 1 Oligonucleotide primers used for amplification by PCR

Gene/locus	Primer name/ direction	Primer sequence	Amplicon size (bp)	Source	Use in PCR[a]
tlh	F2	AGAACTTCATCTTGATGACACTGC	401	This study	M
	R	GCTACTTTCTAGCATTTTCTCTGC		9	
tdh	F	GTAAAGGTCTCTGACTTTTGGAC	269	9	M
	R	TGGAATAGAACCTTCATCTTCACC			
trh	F	CATAACAAACATATGCCCATTTCCG	500	9	M
	R	TTGGCTTCGATATTTTCAGTATCT			
ST36*prp*	F	CGGCTTGAGTTTTCGTCATT	609	This study	S
	R	CCACACCTGCTGGTTATTTAGTTC			
ST36*prp*	F2	TGCGGAATCTGATCTTTATCCTC	1,028	This study	M
	R2	AACTGTTGGGTCTTCGTCTAACC			
ST36*prp*	F3	CCCGAGGCACATCTTCACC	699	This study	M
	R3	TAAACCACTAACATCTTCATCTACC			
ST36*cps*	F1	TTGAGAATTACTTCCGATTATGTAGA	889	This study	M
	R1	TAAACGCATTAGCGAATAGTGC			
ST36*flp*	F1	TGGTTGTGTTTAGAGCAGGG	747	This study	M
	R1	TGTTGGTAATACGATAAGAATGAGA			

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

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 6 h or up to 24 h in HI broth with 3% NaCl, and the cells from 1 ml were pelleted by centrifugation, resuspended in 1 ml of deionized water (diH_2O), 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 multilocus sequence typing [MLST], four of which were confirmed to belong to ST36 based on all seven loci; see Results) and three related environmental strains (referred to here as the reference set), with G61 and F11-3A as standards in each assay. Additionally, 50 environmental isolates from oysters harvested in NH (here referred to as the unknown NH environmental set) or 84 environmental isolates from MA (here referred to as the unknown MA environmental set) recovered on CHROMagar *Vibrio* as purple colonies and cultured on Trypticase soy (T-soy) agar, as previously described (29), were used to further quantify the rate of false positives and the 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 (i.e., 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) determined from duplicate assays on the same sample. Both the F-TLH primer and F2-TLH primer yielded a band of the correct size from each sample (matching that from standards F11-3A and G61). The rate at which negatives were identified as negative (specificity) of the F2-TLH primer was assessed similarly on crude lysates from the unknown NH environmental set (not all of which were *V. parahaemolyticus*), in which the F2-TLH primer yielded an amplicon only from samples that were also amplified by the F-TLH primer.

The accuracy, sensitivity, and specificity of three *prp* primer pairs were first evaluated as a single amplicon assay on controls (MAVP-26, F11-3A, and G61) and then in a four-gene multiplex (with *tlh*, *tdh*, and *trh* primer pairs) (Table 1) on purified DNA of a subset of the reference set (12 ST36 clade strains and 7 nonclade strains) and replicated in three separate trials to identify which primers had the best precision and overall performance. The F2/R2-ST36*prp* and F3/R3-ST36*prp* primer pairs were selected and tested on crude lysates of the complete reference set and the unknown MA environmental set. The accuracy and sensitivity of the *cps* amplification were assessed first on purified DNA from the subset of the reference set used for analysis of *prp* (19 isolates) and then on all 43 isolates identified as ST36 using crude lysates in the 5-gene multiplex assay with the F3/R3-ST36*prp* primer pair, with precision determined by replication (see Results). The range of detection (analytical sensitivity) of the F2-TLH, F3/R3-ST36*prp*, and F/R-ST36*cps* primer pairs was examined in a five-amplicon multiplex assay (with *trh* and *tdh*) on purified and serially diluted DNA from F11-3A as a template. Visualization of all five amplicons from the 1.5-μl PCR product was optimal when between 50 μg and 5 ng of genomic DNA was used as a starting template, with decreased but visible detection of all five amplicons as low as 50 pg.

Nucleotide sequence accession numbers. The raw sequence reads and genome assemblies for the four ST36 genomes used in this study are available at NCBI under BioProject ID PRJNA263814. The nucleotide sequences for the three ORFs from *V. parahaemolyticus* strain MAVP-26 that are orthologous to the diagnostic loci used in this study that were previously annotated as hypothetical proteins in the draft genome of *V. parahaemolyticus* 10329 (AFBW01000001 to AFBW01000033) have also been deposited with NCBI with corresponding annotations and are pathogenesis-related protein (*prp*; KR150771), capsular polysaccharide biosynthesis protein (*cps*; KR150772), and O-antigen flippase (*flp*; KR150773).

