AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Genomic Variation and Evolution of *Vibrio parahaemolyticus* ST36 over the Course of a Transcontinental Epidemic Expansion

Jaime Martinez-Urtaza,[a,b] Ronny van Aerle,[b] Michel Abanto,[a] Julie Haendiges,[c] Robert A. Myers,[c] Joaquin Trinanes,[d,e,f] Craig Baker-Austin,[b] Narjol Gonzalez-Escalona[g]

The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, Somerset, United Kingdom[a]; Centre for Environment, Fisheries and Aquaculture Science (CEFAS), Weymouth, Dorset, United Kingdom[b]; Department of Health and Mental Hygiene, Baltimore, Maryland, USA[c]; Laboratory of Systems, Technological Research Institute, Universidad de Santiago de Compostela, Campus Universitario Sur, Santiago de Compostela, Spain[d]; National Oceanic & Atmospheric Administration, Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida, USA[e]; Rosenstiel School of Marine and Atmospheric Science, University of Miami, Cooperative Institute for Marine and Atmospheric Studies, Miami, Florida, USA[f]; Molecular Methods and Subtyping Branch, Division of Microbiology, Office of Regulatory Science, Center for Food Safety and Applied Nutrition, FDA, College Park, Maryland, USA[g]

ABSTRACT *Vibrio parahaemolyticus* is the leading cause of seafood-related infections with illnesses undergoing a geographic expansion. In this process of expansion, the most fundamental change has been the transition from infections caused by local strains to the surge of pandemic clonal types. Pandemic clone sequence type 3 (ST3) was the only example of transcontinental spreading until 2012, when ST36 was detected outside the region where it is endemic in the U.S. Pacific Northwest causing infections along the U.S. northeast coast and Spain. Here, we used genome-wide analyses to reconstruct the evolutionary history of the *V. parahaemolyticus* ST36 clone over the course of its geographic expansion during the previous 25 years. The origin of this lineage was estimated to be in ~1985. By 1995, a new variant emerged in the region and quickly replaced the old clone, which has not been detected since 2000. The new Pacific Northwest (PNW) lineage was responsible for the first cases associated with this clone outside the Pacific Northwest region. After several introductions into the northeast coast, the new PNW clone differentiated into a highly dynamic group that continues to cause illness on the northeast coast of the United States. Surprisingly, the strains detected in Europe in 2012 diverged from this ancestral group around 2000 and have conserved genetic features present only in the old PNW lineage. Recombination was identified as the major driver of diversification, with some preliminary observations suggesting a trend toward a more specialized lifestyle, which may represent a critical element in the expansion of epidemics under scenarios of coastal warming.

IMPORTANCE *Vibrio parahaemolyticus* and *Vibrio cholerae* represent the only two instances of pandemic expansions of human pathogens originating in the marine environment. However, while the current pandemic of *V. cholerae* emerged more than 50 years ago, the global expansion of *V. parahaemolyticus* is a recent phenomenon. These modern expansions provide an exceptional opportunity to study the evolutionary process of these pathogens at first hand and gain an understanding of the mechanisms shaping the epidemic dynamics of these diseases, in particular, the emergence, dispersal, and successful introduction in new regions facilitating global spreading of infections. In this study, we used genomic analysis to examine the evolutionary divergence that has occurred over the course of the most recent transcon-

Received 17 August 2017 **Accepted** 13 October 2017 **Published** 14 November 2017

Citation Martinez-Urtaza J, van Aerle R, Abanto M, Haendiges J, Myers RA, Trinanes J, Baker-Austin C, Gonzalez-Escalona N. 2017. Genomic variation and evolution of *Vibrio parahaemolyticus* ST36 over the course of a transcontinental epidemic expansion. mBio 8:e01425-17. https://doi.org/10.1128/mBio.01425-17.

Editor Mary Ann Moran, University of Georgia

Address correspondence to Jaime Martinez-Urtaza, jaime.martinez-urtaza@cefas.co.uk.

tinental expansion of a pathogenic *Vibrio*, the spreading of the *V. parahaemolyticus* sequence type 36 clone from the region where it is endemic on the Pacific coast of North America to the east coast of the United States and finally to the west coast of Europe.

KEYWORDS Pacific Northwest, *Vibrio parahaemolyticus*, WGS, climate change, gastroenteritis, seafood

The genus *Vibrio* contains >100 described species, and around a dozen of these have been demonstrated to infect humans (1). Infection is usually initiated by exposure to seawater or consumption of raw or undercooked seafood (2, 3). *Vibrio parahaemolyticus* is a Gram-negative, halophilic bacterium found commonly in temperate and warm estuarine waters worldwide (4). *V. parahaemolyticus* is the most prevalent food poisoning bacterium associated with seafood consumption in many regions globally, typically causing acute gastroenteritis. This bacterium grows preferentially in warm (>15°C), low-salinity (<25 ppt NaCl) marine water (5).

A number of important factors underpin the need for a greater understanding of these food-borne pathogens in an international context. Compared to other major food-borne pathogens, the number of *V. parahaemolyticus* infections is steadily increasing (6). Indeed, according to the U.S. Centers for Disease Control and Prevention (CDC), the average annual incidence of all *Vibrio* infections increased by 85% between 1996 and 2009 (7), with *V. parahaemolyticus* accounting for 52% of those infections and being responsible for a more recent and marked increase in incidence (8). In the United States alone, *V. parahaemolyticus* is estimated to cause around 35,000 human illnesses each year (range, 18,000 to 58,000 cases) (9). Additionally, *V. parahaemolyticus* infections are now being reported in areas with little previous incidence, including South America and northern Europe (6, 10). Although the factors driving the escalation in the number of infections are likely multifactorial, climate warming, in particular, appears to be a substantial contributor to the expansion of pathogenic vibrios, especially in temperate regions (10). Future climate scenarios based on climate modeling suggest that *Vibrio* spp., including *V. parahaemolyticus*, are likely to continue to pose a significant and expanding public health risk.

The most substantial change in the epidemiology of *V. parahaemolyticus* infections over the last 2 decades has been the transition from the dominance of locally restricted strains to the emergence and transcontinental expansion of new clones with pandemic potential. Only two instances of transcontinental expansion of *V. parahaemolyticus* strains have been reported, "pandemic clone" CC3 (serotype O3:K6), which emerged in India in 1996 and subsequently spread around the world (11), and more recently the expansion of sequence type 36 (ST36), which was responsible for numerous large *V. parahaemolyticus* outbreaks in the Pacific Northwest region of the United States over the last 2 decades (12–14). The strains associated with these outbreaks, subsequently termed the Pacific Northwest (PNW) complex (12, 14, 15), appear to be genetically and biochemically distinct, and have a significantly smaller infectious dose than other toxigenic *V. parahaemolyticus* isolates (6).

Prior to 2012, PNW complex strains were restricted to the Pacific Northwest region of the United States and Canada (16). However, illnesses associated with this complex were reported along the northeast coastline of the United States in the spring of 2012 (8, 15) and subsequently in the northwest of Spain in association with a large outbreak of illness in August 2012 (17). The geographic expansion of these strains caused significant economic losses in the shellfish industry in the northeast and caused the largest known food-borne *Vibrio* outbreak reported in Europe (17). A striking observation from characterization of the 2012 outbreak-associated strains, as well as previous clinical isolates of this complex from the United States, was the indistinguishable serotypes (O4:K12 or O4:Kut), pulsotypes, and STs (ST36) (15). This initial observation was noteworthy because it suggested that a single, highly pathogenic clone of *V. parahaemolyticus* had radiated from the Pacific Northwest region and successfully estab-

TABLE 1 Summary of the *V. parahaemolyticus* reference and core genome sizes and the corresponding numbers of genes identified

Strain 10329 genome portion	Length (bp)	No. of genes (chromosome 1/chromosome II)
Reference genome	5,149,046	4,663
Chromosome I	3,316,038	3,002
Chromosome II	1,833,008	1,661
Core genome	4,436,654	4,101 (2,637/1,464)
Nonrecombining core genome	4,396,495	4,053 (2,606/1,447)
Recombining regions	40,159	48 (31/17)
Core genes	3,860,598	4,101 (2,637/1,464)
Nonrecombining core genes	3,810,918	4,053 (2,606/1,447)
Recombining regions	49,680[a]	48 (31/17)

[a]If part of the gene is within a recombinant region, the length of the whole gene is included.

lished itself along the eastern seaboard of the United States and then potentially in western Europe (15). Illnesses associated with this clone were documented from the northeast coast of the United States in 2013, indicating that these strains overwintered in environmental reservoirs (8).

Given the highly pathogenic nature of the PNW complex and the fact that other pathogenic variants of *V. parahaemolyticus* have expanded worldwide (e.g., O3:K6) (11), a clear understanding of the potential source, phylogenetic nature, lineage, and time line of transmission of this group is needed. To this end, we performed a genome-wide analysis of historical and contemporary ST36 *V. parahaemolyticus* from the United States and Europe to gain a comprehensive understanding of the genomic variation and evolutionary process undergone by this group during its geographic expansion and colonization of new areas. These results are critical to understand the particular genetic signatures and evolutionary forces contributing to the expansion of this globally important emerging pathogen.

RESULTS

Genomic analysis of ST36. To investigate the evolutionary changes that characterized the epidemic expansion of *V. parahaemolyticus* ST36, we sequenced 44 strains from different geographic areas on the Pacific and Atlantic coasts of the United States previously recovered from sporadic cases and outbreaks, including the ST36 strains recovered from the single outbreak in the northwest of Spain in 2012 (17) (see Table S1 in the supplemental material). The complete genome of an ST36 strain, 10329, isolated in Washington State in 1998 (12) (complete genome size, 5,149,046 bp; G+C content, 45.3%) was used as the reference genome for most of the analysis. The strain 10329 reference genome consisted of two chromosomes totaling 5,149,046 bp (4,663 genes), with chromosomes I and II being 3,316,038 bp (3,002 genes) and 1,833,008 bp (1,661 genes), respectively (Table 1). The length of the resulting core genome alignment of these 44 genomes and an additional 4 previously sequenced genomes was 4,436,654 bp, whereas the length of the core genes was 3,860,598 bp with a total of 4,101 genes, 2,637 genes on chromosome I and 1,464 on chromosome II (Table 1).

