Journal of Phylogenetics & Evolutionary Biology

Research Article Open Access

Genealogy of the Genome Components in the Highly Homogeneous Pandemic *Vibrio parahaemolyticus* Population

David E Loyola[1], Cristian Yañez[1], Nicolas Plaza[1], Katherine García[2] and Romilio T Espejo[1,3]*

[1]*Centro Nacional de Genómica y Bioinformática, Universidad de Chile, Santiago, Chile*
[2]*Instituto de Ciencias Biomédicas, Universidad Autónoma de Chile, Santiago, Chile*
[3]*Instituto de Nutrición y Tecnología de los Alimentos, Universidad de Chile, Santiago, Chile*

Abstract

Bacterial genomes evolve through two different mechanisms: 1) changes in or occasional loss of ancestral genes, which preserve the founder clonal genealogy or frame, and 2) sporadic gains of new genes via horizontal gene transfer, which introduces DNA with a different genealogy or clonal frame. Evolution has led to the emergence of a pathogenic strain pandemic of *Vibrio parahaemolyticus*, which has propagated globally, causing large outbreaks of seafood-associated diarrhea. The low sequence diversity of the pathogenic strain of *Vibrio parahaemolyticus* genome provides a model that can reveal evolutionary mechanisms that are hidden in bacteria with a greater diversity. Here, we assess the clonal genealogy of the genome components of *Vibrio parahaemolyticus* and identify recombinant segments in 31 isolates obtained worldwide. By comparing whole genomes using a procedure that accounts for at least 98% of the reads obtained from any isolate after high throughput sequencing, we determined that the fraction of the whole genome retaining the founder clonal frame varied from 96.7%-100% of the accountable reads. The fraction in chromosomes with other genealogies varied from 0%-3.3% and in extra-chromosomal elements from 0%-4.2%, with the relative impact of mutation and recombination varying greatly between isolates. The likely causes for this variation are proposed.

Keywords: Bacterial; Genome; Evolution; Phylogeny; Clones; Gene transfer

Introduction

Bacterial genomes evolve through changes or occasional loss of ancestral genes, or via sporadic gains of new genes through horizontal gene transfer (HGT) [1]. Changes in ancestral genes occur during clonal reproduction (reproduction without the exchange of genes with a different clonal genealogy) through a variety of mechanisms including point mutations, genome rearrangement, insertion-deletion, and duplication. However, occasional HGT from bacteria with a different clonal genealogy introduce genes with different clonal frames. The transferred DNA can either recombine with chromosomal DNA or remain as extra chromosomal elements [1]. Recombination in bacteria is, however, non-reciprocal, analogous to gene conversion where the donor contributes only a small contiguous segment of DNA, while the recipient contributes the rest of the genome. DNA that does not recombine may remain as extra chromosomal elements if it can replicate, but in this form such elements are unstable and mobile [2]. These changes generate polymorphisms within the population genome leading to evolution. Understanding the evolution and phylogeny of a bacterial population requires distinguishing between clonal and non-clonal evolution; therefore, the number of clonal genealogies, or clonal frames, in the genome must be distinguished. With the advent of high-throughput sequencing technologies, evolution is now commonly studied by comparing genomes. However, these comparisons usually consider only segments with orthologous genes present in the chromosomes of the strains included in the study. A comprehensive view of evolution in bacteria requires an assessment of the extent and type of changes in the whole genome, including orthologous and non-orthologous genes in both chromosomal and extra-chromosomal elements.

Evolution has led to the emergence of new pathogenic strains. Pandemic *Vibrio parahaemolyticus* first described in 1996 as serotype O3:K6 [3], emerged in Southeast Asia and has propagated globally, causing large outbreaks of seafood-associated diarrhoea [4]. Its low sequence diversity provides a model that can reveal evolutionary mechanisms that are hidden in bacteria with greater sequence diversity because of extensive recombination or genomic reduction [5]. Previous comparisons of some of these genomes showed that the core genomes of most isolates in this population differed by less than 200 single-nucleotide variants (SNVs) per genome of approximately 5,000,000 base pairs (bp). Larger differences between isolates were caused by the presence of segments of DNA with a different clonal genealogy acquired through HGT and gene conversion [6]. Most of the genomic differences between isolates corresponded to the presence of regions that are unique to only one or two isolates acquired via HGT [7]. This study analyzed the evolution of 31 pandemic *V. parahaemolyticus* isolates by assessing the clonal frames of chromosomal and extra-chromosomal genes in the whole genome by including all of the filtered reads obtained after sequencing, in a process we refer to as "read accounting".