RESULTS

Identification of ST36 clade and related strains from among Northeast U.S. clinical and environmental isolates. Although multilocus sequence typing based on seven loci is a widely used method for *V. parahaemolyticus* strain identification (4, 17, 18, 21, 24, 30), it can be cost-prohibitive, and it is not done routinely with

more than a few unique isolates or with all strains of a similar type associated with an outbreak (6, 7). However, using a subset of only four loci is less costly and can sufficiently inform whether strains are related. Additionally, for the loci chosen in this study, the combination of alleles in ST36 is shared only with ST37 and ST39. Only two strains, one of each of these other sequence types, have been reported (18, 30). Analysis of the 7 loci extracted from the draft genome of the single reported ST37 isolate (10290) indicated that this isolate belongs to ST36. Thus, the combination of these four alleles occurs in ST36 isolates only, with the exception of a single ST39 reported isolate, suggesting most isolates with this combination of alleles belong to ST36. We applied this four-locus multilocus sequence analysis (MLSA) approach to examine the relationships of clinical isolates from infections reported in MA, NH, and ME between 2010 and 2013, during which time infections from the ST36 strain were first reported from Atlantic sources (6). A total of 43 isolates were identical to ST36 at these 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 were 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, and it was not traced to a regional source and was not part of the 2013 regional outbreak. MAVP-26, MAVP-36, and MAVP-45 were isolated in 2013, were from the regional outbreak, and were traced to at least two, and potentially three, different shellfish harvesting sites in MA. The analysis of all seven housekeeping loci confirmed that all four of these isolates belong to ST36.

Comparative genomics and identification of loci of potential diagnostic utility. To identify genetic differences that are potentially useful for the 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 prepandemic strain BB22OP (34). Six coding regions in three different genome contigs appeared to be 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 the NCBI *V. parahaemolyticus* genome list (http://www.ncbi.nlm.nih.gov/genome/691). Notably, the NCBI *V. parahaemolyticus* genome list places ST36 strains within genome group 10329, which harbors several sequence types, all of which share ≥92% identity. Loci were considered potentially diagnostic if they (i) were present in virtually every sequenced isolate in the 10329 genome group, (ii) were not frequently present in other distant genome groups, and (iii) 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 the homology of their flanking DNA with that of 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. strain Ex25 (GenBank accession no. YP_003285914.1), was selected as a potential assay locus. This locus is particularly unique in that nucleotide sequence similarity searches querying the nonredundant 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 to 97% identity), and 2 *Vibrio albensis* genomes (90% identity). Sequence similarity searches of the nonredundant database using the translated *prp* locus revealed that 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. strain 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 the 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/691) (Fig. 2). The *prp*, *cps*, and *flp* loci only cooccur in a single clade of closely related strains that belong to ST36, ST59, and one other unknown ST for which there was only one draft genome (vpV223/04) (Fig. 2 and Table 2). A single non-ST36 high-quality genome (MDVP13, of ST678) in the 10329 genome group apparently lacks *prp*, and this genome harbors both *cps* and *flp*; however, based on whole-genome phylogeny, this strain does not group within the same clade as ST36 and ST59 (Table 2 and Fig. 2). 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 (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 *V. 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). The inclusion of just one additional locus (e.g., *cps*)

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

Strain	NCBI genome group[b]	Sequence type	*prp*	*cps*	*flp*	Isolation location[c]	Source[d]	Yr of isolation
vpV223/04	NA	Unk	+	+	+	NA	NA	NA
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[e]	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] The presence (+) or absence (−) of each locus was determined for all high-quality draft genomes. For high-quality genomes that had no publicly identified sequence type, the sequence type was identified using the SRST2 program (21). Unk, sequence type not known due to new sequence type or incomplete sequences at the 7 loci.
[b] NCBI genome groups were determined from http://www.ncbi.nlm.nih.gov/genome/691. NA, not available.
[c] Location of reported infection or isolation by U.S. state.
[d] Source identified as clinical (C) or environmental (E).
[e] Only a partial coding sequence for *cps* was identified from this genome.

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 the NCBI *V. parahaemolyticus* genome list, 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 that an assay utilizing *prp* may be sufficiently accurate for routine screening, but the addition of a second amplicon (*cps*) and requirement for both amplicons would increase the accuracy of identification of ST36-related strains to 100% if all three sequence types within the clade harboring ST36 were included.

