Unique and conserved genome regions in Vibrio harveyi and related species in comparison with the shrimp pathogen Vibrio harveyi CAIM 1792

Iliana Espinoza-Valles, 1 Gary J. Vora, 2 Baochuan Lin, 2
Pimlapas Leekitcharoenphon, 3, 4 Adrián González-Castillo, 1
Dave Ussery, 4, 5 Lone Høj 6 and Bruno Gomez-Gil 1

1 CIAD A.C., Mazatlán Unit for Aquaculture, Mazatlán, Sinaloa, Mexico
2 Center for Bio/Molecular Science & Engineering, Naval Research Laboratory, Washington, DC, USA
3 National Food Institute, Division for Epidemiology and Microbial Genomics, Technical University of Denmark, Kongens Lyngby, Denmark
4 Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Kongens Lyngby, Denmark
5 Comparative Genomics group, Biosciences Division, Oak Ridge National Labs, Oak Ridge, Tennessee, USA
6 Australian Institute of Marine Science, Townsville, Queensland, Australia

Vibrio harveyi CAIM 1792 is a marine bacterial strain that causes mortality in farmed shrimp in north-west Mexico, and the identification of virulence genes in this strain is important for understanding its pathogenicity. The aim of this work was to compare the V. harveyi CAIM 1792 genome with related genome sequences to determine their phylogenetic relationship and explore unique regions in silico that differentiate this strain from other V. harveyi strains. Twenty-one newly sequenced genomes were compared in silico against the CAIM 1792 genome at nucleotidic and predicted proteome levels. The proteome of CAIM 1792 had higher similarity to those of other V. harveyi strains (78 %) than to those of the other closely related species Vibrio owensi (67 %), Vibrio rotiferianus (63 %) and Vibrio campbellii (59 %). Pan-genome ORFans trees showed the best fit with the accepted phylogeny based on DNA–DNA hybridization and multi-locus sequence analysis of 11 concatenated housekeeping genes. SNP analysis clustered 34/38 genomes within their accepted species. The pan genometic and SNP trees showed that V. harveyi is the most conserved of the four species studied and V. campbellii may be divided into at least three subspecies, supported by intergenomic distance analysis. BLASTP atlases were created to identify unique regions among the genomes most related to V. harveyi CAIM 1792; these regions included genes encoding glycosyltransferases, specific type restriction modification systems and a transcriptional regulator, LysR, reported to be involved in virulence, metabolism, quorum sensing and motility.

Abbreviations: DDH, DNA–DNA hybridization; GGDC, genome-to-genome-distance calculator; MLSA, multi-locus sequence analysis.

The GenBank/EMBL/DBJ accession numbers for the genome sequences determined in this study are as follows (strain, acc. no.): 42A, JPTC00000000; 602L, JPTF00000000; 9078-83, JPTEO0000000; CAIM 115, JPTG00000000; CAIM 1500, JPTH00000000; CAIM 198, JPTI00000000; E1, JPTJ00000000; 2415-05, JPTK00000000; 517 4, JPTT00000000; 74F, JPTL00000000; 392 [MAV], PRJNA257174; CAIM 1075, JPTM00000000; CAIM 148, JPTO00000000; CAIM 461, JPTO00000000; CAIM 463, JPTP00000000; CAIM 464, JPTQ00000000; CAIM 606, JPTR00000000; H6, JPTS00000000.

Two supplementary tables and four supplementary figures are available with the online Supplementary Material.
INTRODUCTION

Vibrios are motile, Gram-negative, curved or rod-shaped bacteria with a single polar flagellum (Thompson et al., 2005). Many vibrios have two types of flagellum: one polar flagellum which is produced under liquid conditions and several lateral flagella that are produced in more viscous environments. Within this genus, one of the main groups of species is the Harveyi clade, which contains *Vibrio alginolyticus*, *V. azureus*, *V. campbellii*, *V. harveyi*, *V. jassicida*, *V. mytili*, *V. natriegens*, *V. owensii*, *V. rotiferianus*, *V. parahaemolyticus* and *V. sagamiensis* (Cano-Gomez et al., 2011; Hoffmann et al., 2013; Sawabe et al., 2007; Yoshizawa et al., 2009). The clade includes important opportunistic pathogens of cultured aquatic organisms and strains of medical relevance (Lin et al., 2010; Thompson et al., 2004). The members of the Harveyi clade share a high degree of genetic and phenotypic similarity, which complicates the correct identification and differentiation of species and strains (Cano-Gomez et al., 2009; Sawabe et al., 2007). Misidentifications have previously been reported; strains ATCC BAA-1116, HY01 and 1DA3 were wrongly identified as *V. harveyi*, the first two being *V. campbellii* (Lin et al., 2010). Strain 1DA3, used to describe the novel species *V. communis* (Chimetto et al., 2011), was then reclassified as *V. owensii* since it forms a well-supported clade with LMG 25443 and ATCC 25919 (Urbanczyk et al., 2013).

Reliable identification of *V. harveyi*-related species has been done by multi-locus sequence analysis (MLSA) of housekeeping protein-coding genes (*rpoA*, *pyrH*, *topA*, *fisZ*, *mreB*), while concatenation of only *topA* and *mreB* gene sequences offered similar resolution to that of a five-gene MLSA (Cano-Gomez et al., 2011). However, the inter-species threshold for *V. campbellii*, *V. harveyi*, *V. owensii* and *V. rotiferianus* is above 88 % (Cano-Gomez et al., 2011). Therefore, there is a need to find more accurate biomarkers to discriminate these related species. One approach (Tettelin et al., 2005) relies on the identification of shared genes at three levels: genes found in all members of the investigated group (core), genes found in more than one but not all members (accessory), and genes found in just one member of the group (unique or ORFan). With this approach, it is possible to discover new genetic markers for one species or even one strain in particular.

A case study is *V. harveyi* CAIM 1792, a pathogenic strain reported as the aetiological agent of bright red syndrome in the Pacific white shrimp *Litopenaeus vannamei* (Soto-Rodríguez et al., 2010). The genome of this strain has been sequenced (Espinoza-Valles et al., 2012), and it is now possible to use bioinformatics tools to identify unique and shared genes among strains of this species and of phylogenetically closely related species. In this study, whole-genome sequences of strains of the closely related species *V. harveyi*, *V. owensii*, *V. rotiferianus* and *V. campbellii* were compared to the *V. harveyi* CAIM 1792 genome sequence to reveal the conservation level of genes between these strains and species.

METHODS

Genomic data. The published draft genome sequences of *V. harveyi* CAIM 1792 and various Harveyi clade members were obtained from the NCBI organism genome INSDC database (Karsch-Mizrachi et al., 2012). Twenty-one new draft genome sequences (eleven *V. harveyi*, eight *V. campbellii*, two *V. owensii*, Table 1) were generated using the MiSeq Reagent kit v2 (500-cycle kit) and a MiSeq desktop sequencer (Illumina). *De novo* assembly was performed with Newbler (runAssembly v.2.3, 454 Life Sciences) for each of the new draft genome sequences. Details on *de novo* assembly results from 19 draft genomes are reported in Table S1, available in the online Supplementary Material.