Materials and Method

Samples

The isolates used in this study were described previously (Table 1). The Research Institute of Medical Disease (RIMD) 2210633 genome sequence was used as a reference. VpKX corresponds to the RIMD 2210633 isolate from the Research Institute of Medical Disease in Japan in 2002 [8], which were maintained in our laboratory. Eight isolates corresponded to Chilean isolates whose genomes were previously sequenced, analysed and compared [7]. Twenty isolates corresponded to new isolates whose genome sequences were available in the Sequence Read Archive (SRA) of NCBI and were identified as "parahaemolyticus" and "pandemic", and which were obtained from this database using the

***Corresponding author:** Romilio T Espejo, Centro Nacional de Genómica y Bioinformática, Instituto de Nutrición y Tecnología de los Alimentos, Universidad de Chile, Santiago, Chile, Tel: +56-2978-1411; E-mail: romilio.espejo@gmail.com

Received March 15, 2016; **Accepted** April 18, 2016; **Published** April 26, 2016

Citation: Loyola DE, Yañez C, Plaza N, García K, Espejo RT (2016) Genealogy of the Genome Components in the Highly Homogeneous Pandemic *Vibrio parahaemolyticus* Population. J Phylogen Evolution Biol 4: 165. doi:10.4172/2329-9002.1000165

Sample	Isolation	Sample	Geographic	Biosample reference
alias	date	origin	origin	
RIMD	1996		SEA	[8]
2210633				
VPKX	1996	Stools	SEA	[8]
CHC223.4	2004	Stools	Chile	[23]
INC48.96	1996	Stools	India	
ATC220.98	1998	Stools	Chile	[23]
PMC48.4	2004	Stools	Chile	[23]
PMC14.7	2007	Stools	Chile	[26]
PEC288.1	2001	Stools	Peru	
PMC58.5	2005	Stools	Chile	[25,27]
BAA26.8	2008	Water	Bangladesh	
PMA109.5	2005	Mussel	Chile	[25]
CHC14.1	2001	Stools	Chile	
BAA28.8	2008	Water	Bangladesh	
INC2189.9	2009	Stools	India	
PMC58.7	2007	Stools	Chile	[26]
USC949.6	2006	Stools	USA	
PMA37.5	2005	Mussel	Chile	[25]
PMC81.13	2013	Stools	Chile	
USC86.12	2012	Stools	USA	
BAC603.6	2006	Stools	Bangladesh	
USC87.12	2012	Stools	USA	
ATC210.98	1998	Stools	Chile	[24]
USC85.12	2012	Stools	USA	
USA605.6	2006	Water	USA	
USA861.6	2006	Water	USA	
MOC265.4	2004	Stools	Mozambique	
INC4.97	1997	Stools	India	
INC2640.9	2009	Stools	India	
BAA21.8	2008	Water	Bangladesh	
INC250.98	1998	Stools	India	
INC232.988	1998	Stools	India	

SEA-EA, South-East Asia

Table 1: Properties of the pandemic *V. parahaemolyticus* isolates included in this study.

package SRAdb from R/Bioconductor [9]. One Chilean isolate obtained in 2013 (PMC81.13) was sequenced in our laboratory. The biosample numbers of the isolates included in this study are shown in Table 1S.

Bioinformatics analysis

DNA reads obtained in the SRA of NCBI were transformed into FASTQ files using Fastq-dump from SRA toolkit v2.4.0-1. These files, together with previously obtained FASTQ files, were analyzed for adapter clipping and quality trimming using Trimmomatic v0.32 [10] with a sliding window of 10, a quality threshold of 15 and a minimal sequence length of 35. They were corrected using POLLUX v1.00 for substitutions, insertions, deletions and homopolymers [11]. The filtered reads from each sample were then aligned against the RIMD 2210633 reference genome (GenBank: BA000031.2, length 3,288,558 and BA000032.2 length 1,877,212) using SMALT v0.7.6 with the following settings: wordlen 13, skipstep 13, and minscore 27 (the standard score for Smith-Waterman was match+1, mismatch-2, gap open-4, gap-extension-3), producing a Sequence Alignment Map (SAM) file containing the aligned reads for each isolate. SAM files were processed using Picard Tools v1.96 to convert from SAM to BAM (Binary Alignment Map) (Sam Format Converter module), sorting the BAM files by position (Sort Sam module), adding read groups (AddOrReplaceReadGroups module) and marking read duplicates (MarkDuplicates module). SNVs were then called using FreeBayes v0.9.20 with the following parameters: no-indels, no-mnps, no-complex, ploidy 1, min- base-quality 20, and min-alternate-fraction 0.75. The called SNVs were filtered by the depth of coverage using VCFtools v0.1.12b (1/4 the mean genome alignment depth of coverage). To pinpoint uncovered bases in the reference genome, we used GenomeCoverageBed from the BedTools package. A chromosomal scaffold sequence was subsequently constructed, correcting for SNVs and incorporating N at uncovered positions. A core genome sequence was finally built for each isolate, removing all positions that were uncovered in any isolate from the chromosome scaffold sequence.

Assembly

Unmapped reads were assembled *de novo* using all unmapped reads and flanking reads that mapped within 100 bp on either side of the uncovered segment using target assembly [12]. Only those uncovered segments longer than 100 bp were included. The Velvet assembler v1.2.10 was used for Illumina reads, and the best assembly was selected using VelvetOptimizer with k-mers between 25 and 91. The Mira assembler v4.0.2 assembler was used for Ion Torrent reads with the default parameters. Target-assembled contigs were subsequently filtered based on 3 criteria: 1) contigs without flanking reads, 2) contigs shorter than ¼ of 316 bp and 3) contigs with coverage lower than mean depth coverage of the genome. Those contigs containing elements in the UNIVEC database were also filtered out. *De novo* assembly was performed using both unmapped and mapped reads, with ¼ of filthering for coverage below mean depth coverage of the genome and for sequences in the UNIVEC database. Target and *de novo* assembled contigs were inserted into the chromosome scaffold sequence by aligning against the chromosomal scaffold sequence using MAUVE [13] and then Ns, uncovered bp, together with overlapping sequences were replaced by the contig sequence. Custom scripts were used for all procedures.