A total of 1,081 single nucleotide polymorphisms (SNPs; 725 on chromosome I and 356 on chromosome II) were identified after an analysis of polymorphic sites (Fig. 1; Table S2). Of the 1,081 SNPs, 867 (80%) were identified in coding sequences (CDS; 3,860,598 bp), with 571 on chromosome I and 296 on chromosome II. Synonymous SNPs represented 62% of the total number of SNPs in CDS regions, while missense SNPs accounted for 36%. Start lost, stop lost, and stop gained changes were identified in 1, 3, and 15 SNPs, respectively. The remaining 214 variable sites were found in noncoding regions (576,056 bp), which represented twice the rate of SNPs found in coding regions. The frequency and distribution of SNPs in the core genome and core genes were similar in the two chromosomes (Fig. 1).

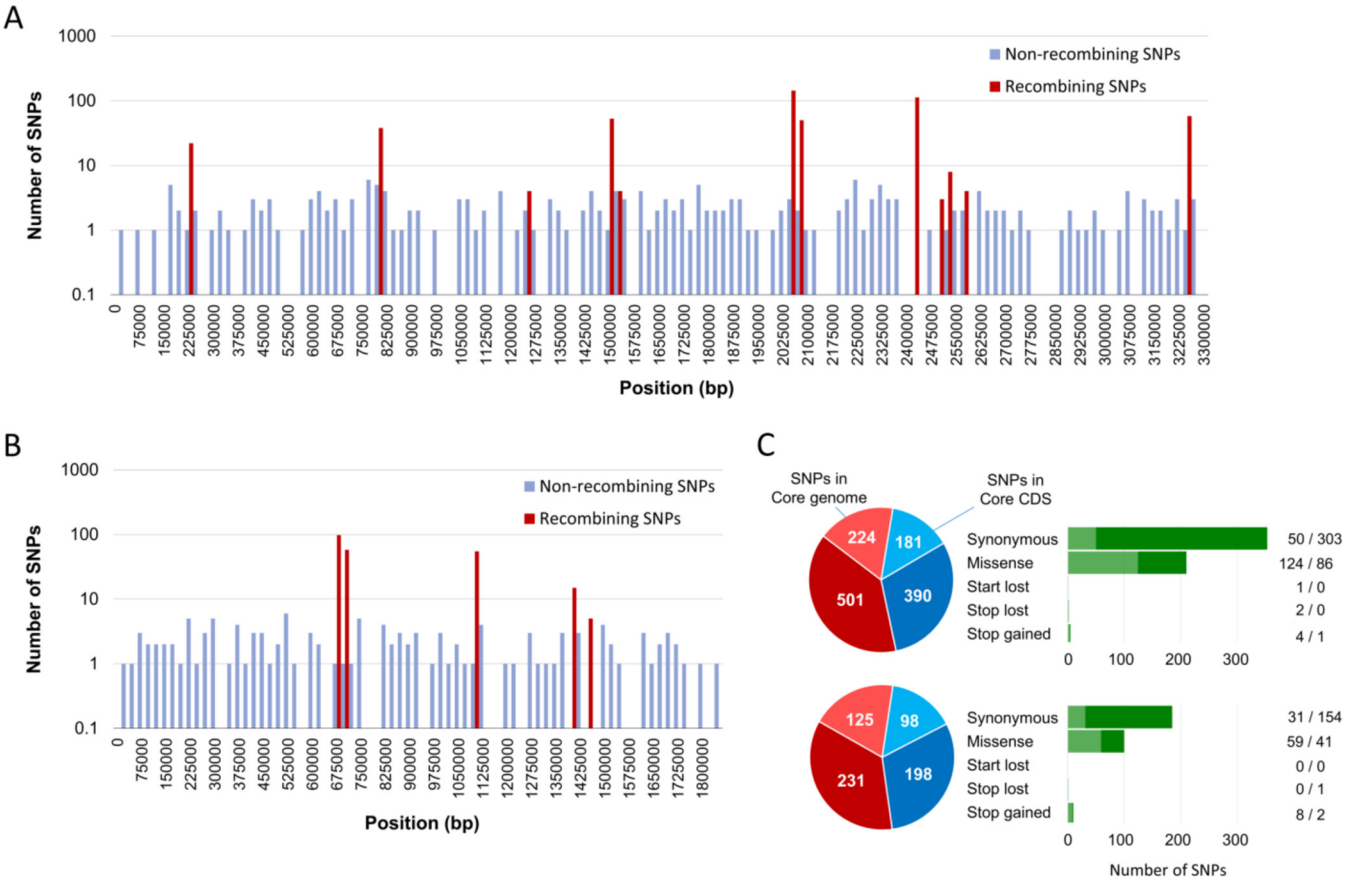

FIG 1 Frequency and distribution of SNPs in the core genome and core genes in the two chromosomes of the 48 *V. parahaemolyticus* strains used in this study. SNPs that were nonrecombinogenic and aroused by recombination are blue and red, respectively. (A) SNPs in chromosome I. (B) SNPs in chromosome II. (C) Summary of the frequencies of the SNPs found in each chromosome and the changes that occurred in the predicted amino acid sequence at that position (synonymous versus nonsynonymous, start or stop codons). Chromosome I, top; chromosome II, bottom. Light red, recombining SNPs; dark red, nonrecombining SNPs in core genome; light blue, recombining SNPs; dark blue, nonrecombining SNPs in CDS regions.

Pangenome analysis with Roary identified 4,407 core genes shared by at least 99% of the strains, 79 soft core genes shared by 95 to 99% of the strains, 174 shell genes shared by 15 to 95% of the strains, and 584 cloud genes that were identified in <15% of the strains. The total number of annotated genes ranged from 4,518 to 4,687 (Fig. 2), resulting in a variation of 169 genes among the 48 genomes examined.

On the basis of pangenome analysis, a distinctive region exclusively in all of the strains from Spain was identified (Fig. 2). This genomic region of 67,308 bp, on chromosome I, contained 53 genes and had a G+C content of ~43.78%. BLAST searches against genomes deposited in the NCBI database revealed that 61% of this region is present in other *V. parahaemolyticus* genomes, including the reference genome, RIMD 2210633 (18). However, the remaining region of ~25 kb was exclusive of ST36 strains from Spain and harbors genes associated with prophages.

Detection of recombination in ST36 genomes. To assess the potential impact of recombination on the evolution of the ST36 group, a preliminary phylogenetic analysis was carried out with the SNPs identified in the genome analysis. Long terminal branches were detected for most of the recent diversification events (Fig. S1). ClonalFrameML was employed to identify recombination events and to infer a more accurate phylogenetic reconstruction without the inflation of genomic divergence generated by recombination.

The average length of recombined fragments (δ) was estimated to be 1,751.8 bp, and the average distance of imports (ν) was 0.017. The ratio of recombination and mutation rates (R/θ) was 0.058, indicating that recombination happened about 20 times