Phylogenetic analysis. Partial sequences of the following housekeeping genes were selected for MLSA: actin-like cytoskeleton protein (*mreB*, 418 bp), cell division protein *fisZ* (463 bp), uridine monophosphate kinase (*pyrH*, 487 bp), glyceraldehyde-3-phosphate dehydrogenase A (*gapA*, 506 bp), DNA topoisomerase 1 (*topA*, 557 bp), DNA replication initiator protein (*rctB*, 605 bp), DNA gyrase B subunit (*gyrB*, 647 bp), DNA recombination repair protein (*recA*, 824 bp), ATP synthase alpha subunit (*atpA*, 1277 bp), RNA polymerase D subunit (*rpoD*, 602 bp) and RNA polymerase B subunit (*rpoB*, 2203 bp). The DNA sequences of these 11 housekeeping genes were extracted from the draft genome contigs (Table S2), aligned in CLUSTAL W (Thompson et al., 1994) with a gap open cost of 16 and gap extend cost of 6.66 and edited in Geneious v.7.05 to obtain equal sizes from the whole set of genomes analysed. *Vibrio cholerae* O1 El Tor N16961 was chosen as the outgroup strain for the dendrograms. Phylogenetic trees were reconstructed using the maximum-likelihood method with Kimura two-parameter (Kimura, 1980) algorithm for nucleotides with MEGA version 6 (Tamura et al., 2011). Specific parameters as equal rates, complete deletion and bootstrap of 1000 were set in the analysis. Species identification for each strain was maintained as published by the original authors.

Predicted proteome similarity. The predicted proteome of each genome was obtained with the program Prodigal v.2.5 (Hyatt et al., 2010). The similarity between proteins was considered significant if it produced 50 % identity for at least 50 % of the length of the longest ORF, called the ‘50/50 rule’ (Vesth et al., 2010). Paralogous genes were obtained from comparing one predicted proteome against itself using the same rule. The estimated proteins found in common between genomes were plotted into a BLAST matrix by a program called blastmatrix, from the CMG-Biotools workbench (Vesth et al., 2013).

Structural properties of *V. harveyi* CAIM 1792 genome and identification of unique and conserved regions in the genome. In order to compare the more closely related genomes to *V. harveyi* CAIM 1792, the predicted proteomes from the four species were arranged from the highest to the lowest percentage of similarity into a BLASTP atlas in accordance with blastmatrix results. The conservation level of each protein-encoding gene was mapped along the two chromosomes by using the ‘50/50 rule’. The study strain *V. harveyi* CAIM 1792 was used as reference proteome. Structural parameters of CAIM 1792 were calculated and normalized to the genomic average values of A+T content and G+C skew expressed as a deviation average (McLean et al., 1998). For each chromosome, the fractions containing local repeats and global repeats of at least 80 % or higher similarity were calculated.
Table 1. Bacterial strains used in this study