Phylogenetic analysis

The clonal frames for the segments containing SNVs in the core genome sequence were assessed using ClonalFrameML [14] by applying the algorithm to data in the maximum likelihood (ML) tree built in the PhyML program with the HKY85 substitution model, together with the core chromosomal sequences, as described previously [13]. To build the Phylogenetic 5.6 Kb tree, the presence of the 5.6 Kb insertion sequence (IS) in each isolate was detected by alignment of the target assembled contigs with the previously identified IS sequence [7]. The ISs identified in 25 isolates were aligned with MAFFT v7 [15] and a maximum likelihood (ML) tree built in the PhyML program with the HKY85 substitution model, using 100 bootstrap replicates and maximum parsimony for starting the tree.

Results

We compared the sequences of 31 pandemic *V. parahaemolyticus* isolates, among which there were fewer than 3,500 SNVs. These isolates were chosen because their low dissimilarity indicated a clonal nature. The main properties of the *V. parahaemolyticus* isolates are shown in Table 1.

Read accounting

Comprehensive comparisons of the whole genomes of isolates require the inclusion of all reads obtained through high-throughput sequencing. For this purpose reads were grouped within sequences likely to be compared; including core chromosomes, isolated chromosomes, and extra-chromosomal elements. As a first step, reads were mapped to the sequences of the two chromosomes of the reference genome RIMD 2210633 [8], and the mapped reads were subsequently used to build the core genome sequence of each isolate, comprised of regions shared by every isolate. The percentage of unmapped reads varied from 0.01%-16.55%. Unmapped reads may correspond to isolated DNA segments with low similarity or which are absent in RIMD 2210633. To account for these reads, they were assembled via target assembly, which consisted of assembling unmapped reads and read mapping on either side of unmapped segments. Using this procedure, reads in contigs with ends overlapping with the reference sequence were inserted in the chromosomal scaffold sequence of the isolate, and were comprised of all reads aligned with RIMD 2210633 [12]. For some isolates, when reads corresponded to highly polymorphic regions (particularly the serotype-coding region), unmapped reads were more easily located within chromosomal sequences using larger *de novo*-assembled contigs of all of the reads (mapped and unmapped reads). These procedures allowed 0%-96% of the unmapped reads to be incorporated into the chromosomal scaffold sequences of the different isolates. The sequence containing the incorporated contigs is referred to as the "chromosomal draft sequence". Unmapped reads in contigs that could not be incorporated into chromosomes because there was no clear overlap with the chromosomal sequence were characterized, and when the information supported it (such as the presence of a plasmid in the isolate [7], close sequence similarity to phages or plasmids, or the existence of other functions related to extra-chromosomal elements found using RAST [17]), they were assigned to extra-chromosomal elements. Reads that could not be assigned to chromosomes or extra-chromosomal elements were designated "unassigned reads". Figure 1 shows the percentages of reads assigned to chromosomes via the mapping and insertion of target contigs and those assigned to extra-chromosomal elements and unassigned reads for each of the isolates. This last group constitutes the portion of genomes whose origin could not be accounted for. Table 2S shows the actual numbers.

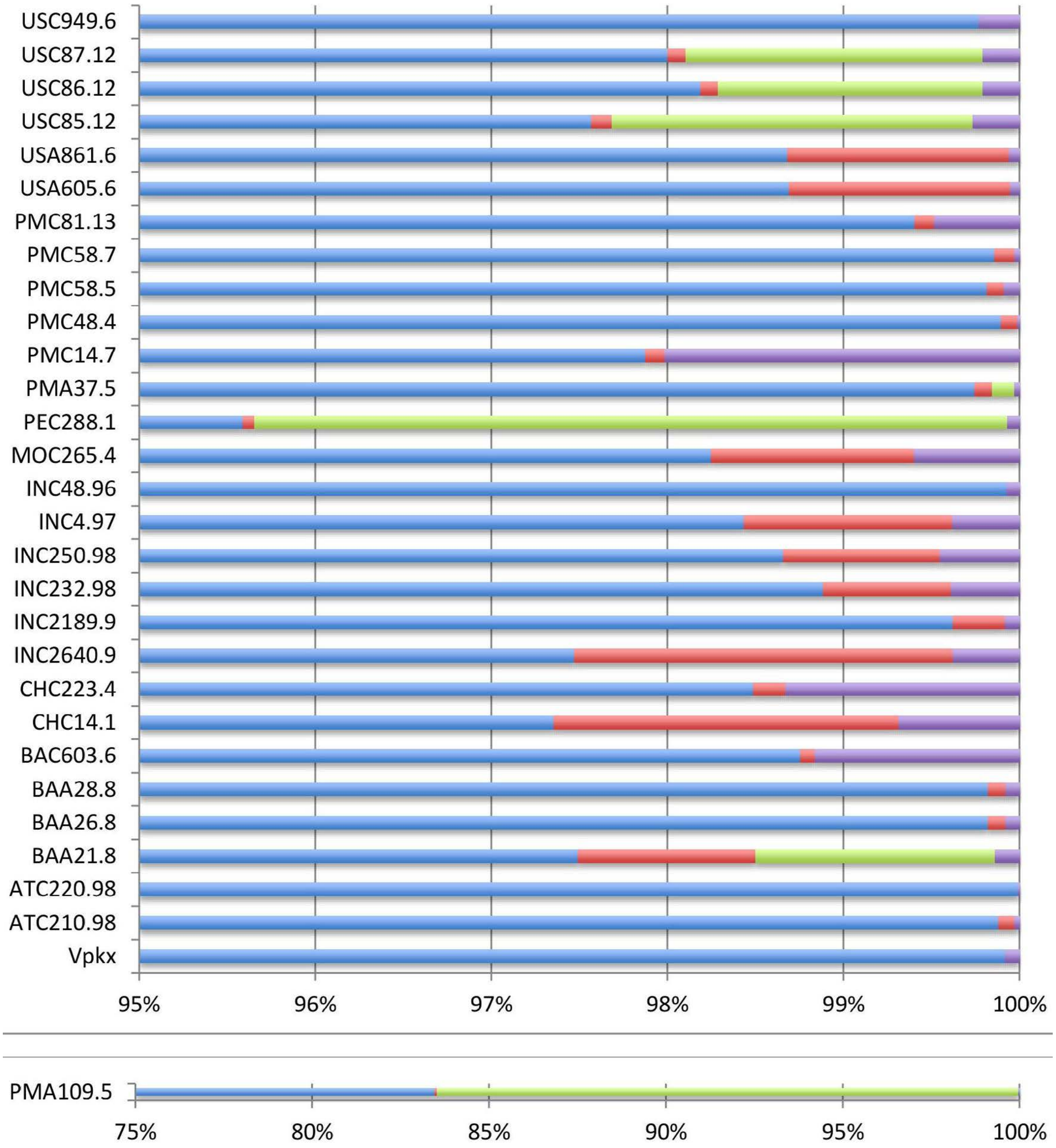
Mutation and recombination in core chromosomes

