

## Dear Author

Here are the proofs of your article.

- You can submit your corrections **online** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- Please return your proof together with the permission to publish confirmation.
- For **fax** submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the journal title, article number, and your name when sending your response via e-mail, fax or regular mail.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style. Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.

### Please note

Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI.

**Further changes are, therefore, not possible.**

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL:

<http://dx.doi.org/10.1007/s00248-009-9596-7>

If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information, go to:

<http://www.springerlink.com>.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us, if you would like to have these documents returned.

The **printed version** will follow in a forthcoming issue.

**To: Springer Customer Support 3**  
**E-mail: CorrAdmin3@spi-bpo.com**  
**Fax: +1-202-3304522**  
**SPi SPi Building, Sacsac Bacong Oriental Negros 6216 Philippines**

**Re:**

**Microbial Ecology DOI 10.1007/s00248-009-9596-7**  
On the Origins of a Vibrio Species  
**Vesth · Wassenaar · Hallin · Snipen · Lagesen · Ussery**

## Permission to publish

I have checked the proofs of my article and

- I have **no corrections**. The article is ready to be published without changes.
- I have **a few corrections**. I am enclosing the following pages:
- I have made **many corrections**. Enclosed is the **complete article**.

**Date / signature:** \_\_\_\_\_

**Metadata of the article that will be visualized in Online**


---

**Please note: Image will appear in color online but will be printed in black and white.**

---

1	Article Title	<b>On the Origins of a <i>Vibrio</i> Species</b>	
2	Article Sub- Title		
3	Article Copyright - Year	<b>The Author(s) 2009 (This will be the copyright line in the final PDF)</b>	
4	Journal Name	Microbial Ecology	
5		Family Name	<b>Ussery</b>
6		Particle	
7		Given Name	<b>David W.</b>
8	Corresponding Author	Suffix	
9		Organization	The Technical University of Denmark
10		Division	Center for Biological Sequence Analysis, Department of Systems Biology
11		Address	Building 208, Kgs. Lyngby 2800, Denmark
12		e-mail	dave@cbs.dtu.dk
13		Family Name	<b>Vesth</b>
14		Particle	
15		Given Name	<b>Tammi</b>
16		Suffix	
17	Author	Organization	The Technical University of Denmark
18		Division	Center for Biological Sequence Analysis, Department of Systems Biology
19		Address	Building 208, Kgs. Lyngby 2800, Denmark
20		e-mail	
21			Family Name
22		Particle	
23		Given Name	<b>Trudy M.</b>
24		Suffix	
25		Organization	The Technical University of Denmark
26	Author	Division	Center for Biological Sequence Analysis, Department of Systems Biology
27		Address	Building 208, Kgs. Lyngby 2800, Denmark
28		Organization	Molecular Microbiology and Genomics Consultants
29		Division	
30		Address	Zotzenheim , Germany
31		e-mail	
32		Family Name	<b>Hallin</b>
33		Particle	

34		Given Name	<b>Peter F.</b>
35		Suffix	
36		Organization	The Technical University of Denmark
37		Division	Center for Biological Sequence Analysis, Department of Systems Biology
38	Author	Address	Building 208, Kgs. Lyngby 2800, Denmark
39		Organization	Novozymes A/S
40		Division	
41		Address	Krogshøjvej 36, Bagsværd 2880, Denmark
42		e-mail	
43		Family Name	<b>Snipen</b>
44		Particle	
45		Given Name	<b>Lars</b>
46		Suffix	
47		Organization	The Technical University of Denmark
48	Author	Division	Center for Biological Sequence Analysis, Department of Systems Biology
49		Address	Building 208, Kgs. Lyngby 2800, Denmark
50		Organization	Norwegian University of Life Sciences
51		Division	Biostatistics, Department of Chemistry, Biotechnology, and Food Sciences
52		Address	Ås , Norway
53		e-mail	
54		Family Name	<b>Lagesen</b>
55		Particle	
56		Given Name	<b>Karin</b>
57		Suffix	
58		Organization	The Technical University of Denmark
59	Author	Division	Center for Biological Sequence Analysis, Department of Systems Biology
60		Address	Building 208, Kgs. Lyngby 2800, Denmark
61		Organization	University of Oslo
62		Division	Centre for Molecular Biology and Neuroscience and Institute of Medical Microbiology
63		Address	Oslo , Norway
64		e-mail	
65		Received	3 July 2009
66	Schedule	Revised	
67		Accepted	17 September 2009
68	Abstract	Thirty-two genome sequences of various <i>Vibrionaceae</i> members are compared, with emphasis on what makes <i>V. cholerae</i> unique. As few as 1,000 gene families are conserved across all the <i>Vibrionaceae</i> genomes analysed; this fraction roughly doubles for gene families conserved within the species <i>V. cholerae</i> . Of these, approximately 200 gene families that cluster on various	