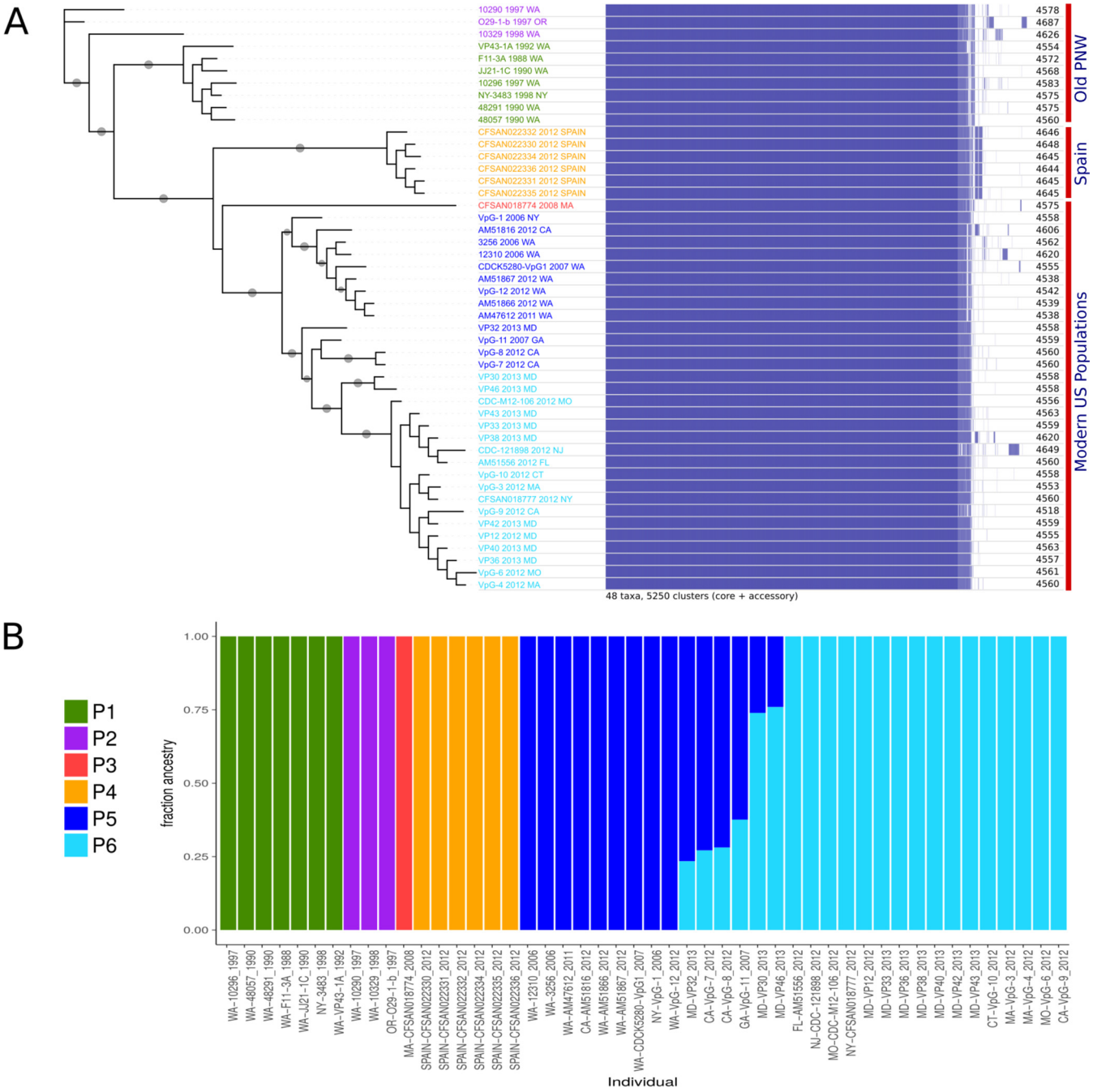

FIG 2 Phylogenetic inference and population structure of the ST36 clone. (A) ML phylogeny (left) inferred from the SNPs identified through analysis after removal of the recombining regions identified with ClonalFrameML and distribution of core and accessory genes determined with Roary for each strain (right), with the total number of genes in the column on the right. Gray circles indicate branches with 100% bootstrap support. (B) Population structure and admixture proportion inferred by using ADMIXTURE showing the six subpopulations identified within the *V. parahaemolyticus* ST36 strains (K value of 6). Bar graph indicating the relative ancestry composition of the data set analyzed, with each color representing one of the genetically differentiated ancestral groups and each vertical colored bar corresponding to one strain. Although there is a clear correspondence between the subpopulations of ST36 and their spatial-temporal distribution, four genomes from the western (VpG-11, VpG-7, VpG-8, and VP32) and two from the eastern (VP30 and VP46) United States demonstrated evidence of admixture between the two modern populations currently existing on the Pacific and Atlantic coasts of the United States, with the fraction of every color in the bar reflecting the proportion of the genome derived from the group represented by that particular color. State name abbreviations used in strain codes: CA, California; CT, Connecticut; FL, Florida; GA, Georgia; MA, Massachusetts; MD, Maryland; MO, Missouri; NJ, New Jersey; NY, New York; OR, Oregon; VA, Virginia; WA, Washington.

less frequently than mutation. However, the effect of recombination in genetic diversification relative to mutation (*r*/*m*) was 1.77, which implies that even though recombination events were approximately twenty-times less frequent than mutation, each recombination event introduced almost twice as many substitutions as mutations. Each

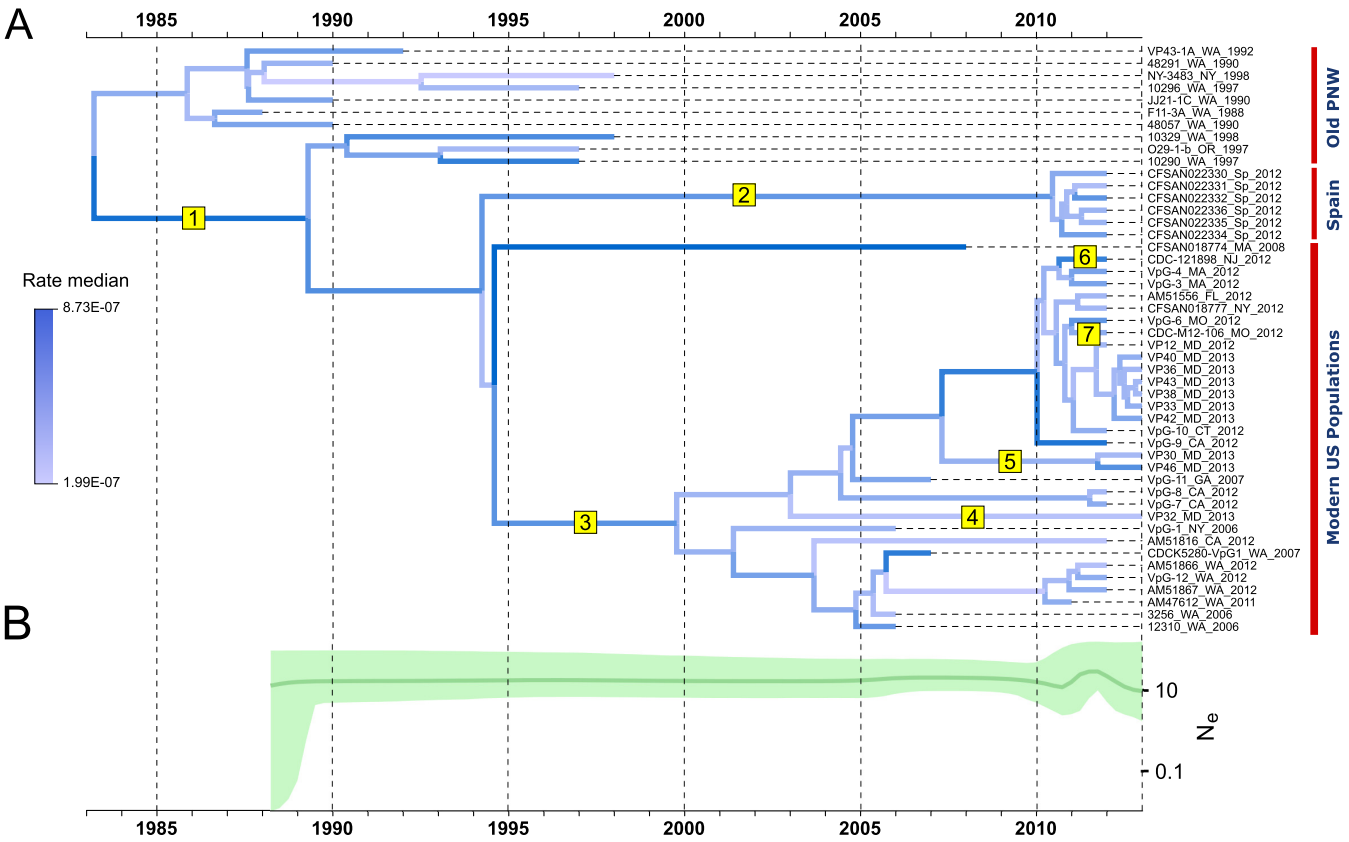

FIG 3 Bayesian analysis of the different lineages of *V. parahaemolyticus* ST36. (A) Markov chain Monte Carlo tree showing the phylogeographic reconstruction throughout the course of the transcontinental expansion of this clone estimated with BEAST (Bayesian skyline demographic model and uncorrelated lognormal molecular clock) analysis with prior temporal information from the date of isolation from the 349 SNPs identified in the core genome of the ST36 strains after the removal of recombination. The color gradient of the branches represents the median substitution rates, and the yellow boxes represent recombination events. (B) Bayesian skyline plot estimated with BEAST showing the demographic changes measured as effective population size (N_e) per generation time. State name abbreviations used in strain codes: CA, California; CT, Connecticut; FL, Florida; GA, Georgia; MA, Massachusetts; MD, Maryland; MO, Missouri; NJ, New Jersey; NY, New York; OR, Oregon; VA, Virginia; WA, Washington.

recombination event introduced, on average, 31 substitutions ($\delta\nu$). The analysis revealed a mutation rate twice as high in chromosome I as in chromosome II, indicating less of an impact of recombination on chromosome II (Table S3). Additionally, the impact of recombination on each specific node and strain was evaluated by using Gubbins; the analysis revealed variable *r*/*m* rates across the phylogenetic tree, with values ranging from 1.5 in the Spanish group to around 25 for modern strains isolated from the northeast coast of the United States (Fig. S2). This supports recombination as a fundamental force driving the emergence of the modern ST36 populations in the United States, with an *r*/*m* rate of 13.

ClonalFrameML identified 20 recombination events on all branches of the clonal genealogy (Table S4; Fig. S1), 16 in chromosome I, and the remaining 4 in chromosome II. Seven imports of long regions ranging from 1.1 to 10 kb were identified on four major nodes of the tree, demonstrating the central contribution of recombination to any of the major diversification processes within the group. Nine of the recombination regions (seven of them in a single strain, CDC_121898) were under 100 bp in length, had no similarities to any sequences in the NCBI nr database, and were not subjected to further analysis. For the remaining 11 insertion sequences with BLAST hits, one region of strain CDC_121898 (367 bp) provided the highest sequence similarity to *Vibrio vulnificus* (94.60% identical sites). All other sequences had the highest percentage of sites identical to those of *V. parahaemolyticus*; most of the similarities were relatively strong (≥97% identical sites).

The first hot spot was found in the first node of diversification (node 1 in Fig. 3; node

92 in Table S4) and spanned positions 666392 to 668923 of chromosome I (2,531 bp). This region corresponded to three genes (Table S4) with 98.5% of the sites identical to those of *V. parahaemolyticus* UCM-V493 (19). The second hot spot, spanning positions 3816980 to 3819912 (2,932 bp) of chromosome II, was found exclusively in the strains identified in Spain, (node 2 in Fig. 3; node 83 in Table S4) and corresponded to genes involved in the ABC transporter substrate-binding protein (Table S4). A total of 98.9% of its sites are identical to those of *V. parahaemolyticus* RIMD 2210633 (20). The third hot spot, spanning positions 3445960 to 3456047 of chromosome II, was identified in the node driving the modern populations of the ST36 isolates prevailing on the Atlantic and Pacific coasts of the United States (node 3 in Fig. 3; node 83 in Table S4); this represented the largest region identified by ClonalFrameML (10,087 bp), with 15 proteins and the highest BLAST similarity to *V. parahaemolyticus* O1:K33 strain CDC_K4557 (98.8% identical).