<table>
<thead>
<tr>
<th>Species</th>
<th>Strain</th>
<th>INSDC acc. no.</th>
<th>Size</th>
<th>Contigs</th>
<th>% Similar 1792</th>
<th>Isolation source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>V. campbellii</td>
<td>ATCC BAA-1116</td>
<td>CP000789-91</td>
<td>6.06</td>
<td>3</td>
<td>54.9</td>
<td>Artemia nauplii</td>
<td>Bassler et al. (1997)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>DS40M4</td>
<td>AGIE00000000</td>
<td>5.12</td>
<td>121</td>
<td>59.4</td>
<td>Seawater</td>
<td>Dias et al. (2012)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>CAIM 519T*</td>
<td>AMDG00000000</td>
<td>5.07</td>
<td>213</td>
<td>59.6</td>
<td>Seawater</td>
<td>Karsch-Mizrachi et al. (2012)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>200612B</td>
<td>BANY00000000</td>
<td>5.72</td>
<td>229</td>
<td>65.6</td>
<td>Seawater</td>
<td>Urbanczyk et al. (2013)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>PEL22A</td>
<td>JPTF00000000</td>
<td>5.32</td>
<td>246</td>
<td>68.5</td>
<td>Seawater</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>602L*</td>
<td>JPTT00000000</td>
<td>5.22</td>
<td>371</td>
<td>67.3</td>
<td>White grunt (Haemulon plumieri) gut</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>CAIM 115*</td>
<td>JPTG00000000</td>
<td>5.72</td>
<td>502</td>
<td>65.6</td>
<td>Shrimp ( lithopeneus sp.) haemolymph</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>HY01</td>
<td>AAWP00000000</td>
<td>5.40</td>
<td>554</td>
<td>59.3</td>
<td>Dead shrimp with vibrosis</td>
<td>Rattanama et al. (2009)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>CAIM 1500*</td>
<td>JPTH00000000</td>
<td>5.16</td>
<td>791</td>
<td>65.4</td>
<td>Snapper ( Lampris guttatus) liver</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>CAIM 198*</td>
<td>JPT00000000</td>
<td>5.37</td>
<td>1176</td>
<td>60.9</td>
<td>Shrimp ( lithopeneus sp.) hepatopectenes</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>42A*</td>
<td>JPTC00000000</td>
<td>5.21</td>
<td>1581</td>
<td>60.3</td>
<td>Healthy coral ( Musisimila hispida)</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. campbellii</td>
<td>9078-83*</td>
<td>JPTE00000000</td>
<td>5.41</td>
<td>2699</td>
<td>49.3</td>
<td>Research isolate</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>VHJR7</td>
<td>CAUU00000000</td>
<td>5.92</td>
<td>70</td>
<td>81.9</td>
<td>Heart tissue of Lates calcarifer</td>
<td>Karsch-Mizrachi et al. (2012)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>E385</td>
<td>AYKI00000000</td>
<td>6.35</td>
<td>84</td>
<td>75.5</td>
<td>Diseased cage-cultured grouper</td>
<td>Yu et al. (2013)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>CAIM 1792</td>
<td>AHHQ00000000</td>
<td>5.84</td>
<td>86</td>
<td>–</td>
<td>Diseased shrimp ( lithopeneus vannamei) lesion</td>
<td>Soto-Rodriguez et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>AOI31</td>
<td>AOMR00000000</td>
<td>6.12</td>
<td>103</td>
<td>77.7</td>
<td>Diseased juvenile giant grouper ( Epinephelus lanceolatus)</td>
<td>NCBI</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>HENC-01</td>
<td>AJSQ00000000</td>
<td>5.88</td>
<td>139</td>
<td>84.3</td>
<td>Unknown</td>
<td>Karsch-Mizrachi et al. (2012)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>ZJ0603</td>
<td>AKIH00000000</td>
<td>6.86</td>
<td>206</td>
<td>72.9</td>
<td>Diseased orange-spotted grouper ( Epinephelus coïoides)</td>
<td>Huang et al. (2012)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>ATCC 14126T*</td>
<td>PRJNA173805</td>
<td>5.61</td>
<td>253</td>
<td>83.8</td>
<td>Dead amphipod ( Talorchestia sp.), juvenile shrimp</td>
<td>Johnson &amp; Shunk (1936)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>CAIM 1075*</td>
<td>JPTM00000000</td>
<td>5.64</td>
<td>441</td>
<td>80.2</td>
<td>Oyster ( Crassostrea gigas)</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>CAIM 517*</td>
<td>JPTT00000000</td>
<td>5.65</td>
<td>529</td>
<td>80.6</td>
<td>Brown shark ( Caracharinus plumbeus) kidney</td>
<td>Grimes et al. (1984); Pedersen et al. (1998)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>CAIM 464*</td>
<td>JPTQ00000000</td>
<td>5.90</td>
<td>547</td>
<td>78.8</td>
<td>Turbot ( Scophthalmus maximus)</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>CAIM 148*</td>
<td>JPTN00000000</td>
<td>5.67</td>
<td>559</td>
<td>82.5</td>
<td>Diseased shrimp ( Penaecus sp.) haemolymph</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>2415-05*</td>
<td>JPTK00000000</td>
<td>5.50</td>
<td>667</td>
<td>78.0</td>
<td>Human blood</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>H6*</td>
<td>JPT50000000</td>
<td>5.58</td>
<td>730</td>
<td>78.8</td>
<td>Squirrelfish ( Holocentrus sp.) gut</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>74F*</td>
<td>JPTL00000000</td>
<td>5.79</td>
<td>877</td>
<td>76.5</td>
<td>Diseased coral ( Musisimila braziliensis)</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>CAIM 606*</td>
<td>JPTR00000000</td>
<td>6.15</td>
<td>909</td>
<td>71.9</td>
<td>Japanese horse mackerel ( Trachurus japonicus)</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>CAIM 461*</td>
<td>JPT00000000</td>
<td>5.78</td>
<td>1192</td>
<td>73.2</td>
<td>Shark tank water</td>
<td>Lin et al. (2010)</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>392 [MAV]*</td>
<td>PRJNA257174</td>
<td>5.51</td>
<td>1566</td>
<td>71.3</td>
<td>Unknown</td>
<td>Reichelt &amp; Baumann (1973)</td>
</tr>
<tr>
<td>V. owensii</td>
<td>1D3A</td>
<td>ACZC00000000</td>
<td>5.93</td>
<td>140</td>
<td>68.3</td>
<td>Diseased corals ( M. hispida and Phylogogia dilatata)</td>
<td>Thompson et al. (2009)</td>
</tr>
<tr>
<td>V. owensii</td>
<td>ATCC 25919</td>
<td>BANZ00000000</td>
<td>5.80</td>
<td>174</td>
<td>68.7</td>
<td>Seawater</td>
<td>Urbanczyk et al. (2013)</td>
</tr>
<tr>
<td>V. owensii</td>
<td>47666-1*</td>
<td>JPRC00000000</td>
<td>5.83</td>
<td>55</td>
<td>68.4</td>
<td>Diseased giant tiger prawn ( Penaecus monodon) larvae</td>
<td>Harris &amp; Owens (1999)</td>
</tr>
<tr>
<td>V. owensii</td>
<td>ATCC 25430</td>
<td>BAOE00000000</td>
<td>5.82</td>
<td>341</td>
<td>69.2</td>
<td>Healthy coral ( M. hispida)</td>
<td>Chimetto et al. (2011)</td>
</tr>
<tr>
<td>V. owensii</td>
<td>DY05*</td>
<td>JPRD00000000</td>
<td>6.34</td>
<td>119</td>
<td>60.3</td>
<td>Diseased spiny lobster ( Panulirus ornatus) larvae</td>
<td>Cano-Gomez et al. (2010)</td>
</tr>
<tr>
<td>V. rotiferianus</td>
<td>DAT722</td>
<td>AFAJ00000000</td>
<td>5.33</td>
<td>79</td>
<td>64.4</td>
<td>Mud crab larvae aquaculture tank</td>
<td>Roy Chowdhury et al. (2011)</td>
</tr>
<tr>
<td>V. rotiferianus</td>
<td>CAIM 577T</td>
<td>BAOI00000000</td>
<td>5.33</td>
<td>133</td>
<td>61.0</td>
<td>Cultures of rotifer ( Brachionus plicatilis)</td>
<td>Gomez-Gil et al. (2003)</td>
</tr>
</tbody>
</table>

Bold letters indicate demonstrated pathogenic strains. *: Type strain. CAIM, Collection of Aquatic Important Micro-organisms (http://www.ciam.mx/caim); ATCC, American Type Culture Collection. CAIM 517=ATCC 35084, DY05* = CAIM 1854=ATCC 25443, CAIM 577T = LMG 21460*. % Similar 1792, percentage of similarity against V. harveyi CAIM 1792 in amino acid sequence. Species names as reported in the corresponding publications. NCBI, genome obtained from the NCBI database, without published reference.

*21 New genomes.
The structural topology of the DNA chain or intrinsic curvature was calculated on the basis of dinucleotide models using the CURVATURE program (Bolshoy et al., 1991; Shipgelman et al., 1993), since local topology affects binding to specific regulatory proteins and subsequent protein-induced bending (Hadjirengiskou, & Koehler, 2008). The stacking energy was calculated from dinucleotide values according to Ornstein et al. (1998). The trinucleotide position preference parameter was calculated on the basis of the preferential location of sequences within nucleosomal core sequences as described by Satchwell et al. (1986), as absolute values (Baldi et al., 1996) from close to zero to 0.28 (flexible to rigid). This measure is considered as the more sensitive since very few trinucleotides have values close to zero that correlate with local DNA properties and indicate the likeness of the DNA molecule to be tightly wrapped or more likely to be highly expressed (Ussery et al., 2009).

**Pangenomic clustering.** Sequences were grouped into gene families based on sequence similarity. Two sequences were placed in the same gene family if there were significant alignments between them when either sequence was used as query, with a significant alignment defined as one spanning at least 50 % of the length of the query sequence and containing at least 50 % identity (Tettelin et al., 2005; Vesth et al., 2010). The pan- and core-genome calculations were used to extract subsets of genes for each species using the program specificGenes (Vesth et al., 2013), with the BLAST output from the pancore program serving as the input (Vesth et al., 2013).

A hierarchical clustering using relative Manhattan distances was performed for the complete pan-genome (Snipen & Ussery, 2010). In addition, three weighting schemes were applied; the first gave a weight of 1 to all gene families (core) except the ORFans, i.e. those gene families only present in one genome, which received a weight of 0. The second gave a gradually higher weight to the gene families found in the majority of the genomes, the shell. The third represented the opposite strategy of the shell, emphasizing the cloud or accessory genome (Snipen & Ussery, 2010). Standard settings for BLAST were used and the E-value cut-off was set to 10^{-7}. Bootstrap values were computed for each inner node by resampling gene families and reclustering these data.