Analysis of the distribution of *prp* and *cps* in clinical and environmental isolates from the northeastern United States using multiplex PCR. To examine the utility of loci identified by whole-genome comparisons for strain identification not only *in silico*, but also *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 and Table 1; see also Materials and 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 pair yielded bands of a predicted size for all amplicons in F11-3A (data not shown).

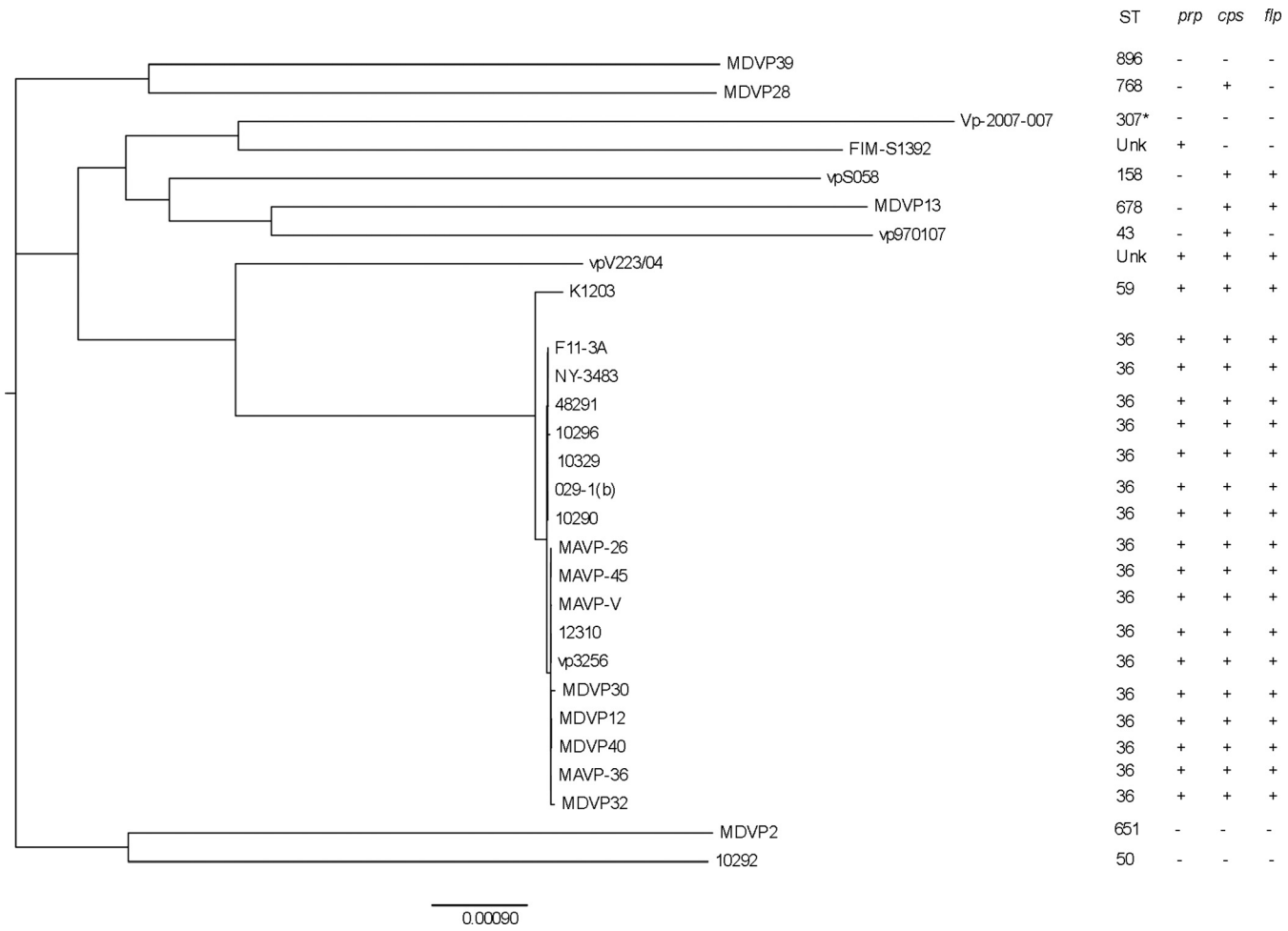

FIG 2 Distribution of potentially diagnostic loci in ST36 and related draft genomes. Genome sequence alignment-based phylogenies using 10290, 10329, and 12310 as references were reconstructed using RealPhy version 1.09 with a representative subset of sequenced isolates where the merged alignment represents 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 the NCBI *V. parahaemolyticus* genome list 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).