Two other large regions of 8,862 and 4,692 bp, on chromosome I, were identified in the node of divergence of strains Vp_30 and Vp_46, which were isolated from cases in Maryland in 2013 (node 5 in Fig. 3; node 49 in Table S4). The first one, spanning positions 2070377 to 2079239, was 98.7% identical to both *V. parahaemolyticus* O1:K33 strain CDC_K4557 and strain FORC_006 (21), with slight differences in the six coding regions and gene products. The second large region, spanning positions 2756482 to 2761174, was 99.3% identical to *V. parahaemolyticus* O1:K33 strain CDC_K4557, with the presence of putative proteins that have molecular functions related to DNA binding (peptidases), endonuclease, metalloendopeptidase, and *N*(6)-L-threonylcarbamoyladenine synthase activities (Tables S4).

A total of 11 hot spots were identified in strain CDC-121898_2012_NJ, which was isolated in New Jersey in 2012. No BLAST results were available for seven recombination regions. One region was identified as a type 1 secretion system-secreted agglutinin (RTX) 97% identical to *V. parahaemolyticus* UCM-V493 (19). Although this insertion was only 134 bp in length, covering a small proportion of the coding region (1,116 bp), functional annotations reveal that this protein was involved in cell communication. Three recombination regions were identified as hypothetical proteins, although one region was most similar to *V. vulnificus* (94.60% of sites identical; Table S4). Finally, two regions were identified in strain VP32_2013_MD, one 97.90% identical to *V. parahaemolyticus* O1:K33 strain CDC_K4557 and a second 97.9 to 98.6% identical to *V. parahaemolyticus* RIMD 2210633 (Table S4).

Phylogenetic inference. Preliminary maximum-likelihood (ML) phylogenetic reconstruction of the core genome with RAxML provided strong evidence of recombination, as evidenced by the relative lengths of branches in the reconstructed phylogeny (Fig. S3). As expected, the accuracy of the tree's topology was impacted by recombination of ST36 observed as distortion in the lengths of the phylogenetic tree but not by alteration of branch topology. The topology of the tree inferred by using complete core genomes resulted in long internal branches and shorter terminal branches with the reverse situation when the recombination was removed. This situation was particularly relevant where large importations resulted in an upwardly biased inference of branch length. As distortion of branch length has been widely recognized as a likely contributor to inaccurate inference of demography and molecular clocks when phylogenetic methods are applied to recombining populations (22), only core genomic regions with recombining sites removed were used for any further analysis. A total of 732 SNPs associated with recombination events were identified, which contributed to 68% of the variation identified in both chromosomes (4,436,654 bp) (Table 2). The remaining 349 SNPs were detected in nonrecombining regions and ultimately used for phylogenetic inference.

Branch topology was robustly reconstructed from whole genomes by ML phylogenetic methods and, even in the presence of recombination, showed a solid grouping with the existence of three groups that were clearly differentiated according to the origin of the isolates: old PNW, modern United States, and Spain (Fig. 2). Strains isolated

TABLE 2 SNPs identified across the two chromosomes of *V. parahaemolyticus* ST36 strains

Genome portion	Total no. of SNPs	No. of SNPs in:					
		Chromosome I			Chromosome II		
		All	Recombining	Nonrecombining	All	Recombining	Nonrecombining
Core genome	1,081	725	501	224	356	231	125
CDS	867	571	390	181	296	198	98
Synonymous SNPs	538	353	303	50	185	154	31
Missense SNPs	310	210	86	124	100	41	59
Start lost SNPs	1	1	0	1	0	0	0
Stop lost SNPs	3	2	0	2	1	1	0
Stop gained SNPs	15	5	1	4	10	2	8

from the Pacific Northwest before 2000 were clearly differentiated from the rest of the strains in a single group from which the Spain and modern U.S. strains diverged.

Population structure. To assess the level of genetic differentiation between the subpopulations in each geographic region, we investigated the population structure and ancestry relationship by using different numbers of populations (using K values ranging from 1 to 9) with ADMIXTURE (Fig. 2). A K value of 6 showed the lowest cross-validation error and was finally selected for the analysis. Outputs from ADMIXTURE identified six subpopulations within the *V. parahaemolyticus* ST36 strains with a clear correspondence between the subpopulations and their spatial-temporal distribution. Strains identified in the Pacific Northwest before 2000 were split into two different groups, and the modern U.S. population was clearly segregated into two subpopulations that closely corresponded to their origins (Pacific and Atlantic regions). Four strains from the Pacific region (VpG-11, VpG-7, VpG-8, and VP32) and two from the Atlantic region (VP30 and VP46) showed evidence of admixture between the two modern populations in the United States, with fractions of their genome belonging to each subpopulation.

Genome size, gene number, and genetic diversity within subpopulations. A trend toward gene number reduction was observed over the course of evolution. Larger genomes and higher gene numbers were observed in strains from the old subpopulation P2 (a median of 5,149,046 bp and 4,665 CDS) identified in the Pacific Northwest in the 1990s, whereas shorter genomes were observed in modern populations from the United States (Fig. S4). The effective reduction was observed in subpopulations 5 and 6 (Fig. S4), comprising the modern U.S. strains, and the two subpopulations had similar genome sizes and gene numbers (a median of 5,108,489 bp and 4,597 CDS and a median of 5,111,360 bp and 4,596 CDS, respectively). Conversely, recent strains from Spain retained the ancestral trait of large genomes observed in the extinct PNW group, which tightens the link between those populations and bolstered the notion that the genome reduction arose only in the course of evolution of modern ST36 populations in the United States.

Genetic diversity estimated for each of the *V. parahaemolyticus* ST36 subpopulations identified by ADMIXTURE (Fig. S5) showed variable levels of nucleotide diversity within each subpopulation, with the higher values in the dominant old (P2) and modern (P5) PNW subpopulations. The lowest genetic diversity was in the Spain and modern U.S. populations. These low levels of genetic variation may have occurred as a result of the founder effect after a recent introduction into these regions. However, while the diversity of the subpopulation of Spain remained extremely low, the subpopulation from the northeast of the United States showed rapid diversification and effective divergence from the original group. Examination of the environmental conditions on the U.S. east coast revealed a warming trend over the period of the study and identification of a step change with one regime shift in 2010 that resulted in an abrupt sea surface temperature (SST) change of 2.2°C (from 13.8 to 16°C) that corresponded to the radiation of ST36 in the U.S. northeast (Fig. 4).

Molecular clock, evolutionary rates, and phylodynamics of transmission. Phylogenetic trees inferred with sequences of the core genome were analyzed to identify

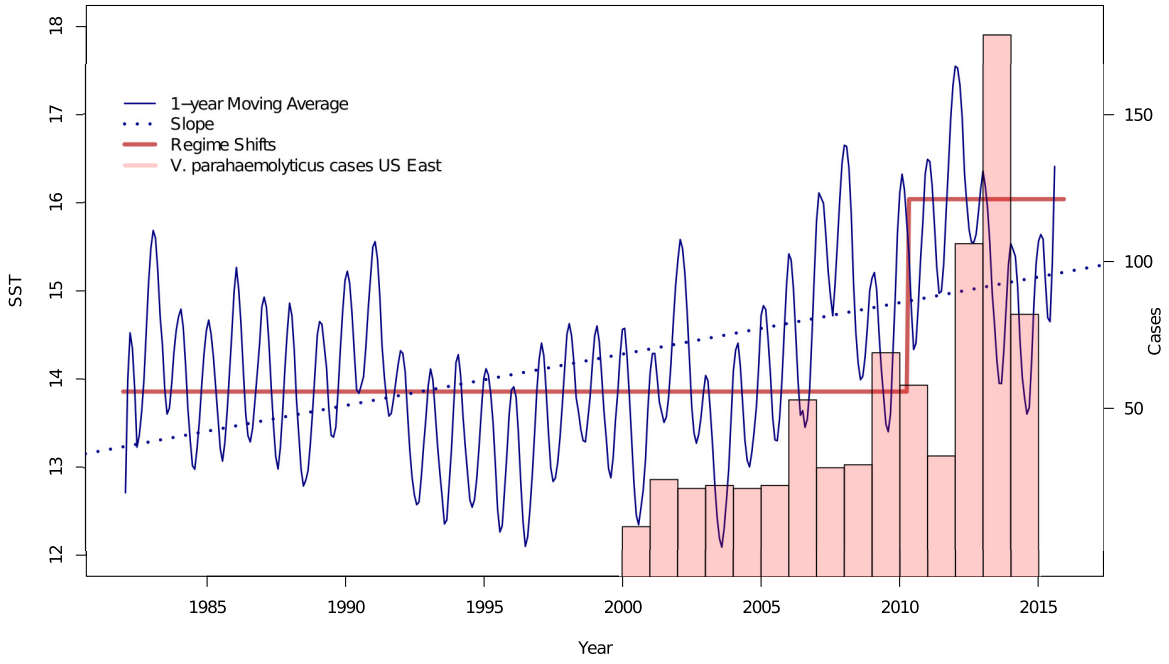

FIG 4 Sea surface temperature (SST) trend estimated by using daily SST data from the Chesapeake Bay over the study period. Analysis of mean SST records shows the existence of a step change rather than a linear pattern with one regime shift over this period in 2010 resulting in abrupt changes in SST with a warming of 2.2°C (from 13.8 to 16°C) in close correspondence to the transition in the epidemiological pattern of *V. parahaemolyticus* infections along the Atlantic coast of the United States with an increase in the number of cases reported to the Cholera and Other Vibrio Illness Surveillance (COVIS) system (U.S. CDC) and the parallel to the process of radiation and diversification of the ST36 population in the region.

trees that maximized the correlation of root-to-tip distance with the sampling date. The best fit was identified for sequence data from the core genome, with a correlation coefficient (*r*) of 0.86 and a variance (R^2) of 0.74 for the dated tips. The slope of the regression for the core genome, a proxy for the rate of evolution, was estimated as 4.93E-7 mutations per site per year, which agreed with previously published data for *Vibrio cholerae* (23) and is equivalent to 3.3 mutations per genome per year (24). The estimate of the time of the most recent common ancestor (TMRCA) of the ST36 strains analyzed (*x* intercept) was 1980.