Core chromosome sequences, containing only those positions covered by reads from every isolate were employed for the preliminary analysis of clonal phylogeny. The lengths of the sequences of both core chromosomes totaled 5,086,167 bp, which was 79,603 bp shorter than the length of the two RIMD 2210633 chromosomes used as a reference. Most (95.3%) of the non-covered positions were in segments longer than 100 bp, and predominantly located in regions related to O:K antigen type determinants, phages, integrons and super-integrons. The number of SNVs in the core genome varied from 11 to 2919 between isolates. The clonal frames of segments containing these SNVs were determined using ClonalFrameML [13], which distinguishes between the effects of mutation and recombination on genetic data. By applying the ClonalFrameML algorithm to data in the ML tree and the core chromosomal sequences, we assessed non-pandemic clonal segments and calculated the ClonalFrameML tree, which only accounts for substitutions introduced by mutation (Figure 2). The number of SNVs within the pandemic clonal frame segments varied from 4 to 399. The mean parameters θ (rate were of estimated recombination/rate to be R/ of mutation) =0.203, δ (mean of the exponential distribution modelling the length of the recombinant segments) 1,808 ν (rate bp, of and nucleotide differences in recombinant stretches) =1.74 × 10^{-2}. Using these values the ratio of effects of recombination and mutation (r/m), in the whole population, was 6.4. Extensive variation was observed between the branches of the tree, as shown in Figure 2. Sixteen isolates showed no recombinant segments, while 15 exhibited recombinant segments totaling 65,317 bp. Forty-two recombinant or non-pandemic clonal segments with sizes ranging from 4-41,941 bp were detected in these 15 isolates. BLAST [18] analysis of the 26 recombinant segments longer than 50 bp showed eleven segments sharing a higher identity and coverage with *V. parahaemolyticus* strains other than the pandemic strain (Table 3S). A small recombinant segment of 348 bp containing part of the gene tdh coding for thermostable direct haemolysin associated with pathogenicity was identified in isolate USC87.12. Large recombinant regions were found in four isolates within the region related to O:K antigen type determinants (Table 3S), although most of this region was absent in the core chromosome because some isolates did not contain reads that aligned with RIMD 2210633 in this region.

Non-core regions of the chromosomes

The above comparison of core chromosomes did not include reads present in non-core regions of the isolates. To include reads in the chromosomes that were absent in the sequences obtained through mapping in the comparison of isolates, reads that did not map to RIMD 2210633 chromosomes were assembled via target assembly [12]. Contigs with ends that overlapped with the reference sequence were inserted in the chromosomal scaffold sequence of the isolate obtained after alignment with RIMD 2210633 at sites of RIMD 2210633 that were not covered by the reads from the isolate. Forty-nine target contigs were inserted at different positions in the chromosomes; twenty five corresponded to a 5.6 Kb IS inserted next to a tmRNA in tandem with another IS of similar size described in every isolate previously studied [6,7]. This second IS was present in every isolate for except RIMD 2210633 and 4 other isolates. In one isolate (CHC14.1), this IS was replaced by two larger IS contigs containing an approximately 28 Kb insertion showing closest similarity with a region in *V. parahaemolyticus* strain FORC_008 encoding a putative temperate phage. Comparison of the two ISs between isolates showed that while the IS present in every isolate did not exhibit sequence differences, the second IS found in most but not all isolates, was present in up to 15 SNVs, an unexpected mutation rate for segments of the pandemic clonal frame. Its closest match determined using BLASTn was to *V. harveyi* ATCC33843 (55% query coverage with 74% identity). Phylogenetic analysis based on this insertion sequence (Figure 3) indicated the existence of different insertion events and suggested that isolates from Chile and Peru are derived from a common ancestor that underwent the same insertion event.