**SNP analysis.** Since the whole-genome sequence data are available for each strain analyzed here, a phylogenetic tree was developed on the basis of concatenated SNPs using the online server snpTree-1.1, available at http://www.cge.cbs.dtu.dk/services/snpTree/ (Leekitcharoennphon et al., 2012).

**Genome-to-genome distance analysis.** The pan-genome trees in some cases split strains that were identified as belonging to the same species by MLSA. To test whether these strains could represent other species or subspecies, a few were selected for analysis by the genome-to-genome-distance calculator (GGDC) available at http://www.ggc.dsmz.de/distcalc2.php. The intergenomic distances used for DNA-DNA hybridization (DDH) prediction were resampled by bootstrapping and submitted (Meier-Kolthoff et al., 2013). The GGDC analysis needed only about 20 % of the genome to get the same result as with the full genome, and produced a DDH estimate with both a modelled confidence interval and a bootstrap-based confidence interval. The model-based confidence interval was used for reporting the probabilities that the two species being compared belong to the same species (DDH ≥ 70 %) and subspecies (DDH ≥ 79 %).

**RESULTS**

**Phylogenetic analysis.**

MLSA has been proposed as a valuable technique for the identification and classification of vibrio (Hoffmann et al., 2013; Pascual et al., 2010; Sawabe et al., 2013) and for V. harveyi-like species (Cano-Gomez et al., 2011; Lin et al., 2010; Rivera-Posada et al., 2011; Thompson et al., 2007). A dendrogram with concatenated sequences (Fig. 1) was generated because the analysis of concatenated housekeeping gene sequences provides a greater power of taxonomic resolution and minimizes the weight of recombination events (Pascual et al., 2010). Contig numbers where the partial gene sequences were identified for the MLSA analysis are listed in Table S2.

**Predicted proteome similarity.**

The highest similarity in amino acid sequence to V. harveyi CAIM 1792 (Fig. S1) was found in the V. harveyi genomes HENC-01 (84.3 %), ATCC 14126T (83.8 %), CAIM 148 (82.5 %), VHJR07 (81.9 %) and CAIM 517 (80.6 %). Less than 78 % similarity was related (R^2=0.90325) to a high number of contigs (>550) for this species (Table 1). Genes of closely related species showed different proteomic similarity values against V. harveyi CAIM 1792 (Fig. S2), with a mean similarity of 67.0 % with V. owensii and 62.7 % with V. rotiferianus, V. campbellii shared a mean of 58.9 %, where strain 602L harboured the highest similarity (68.5 %) with CAIM 1792. In the V. campbellii genomes, there was no relation between the number of proteins that matched the ‘50/50 rule’ and the number of contigs (shared protein hits/contigs) that were found when comparing the strains against each other. An obvious example of this genetic diversity can be found in V. campbellii PEL22A, which had the lowest shared protein hits (27.3–33.7 %) towards all other V. campbellii strains tested (Fig. S1).

Paralogous genes in CAIM 1792 constituted only 3.3 % in the genome, which is near the average (3.2 %) of the whole set of analysed genomes. The lowest value was seen in V. campbellii 602L with only 2.4 %, while the maximum value of 6.9 % was obtained in the atypical V. campbellii PEL22A.

**Pangenomic clustering.**

A pan–core plot (Fig. 2) based on gene content frequency in each genome allowed visualization and calculation of the pan-genome of the four species. The core genome represented 7.9 % (1615 CDS) of the total sequences from the pan-genome (20372 CDS), while the shell and cloud genomes contained 92.0 % (18757 CDS) of the pan-genome. The numbers of shared gene families calculated for each possible species combination are represented in a Venn diagram (Fig. S3). The species V. harveyi shared the most with V. owensii (3714), closely followed by V. rotiferianus (3541), with the lowest number found for V. campbellii (1658).

The number of gene families within the core genome of each species was calculated as a way to measure the level of conservation of each species. However, few gene families from the core genome were species-specific. For V. harveyi,
140 gene families were found exclusively for this species, and for V. owensii, V. campbellii and V. rotiferianus the corresponding numbers were 60, 28 and 177, respectively (Fig. S3).

In the pangenomic trees, the four species were clustered according to their gene family content at the protein level from core genes (Fig. 3a), ORFs found in only one genome (ORFans, Fig. 3b), moderately conserved genes
which is in agreement with the BLASTP matrix results. B392 [MAV] as the most distant strain for this species, V. harveyi to their evolutionary distance, with strain V. harveyi PEL22A, seen as outgroups. The strains in the main omes clustered together except genomes HY01 and PEL22A. The strains in the shell tree (Fig. 4a), most of the V. campbellii genomes clustered together except genomes HY01 and PEL22A, seen as outgroups. The strains in the main V. harveyi cluster had a hierarchical ordering according to their evolutionary distance, with strain V. harveyi B392 [MAV] as the most distant strain for this species, which is in agreement with the BLASTP matrix results. The most evolutionarily distant genomes were V. campbellii HY01 and PEL22A. V. owensii clustered with V. harveyi, revealing again the close relationship between V. owensii and V. harveyi. The V. campbellii genomes in the shell tree Fig. 4(a) are displayed in two clusters: (i) those strains most closely related to V. harveyi and V. owensii (602L, CAIM 115, 200612B, CAIM 198, CAIM 1500, E1, 42A and 9078) and (ii) the V. campbellii type strain CAIM 519\(^3\) and ATCC BAA-1116. The less related strains HY01 and PEL22A appeared as outgroups.

In the cloud tree (Fig. 4b), most genomes clustered according to species-specific gene content except for three outgroups: (i) the least similar V. campbellii PEL22A, (ii) the most fragmented and gene-rich V. campbellii 9078-23 genome (2699 contigs) and (iii) V. owensii DY05, the largest genome in its species (6.3 Mb) and the fourth largest in predicted genes from the four species. Interestingly, the species clustering with the ORFans tree was the most similar to the accepted DDH/MLSA-based phylogeny, with species differentiation obtained at a relative Manhattan distance similarity of 0.2, with the exception of V. campbellii PEL22A.

**SNP analysis**

Two main branches were obtained in the SNP tree (Fig. 5). The first branch clustered V. harveyi isolates and the second clustered the V. campbellii, V. owensii and V. rotiferianus

---

**Fig. 2.** Pan–core plot of 38 whole-genome sequences from V. campbellii, V. rotiferianus, V. owensii and V. harveyi. V. harveyi B392 = V. harveyi 392 [MAV].
Fig. 3. Pan-genomic dendrograms by the unweighted pair group method with arithmetic mean (UPGMA) method of whole-genome sequences from V. harveyi, V. owensii, V. rotiferianus and V. campbellii. Alignments of their gene content were...