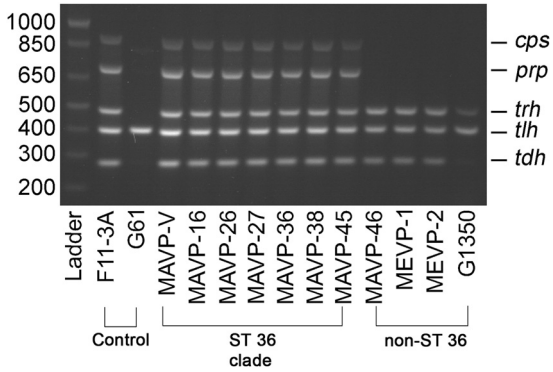

FIG 3 Improved multiplex PCR assay for identification of ST36 *V. parahaemolyticus*. The presence of virulence-associated *tdh* and *trh* amplicons, strain-associated *prp* (using F3/R3ST36*prp* primers) and *cps* amplicons, and the species-specific marker *tlh* on seven Northeast ST36 clade members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

The above-described assays were applied to the reference set of 94 clinical isolates, 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 be associated only 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 pair produced amplicons from all four diagnostic loci, including *prp*, in all 43 isolates that grouped within the ST36 clade, and they did so reproducibly in duplicated assays (data not shown). Furthermore, the *prp* amplicon was not detected in any other

clinical or environmental isolate from the Northeast, including the six isolates identified as being most closely related to the ST36 clade (Fig. 1 and 3). Thus, the four-amplicon assay using either the F2/R2-ST36*prp* or F3/R3-ST36*prp* primer pair was 100% specific, accurate, and precise with both purified DNA and freshly prepared crude lysates (see Materials and Methods). When the *cps* locus primers were included in a five-locus multiplex assay in replicated assays with either purified DNA or crude lysate including the F3/R3-ST36*prp* primer pair, 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 in 2012 attributed to a nonnative ST36 strain of *V. parahaemolyticus* (6), with widespread infections in 2013 over multiple states, indicates that 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 are 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 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 indicate that the *prp* locus has a very narrow distribution and is conserved in ST36; therefore, it may be used for strain typing (Fig. 3). Furthermore, the cooccurrence 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 and Fig. 2), suggesting that the combined presence of two loci can 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 that *prp* is exclusively and always detected in ST36 clade strains (see Results and Fig. 3). This suggests that the locus can accurately distinguish these strains from their close relatives. Although not surveyed as broadly, *cps* was detected in each of the ST36 clade strains and in none of the environmental strains (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 sets. However, the high performance of any PCR-based assay requires quality samples with an 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 the 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 nondetection and false detection is still possible. Confirmation could be done by additional genotyping, such as for the *flp* locus (Fig. 2 and 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 as ST36 by this method that are traced to regions that currently are not known to contain these as resident pathogens, some additional analysis would be warranted. Since *V. parahaemolyticus* is known to undergo recombination (4, 17) that might result in the mobilization of 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 to a regional collection only, the threat by the Pacific-native ST36 strain is not limited to the Northeast, as outbreaks have also occurred in the mid-Atlantic U.S. coast and Spain (6), suggesting that this clonal complex of strains may be spreading more broadly. We anticipate that the method will help determine the extent of the geographic expansion of these strains beyond the Northeast, the establishment of stable local populations, and the seasonal dynamics of these strains, thereby aiding in the 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 specifically thank Associate Commissioner S. Condon and K. Foley of the Massachusetts Department of Public Health, M. Hickey and C. Schillaci from the Massachusetts Department of Marine Fisheries, J. C. Mahoney, J. K. Kanwit of the Maine Department of Marine Resources, 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 was 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 no. 233555), and NH00625 (accession no. 1004199). Additional funding provided by the National Oceanic and Atmospheric Administration College Sea Grant program and grants R/CE-137, R/SSS-2, and R/HCE-3. Support was also provided through the National Institutes of Health grant 1R03AI081102-01, the National Science Foundation EPSCoR grant IIA-1330641, and the National Science Foundation grant DBI 1229361 NSF MRI. This is scientific contribution number 2582 for the New Hampshire Agricultural Experiment Station.