BLAST analysis of the 22 other inserted target contigs showed 12 segments sharing higher identity and coverage with bacterial spp. other than *V. parahaemolyticus* (Table 2S). Isolates USA861.6, USA605.6, and INC4.97, which cluster together according to the IS (Figure 3), shared the same recombinant 57 Kb segment. A contig similar to this 57 Kb sequence was also present in MOC265.4, which forms a different cluster according to the IS but clusters together with the other three by ML (Figure 2). In three isolates in which reads corresponded to highly polymorphic regions (particularly the serotype-coding region) with many small target assembled contigs, unmapped reads were more easily located in the chromosomal scaffold sequences using larger de novo-assembled contigs of all of the reads (both mapped and unmapped reads) (Table 2S). Overall, these two procedures allowed 0%-96% of the reads that did not map to the RIMD 2210633 sequence to be incorporated into the chromosomal scaffold sequences of the different isolates (Figure 2 and Table 2S), generating the chromosomal draft sequences of each isolate.

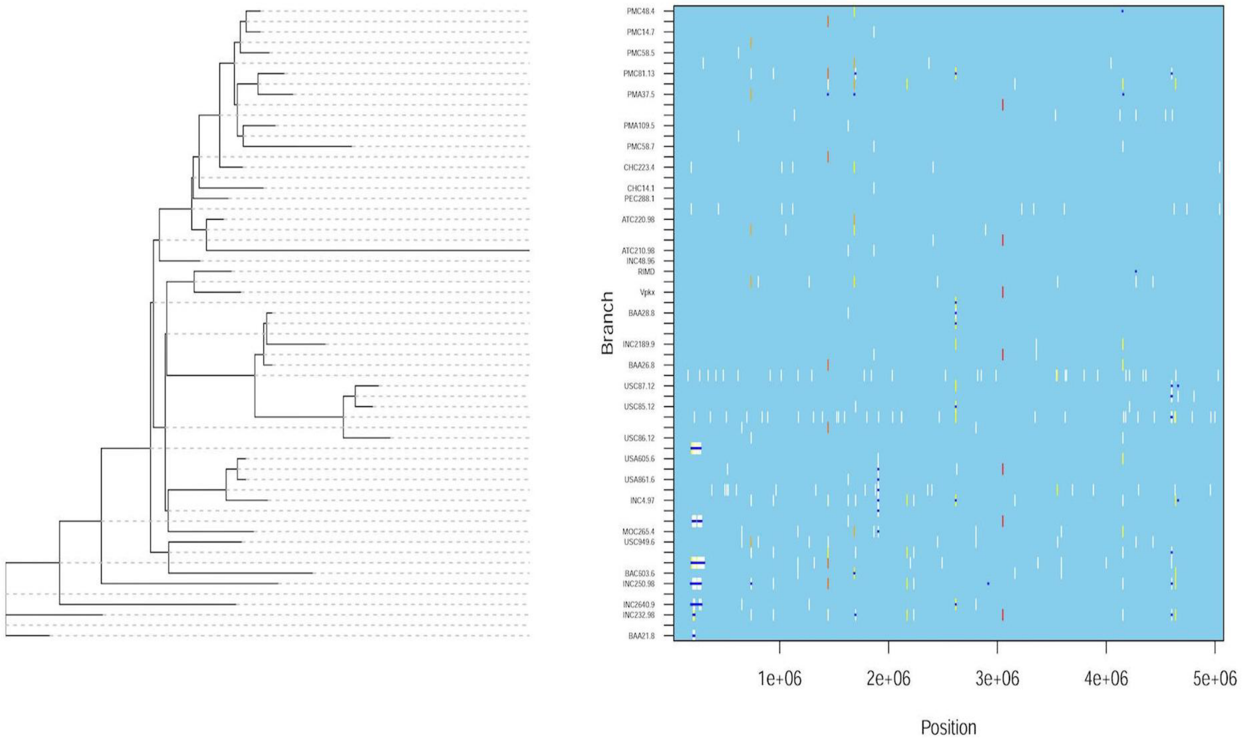


Blue: Aligned with the reference genome RIMD 2210633. Orange: Present in target assembly contigs inserted in the chromosomes. Green: In contigs or scaffolds classified as extra-chromosomal elements. Violet: Not assigned

Figure 1: Graphic representation of the percentages of reads ascribed to different parts of the genome

Extra-chromosomal elements

Reads that did not map to the final chromosomal sequences determined after the insertion of contigs obtained through de novo and target assembly may consist of reads that are part of extra-chromosomal elements or chromosomal segments that were not incorporated into the final assembly because overlapping sequences were not detected. For further analysis of these reads, those that did not map to the chromosomal draft sequence obtained for the corresponding isolates were assembled de novo, and the resulting contigs were analyzed with BLAST and annotated with RAST [17] to identify close relatives and functions that might indicate their possible nature and origin. Using this procedure, some contigs were identified as extra-chromosomal elements corresponding to plasmids or phages. These elements included extra-chromosomal elements previously identified in Chilean isolates [7] consisting of the linear plasmid prophage VP58.5 [19] and two other large plasmids, one with a length of ~ 82.0 Kb that was very similar to plasmid p0908 and similar to the enterobacteria phage P1 found in *Vibrio* spp. and another of 85.8 Kb with 99% identity to the plasmid pVPUCMV of the environmental strain *V. parahaemolyticus* UCM-V493 [20]. None of these plasmids was detected in the newly analyzed isolates, but three other extra-chromosomal elements were identified: 1) a large 92.6 Kb element found in an isolate from Peru (PEC288.1) that did not show any considerable relationship to sequences in the database, with the exception of sharing 1% of its sequence with 70% identity with transposons reported in plasmids [21], and which resembled a bacteriophage based on 31 genes that