locations of the genome are not found in other sequenced *Vibrionaceae*; these are possibly unique to the *V. cholerae* species. By comparing gene family content of the analysed genomes, the relatedness to a particular species is identified for two unspiciated genomes. Conversely, two genomes presumably belonging to the same species have suspiciously dissimilar gene family content. We are able to identify a number of genes that are conserved in, and unique to, *V. cholerae*. Some of these genes may be crucial to the niche adaptation of this species.

---

69 Keywords  
separated by ' - '

---

70 Foot note  
information

---

4 **On the Origins of a *Vibrio* Species**5 **Tammi Vesth · Trudy M. Wassenaar · Peter F. Hallin ·**  
6 **Lars Snipen · Karin Lagesen · David W. Ussery**7 Received: 3 July 2009 / Accepted: 17 September 2009  
8 © The Author(s) 2009. This article is published with open access at Springerlink.com9 **Abstract** Thirty-two genome sequences of various *Vibrio-*  
10 *naceae* members are compared, with emphasis on what  
11 makes *V. cholerae* unique. As few as 1,000 gene families  
12 are conserved across all the *Vibrionaceae* genomes ana-  
13 lysed; this fraction roughly doubles for gene families  
14 conserved within the species *V. cholerae*. Of these,  
15 approximately 200 gene families that cluster on various  
16 locations of the genome are not found in other sequenced  
17 *Vibrionaceae*; these are possibly unique to the *V. cholerae*  
18 species. By comparing gene family content of the analysed  
19 genomes, the relatedness to a particular species is identified  
20 for two unspiciated genomes. Conversely, two genomes  
21presumably belonging to the same species have suspicious- 22  
ly dissimilar gene family content. We are able to identify a 23  
number of genes that are conserved in, and unique to, *V.* 24  
*cholerae*. Some of these genes may be crucial to the niche 25  
adaptation of this species. 26**Introduction** 28The species concept for bacteria has long been under siege 29  
from several angles, and now with thousands of bacterial 30  
genomes being sequenced, the disputes have intensified [8]. 31  
One frequently used definition of a bacterial species is “a 32  
category that circumscribes a (preferably) genomically 33  
coherent group of individual isolates/strains sharing a high 34  
degree of similarity in (many) independent features, 35  
comparatively tested under highly standardized conditions” 36  
[12]. Such independent features are usually phenotypes that 37  
can easily be tested. For a new species to be defined, 38  
amongst other criteria, inter-species DNA–DNA hybrid- 39  
isation has to be below 70%, although this rule is not 40  
without its limitations [18]. In the late 1970s and 1980s, the 41  
16S rRNA gene sequence was introduced as a molecular 42  
clock that could be used to infer phylogenetic relationships 43  
[51]. Ideally, isolates belonging to the same species have 44  
identical or nearly identical 16S rRNA genes, and these 45  
differ from isolates belonging to different species [32, 45]. 46  
In practice, this is not always the case. Examples exist of 47  
different species sharing identical rRNA genes (for 48  
instance, *E. coli* and *Shigella* [37] that are even placed in 49  
different genera); in addition, isolates of one species can 50  
have different rRNA genes beyond the 97% that is 51  
considered to demarcate species [4]. Lateral transfer of 52  
genetic material (to which ribosomal genes are believed to 53  
be resistant) destroys the phylogenetic relationship, so that 54T. Vesth · T. M. Wassenaar · P. F. Hallin · L. Snipen ·  
K. Lagesen · D. W. Ussery (✉)  
Center for Biological Sequence Analysis,  
Department of Systems Biology,  
The Technical University of Denmark,  
Building 208,  
2800 Kgs. Lyngby, Denmark  
e-mail: dave@cbs.dtu.dkT. M. Wassenaar  
Molecular Microbiology and Genomics Consultants,  
Zotzenheim, GermanyP. F. Hallin  
Novozymes A/S,  
Krogshøjvej 36,  
2880 Bagsværd, DenmarkL. Snipen  
Biostatistics, Department of Chemistry, Biotechnology,  
and Food Sciences, Norwegian University of Life Sciences,  
Ås, NorwayK. Lagesen  
Centre for Molecular Biology and Neuroscience and Institute  
of Medical Microbiology, University of Oslo,  
Oslo, Norway