A Bayesian analysis of the core genome sequences was then performed with BEAST by using strict and relaxed molecular clock analyses to reconstruct the evolution and phylogeography of the ST36 clone throughout the course of its geographic expansion. After running different combinations of demographic and molecular clock models, a Bayesian skyline demographic model and uncorrelated lognormal molecular clock were finally selected as the best demographic model for this data set according to path sampling (PS)/stepping-stone sampling (SS) values (Table S5).

The TMRCA of the group was around 1983 (Table S5) and rapidly diverged in two groups that composed the old PNW lineage that was the sole population belonging to this group in the Pacific Northwest region until the middle of the 1990s (Fig. 3). By 1995, a first diversification event occurred, driving the group detected in Spain in 2012 and then the emergence of the modern U.S. lineage, which completely replaced the old PNW lineage (Fig. 3). Strains belonging to the old PNW lineage were not identified after 2000, with the exception of the related strains from Spain 12 years later. The ST36 populations prevailing today in the United States diverged from a common ancestor by 2000 and quickly evolved into the modern lineages prevalent today on the Pacific and Atlantic coasts of the United States; this group showed very active diversification, with the emergence of a substantial number of genetic variants in only 3 years. All of the major diversification events, emergence of new lineages, and lineage replacement of the different groups were concurrent with frequent homologous recombination identified by ClonalFrameML (Fig. 3).

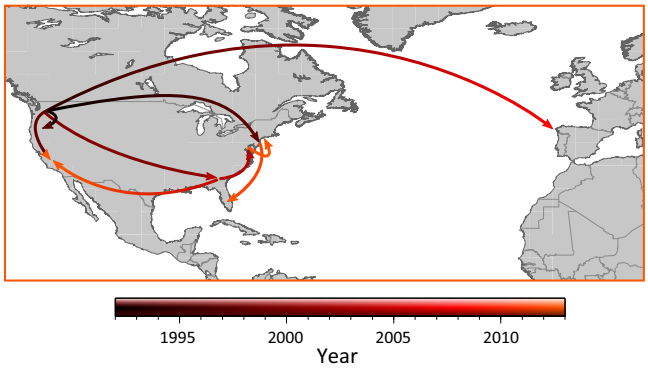

FIG 5 Spatial phylogenetic reconstruction of *V. parahaemolyticus* ST36 evolutionary dynamics inferred with SPREAD. A BEAST Bayesian phylogeographic reconstruction incorporating discrete spatial-temporal diffusion was used to visualize the phylodynamics of transmission between the original regions of emergence of this clone (PNW) and the geographic locations where different lineages were detected over the course of the expansion.

Despite the active dynamics within the ST36 group, the effective population size inferred by BEAST showed an almost constant value throughout the period, with the exception of a slight rise after 2010, concurring with the emergence and diversification of numerous strains in both the Pacific Northwest and northeast of the United States (Fig. 3B). The mean evolutionary rate in the core genome of the ST36 population estimated with BEAST was 4.407E-7 nucleotide substitutions per site per year with a 95% highest posterior density interval of 2.332E-7, 6.128E-7. The mean coefficient of variation among branch rates was 0.6156, with evolutionary rates ranging from 2.028E-7 to 6.853E-7 nucleotide substitutions per site per year (Fig. 3), which resulted in 0.87 to 3.84 mutations per genome per year. A local clock model analysis with BEAST was then performed to investigate the evolutionary rates within all of the ST36 lineages identified by ADMIXTURE. A fixed local clock under a demographic model of constant population size was used, revealing a substantial variation in the evolutionary rate between the different lineages (Fig. S5). The highest mean rates were identified in P2 from the old PNW population (5.14E-7) and the group from Spain (4.83E-7), whereas the modern populations from the United States showed lower mutation rates, with mean rates of 2.56E-7 and 3.31E-7 in P5 (PNW) and P6 (Atlantic), respectively. The lowest rate was found in the oldest population from the Pacific Northwest (P1), with a mean rate of 1.91E-7 in this group.

A Bayesian phylogeographic reconstruction incorporating a discrete model of spatial-temporal diffusion was used to visualize the phylodynamics of transmission between geographic locations with SPREAD (Fig. 5). The map showed several waves of dissemination of the old PNW strains in the 1990s across the United States with different events of transference between the Pacific and Atlantic coasts of the United States. However, the introduction of the ST36 clone to the U.S. Atlantic coast was not successful until the emergence of the modern U.S. ST36 lineage (after 2000) when, after several introductory events, it became resident in the area. Once introduced into the region, the modern U.S. ST36 lineage initiated a process of differentiation and evolved, originating a new lineage unique to this area, which has shown a very active process of diversification over recent years. The group identified in Spain in 2012 diverged from the old PNW population in the late 1990s and from that point was not reported until its sudden emergence in the course of the 2012 Spanish outbreak.

DISCUSSION

The number of reported cases of *V. parahaemolyticus* infection has increased steadily over the previous 2 decades as a result of expansion on a global scale (10, 25). In addition to the environmental factors driving this expansion, the emergence and transcontinental dissemination of some particular genetic variants of *V. parahaemolyti-*

cus are contributing to this process (15, 17, 26–28). Classical typing techniques applied to the investigation of outbreaks (initially repetitive element PCR and pulsed-field gel electrophoresis and later multilocus sequence typing [MLST]) provided insights into the potential sources and origins of these new variants, documenting the first connections between populations implicated in outbreaks across large geographic distances (11, 12, 29–31). This situation was particularly relevant for understanding the expansion of *V. parahaemolyticus* caused by the O3:K6 pandemic clone (11, 32). The application of whole-genome sequencing for the study of pathogenic *V. parahaemolyticus* populations was crucial to determine that the pandemic clone was not the only group that underwent transoceanic dispersal. Also, almost all of the major *V. parahaemolyticus* outbreaks identified in Peru and Chile over the last 25 years had been associated with the introduction of new genetic variants typically originating in Asia (33).

Following this precedent, we used genome-wide analysis to investigate a more recent instance of transcontinental spreading of a highly pathogenic *V. parahaemolyticus* group, ST36. This group, primarily identified by MLST (12), was initially reported only from the Pacific Northwest. Over the last 6 years, its detection along the northeast coast of the United States has been associated with a rise in the number of cases in the region (8). In addition, ST36 was identified for the first, and only, time outside North America in the northwest of Spain, where it caused a large outbreak in 2012 (15, 17). The subsequent emergence of this clone in Europe triggered concern about the potential implications of the transcontinental spreading of a second *V. parahaemolyticus* strain and the opportunities for a new pandemic expansion. Furthermore, recent studies have suggested the existence of a new population of ST36 prevailing among illnesses in the northeast of the United States (8, 34), which have introduced an additional level of uncertainty about the evolutionary history of this group.

The present study provided an exceptional opportunity to investigate the evolution of *V. parahaemolyticus* populations in the course of epidemic expansion. The distribution of the ancestral lineages of ST36 was restricted to the Pacific Northwest, and there is no record of possible introductions to any other region. It was not until the emergence of the modern lineage of this clone by 1995 that it showed effective dispersal, particularly after 2000. This modern lineage from the Pacific Northwest was repeatedly introduced into the east coast of the United States until it became endemic to the area by 2008, when it initiated a differentiation process leading to the emergence of the modern U.S. northeast population, which was responsible for large outbreaks of illness from 2013 onward. Our results identified recombination as the major source of genomic variation with a critical contribution to the major processes of diversification within the ST36 group and clear implications in the evolution of the modern lineages in Spain and the United States. In particular, recombination was of crucial importance in the emergence and diversification of the modern populations in the United States. Homologous recombination has been previously identified as a major evolutionary driver in *V. parahaemolyticus*, with a high level of recombination in environmental strains ($r/m = 39.8$) (18) and more moderate levels in disease-related populations (12). Here we demonstrated that most of the genetic divergence within this ST36 clonal population occurred by recombination, which introduced almost twice as many substitutions as mutations. Furthermore, a fine-tuned analysis of recombination rates for each node revealed that recombination played a fundamental role in the evolution of the modern lineages, reaching r/m rates of 13 (overall) and around 25 in particular subpopulations undergoing high diversification. These data stress the critical importance of recombination not only as a source of variation among the highly diverse environmental populations but also within the major clonal populations that emerged from them (12) as a major driver of the emergence of new pathogenic variants within the population.