Reconstructed substitutions (white vertical bars) are shown for each branch of the ML tree. Dark blue horizontal bars indicate recombination events detected through the analysis. X-axis: position in chromosomes 1 and 2 in tandem.

Figure 2: ClonalFrameML analysis of recombination in 30 genomes mapped to RIMD 2210633.

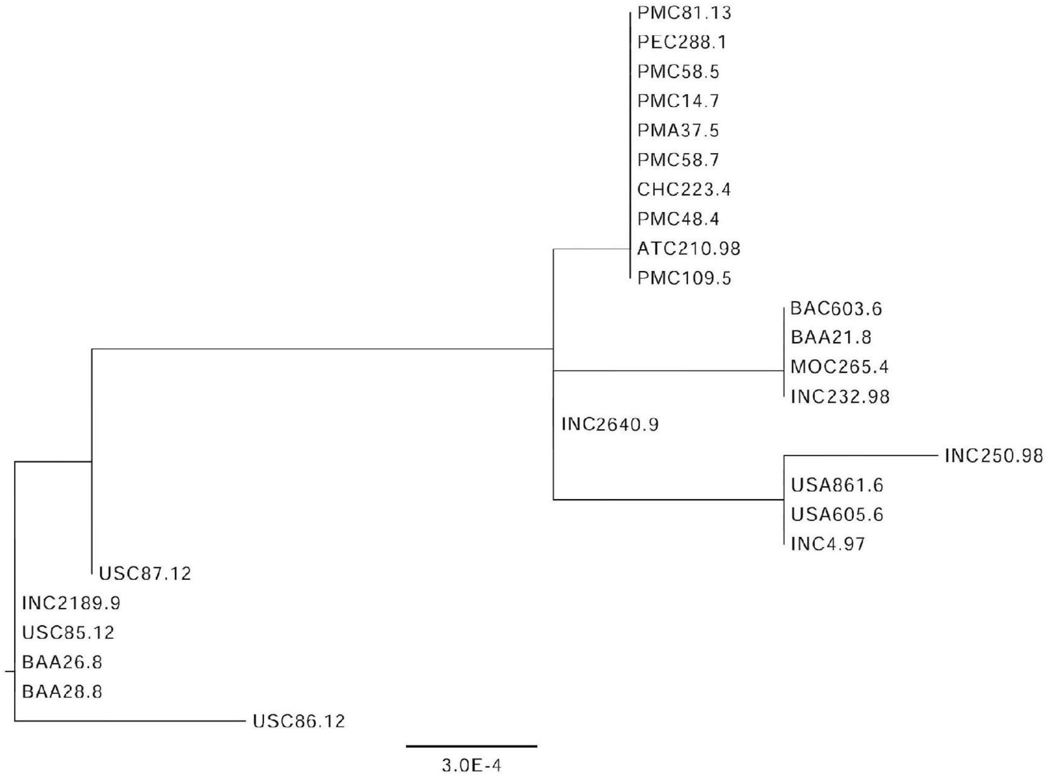


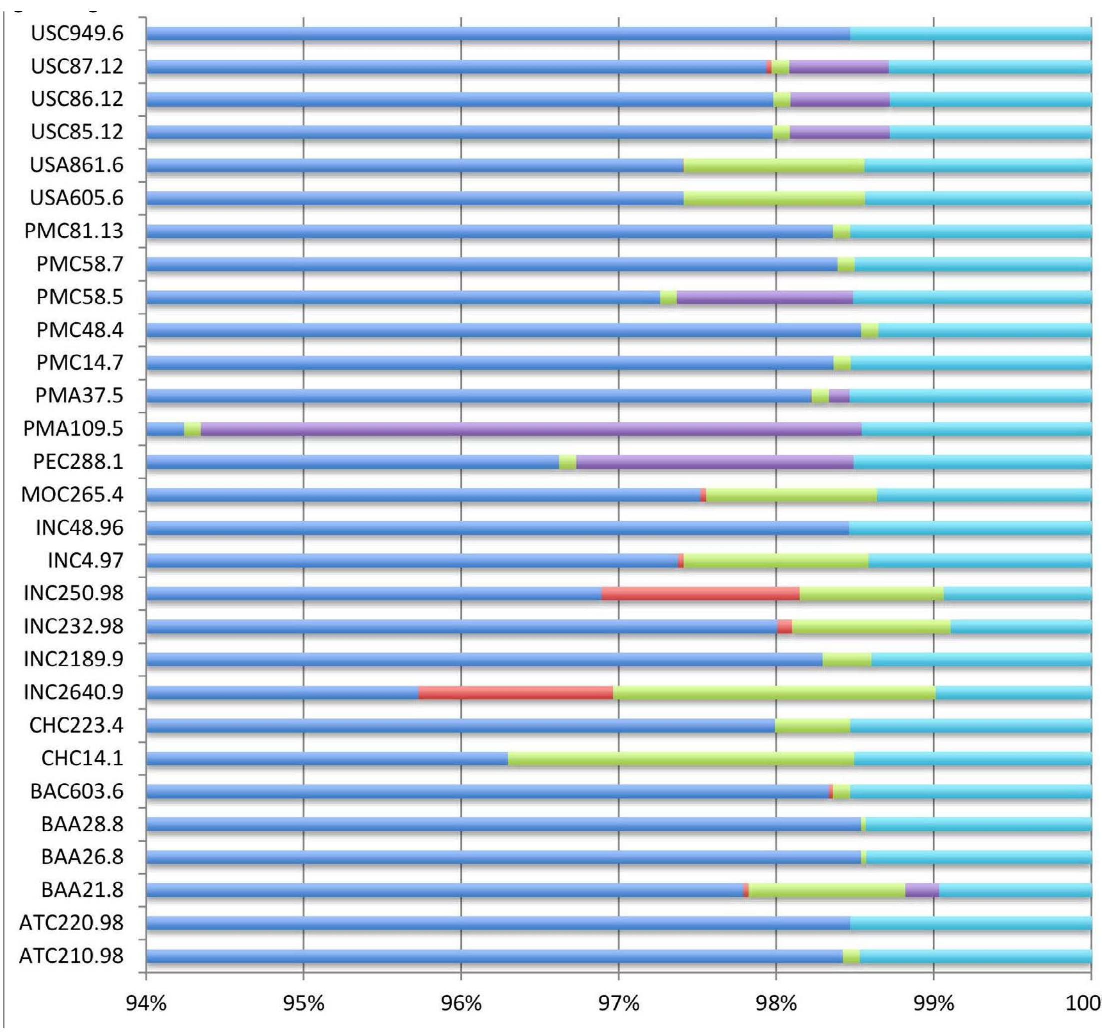
Figure 3: Maximum likelihood tree of pandemic *V. parahaemolyticus* isolates based on the sequence of the 5.6 Kb insertions.

may encode structural phage proteins according to RAST; 2) a 11.1 Kb plasmid found in an isolate obtained in Bangladesh, which is related to a plasmid detected in *V. nigripulchritudo* responsible for mass mortality of shrimp in New Caledonia and in *V. shilonii*, a coral pathogen [22]; and 3) a 27.8 Kb plasmid with functions related to a type IV secretion system complex found in three isolates originating in the USA, in which 34% of the sequence had 68% similarity with a 45.8 Kb plasmid in V. fischeri that is homologous to pES100 (common among other symbiotic strains of *V. fischeri* [23] and a 5,090 bp DNA with genes potentially related to DNA mobilization.

Discussion

The procedure described here allowed us to perform a comprehensive comparison of the genomes of isolates from the pandemic *V. parahaemolyticus* population, assessing either the clonal frames (or the genealogical origin of the DNA segments making up the genomes) and the fraction of the genome that was not included in the comparison. Using the read accounting procedure, we determined that the fraction of reads that were not included in the comparison ranged from 0.01%-2% for each isolate. Because 