55 phylogenies based on alternative housekeeping genes can  
 56 differ from a 16S rRNA tree and frequently are not even in  
 57 accordance to each other. Such observations question the  
 58 validity of a phylogenetic tree as the most suitable model  
 59 for bacterial ancestry, when multiple genetic transfers  
 60 would produce a network-like evolutionary structure [6].  
 61 On the other hand, it is observed that lateral gene transfer is  
 62 most frequent between genetically related members sharing  
 63 a similar base content and occupying the same ecological  
 64 niche [29]. Nevertheless, a core of genes can be recognised  
 65 that produce coherent phylogenetic trees, though these may  
 66 not represent the species' complete evolutionary history as  
 67 they comprise only a minor fraction of the genetic content  
 68 of the organism [35].

69 Whether a tree or a network is more accurate to describe  
 70 phylogeny, in either case bacterial species may be consid-  
 71 ered as a cloud of isolates having a higher level of genetic  
 72 similarity to each other than to organisms belonging to a  
 73 different species. When such clouds have fuzzy and  
 74 overlapping borders, the species concept falls apart but that  
 75 will only apply to certain cases [7]. Since 16S rRNA genes  
 76 are not informative on the level of diversity within a  
 77 species, the 'density' of a cloud of isolates making up a  
 78 species cannot be determined by this gene. Those genes  
 79 shared by all isolates belonging to one species comprise the  
 80 core genome of that species [39], and the degree of  
 81 diversity in the remaining non-core genes determines the  
 82 density of the species cloud.

83 We hypothesised that certain genes can be recognised as  
 84 specific to a particular species, to be conserved in that  
 85 species but not present in related species. We tested our  
 86 hypothesis with complete genome sequences of the bacte-  
 87 rial family *Vibrionaceae*, which belong to the  $\gamma$ -  
 88 Proteobacteria and comprises eight genera. Most available  
 89 genome sequences belong to the genus *Vibrio*. This genus  
 90 contains 51 recognised species [10, 47] which are mainly  
 91 found in marine environments, frequently living in associ-  
 92 ation with marine organisms such as corals, fish, squid or  
 93 zooplankton. Most of them are symbionts and only a few  
 94 are human pathogens, notably particular serotypes of *V.*  
 95 *cholerae* producing cholera, *Vibrio parahaemolyticus*  
 96 (causing gastroenteritis) and *Vi vulnificus* (causing wound  
 97 infections) [47]. Other *Vibrionaceae*, including *V. vulnifi-*  
 98 *cus*, *Aliivibrio salmonicida* and *V. harveyi*, are fish or  
 99 shellfish pathogens and have major economic impact.  
 100 *Photobacterium profundum*, representing another genus  
 101 within the *Vibrionaceae*, was also included.

102 The gene content of 32 available sequenced *Vibriona-*  
 103 *ceae* genomes was compared and the results were analysed  
 104 in various ways. The data allowed us to identify possible *V.*  
 105 *cholerae*-specific genes, since this species was represented  
 106 by 18 genomes that was a sufficient number to test  
 107 conservation both within the species and across species.

We found that a two-component signal transduction  
 pathway is uniquely conserved in *V. cholerae* but is not  
 found outside this species. Our findings further indicated  
 that possibly a relatively small set of genes could confer  
 niche specialisation allowing *V. cholerae* to be adopted to a  
 unique environment, so that over time *V. cholerae* have  
 become a distinct species.

## Materials and Methods

### Genomes and Gene Annotations Used

Publicly available genome sequences of *Vibrionaceae* were  
 selected that were provided in less than 300 contigs and in  
 which full-length 16S rRNA sequence could be found using  
 the rRNA gene finder RNAmmer [19]. The 32 genome  
 sequences included are shown in Table 1.

The gene annotations as provided in GenBank were  
 used, except for those genomes marked "Easygene" in  
 Table 1 where protein annotation was not available in the  
 RefSeq file at the time of analysis, and we used EasyGene  
 [20] to identify the genes. As a control, an available  
 GenBank annotation was compared to a generated Easy-  
 gene annotation to confirm that the number of identified  
 genes was comparable.