Another relevant aspect of the evolution of this clone was the diversity in the mutation rates found across lineages. The highest evolutionary rate was found in the old PNW lineage, which also showed a higher level of diversity and the largest genomes among the strains analyzed. These particular characteristics were uniquely retained by

the strains from Spain, which tighten the links between these populations. Moreover, the present study provided a unique perspective on the evolutionary changes that occurred within a single population of *V. parahaemolyticus* in the extremely infrequent process of transition from a locally adapted clone to an epidemic clone undergoing a transcontinental pandemic expansion event. The modern populations from the United States, both western and eastern lineages, showed lower evolutionary rates and smaller genomes than their ancestral lineages, where almost all of the processes of diversification and evolution were driven by recombination. Although this needs to be examined in further detail, a first analysis suggests that the gene number reduction and lower mutation rate could be associated with a more specialized lifestyle as a result of niche adaptation. Genome reduction has been observed in many bacterial lineages in their process of specialization to new environments (35). This pattern of genome shrinkage has been recently documented in other free-living marine organisms, such as *Prochlorococcus* (36), which has undergone a genome reduction as a result of adaptation to the environment. We assume that a similar process may occur in the modern lineage of ST36 evolving through genome reduction resulting from specialization to narrow ecologic niches, limiting its versatility and survival under changing conditions. In terms of colonization, a highly specialized population may lead to a higher rate of survival over the dispersal and also a higher rate of success in the introduction into new areas. Recent experimental observations have revealed a link between genome reduction and a growth rate decrease in bacteria (37). Similar circumstances may have occurred over the evolution of ST36, where multiple genomic deletions may lead to decreases in the growth rate of modern lineages of this clone, reducing the mutation rate because of a lower number of cell divisions. Although the ecologic implications of this evolutionary pattern need to be explored further, it would be important to analyze other *V. parahaemolyticus* clones undergoing similar processes of geographic expansion to assess whether this is a common strategy in the evolution of major epidemic clones. Finally, the exceptional warming trend and regime shift (from 13.8 to 16°C) identified in the northeast region of the United States coinciding with the expansion of the ST36 populations in the area (Fig. 4) may be the definitive factors contributing to the adaptation of these populations and fostering the growth of populations and interactions between them.

V. parahaemolyticus infections are currently undergoing a process of geographic expansion, reaching new regions and typically associated with the introduction of strains originated from a remote area. Despite numerous studies reporting these particular patterns of spreading (e.g., reference 5), little is known about the mechanisms and biological strategies used by this organism over the process of dispersal. The release of ballast water transported by cargo ships has been identified as one of the potential vehicles of dispersal and sources of introduction of foreign *Vibrio* strains (38) and has been associated with outbreaks occurring in areas in close proximity to important international ports (e.g., references 39 and 40). Movement of oceanic waters was also documented as a mechanism of dispersal in some instances where the emergence and onset of infections correlated with the intrusion of warm oceanic waters into the region (6, 32). The decrease in the extent of sea ice observed in the Arctic over the last 2 decades has potentially activated a new route for ship traffic through the Bering Strait, allowing an effective connection between the west and east coasts of the United States and the potential dispersal of *Vibrio* populations. In a similar context, the melting of Arctic sea ice is removing the physical boundaries between the Pacific and Atlantic Oceans, opening a natural route for the migration of plankton species between both coasts of North America documented over recent years (41, 42). Without ruling out these two alternatives, it seems unlikely that these natural processes could have provided the opportunity for recurrent introductions of ST36 populations on the east coast of the United States. Furthermore, the presence of ST36 strains in the northwest of Spain represents an additional obstacle to the identification of a single mechanism for the dispersal of this clone. As an additional alternative, the global trade of shellfish may have also been a contributor to the dispersal of *V. parahaemolyticus*

populations. Recent genetic studies tracking the global distribution and introduction of Manila clams in Europe have identified the origin of clam populations introduced into the northwest of Spain in the Pacific coast of Canada with frequent importations of clams from British Columbia in Canada over the end of the 1990s and the beginning of the 2000s (43, 44).

Fine-resolution genome-wide analysis of ST36 strains over the course of geographic expansion has facilitated a better understanding of the evolution of this clone over the process of dispersal and introduction in areas of the United States and Spain. A similar approach applied to the study of other clonal groups undergoing similar processes of cross-continental expansion could help to assess whether the evolutionary patterns identified here are shared by other pathogenic *V. parahaemolyticus* strains in their transition from local distribution to the status of an epidemic clone with a global impact. Furthermore, a more extensive analysis combining disciplines such as evolution, climate science, and oceanography will provide new insights into the complex interactions between these populations and the variable ecologic conditions of their surrounding environments over the process of diversification, aspects that are critical to an understanding of the basis of the mechanisms driving the evolution of novel pathogenic clones and the initiation of geographic expansion and epidemic radiation.

MATERIALS AND METHODS

Bacterial strains and DNA extraction. The 44 *V. parahaemolyticus* strains sequenced in this study are listed in Table S1. Data from four additional ST36 strains (10296, 12310, 3256, and 10329) previously sequenced were retrieved from NCBI for use in the genomic comparison (Table S1). These 48 strains were selected on the basis of geography (representing both the Pacific and Atlantic coasts of the United States) and association with sporadic illnesses and outbreaks. Six of the ST36 strains (G25, G30, G31, G37, G36, and G35) represent the single outbreak in Galicia (northwest of Spain) in 2012 (17).

All 44 strains sequenced in this study were retrieved from storage (−80°C freezer), transferred to Luria-Bertani (LB) medium with 3% NaCl, and incubated at 37°C with shaking at 250 rpm. Genomic DNA was extracted from overnight cultures with the DNeasy Blood and Tissue kit (Qiagen, Valencia, CA).

Genome sequencing. The genomes of 43 strains were sequenced by MiSeq (Illumina) with a minimum coverage of 40× to 120×. Libraries were prepared with the Nextera XT DNA sample preparation kit (Illumina). One isolate was sequenced on the Pacific Biosciences (PacBio) RS II platform by the Institute for Genome Sciences, University of Maryland School of Medicine (Baltimore, MD). The continuous long-read data were *de novo* assembled by the PacBio hierarchical genome assembly process (HGAP version 2.0) by using default parameters. The assembled sequences were annotated by using the NCBI Prokaryotic Genome Automatic Annotation Pipeline (45) and subsequently deposited at DDBJ/EMBL/GenBank. The closed genome sequence of strain 10239 was sequenced with 100× coverage.

Sequence processing, genome assembly, and core genome. Reads were quality trimmed with Trimmomatic v0.32 (46). The first 20 bases were removed from each read and a 4-base-wide sliding window was used to cut when the average Phred quality score per base was <15. Reads of <50 bp were removed from the data set. Draft genomes were assembled *de novo* for each strain with the A5-miseq pipeline v20140604 (47) and annotated with Prokka v1.11 (48). The core genome was produced with Harvest v1.0.1 (49) by using the ST36 (strain 10329; PacBio) genome as the reference. Sites with gaps in the multiple-genome alignment of the ST36 strains were removed with trimAl v1.4 (50). The core genome of the reference strain was annotated with Prokka v1.11 (48), and the coordinates of the predicted coding regions were used to extract the corresponding regions from the core genome of all other strains. A core gene alignment was created for each strain by concatenating all of the predicted genes.

Pangenome analysis. Before running the pangenome analysis, contamination in the genomes was assessed with acdc (51). Genomes with unusual gene content or genome size identified during the pangenome analysis and suspected of possible contamination were analyzed to identify contamination and remove any suspected instances where contaminant sequences were identified. Of all the genomes, only two had an unusual genome size or unusual gene content (VP-143A and Vp-G9), and contaminant sequences were identified and filtered out. Pangenome analysis was carried out with Roary (52). A plot summarizing the core and accessory genes was produced with the roary2svg.pl script.

SNP calling and phylogenetic inference. SNPs were called with Harvest v1.0.1 (49) and annotated with SnpEff (53). For each chromosome, SNPs in recombining, nonrecombining, and coding regions were determined and SNP frequencies were plotted across both chromosomes. Phylogenetic inference by ML was performed on both the core genome and core gene data sets with RAxML v8.1.15 (54) and the GTRGAMMA model (1,000 bootstrap replicates). Subsequent searches for the best trees were conducted by using the GTRCAT model approximation.

Recombination testing. Recombination analysis of the core genome data set was performed with ClonalFrameML (22) and by using the best ML tree produced by RAxML as the starting tree. ClonalFrameML estimated the ratio of recombination to mutation rates (R/θ), the mean length of recombination events (δ), and the average distance between events (ν). Identified recombining regions were removed from the core genome data set with BEDTools (55) to create a nonrecombining core genome

data set. Similarly, all genes that were part of recombining regions were discarded from the core genes, creating a nonrecombining core gene data set. To confirm the results obtained by ClonalFrameML and explore the *r*/*m* rate for each node and strain, we used Gubbins (56), which identifies loci containing high densities of base substitutions and constructs an ML phylogeny based on SNPs identified outside the regions spotted as undergoing recombination.

Population structure. ADMIXTURE (57) was used to gain insights into the population structure and ancestral relationship of the *V. parahaemolyticus* ST36 group. Briefly, VCFtools (58) was used to convert the SNPs extracted from the genome alignment without recombination (see above) to PLINK format v1.07 (59), producing PED (which describes the individuals and genetic data) and MAP (describes the 349 SNPs, including their positions) files. SHAPEIT v1 (60) was used with default parameters to phase the data. The haplotype data obtained were subsequently used to estimate the ML of individual ancestries by ADMIXTURE. Different numbers of populations (K values of 1 to 9) were evaluated, and a K value of 6 was chosen as a sensible modeling choice (exhibited less cross-validation error than other K values). The output containing the ancestry fractions (Q) and allele frequency of the inferred ancestral population (P) was plotted with an R script (https://github.com/zeeev/ZevRTricks/blob/master/Addmixture2.plots.R).