Downloaded from www.microbiologyresearch.org by...
IP:  54.191.40.80
On: Sat, 12 Aug 2017 00:02:07
The SNP analysis clustered some sister strains such as V. harveyi CAIM 463 and CAIM 464, V. campbellii DS40M4 and CAIM 519\(^T\), V. campbellii 602L and CAIM 115, and V. rotiferianus genomes as observed before in pan-genome trees (Figs 5 and 6).

**In silico identification of gene families exclusive to V. harveyi CAIM 1792**

Gene families were calculated for CAIM 1792 from the pan-genome subsets, visualized in the pan–core plot (Fig. 2). The V. harveyi CAIM 1792 genome was set as last in the calculation, preceded by the more related genomes (according to blastmatrix results) to obtain both core and unique gene families. A total of 186 genes clustered in 177 gene families were identified in the core genome, and only 80 hits were exclusive to CAIM 1792. Identification of these 80 exclusive gene families was done by BLASTP searches in the UniRef90 database (E-value 10\(^{-3}\)), and just 24 gene families could be identified, including ABC transporters (2), antidote and killer protein (1), glycosyltransferases (4), GspD family type I site-specific DNase, integral membrane protein, LysR family transcriptional regulator, phage integrase, plasmid partition protein ParB, putative DGQHR domain protein, putative DNA mismatch repair protein, putative replication protein, putative restriction endonuclease, restriction modification system DNA specificity domain-containing protein, transposase IS3/IS911, type I and II restriction-modification system (2), type II restriction enzyme and UBA/THIF-type NAD/FAD binding protein (2).

**Structural properties of the V. harveyi CAIM 1792 genome and identification of unique regions**

Based on differences in the amino acid sequence conservation of the 37 genomes against V. harveyi CAIM 1792, CDS that were exclusively found in chromosomes I and II from CAIM 1792 could be distinguished (gaps in Fig. 6, Table 2). Within both chromosomes, nine rRNA operons were identified as global direct and inverted repeats and all of these were located near the respective origin of replication in Fig. 6. Direct repeats were found in gaps in chromosome I: \(f\) and chromosome II: \(k\) (see Table 2 for a detailed description). Indirect repeats occurred at the very end of the genome encoding TPR, pentapeptide repeats and hypothetical proteins. The position preference of nucleotides in both chromosomes was related to its trend to be highly expressed (this an interpretation of the arrangement of genes that likely will lead to high expression levels) in specific regions and gaps. In chromosome II, a potentially high expression level was observed around two conserved regions \(g\) and \(o\) corresponding to a RTX toxin cluster including a membrane protein (Omp), an agglutinin or outer membrane component (Agg), a putative RTX toxin and peptidoglycan-associated lipoprotein (Lap). In chromosome II, two large RTX toxins (14.8 kb and 17.6 kb in size) were present in some of the 19 strains harbouring this important virulence factor, two are demonstrated crustacean pathogens (V. owensii DY05\(^T\), 47666-1), four were isolated from diseased hosts (V. harveyi ZJ0603, AOD131, ATCC 14126\(^T\); V. owensii 1DA3), one was isolated from a healthy host (V. owensii LMG 25430), six were isolated from hosts of unknown health status (V. harveyi CAIM 606, CAIM 463, CAIM 1075; V. campbellii E1, CAIM 115, CAIM 198), one was isolated from an unknown source (V. harveyi HENC-01) and five were isolated from seawater (V. owensii ATCC 25919; V. campbellii CAIM 519\(^T\), 200612B, 602L, PEL22A).

In chromosome I, an incomplete \(f_{237}\) prophage was detected (gap \(d\)); this prophage is commonly found in other vibrios (Chen et al., 2011; Das et al., 2011; Gomez-Gil et al., 2014; Lan et al., 2009), but in this case was found to lack the genes for the cholera enterotoxin (ctxAB). This prophage might not be highly expressed due to the high stacking energy values observed in its DNA sequence. Regarding DNA structural features, the rRNAs and gap XII (RTX) were identified as being less flexible in accordance with their conservation level. Gap XII is also a high stacking energy region suggesting less likelihood of melting and greater stability.

Gaps \(a\), \(b\), \(f\), \(g\), \(h\) and \(m\) (Fig. 6, Table 2) in both chromosome I and II from CAIM 1792 have an intrinsic curvature that confers more flexibility to these regions. The most evident is gap \(f\) that corresponds to an integrase consisting mostly of hypothetical proteins and toxins, such as HigB toxin, HigA (anti-toxin to HigB), death on curing proteins (Doc toxins) and an integrase integrase (IntI4). Gap \(b\) encodes three non-conserved glycosyltransferases involved in sugar transfer from the cytosol prior to secretion; these genes have been observed in V. fischeri MJ11 as part of an exopolysaccharide biosynthesis cluster, and one of them was recently reported to be involved in cell wall biosynthesis (Zhang et al., 2014). Gap \(g\) encodes a transcription termination factor Rho and a thioredoxin (Trx), as part of an antioxidant system involved in the defence against oxidative stress through disulfide reductase activity found in many bacteria outside of the Harveyi clade (Lu & Holmgren, 2014).

The DNA A+T content in the whole BLASTP atlas ranged from 20 to 80 mol% with a mean of 50 mol%. There are regions (Fig. 6, gaps \(f\), \(g\)) in the innermost circle that are much more A+T-rich than the rest of the chromosome, resulting in a less stable genome structure.
I. Espinoza-Valles and others

(a)

(b)

Microbiology 161
Besides, it was found that these gaps encode outer-membrane proteins, according to the prediction of transmembrane helices in proteins (TMHMM Server v. 2.0).

The identification of insertions was achieved by the G+C skew approach using the BLASTP atlas. One transposase and two enterobactin VctA receptors that transport enterobactin across the outer membrane (Mey et al., 2002) were identified towards the end of chromosome II. The receptors are present in two genomes isolated from diseased marine organisms: V. harveyi CAIM 148 and ZJ0603, from an unknown health status oyster genome: V. harveyi 1075, and from an environmental source isolate from Haiti: HENC-01.