### Ribosomal RNA Analysis

RNAmmer [19] was used to identify 16S rRNA sequences  
 within the 32 genomes. Sequences were considered reliable  
 if they were between 1,400 and 1,700 nucleotides long and  
 had an RNAmmer score above 1,800. In cases where the  
 program found multiple and variable 16S sequences within  
 a genome, one of these (with satisfactory RNAmmer  
 scores) was arbitrarily chosen. The sequences were aligned  
 using PRANK [23, 24], and the program MEGA4 was used  
 to elucidate a phylogenetic tree [46]. Within MEGA4, the  
 tree was created using the Neighbor-Joining method with  
 the uniform rate Jukes–Cantor distance measure and the  
 complete-delete option. Five hundred resamplings were  
 done to find the bootstrap values.

### Pan-Genome Family Clustering

Clustering based on shared gene families from the *Vibrio*  
 pan-genome was constructed, based on BLASTP similarity  
 using default settings. A BLASTP hit was considered  
 significant if the alignment produced at least 50% identity  
 for at least 50% of the length of the longest gene (either  
 query or subject). Using this criterion, each pair of genes  
 producing a significant reciprocal best hit was scored as  
 belonging to the same gene family. A genome matrix was

Origins of *V. cholerae*

t1.1 **Table 1** *Vibrionaceae* genomes used in this analysis

t1.2	GPID	Organism	Contigs	Accession/GenBank	Status	No. of genes	Ref.
t1.3	36	<i>V. cholerae</i> N16961 <sup>a</sup>	2	AE003852.1	Fully sequenced	3,828	[15]
t1.4	15667	<i>V. cholerae</i> O395 TIGR <sup>a</sup>	2	CP000626.1	Fully sequenced	3,875	[11]
t1.5	32853	<i>V. cholerae</i> O395 TEDA <sup>a</sup>	2	CP001235.1	Fully sequenced	3,934	[50]
t1.6	33555	<i>V. cholerae</i> MJ-1236 <sup>a</sup>	2	CP001485.1	Fully sequenced	3,774	[31]
t1.7	15666	<i>V. cholerae</i> MO10 <sup>a</sup>	153	NZ_AAKF00000000	Unfinished (Easygene)	3,421	[5]
t1.8	15670	<i>V. cholerae</i> V52 <sup>a</sup>	268	NZ_AAKJ00000000	Unfinished (NCBI)	3,815	[16]
t1.9	33559	<i>V. cholerae</i> BX330286 <sup>a</sup>	8	NZ_ACIA00000000	Unfinished (NCBI)	3,632	[31]
t1.10	33557	<i>V. cholerae</i> B33 <sup>a</sup>	17	NZ_ACHZ00000000	Unfinished (NCBI)	3,748	[31]
t1.11	33553	<i>V. cholerae</i> RC9 <sup>a</sup>	11	NZ_ACHX00000000	Unfinished (NCBI)	3,811	[31]
t1.12	32851	<i>V. cholerae</i> M66-2	2	CP001233.1	Fully sequenced	3,693	[50]
t1.13	18495	<i>V. cholerae</i> MZO-2	162	NZ_AAWF00000000	Unfinished (NCBI)	3,425	[16]
t1.14	18265	<i>V. cholerae</i> 1587	254	NZ_AAUR00000000	Unfinished (NCBI)	3,758	[16]
t1.15	18253	<i>V. cholerae</i> 2740-80	257	NZ_AAUT00000000	Unfinished (NCBI)	3,771	[16]
t1.16	17723	<i>V. cholerae</i> AM-19226	154	NZ_AATY00000000	Unfinished (Easygene)	3,407	[33]
t1.17	33561	<i>V. cholerae</i> 12129	12	NZ_ACFQ00000000	Unfinished (NCBI)	3,574	[31]
t1.18	33549	<i>V. cholerae</i> VL426	5	NZ_ACHV00000000	Unfinished (NCBI)	3,461	[31]
t1.19	33579	<i>V. cholerae</i> TM 11079-80	35	NZ_ACHW00000000	Unfinished (NCBI)	3,621	[31]
t1.20	33551	<i>V. cholerae</i> TMA 21	20	NZ_ACHY00000000	Unfinished (NCBI)	3,600	[31]
t1.21	13564	<i>V. campbellii</i> AND4	143	NZ_ABGR00000000	Unfinished (NCBI)	3,935	[13]
t1.22	19857	<i>V. harveyi</i> BAA-1116	3	CP000789.1	Fully sequenced	6,064	[1]
t1.23	349	<i>V. vulnificus</i> CMCP6	2	AE016795.2	Fully sequenced	4,538	[38]
t1.24	1430	<i>V. vulnificus</i> YJ016	3	BA000037.2	Fully sequenced	5,028	[3]
t1.25	19397	<i>V. shilonii</i> AK1	158	NZ_ABCH00000000	Unfinished (NCBI)	5,360	[41]
t1.26	15693	<i>Vibrio</i> sp. Ex25	222	NZ_AAKK00000000	Unfinished (Easygene)	4,004	[16]
t1.27	13616	<i>Vibrio</i> sp. MED222	99	NZ_AAND00000000	Unfinished (NCBI)	4,590	[36]
t1.28	32815	<i>V. splendidus</i> LGP32	2	FM954973.1	Fully sequenced	4,434	[27]
t1.29	19395	<i>V. parahaemolyticus</i> 16	78	NZ_ACCV00000000	Unfinished (Easygene)	3,780	[9]
t1.30	360	<i>V. parahaemolyticus</i> 2210633	2	BA000031.2	Fully sequenced	4,832	[25]
t1.31	12986	<i>A. fischeri</i> ES114	3	CP000020.1	Fully sequenced	3,823	[42]
t1.32	19393	<i>A. fischeri</i> MJ11	3	CP001133.1	Fully sequenced	4,039	[26]
t1.33	30703	<i>A. salmonicida</i> LFI1238	6	FM178379.1	Fully sequenced	4,284	[17]
t1.34	13128	<i>P. profundum</i> SS9	3	CR354531.1	Fully sequenced	5,480	[49]