Genetic diversity. The genetic diversity of the strains was calculated from within and between the six populations previously identified by ADMIXTURE analysis. The mean nucleotide diversity of populations with more than two genomes and the mean interpopulation diversity (previously identified by ADMIXTURE analysis) were calculated with MEGA7 (61) by using the core genome alignment without recombination regions, 100 bootstrap replications for variance estimation, and default settings. The plots were generated with ggplot2 (http://ggplot2.org/) and plotly (https://plot.ly/).

Bayesian phylodynamic analysis. The temporal signal in the data and how well molecular phylogenies conform to a molecular clock were initially explored with Path-o-gen v1.4 (62) (renamed Tempest; http://tree.bio.ed.ac.uk/software/tempest/) by regression analysis of the root-to-tip distance over time. Best-fitting root analysis identified the tree and the root of the tree that gave the best fit to the hypothesis of a constant rate of evolution. The spatial dynamics of ST36 were constructed by a Bayesian discrete phylogeographic approach in BEAST v1.8.1 (63) on the basis of the nonrecombining core genome sequences of 48 samples isolated at different times (1988 to 2013) at different locations on the Pacific Northwest or Atlantic coast of the United States and in Spain (Table S1). A total of 349 SNPs and 4,396,495 bp of nonpolymorphic sites were used for Bayesian inference by using the Hasegawa-Kishino-Yano (HKY) nucleotide substitution model, accounting for site heterogeneity with a gamma distribution (four categories). Different demographic (constant population size, exponentially growing population, Gaussian Markov random field, Bayesian skyride, and Bayesian skyline plot) and molecular clock (strict, random, and uncorrelated lognormal) models were run. For each demographic-molecular clock combination, the harmonic mean estimator (64), posterior simulation-based analogue of Akaike's information criterion (65), and PS and SS values (66) were calculated to select the best demographic model for this data set. The selection of the demographic model was based on the comparison of Bayes factors after thermodynamic integration to compute the marginal likelihood of each model by PS and SS methods. Each model was run for 100,000,000 states, with a sample frequency of 10,000, to check for convergence in the data set. A final target tree was generated with the TreeAnnotator utility in BEAST (burn-in of 10,000,000 states), and all effective sample size (ESS) values for the model parameters in the runs were >200.

Phylogenetic analysis and evolutionary rate estimation based on local clock. To investigate the evolutionary rates within all of the ST36 lineages identified by ADMIXTURE, we performed local clock model analyses with BEAST v1.8 (62). A fixed local clock under a demographic model of constant population size was selected as the model with the fewest parameters to prevent overfitting, as suggested by Ho and Duchêne (67). We used as the input the core genome alignment (4,396,497 bp) without recombinant regions. The strains were classified as different taxa on the basis of the six populations identified by ADMIXTURE. Chains were run for 50 million iterations and sampled every 1,000 generations. Convergence of parameters was confirmed by calculating the ESS with Tracer v1.6.1 (http://beast.community/tracer) and excluding an initial 10% of the burn-in for each run. All parameter estimates showed an ESS of >200. The maximum-credibility trees were summarized with TreeAnnotator and visualized with FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree).

Spatial phylogenetic reconstruction. SPREAD v1.0.6 (68) was used to analyze and visualize phylogeographic reconstructions resulting from Bayesian inference of spatiotemporal diffusion by using the outputs from the Bayesian phylogeographic analysis. This software mapped phylogenies annotated with discrete spatial information on the ST36 genomes and exported high-dimensional posterior summaries to keyhole markup language for animation of the spatial diffusion through time in virtual-globe software such as Google Earth (https://www.google.com/earth/).

Analysis of SST trends on the east coast of the United States. Trends in mean SSTs were estimated by using daily SST data from a coastal area limited by the coordinates 37.75 to 39.25°N and 75.5 to 76.5°W. The mean SST data were obtained from the Optimum Interpolation SST 1/4° daily data set, which extends from 1982 to the present and is distributed by NOAA/National Centers for Environmental Information. This data set combines satellite retrievals and *in situ* SST data from ships and buoys. We used these analyzed fields to estimate the trends in the region of interest, detect possible significant regime shifts, and study the habitat suitability of *V. parahaemolyticus* in the region. Regime shift, defined as rapid reorganizations of ecosystems from one relatively stable state to another, was investigated with the Sequential Regime Shift Detection software (http://www.beringclimate.noaa.gov/regimes/). This program detects statistically significant shifts in the mean level and the magnitude of fluctuations in time series taking autocorrelation into account (69).

Accession number(s). The draft genome sequences of all 44 *V. parahaemolyticus* strains used in our study are available in GenBank under the accession numbers listed in Table S1. The genome sequence of strain 10329 is available in GenBank under accession number JWSS00000000.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.01425-17.

FIG S1, PDF file, 0.1 MB.
FIG S2, PDF file, 0.1 MB.
FIG S3, PDF file, 0.1 MB.
FIG S4, PDF file, 0.1 MB.
FIG S5, PDF file, 0.7 MB.
TABLE S1, DOCX file, 0.02 MB.
TABLE S2, DOCX file, 0.01 MB.
TABLE S3, DOCX file, 0.01 MB.
TABLE S4, DOCX file, 0.02 MB.
TABLE S5, DOCX file, 0.02 MB.

ACKNOWLEDGMENTS

We thank Cheryl Tarr (U.S. CDC) and Jessica Jones (U.S. FDA) for providing some of the strains included in this study and also for insightful comments on the manuscript. We thank Sam Sheppard and MRC Cloud Infrastructure for Microbial Bioinformatics (CLIMB) for support with computationally intensive tasks.

M. Abanto and J. Martinez-Urtaza were funded by Natural Environment Research Council (NERC) project NE/P004121/1. J. Trinanes was partially financially supported by NOAA/CoastWatch and NOAA/AOML. C. Baker-Austin and R. van Aerle acknowledge funding from Cefas Seedcorn (Project DP367). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. Austin B, Austin D, Sutherland R, Thompson F, Swings J. 2005. Pathogenicity of vibrios to rainbow trout (Oncorhynchus mykiss, Walbaum) and Artemia nauplii. Environ Microbiol 7:1488–1495. https://doi.org/10.1111/j.1462-2920.2005.00847.x.
2. Altekruse SF, Bishop RD, Baldy LM, Thompson SG, Wilson SA, Ray BJ, Griffin PM. 2000. Vibrio gastroenteritis in the US Gulf of Mexico region: the role of raw oysters. Epidemiol Infect 124:489–495. https://doi.org/10.1017/S0950268899003714.
3. Dechet AM, Yu PA, Koram N, Painter J. 2008. Nonfoodborne Vibrio infections: an important cause of morbidity and mortality in the United States, 1997–2006. Clin Infect Dis 46:970–976. https://doi.org/10.1086/529148.
4. DePaola A, Nordstrom JL, Bowers JC, Wells JG, Cook DW. 2003. Seasonal abundance of total and pathogenic *Vibrio parahaemolyticus* in Alabama oysters. Appl Environ Microbiol 69:1521–1526. https://doi.org/10.1128/AEM.69.3.1521-1526.2003.
5. Baker-Austin C, Stockley L, Rangdale R, Martinez-Urtaza J. 2010. Environmental occurrence and clinical impact of Vibrio vulnificus and Vibrio parahaemolyticus: a European perspective. Environ Microbiol Rep 2:7–18. https://doi.org/10.1111/j.1758-2229.2009.00096.x.
6. Martinez-Urtaza J, Bowers JC, Trinanes J, DePaola A. 2010. Climate anomalies and the increasing risk of Vibrio parahaemolyticus and Vibrio vulnificus illnesses. Food Res Int 43:1780–1790. https://doi.org/10.1016/j.foodres.2010.04.001.
7. Centers for Disease Control and Prevention (CDC). 2010. Preliminary FoodNet data on the incidence of infection with pathogens transmitted commonly through food—10 states, 2009. MMWR Morb Mortal Wkly Rep 59:418–422. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5914a2.htm.
8. 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. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6315a6.htm.
9. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM. 2011. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis 17:7–15. https://doi.org/10.3201/eid1701.P11101.
10. Baker-Austin C, Trinanes JA, Taylor NGH, Hartnell R, Siitonen A, Martinez-Urtaza J. 2013. Emerging Vibrio risk at high latitudes in response to ocean warming. Nat Clim Chang 3:73–77. https://doi.org/10.1038/nclimate1628.
11. 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. https://doi.org/10.1128/CMR.00025-06.
12. González-Escalona N, Martinez-Urtaza J, Romero J, Espejo RT, Jaykus LA, DePaola A. 2008. Determination of molecular phylogenetics of Vibrio parahaemolyticus strains by multilocus sequence typing. J Bacteriol 190:2831–2840. https://doi.org/10.1128/JB.01808-07.
13. Paranjpye R, Hamel OS, Stojanovski A, Liermann M. 2012. Genetic diversity of clinical and environmental *Vibrio parahaemolyticus* strains from the Pacific Northwest. Appl Environ Microbiol 78:8631–8638. https://doi.org/10.1128/AEM.01531-12.
14. 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. https://doi.org/10.1371/journal.pone.0055726.
15. 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. https://doi.org/10.1056/NEJMc1305535.
16. Banerjee SK, Kearney AK, Nadon CA, Peterson CL, Tyler K, Bakouche L, Clark CG, Hoang LD, 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. https://doi.org/10.1128/JCM.03047-13.