Differences among the species V. harveyi, V. owensii, V. rotiferianus and V. campbellii are observed in Fig. 6 in gaps ‘I–XV’ (chromosome I: I–VII; chromosome II: VIII–XV), showing a higher number of non-conserved proteins or gaps in chromosome II than in chromosome I. Iron sulfur (Fe–S) family genes (sufABC,E and sufS) can be found in chromosome I (Fig. 6 in green, gap I); these are required for critical biochemical pathways, including respiration and nitrogen fixation (Ayala-Castro et al., 2008), and participate in the assembly of cysteine desulforase (SufD) proteins also involved in oxidative stress defence (Rybniker et al., 2014). The cysteine desulforase (SufS) encodes l-alanine ligase or selenocysteine lyase present in V. harveyi. Other differences include operons coding for iron uptake (MotA/TolB/ExxB; Fig. 6 gap Ila), type 3 secretion systems (T3SS, YscQ–U; Fig. 6 gap IIa), a wide colonization island (TadA–GVZ; Fig. 6 gap IV) including hydrolases, virB11 and genes encoding type 2/4 secretion systems (T2/4SS, RcpA/CpaC, RcpC/CpaB; Fig. 6, gap IV). The differing regions of chromosome II among species (Fig. 6, in green) harbour a capsular polysaccharide operon (gap XI), and a large widespread colonization island that encodes pilus assembly proteins (T2/4SS) with several hypothetical proteins (gap XV). Neither gap is present in the V. owensii and V. rotiferianus genomes. A second wide colonization island, T2/T4SS, associated with pilus assembly and hypothetical proteins (gap XIII), was identified also only in V. harveyi genomes. Specific iron uptake operons (gaps IIIa, VIII) were present only in V. harveyi genomes. V. rotiferianus and V. harveyi only share the iron uptake region coding for TonB haem receptor and vibrioferrin (gap IX) with 96.5 % pairwise identity. The V. rotiferianus genomes analysed here differ from those of V. harveyi mainly in regard to the absence of T3SS CDS YscDQ–U (gaps IIIa and IIIb) related to the T3SS of Yersinia pestis (Ross & Plano, 2011). The specific T3SS genes (YscQ–U) found in CAIM 1792 are reported to conform near half of the conserved proteins that constitute the basal structure of the machinery in most of the known T3SS systems of bacteria (Troisfontaines & Cornelis, 2005).

An operon-encoding citrate metabolism transport and regulation (Fig. 6, gap V) was identified in all V. harveyi genomes. Interestingly, citrate utilization has been found as a discriminatory characteristic among isolates identified as V. harveyi, V. campbellii, V. owensii and V. rotiferianus (Chimetto et al., 2011; Rivera-Posada et al., 2011). This operon is similar to the one found in Photobacterium profundum and in several toxicogenic V. cholerae strains (Vesth et al., 2010), and it contains from five to seven CDS including, among others, genes encoding citrate lyase. BLASTX searches resulted in only one hit, with V. paraaerolyticus O1:K33 strain CDC K4557 sharing 98 % similarity as a whole. However, this CDS is not reported to be involved in the citrate cycle first carbon oxidation pathway [PATH:va00020] that converts oxaloacetate into citrate and isocitrate into 2-oxoglutarate. Our results support the usefulness of using whole-genome sequences to identify characteristic phenotypes of vibrio. This was recently suggested by Amaral and co-workers (2014) based on their observation that differentiation of V. cholerae from V. mimicus by fermentation of sucrose and D-mannose phenotypes is due to the absence of three genes in V. mimicus.

**DISCUSSION**

**Phylogenetic analysis and predicted proteome similarity**

The MLSA with 11 concatenated genes allowed good discrimination of the highly related species V. harveyi, V. owensii, V. campbellii and V. rotiferianus. Meanwhile, comparisons of the predicted proteomes of these species also revealed large conserved regions that likely account for the difficulty in differentiating these sister species by traditional phenotypic analyses (Lukjancenko et al., 2012). The V. harveyi strains with the highest protein similarity to CAIM 1792 included strains isolated from diseased organisms (CAIM 148, ATCC 14126T, E385, ZJ0603) and also those of unknown pathogenicity (HENC-01, VHJR7, CAIM 517). The protein similarity percentages of the V. harveyi strains analysed showed that V. harveyi is a conserved species. This is in contrast to the V. campbellii genomes, which showed large genetic
Fig. 5. Maximum-likelihood tree from the SNP alignment with minimum distance between SNP of 5 bp with default parameters by the snpTree-1.1 server. *V. campbellii* ATCC BAA-1116 was used as a reference genome. *V. harveyi* B392 = *V. harveyi* 392 [MAV]. Bar, substitutions per site. Numbers at nodes denote level of bootstrap support as percentages of 1000 replications.
Fig. 6. BLASTP atlases of chromosomes I (upper) and II (lower). Inner circles represent the structural properties of *V. harveyi* CAIM 1792 DNA. BLASTP lanes (coloured circles) represent genomes in order from the genome most closely related to CAIM 1792 (inner) to the least related (outer circle): dark blue, *V. harveyi*; brown, *V. owensii*; green, *V. rotiferianus*; purple, *V. campbellii*. For names of the genomes see Fig. S4.
diversity, for which there could be two explanations: the sequencing quality of genomes (related to number of contigs) and/or the nature of the genome itself (number of proteins from each strain) conferring more diversity within this species.