GPID genome project identifier at NCBI. Contigs the number of contiguous sequences, which for a completely sequenced genome is at least two (for two chromosomes) and can be up to six when plasmids are present. Unfinished sequences are represented by multiple contigs per chromosome

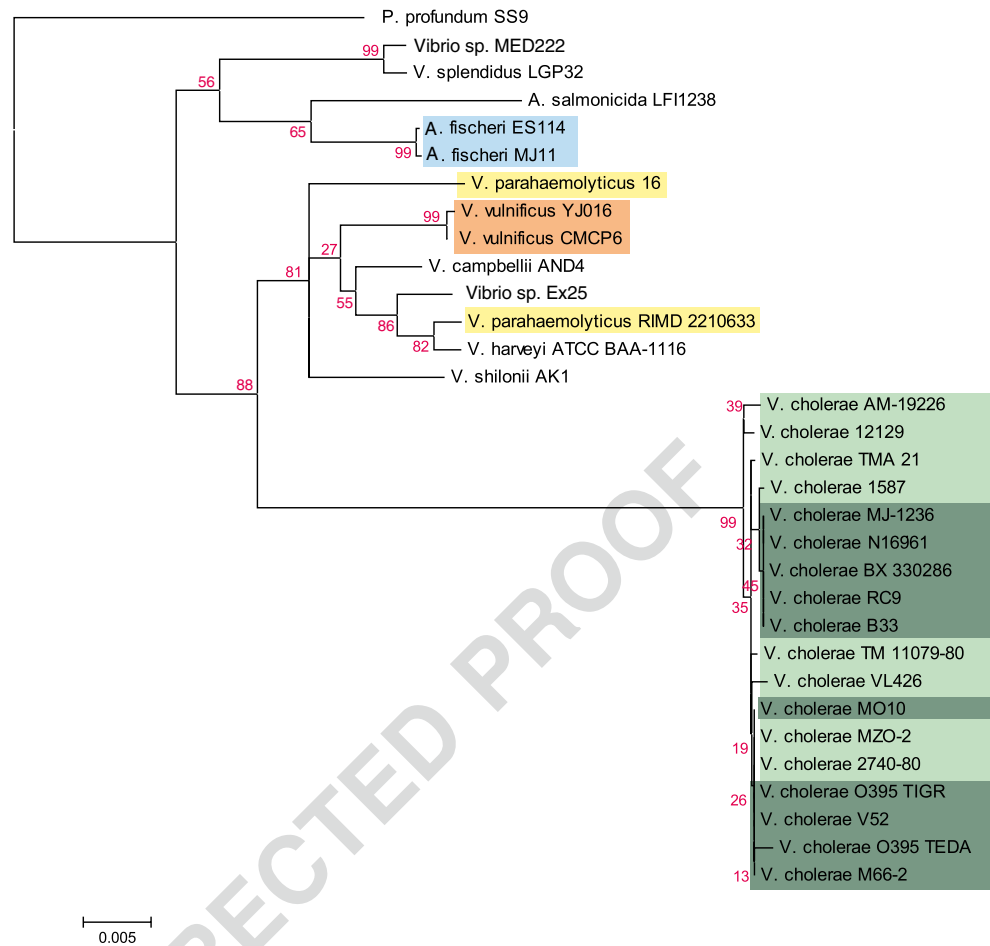
<sup>a</sup> Strains containing the genes encoding the cholera enterotoxin subunits are indicated