REFERENCES

1. **Daniels NA, MacKinnon L, Bishop R, Altekruse S, Ray B, Hammond RM, Thompson S, Wilson S, Bean NH, Griffin PM, Slutsker L.** 2000. *Vibrio parahaemolyticus* infections in the United States, 1973–1998. J Infect Dis **181:**1661–1666. http://dx.doi.org/10.1086/315459.
2. **Altekruse S, Bishop R, Baldy L, Thompson S, Wilson S, Ray B, Griffin PM.** 2000. *Vibrio* gastroenteritis in the U.S. Gulf of Mexico region: the role of raw oysters. Epidemiol Infect **124:**489–495. http://dx.doi.org/10.1017/S0950268899003714.
3. **Johnson CN, Bowers JC, Griffitt KJ, Molina V, Clostio RW, Pei S, Laws E, Paranjpye RN, Strom MS, Chen A, Hasan NA, Hug A, Noriea NF,**

III, Grimes DJ, Colwell RR. 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. http://dx.doi.org/10.1128/AEM.01296-12.

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:**e55726 http://dx.doi.org/10.1371/journal.pone.0055726.
5. **Johnson CN, Flowers AR, Noriea NF, III, Zimmerman AM, Bowers JC, DePaola A, Grimes DJ.** 2010. Relationships between environmental factors and pathogenic vibrios in the northern Gulf of Mexico. Appl Environ Microbiol **76:**7076–7084. http://dx.doi.org/10.1128/AEM.00697-10.
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. http://dx.doi.org/10.1056/NEJMc1305535.
7. **Newton AE, Garrett N, Stroika SG, Halpin JL, Turnsek M, Mody RK.** 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. http://dx.doi.org/10.1128/CMR.00025-06.
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. http://dx.doi.org/10.1128/AEM.70.12.7436-7444.2004.
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. http://dx.doi.org/10.1128/AEM.02895-13.
11. **Gutierrez West CK, 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. http://dx.doi.org/10.1128/AEM.03792-12.
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. http://dx.doi.org/10.1097/00013542-199304000-00006.
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. http://dx.doi.org/10.1128/IAI.01051-09.
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 **50:**2343–2352. http://dx.doi.org/10.1128/JCM.00196-12.
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. http://dx.doi.org/10.1128/JCM.03047-13.
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. http://dx.doi.org/10.1128/AEM.01450-10.
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. http://dx.doi.org/10.1128/AEM.07794-11.
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. http://dx.doi.org/10.1128/JB.01808-07.
19. **Ausubel F, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K.** 1990. Current protocols in molecular biology. 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 Biol Evol **31:**1077–1088. http://dx.doi.org/10.1093/molbev/msu088.
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. http://dx.doi.org/10.1186/1471-2164-13-338.
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. http://dx.doi.org/10.1093/molbev/mst197.
23. **Tritt A, Eisen JA, Facciotti MT, Darling AE.** 2012. An integrated pipeline for *de novo* assembly of microbial genomes. PLoS One **7:**e42304 http://dx.doi.org/10.1371/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. http://dx.doi.org/10.1186/1471-2164-12-402.
25. **Darling AC, Mau B, Blattner FR, Perna NT.** 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res **14:**1394–1403. http://dx.doi.org/10.1101/gr.2289704.
26. **Seemann T.** 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30:**2068–2069. http://dx.doi.org/10.1093/bioinformatics/btu153.
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. http://dx.doi.org/10.1093/sysbio/syq010.
28. **Rambaut A.** 2012. FigTree v1.4. http://tree.bio.ed.ac.uk/software/figtree/.
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 cholerae* population related to toxigenic isolates. Appl Environ Microbiol **77:**7568–7575. http://dx.doi.org/10.1128/AEM.00378-11.
30. **Jolley KA, Chan M-S, Maiden MC.** 2004. mlstdbNet–distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics **5:**86. http://dx.doi.org/10.1186/1471-2105-5-86.
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. http://dx.doi.org/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. http://dx.doi.org/10.1128/JB.05044-11.
33. **Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, Iijima Y, Najima M, Nakano M, Yamashita A, Kubota Y, Kimura S, Yasunaga T, Honda T, Shinagawa H, Hattori M, Iida T.** 2003. Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of *V. cholerae*. Lancet **361:**743–749. http://dx.doi.org/10.1016/S0140-6736(03)12659-1.
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 Announc **1**(1):e00002-12. http://dx.doi.org/10.1128/genomeA.00002-12.
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 pandemic *Vibrio 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. http://dx.doi.org/10.1128/JCM.41.10.4676-4682.2003.
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. http://dx.doi.org/10.1128/AEM.69.4.2194-2200.2003.