<table>
<thead>
<tr>
<th>Gap</th>
<th>Position (Mb)</th>
<th>CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique regions of <em>V. harveyi</em> CAIM 1792</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>0.070–0.099</td>
<td>Hypothetical proteins, transposase and phage integrase</td>
</tr>
<tr>
<td>b</td>
<td>0.251–0.260</td>
<td>Putative glycosyltransferases</td>
</tr>
<tr>
<td>c</td>
<td>0.948–0.956</td>
<td>Hypothetical proteins</td>
</tr>
<tr>
<td>d</td>
<td>1.720–1.723</td>
<td>f237 prophage and hypothetical proteins</td>
</tr>
<tr>
<td>e</td>
<td>1.853–1.856</td>
<td>Multidrug efflux pump, cAMP-binding protein and hypothetical protein</td>
</tr>
<tr>
<td>f</td>
<td>2.048–2.135</td>
<td>Integrin, HigAB system (30 kb)</td>
</tr>
<tr>
<td>g</td>
<td>3.387–3.402</td>
<td>Transcription termination factor Rho, thioredoxin and hypothetical proteins</td>
</tr>
<tr>
<td>Chromosome II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>0.277–0.285</td>
<td>Hypothetical proteins</td>
</tr>
<tr>
<td>i</td>
<td>0.437–0.438</td>
<td>Hypothetical proteins and transposase</td>
</tr>
<tr>
<td>j</td>
<td>0.492–0.496</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>k</td>
<td>0.852–0.862</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>l</td>
<td>1.359–1.369</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>m</td>
<td>1.414–1.416</td>
<td>Putative pore-forming cytoxin integrase and hypothetical protein</td>
</tr>
<tr>
<td>n</td>
<td>1.722–1.741</td>
<td>Hypothetical proteins, ß-alanyl-ß-alanine carboxypeptidase, transcriptional regulator, glutathione S-transferase, dihydrofolate reductase, urease domain protein, integrase and hypothetical transcriptional regulator (LysR)</td>
</tr>
<tr>
<td>Conserved regions of <em>V. harveyi</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.649–0.655</td>
<td>Iron sulfur cluster family SufA–E, SufS genes</td>
</tr>
<tr>
<td>II</td>
<td>0.910–0.930</td>
<td>Hypothetical protein, oxoacyl-acyl synthases, reductases, dehydratases, transferases</td>
</tr>
<tr>
<td>IIIa</td>
<td>1.166–1.200</td>
<td>Two SST3 operons (9.47 and 17.64 kb) and hypothetical proteins</td>
</tr>
<tr>
<td></td>
<td>1.210–1.213</td>
<td>TonB-dependent haem and haemoglobin receptor (HutA)</td>
</tr>
<tr>
<td>IIIb</td>
<td>1.230–1.245</td>
<td>Hypothetical proteins, DNA polymerase III epsilon subunit, signal-transduction proteins with cAMP- and CBS-binding domains, short-chain dehydrogenase/reductase SDR, Na+ -driven multidrug efflux pump</td>
</tr>
<tr>
<td>IV</td>
<td>1.800–1.810</td>
<td>Colonization island: TadA-GVZ, Vir B11, T2/4SS genes (RcpA/CpaC, RcpC/ CpaB)</td>
</tr>
<tr>
<td>V</td>
<td>1.911–1.916</td>
<td>Oxaloacetate, citrate and hypothetical protein</td>
</tr>
<tr>
<td>VI</td>
<td>2.961–2.974</td>
<td>Hypothetical protein, lamda phage, integrase and regulator phage integrase</td>
</tr>
<tr>
<td>VII</td>
<td>3.103–952</td>
<td>ABC-type Fe3+-hydroxamate transport system, periplasmic component and putative inner membrane protein</td>
</tr>
<tr>
<td>VIII</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IX</td>
<td>3.104 938</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>0.001–0.015</td>
<td>Iron uptake operon: proton channel family protein (ExbB/TolQ/MotA), biopolymer transport protein (ExbD/TolR) and ferric siderophore transport system periplasmic binding protein (TonB)</td>
</tr>
<tr>
<td>XI</td>
<td>0.362–0.365</td>
<td>Periplasmic protein p19 involved in high-affinity Fe2+ transport, putative high-affinity iron permease, integral protein and hypothetical proteins</td>
</tr>
<tr>
<td>XII</td>
<td>0.508–0.512</td>
<td>Hypothetical proteins and glycosyltransferase</td>
</tr>
<tr>
<td>XIII</td>
<td>0.557–0.570</td>
<td>Cps, maltodextrin ABC transporters</td>
</tr>
<tr>
<td>XIV</td>
<td>0.714–0.731</td>
<td>RTX toxins and related Ca2+ -binding proteins</td>
</tr>
<tr>
<td>XV</td>
<td>1.043–1.045</td>
<td>RTX toxins and related Ca2+ -binding proteins</td>
</tr>
<tr>
<td>XVI</td>
<td>1.311–1.327</td>
<td>Vibrioferrin synthesis operon and hypothetical proteins</td>
</tr>
<tr>
<td>XVII</td>
<td>1.509–1.521</td>
<td>Colonization island: SST2/4 pilus assembly proteins and hypothetical proteins. Vibriolysin and T4SS</td>
</tr>
</tbody>
</table>

**Structural properties of the *V. harveyi* CAIM 1792 genome and identification of unique regions**

A BLASTP atlas using CAIM 1792 as the reference strain located both unique and intraspecific differences among...
the genomes analysed. Since the BLASTP atlas approach is highly influenced by the selected reference, it is likely that additional intraspecific regions would be identified among V. harveyi if other strains were used as the reference. Structural properties of the genome of V. harveyi CAIM 1792 allowed the identification of global direct repeats, as seen in other bacteria such as Photobacterium luminescens and mycobacteria that act as commensals or pathogens for other organisms. Approximately less than 3 % of bacterial genomes are composed of global repeats, and these are particular for each phylum (Ussery et al., 2009).

Gene duplication in the CAIM 1792 genome (3.3 %) is similar to that in the predicted proteomes analysed, and no pattern that could differentiate among species was found. Gene duplication might be strain specific either due to its nature or due to sequence or assembly quality. In Escherichia coli K-12, paralogous genes vary from 7.5 to 11.9 % denoting, similarly, a strain-specific feature since gene duplication is one of the main sources of functional divergence in organisms (Babu et al., 2004; Conant & Wolfe, 2008; Gelfand, 2006; Teichmann & Babu, 2004). In addition, gene duplication is believed to be a major driving force for creating new genes in genomes (Brenner et al., 1995; Teichmann et al., 1998). Examples of gene duplication previously found in Vibrio genomes are the haemolysin genes vhh, identified in the virulent isolate V. harveyi VIB 645 (Zhang et al., 2001), or multiple copies of the cholera toxin (ctx) in V. cholerae El Tor strains arranged on large tandem repeats (Mekalanos, 1983). In CAIM 1792, the rtx toxin genes are arranged in large tandem repeats (Fig. 6, gaps XII and XIII) (Espinoza-Valles et al., 2012).

One inserted region encoding enterobactin receptors is proposed to be part of a pathogenicity island promoting host tissue damage, and thus aiding in the acquisition of iron and other nutrients as seen in V. cholerae (Ogierman et al., 1997).

**Membrane proteins, acyltransferases and two colonization islands coding for T2/4SS from V. harveyi are absent in V. owensii genomes**

Among the main differences between V. harveyi and its most closely related species V. owensii were genes encoding membrane proteins and acyltransferases in chromosome I (Fig. 6, gap II). These participate in fatty acid biosynthesis (Hoang et al., 2002) as observed in V. cholerae serotype O1, Vibrio vulnificus YJ016, Vibrio nigrisplicruditus SFn1, Vibrio anguillarum 775 and V. alginolyticus ATCC 17749† (UniProt Consortium, 2014). These acyltransferases have also been identified in Pseudomonas aeruginosa and Pseudomonas syringae, where they regulate virulence factors such as siderophores via N-acylhomoserine lactone and fatty acid synthesis (Stintzi et al., 1998; Taguchi et al., 2006, 2010). The colonization islands coding for T2/4SS, are absent in V. owensii genomes (Fig. 6, gap IV and XV). In contrast, V. rotiferianus shares one wide colonization island related to T2/4SS with CAIM 1792 and specific iron uptake siderophores (gap IV).

**A siderophore complex TonB, ToiR, ToIQ is intra-specific for V. harveyi**

Genes identified in this study as specific for V. harveyi species, such as the siderophore complex ExbB/ToIQ/ MotA, ExbD/ToIQ and TonB, are found partially in other species outside of the Harveyi clade. ExbB/ToIQ is found in E. coli, playing roles in signal transduction between cytoplasm and periplasm and the transition from Exb homodimers to homotetramers (Baker & Postle, 2013). In V. alginolyticus, MotA may be strain-specific (Li et al., 2011). However, in V. harveyi these three genes were highly conserved with 99.7, 99.6 and 98.9 % nucleotide sequence pairwise identity, respectively. Additionally, the siderophore vibrioferrin gene (psuA) may be useful for discrimination of V. harveyi from related strains since this gene is relatively more conserved in V. harveyi (93.9±14.9 %) than in other species of the Harveyi clade: V. rotiferianus (79.7±2.2 %), V. parahaemolyticus (76.7±0.2 %) and V. alginolyticus (76.4±0.3 %).