153 constructed, containing one row for each genome and one 164  
 154 column for each gene family. Cell (*i*, *j*) in this matrix is 1 if 165  
 155 genome *i* has a member in gene family *j*, 0 otherwise. A 166  
 156 hierarchical clustering, with average linkage based on the  
 157 Manhattan distance between genomes was then performed. Pan- and Core Genome Analysis 167  
 158 Two trees were made, one with more weight given to gene  
 159 families present in most (90%, or between 27 and 30) 168  
 160 *Vibrio* genomes (“stabilome”), and the other with more 169  
 161 weight given to gene families present in only a few (two, 170  
 162 three, or four) genomes (“mobilome”). Thus, the original 171  
 163 Boolean matrix is now scaled differently, depending on the 172  
 number of genomes in each gene family [44]. For both  
 trees, singletons (families which are only found in one  
 genome) have been excluded.

173	in the second genome were recorded and the accumulative	senting at least six species is recognised, and within this	218
174	number of gene families (as defined above) now recognised	cluster the two <i>V. parahaemolyticus</i> genes are not found on	219
175	in total was plotted for the pan-genome. The number of	the same branch. A third cluster, a bit further removed,	220
176	gene families with at least one representative gene in both	includes <i>Aliivibrio fischeri</i> and <i>A. almonidica</i> as well as <i>V.</i>	221
177	genomes was plotted for the core genome. A running total	<i>splendidus</i> and <i>Vibrio</i> species MED 222; the gene of	222
178	is plotted for the pan-genome which increases as more	<i>Photobacterium profundum</i> is the most distant.	223
179	genomes are added, whilst the core genome representing		
180	conserved gene families slowly decreases with the addition		
181	of more genomes.		
182	Whole-Genome BLAST Analysis and Construction		
183	of a BLAST Matrix		
184	The predicted genes of every genome (annotated or found	Pan-Genome Family Trees	224
185	by Easygene) were translated and every gene was compared,		
186	by BLASTP against every other genome and its own genome.	Starting with a database containing the total set of all <i>Vibrio</i>	225
187	In the latter case, the hit to self was ignored. The	gene families, a profile of matching gene families was	226
188	50/50 rule for BLAST hits as described above was used. If	constructed for each individual genome. This was stored as	227
189	these requirements were met, genes were combined in a	a matrix, containing a column for each gene families, and a	228
190	gene family. The BLAST results were visualised in a	row for each genome. The rows contain a 0 or 1	229
191	BLAST matrix [2], which summarises the results of	representing the presence or absence of the gene family.	230
192	genomic pairwise comparisons and reports, both as per-	This matrix was weighted to emphasise either the genes	231
193	centage and as absolute numbers, the number of reciprocal	found in most genomes (the “stabilome”) or in only a few	232
194	BLAST hits as a fraction of the total number of gene	genomes (the “mobilome”); from these weighted matrices,	233
195	families found in the two genomes. For easier visual	clustering of gene families yielded the resulting trees shown	234
196	inspection, the cells in the matrix are coloured darker as	in Fig. 2. Shorter distances represent genomes with many	235
197	the fraction of similarity increases. Hits identified within a	gene families in common, and larger distances reflect	236
198	genome are differently coloured.	genomes with fewer gene families in common. As	237
199	BLAST Atlas	expected, in both trees, genomes from the same species	238
200	BLAST results were also visualised in a BLAST atlas, this	cluster together, whereby the depth of resolution within a	239
201	time visualising, for all genes in the reference genome <i>V</i>	species is considerably better than can be seen in the 16S	240
202	<i>cholerae</i> N16961, their best hit in all other genomes, again	rRNA tree in Fig. 1. Similarity between the unspiciated	241
203	with a threshold of 50% identity over at least 50% of the	<i>Vibrio</i> isolate MED222 and <i>V. splendidus</i> is suggested by	242
204	length of the query protein. The atlas displays the hits as	their close clustering; this is a connection also suggested by	243
205	they are located in the reference strain [14]. The BLAST	others [21]. Note that the unspiciated <i>Vibrio</i> isolate Ex25	244
206	scores obtained for each queried gene is plotted, so that	and <i>V. parahaemolyticus</i> 2210633 cluster together in the	245
207	conserved and variable regions are located with respect to	mobilome tree, but are more distant in the stabilome. This	246
208	the reference genome. Note that genes absent in the	implies that the genes shared between these two genomes	247
209	reference genome are not shown in the lanes of the query	are less common genes within the <i>Vibrio</i> genomes	248
210	genomes.	examined here. As already indicated by the 16S rRNA	249
211	<b>Results</b>	tree, the two <i>V. parahaemolyticus</i> isolates are quite	250
212	Ribosomal RNA Analysis	dissimilar, and appear on separate branches. The <i>Aliivibrio</i>	251
213	A phylogenetic tree based on the 16S rRNA gene extracted	cluster is placed within <i>Vibrio</i> genomes in both the	252
214	from the 32 analysed <i>Vibrionaceae</i> genomes is shown in	stabilome and the mobilome, as was the case for their 16S	253
215	Fig. 1. The 18 <i>V. cholerae</i> genomes build a tight subcluster,	rRNA gene. <i>P. profundum</i> is not such an outlier as in the	254
216	quite distanced from the other species. Above this in the	16S rRNA tree, and in the stabilome. It is even positioned	255
217	figure, another subcluster comprising eight genomes repre-	close to the <i>Aliivibrio</i> genomes. Zooming in at the genomes	256
		of <i>V. cholerae</i> , a division into two subclusters can be seen;	257
		these clusters correspond to environmental vs. clinical	258
		isolates (with the exception of V52 in the stabilome).	259
		Pan- and Core Genome Plot	260
		BLAST results were analysed to construct a pan-genome,	261
		which is a hypothetical collection of all the gene families	262
		that are found in the investigated genomes [28]. The core	263
		genome was constructed from all gene families that were	264
		represented at least once in every genome. Thus, the gene	265
		families conserved in all genomes represent their core	266