17. Martinez-Urtaza J, Powell A, Jansa J, Rey JL, Montero OP, Campello MG, López MJ, Pousa A, Valles MJ, Trinanes J, Hervio-Heath D, Keay W, Bayley A, Hartnell R, Baker-Austin C. 2016. Epidemiological investigation of a foodborne outbreak in Spain associated with US West Coast genotypes of Vibrio parahaemolyticus. Springerplus 5:87. https://doi.org/10.1186/s40064-016-1728-1.
18. Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199–208. https://doi.org/10.1038/ismej.2008.93.
19. Kalburge SS, Polson SW, Boyd Crotty K, Katz L, Turnsek M, Tarr CL, Martinez-Urtaza J, Boyd EF. 2014. Complete genome sequence of *Vibrio parahaemolyticus* environmental strain UCM-V493. Genome Announc 2:e00159-14. https://doi.org/10.1128/genomeA.00159-14.
20. 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. https://doi.org/10.1016/S0140-6736(03)12659-1.
21. Ahn S, Chung HY, Lim S, Kim K, Kim S, Na EJ, Caetano-Anolles K, Lee JH, Ryu S, Choi SH, Kim H. 2016. Complete genome of Vibrio parahaemolyticus FORC014 isolated from the toothfish. Gut Pathog 8:59. https://doi.org/10.1186/s13099-016-0134-0.
22. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041. https://doi.org/10.1371/journal.pcbi.1004041.
23. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 13:601–612. https://doi.org/10.1038/nrg3226.
24. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462–465. https://doi.org/10.1038/nature10392.
25. Vezzulli L, Grande C, Reid PC, Hélaouët P, Edwards M, Höfle MG, Brettar I, Colwell RR, Pruzzo C. 2016. Climate influence on Vibrio and associated human diseases during the past half-century in the coastal North Atlantic. Proc Natl Acad Sci U S A 113:E5062–E5071. https://doi.org/10.1073/pnas.1609157113.
26. González-Escalona N, Gavilan RG, Brown EW, Martinez-Urtaza J. 2015. Transoceanic spreading of pathogenic strains of Vibrio parahaemolyticus with distinctive genetic signatures in the recA gene. PLoS One 10:e0117485. https://doi.org/10.1371/journal.pone.0117485.
27. Gonzalez-Escalona N, Gavilan RG, Toro M, Zamudio ML, Martinez-Urtaza J. 2016. Outbreak of Vibrio parahaemolyticus sequence type 120, Peru, 2009. Emerg Infect Dis 22:1235–1237. https://doi.org/10.3201/eid2207.151896.
28. Harth E, Matsuda L, Hernández C, Rioseco ML, Romero J, González-Escalona N, Martínez-Urtaza J, Espejo RT. 2009. Epidemiology of Vibrio parahaemolyticus outbreaks, southern Chile. Emerg Infect Dis 15:163–168. https://doi.org/10.3201/eid1502.071269.
29. Ansaruzzaman M, Lucas M, Deen JL, Bhuiyan NA, Wang XY, Safa A, Sultana M, Chowdhury A, Nair GB, Sack DA, von Seidlein L, Puri MK, Ali M, Chaignat CL, Clemens JD, Barreto A. 2005. Pandemic serovars (O3:K6 and O4:K68) of *Vibrio parahaemolyticus* associated with diarrhea in Mozambique: spread of the pandemic into the African continent. J Clin Microbiol 43:2559–2562. https://doi.org/10.1128/JCM.43.6.2559-2562.2005.
30. Gil AI, Miranda H, Lanata CF, Prada A, Hall ER, Barreno CM, Nusrin S, Bhuiyan NA, Sack DA, Nair GB. 2007. O3:K6 serotype of Vibrio parahaemolyticus identical to the global pandemic clone associated with diarrhea in Peru. Int J Infect Dis 11:324–328. https://doi.org/10.1016/j.ijid.2006.08.003.
31. Okuda J, Ishibashi M, Hayakawa E, Nishino T, Takeda Y, Mukhopadhyay AK, Garg S, Bhattacharya SK, Nair GB, Nishibuchi M. 1997. Emergence of a unique O3:K6 clone of *Vibrio parahaemolyticus* in Calcutta, India, and isolation of strains from the same clonal group from Southeast Asian travelers arriving in Japan. J Clin Microbiol 35:3150–3155.
32. Martinez-Urtaza J, Huapaya B, Gavilan RG, Blanco-Abad V, Ansede-Bermejo J, Cadarso-Suarez C, Figueiras A, Trinanes J. 2008. Emergence of Asiatic vibrio diseases in South America in phase with el Nino. Epidemiology 19:829–837. https://doi.org/10.1097/EDE.0b013e3181883d43.
33. Martinez-Urtaza J, Trinanes J, Gonzalez-Escalona N, Baker-Austin C. 2016. Is el Nino a long-distance corridor for waterborne disease? Nat Microbiol 1:16018. https://doi.org/10.1038/nmicrobiol.2016.18.
34. 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. https://doi.org/10.3389/fmicb.2015.00272.
35. Lee MC, Marx CJ. 2012. Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genet 8:e1002651. https://doi.org/10.1371/journal.pgen.1002651.
36. Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik BP, Paulsen IT, de Marsac NT, Wincker P, Dossat C, Ferriera S, Johnson J, Post AF, Hess WR, Partensky F. 2008. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol 9:R90. https://doi.org/10.1186/gb-2008-9-5-r90.
37. Kurokawa M, Seno S, Matsuda H, Ying BW. 2016. Correlation between genome reduction and bacterial growth. DNA Res 23:517–525. https://doi.org/10.1093/dnares/dsw035.
38. DePaola A, Capers GM, Motes ML, Olsvik O, Fields PI, Wells J, Wachsmuth IK, Cebula TA, Koch WH, Khambaty F, et al. 1992. Isolation of Latin American epidemic strain of Vibrio cholerae O1 from US Gulf Coast. Lancet 339:624. https://doi.org/10.1016/0140-6736(92)90917-R.
39. Martinez-Urtaza J, Simental L, Velasco D, DePaola A, Ishibashi M, Nakaguchi Y, Nishibuchi M, Carrera-Flores D, Rey-Alvarez C, Pousa A. 2005. Pandemic Vibrio parahalemolyticus O3:K6, Europe. Emerg Infect Dis 11:1319–1320. https://doi.org/10.3201/eid1108.050322.
40. DePaola A, Kaysner CA, Bowers J, Cook DW. 2000. Environmental investigations of *Vibrio parahaemolyticus* in oysters after outbreaks in Washington, Texas, and New York (1997 and 1998). Appl Environ Microbiol 66:4649–4654. https://doi.org/10.1128/AEM.66.11.4649-4654.2000.
41. Wassmann P, Duarte CM, Agustí S, Sejr MK. 2011. Footprints of climate change in the Arctic marine ecosystem. Glob Chang Biol 17:1235–1249. https://doi.org/10.1111/j.1365-2486.2010.02311.x.
42. Wisz MS, Broennimann O, Grønkjær P, Møller PR, Olsen SM, Swingedouw D, Hedeholm RB, Nielsen EE, Guisan A, Pellissier L. 2015. Arctic warming will promote Atlantic-Pacific fish interchange. Nat Clim Chang 5:261–265. https://doi.org/10.1038/nclimate2500.
43. Chiesa S, Lucentini L, Freitas R, Marzano FN, Breda S, Figueira E, Caill-Milly N, Herbert RJH, Soares AMVM, Argese E. 2017. A history of invasion: COI phylogeny of Manila clam Ruditapes philippinarum in Europe. Fish Res 186:25–35. https://doi.org/10.1016/j.fishres.2016.07.024.
44. Cordero D, Delgado M, Liu BZ, Ruesink J, Saavedra C. 2017. Population genetics of the Manila clam (Ruditapes philippinarum) introduced in North America and Europe. Sci Rep 7:39745. https://doi.org/10.1038/srep39745.
45. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, Kiryutin B, O'Neill K, Resch W, Resenchuk S, Schafer S, Tolstoy I, Tatusova T. 2009. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Res 37:D216–D223. https://doi.org/10.1093/nar/gkn734.
46. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
47. Coil D, Jospin G, Darling AE. 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics 31:587–589. https://doi.org/10.1093/bioinformatics/btu661.
48. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
49. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15:524. https://doi.org/10.1186/PREACCEPT-2573980311437212.
50. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
51. Lux M, Krüger J, Rinke C, Maus I, Schlüter A, Woyke T, Sczyrba A, Hammer B. 2016. acdc—Automated Contamination Detection and Confidence estimation for single-cell genome data. BMC Bioinformatics 17:543. https://doi.org/10.1186/s12859-016-1397-7.
52. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

53. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly 6:80–92. https://doi.org/10.4161/fly.19695.
54. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
55. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
56. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196.
57. Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664. https://doi.org/10.1101/gr.094052.109.
58. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
59. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575. https://doi.org/10.1086/519795.
60. Delaneau O, Marchini J, Zagury JF. 2011. A linear complexity phasing method for thousands of genomes. Nat Methods 9:179–181. https://doi.org/10.1038/nmeth.1785.
61. Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054.
62. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214. https://doi.org/10.1186/1471-2148-7-214.
63. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973. https://doi.org/10.1093/molbev/mss075.
64. Newton MA, Raftery AE. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J R Stat Soc B Stat Methodol 56:3–48. http://www2.stat.duke.edu/~scs/Courses/Stat376/Papers/NormConstants/NewtonRaftery1994.pdf.
65. Raftery A, Newton M, Satagopan J, Krivitsky P. 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity, p 1-45. *In* Bernardo JM, Bayarri MJ, Berger JO, Dawd AP, Heckerman D, Smith AFM, West M (ed), Bayesian statistics 8. Oxford Science Publications, New York, NY.
66. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol 29:2157–2167. https://doi.org/10.1093/molbev/mss084.
67. Ho SY, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol 23:5947–5965. https://doi.org/10.1111/mec.12953.
68. Bielejec F, Rambaut A, Suchard MA, Lemey P. 2011. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics 27:2910–2912. https://doi.org/10.1093/bioinformatics/btr481.
69. Rodionov SN. 2006. Use of prewhitening in climate regime shift detection. Geophys Res Lett 33:L22603. https://doi.org/10.1029/2006GL025904.