**Exclusive regions in V. harveyi CAIM 1792 are related to pathogenic strains outside of the Harveyi clade**

Some findings that could explain the pathogenicity in CAIM 1792 were observed in gaps that appeared exclusive to this strain in the BLASTP atlas. For example, some gaps were found to encode proteins present in bacteria outside of the Harveyi clade. For example, CDS in gap g encodes a transcription termination factor Rho and a thioredoxin derived from related strains since these three genes were highly conserved with 99.7, 99.6 and 98.9 % nucleotide sequence pairwise identity, respectively. In V. harveyi (93.9±14.9 %) than in other species of exclusive regions in the Harveyi clade: V. rotiferianus (79.7±2.2 %), V. parahaemolyticus (76.7±0.2 %) and V. alginolyticus (76.4±0.3 %).

**Pangenomic clustering was better defined with the strain-specific (accessory) genes**

In pangenomic plots, the plotted distance between two genomes shows the proportion of gene families where their present/absent status differs (Snipen & Ussery, 2010). Some genomes did not cluster together in the shell tree; nevertheless they clustered together in the cloud tree. The latter was reconstructed using an gradual weighting scheme opposite to that for the shell tree, with increasing weight for gene families present in fewer genomes (Snipen & Ussery, 2010). Hence, when considering the shared gene content of V. harveyi, V. owensii, V. campbellii and V. rotiferianus, the strains did not cluster according to their species as defined by DDH/MLSA. However, the four species were better defined when the rarely occurring, more strain-specific (accessory) genes (Fig. 4b) and ORFans (Fig. 3b) were used. One study with E. coli and Shigella genomes (Yu & Stoltzfus, 2012) suggested that ORFans encode real genes whose protein-coding capacity is conserved.

---

http://mic.sgmjournals.org
reflecting selection against non-synonymous mutations. On the other hand, they might be rapidly evolving genes responsible for lineage-specific adaptation; or they might be conserved genes responsible for lineage-specific features of morphology or physiology (Khalturin et al., 2008).

In general, the species discrimination with the MLSA nucleotide model (Fig. 1) is very similar to the one obtained in the pangenomic ORFans tree. Whole-genome trees are not unlike many other gene-content trees, with the emphasis on describing functional differences between closely related genomes, within a species or genus (Snipe & Ussey, 2010). Variants of such trees have previously been used successfully for differentiation of clinical and environmental strains of Vibrio (Vesth et al., 2010), E. coli (Lukjancenko et al., 2012), Staphylococcus aureus (Snipe & Ussey, 2010), Salmonella enterica (Jacobsen et al., 2011) and Bifidobacterium and Lactobacillus (Lukjancenko et al., 2012).

V. harveyi was the most conserved species in SNP analysis

The automated SNP analysis represented a good opportunity to compare and evaluate the accuracy and reliability of housekeeping genes and SNPs to delineate species of the Harveyi clade. SNP analysis separated the 38 genomes into their expected species, with the exception of two V. harveyi (2415-05 and E385) and two V. owensii genomes (47666-1 and DY051). It is possible that the unexpected clustering of these four strains is related to the wide geographical origin and source range of the strains included in the analysis, as observed previously for other species (Leekitcharoenphon et al., 2012). It is therefore suggested that SNP analysis continues in the future when complete reference genomes are available for V. harveyi and V. owensii.

V. campbellii comprises at least three subspecies

The GGDC analysis showed that V. campbellii PEL22A, V. campbellii CAIM 198 and V. campbellii 200612B represent subspecies of V. campbellii, as defined by the type strain 5191. Comparison of V. campbellii PEL22A and 5191 resulted in a DDH estimate of 73.60 ± 2.92 %, with a probability of being the same species of 83.97 %; however, the probability that the two strains belong to the same subspecies was only 35.5 %. V. campbellii CAIM 198 and V. campbellii 200612B both yielded DDH estimates near the threshold of bacterial species delimitation (< 70 %), with 68.20 ± 2.92 % and 69.50 ± 2.92 %, respectively. The likelihood that these strains belong to the same species as 5191 was estimated to 74.85 % and 77.36 %, respectively, while the probability of belonging to the same subspecies was only 30 %. Furthermore, while the MLSA and ORFans trees supported the notion that the three strains belong to V. campbellii, the core-genome and moderately conserved gene (shell) trees supported the existence of at least three V. campbellii subspecies.

Glycosyltransferases and restriction modification systems are specific for V. harveyi CAIM 1792

At the species level, unique genes have been reported to be scarce and mostly with unknown function. For example, a recent study of unique core genomes in Vibrionaceae reported that V. cholerae and V. parahaemolyticus have 12 and 124 unique genes, respectively (Kahlke et al., 2012). A comparable number was obtained in this study for V. harveyi with 94 unique gene families (data not shown). At the strain level, 80 unique gene families were found for CAIM 1792, of which only 24 have a known function. The genes related to pathogenic activity are few, such as a hypothetical regulator, LysR that regulates a diverse set of genes, including some involved in virulence, metabolism, quorum sensing and motility (Maddocks & Oyston, 2008). Other examples are the glycosyltransferases, which probably share the function of cell wall biosynthesis as in V. fischeri MJ11 (Zhang et al., 2014), since one conserved neighbour protein (phospholipid-lipopolysaccharide ABC transporter CDS) is in the same operon. In addition to the characterized proteins set, three types of restriction modification system (I, II and III) that provide a defence mechanism against invading viruses (Krüger & Bickle, 1983) are of special interest due to the DNA specificity domains in the restriction enzymes (hsdSRM).

This study found unique gene families for V. harveyi CAIM 1792 within the Harveyi clade, related to virulence factors, with evidence of mobile elements. Moreover, the conserved regions of its genome allowed the identification of wide colonization islands, a siderophore complex and an operon encoding citrate metabolism transport and regulation.

This study shows that accessory genes strongly supported the phylogenetic relationships found with MLSA, and hence suggests that accessory genes could be used to aid in the identification of V. harveyi, V. campbellii, V. owensii and V. rotiferianus. Furthermore, these genes may also be important in the diagnosis of disease-causing agents in aquaculture. These results are comparable to published phylogenomic and phylogenetic analyses and provide species differentiation based on less conserved genetic regions at the genomic level.

ACKNOWLEDGEMENTS

We would like to thank Drs Zheng Wang and W. Judson Hervey for their bioinformatic assistance. This work was supported by CONACYT CB-2009-01 132328 (I. E.-V. and B. G.-G.) and the Office of Naval Research via US Naval Research Laboratory core funds (G. J. V.).

REFERENCES


Leekitcharoenphon, P., Kaas, R. S., Thomsen, M. C. F., Friis, C., Rasmussen, S. & Aarestrup, F. M. (2012). SNPtree—a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics 13 (Suppl. 7), S6.


Edited by Paul O'Toole