Origins of *V. cholerae*

**Figure 1** Phylogenetic tree of the 16S rRNA gene extracted from 32 sequenced *Vibrio* genomes listed in Table 1. Environmental *V. cholerae* lacking the cholera enterotoxin genes are highlighted in *bright green*, whilst pathogenic *V. cholerae* genomes are in *dark green*. Further colouring was used for species for which two genomes are represented



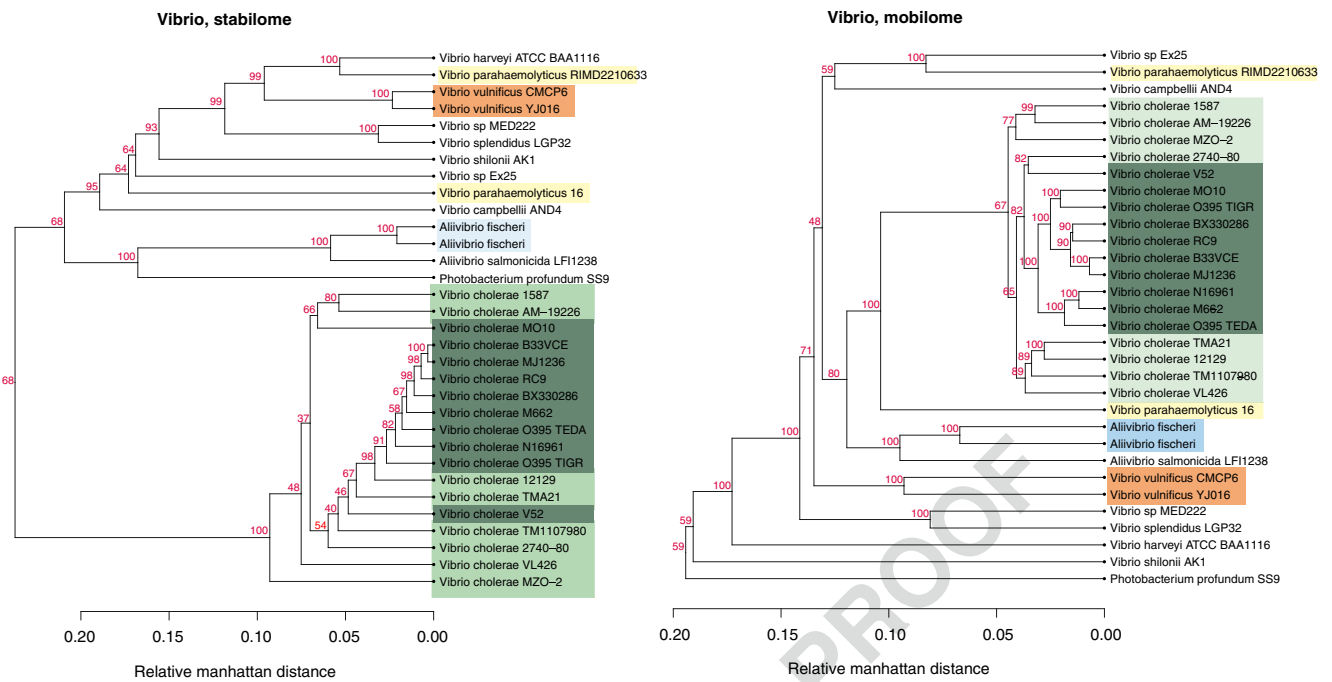
267 genome; adding the remaining gene families produces the  
 268 pan-genome. The resulting pan- and core genome plot is  
 269 shown in Fig. 3. The genomes start with the documented  
 270 clinical isolates of *V. cholerae* and then follow the order  
 271 suggested by the pan-genome family clustering (Fig. 2),  
 272 although genomes from the same species were kept  
 273 together (the two *V. parahaemolyticus* genomes were split  
 274 in the trees). As more genomes are added in the plot, the  
 275 number of gene families in the pan-genome (blue line)  
 276 increases, and the number of conserved gene families (red  
 277 line) in the core genome decreases, albeit at a lower rate.  
 278 This is because every genome can add many novel (and  
 279 frequently different) genes to the pan-genome but only  
 280 decreases the core genome with a few genes that are absent  
 281 in that particular strain but that were conserved in the  
 282 previously analysed genomes. The pan-genome curve  
 283 increases with a relative steep slope when a novel species  
 284 is added, as is obvious when a *V. parahaemolyticus* genome  
 285 is added after the last *V. cholerae*. A stable plateau can be  
 286 seen for the pan-genome of *V. cholerae* around 6,500 genes.  
 287 Nevertheless, a small increase occurs when adding *V.*  
 288 *cholerae* 11587; this is caused by the difference between  
 289 the two subclusters of *V. cholerae* seen in Fig. 2. *V.*  
 290 *cholerae* strain 2740-80 behaves atypical in all the figures