1% of reads would correspond to a DNA segment of approximately 50,000 bp, considering a depth coverage equal to the mean coverage observed for the chromosomes, the fraction of reads left unaccounted for would constitute a small fraction of the genome.

The relative amounts of DNA in the pandemic clonal frame, the non-pandemic clonal frame in chromosomes and the non-pandemic clonal frame in extra-chromosomal elements were highly variable between isolates (Figure 4 and Table 4S). The fraction of bp remaining from the ancestral pandemic *V. parahaemolyticus* (belonging to the pandemic clonal frame) varied from 94.2%-98.5%. The assessed percentage of bp in other clonal frames located in the chromosome gained through HGT and recombination was highly variable, ranging from 0.0%-3.3%. The percentage of bp in extra-chromosomal elements corresponding to DNA gained via HGT that did not recombine with the chromosome but was able to replicate independently and be maintained during bacterial growth varied from 0.0%-4.2%. The percentage of bp in the genome that could not be assigned to any of these classes (or accounted for) was fairly consistent at 1.0%-1.5%, which corresponds to those segments in the isolate genomes that mapped to the reference genome but were present in only some of the isolates.



Blue: Pandemic *V. parahaemolyticus* clonal frame. Orange: Non-pandemic clonal frames in core chromosomes. Green: Non-pandemic clonal frames in non-core chromosomes. Violet: Non-pandemic clonal frames in extra-chromosomal elements. Light blue: Not ascribed.

Figure 4: Graphical representation of the percentages of base pairs ascribed to regions in different clonal frames or genealogies.

The assessment of the relative impact of mutation and recombination varied greatly according to the elements of the genome included in the assessment. When only core elements were considered the r/m was 6.4. Assessment of the r/m including both non-core recombinant and extra-chromosomal segments raised the value by several orders of magnitude. However, this last assessment is problematic because in the absence of alleles for these segments the rate of nucleotide differences between "recombinant" stretches cannot be assessed. The large contribution of DNA with different clonal genealogy to the diversification of the genomes suggests that the emergence of new pathogens is primarily caused by HGT. However, HGT might depend on the vicissitudes of the life of each bacterium, as exemplified by the presence of isolates with exclusively pandemic clonal frame DNA, such as ATC220.98 and INC48.96, and isolates with more than 100,000 bp DNA of non-pandemic clonal frames, probably acquired by HGT (Figure 3 and Table 4S). In some isolates, these new DNA segments were in chromosomes (CHC223.4, INC2640.9, INC250.98) implying actual recombination or gene conversion while in others they were in extra-chromosomal elements (PMA109.5). Given the high genome sequence similarity between isolates of this highly clonal population, determining the number of mutations and recombination events suffered by each isolate is unpredictable, and depends on the normal physical and biological environment of the bacteria. The availability of nutrients will determine the rate of duplication, and the consequential rate of mutation, the presence of free DNA, bacteriophages or other bacteria able to contribute DNA with potential for gene flow will determine the rate of recombination or the gain of DNA by H.

Acknowledgement

This work was supported by grant 1140732 from the Comisión Nacional de Investigación Científica y Tecnológica de Chile (CONICYT). We thank María Luisa Rioseco from the Hospital Regional de Puerto Montt for providing isolate PMC81.13.

References

1. Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol 12: 263-273.

2. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3: 722-732.

3. Okuda J, Ishibashi M, Hayakawa E, Nishino T, Takeda Y, et al. (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 12: 3150-3155.

4. Matsumoto C, Okuda J, Ishibashi M, Iwanaga M, Garg P, et al. (2000) Pandemic spread of an O3:K6 clone of Vibrio parahaemolyticus and emergence of related strains evidence by arbitrarily primed pcr and toxRS sequence analyses. J Clin Microbiol 2: 578-585.

5. Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6: 431-440.

6. Chen Y, Stine OC, Badger JH, Gil AI, Nair BG, et al. (2011) Comparative genomic analysis of Vibrio parahaemolyticus: serotype conversion and virulence. BMC Genomics 294.

7. Loyola DE, Navarro C, Uribe P, Garcia K, Mella C, et al. (2015) Genome diversification within a clonal population of pandemic Vibrio parahaemolyticus seems to depend on the life circumstances of each individual bacteria. BMC Genomics 16:176.

8. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, et al. (2003) Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae. Lancet 361: 743-749.

9. Zhu Y, Stephens RM, Meltzer PS, Davis SR (2013) SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics 14: 19.

10. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

11. Marinier E, Brown DG, McConkey BJ (2015) Pollux: platform independent error correction of single and mixed genomes. BMC Bioinformatics 15: 435-436.

12. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, et al. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res 18: 2024-2033.

13. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomics sequence with rearrangements. Genome Res 14: 1394-1403.

14. Didelot X, Wilson DJ (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11: e1004041.

15. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-780.

16. Leinomen R, Sugawara H, Shumway M (2011) The sequence read archive. Nucleic Acids Res 39: D19-D21.

17. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, et al. (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res D206-D214.

18. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

19. Zabala B, Hammerl JA, Espejo RT, Hertwig S (2009) The linear plasmid prophage Vp58.5 of Vibrio parahaemolyticus is closely related to the integrating phage VHML and constitutes a new incompatibility group of telomere phages. J Virol 83: 9313-9320.

20. Kalburge SS, Polson SW, Crotty KB, Katz L, Turnsek M, et al. (2014) Complete Genome Sequence of Vibrio parahaemolyticus Environmental Strain UCM-V493. Genome Announc 2: e00159-14

21. Lee CT, Amaro C, Wu KM, Valiente E, Chang YF, et al. (2008) A common virulence plasmid in biotype 2 Vibrio vulnificus and its dissemination aided by a conjugal plasmid. J Bacteriol 190: 1638-1648.

22. Sobrinho P de SC, Destro MT, Franco BD, Landgraf M (2010) Correlation between environmental factors and prevalence of Vibrio parahaemolyticus in oysters harvested in the southern coastal area of Sao Paulo State, Brazil. Appl Environ Microbiol 76: 1290-1293.

23. Ruby EG, Urbanowski M, Campbell J, Dunn A, Faini M, et al. (2005) Complete genome sequence of Vibrio fischeri: a symbiotic bacterium with pathogenic congeners. Proc Natl Acad Sci USA 102: 3004-3009.

24. González-Escalona N, Cachicas V, Acevedo C, Rioseco ML, Vergara JA, et al. (2005) parahaemolyticus diarrea, chile, 1998 and 2004. Emerging Infectious Diseases 11:129-131.

25. Fuenzalida L, Armijo L, Zabala B, Hernández C, Rioseco ML, et al. (2007) Vibrio parahaemolyticus strains isolated during investigation of the summer 2006 seafood related diarrhea outbreaks in two regions of Chile. Int J Food Microbiol 117: 270-275.

26. Harth E, Matsuda L, Hernández C, Rioseco ML, Romero J, et al. (2009) Epidemiology of Vibrio parahaemolyticus outbreaks, southern Chile. Emerg Infect Dis 15:163-168.

27. Zabala B, García K, Espejo RT (2009) Enhancement of Vibrio parahaemolyticus O3:K6 pandemic strain ultraviolet light sensitivity due to natural lysogenization by a telomeric phage. Appl Environ Microbiol 75:1697-1702.