shown; although documented as an environmental isolate, it  
 appears closer to the clinical isolates, in terms of overall  
 genomic properties.

When the first genome of *A. fischeri* is added, which is  
 not a member of the *Vibrio* genus, it does not add  
 significantly more novel genes to the pan-genome than  
*Vibrio* genomes did. This contrasts with *P. profundum*  
 which produces a sharp increase in the pan-genome, as  
 does, interestingly, *V. shilonii*. Note that there are approx-  
 imately 20,200 total gene families within the 32 sequenced  
*Vibrionaceae* genomes, whereas the core genome decreases  
 to approximately 1,000 gene families.

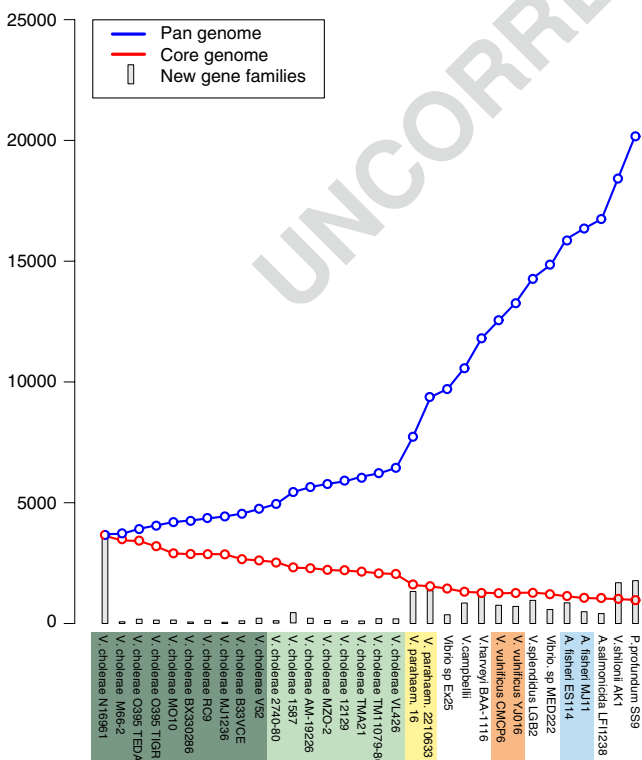
BLAST Comparison Visualised in a BLAST Matrix

A BLAST matrix provides a visual overview of reciprocal  
 pairwise whole-genome comparisons, as shown in Fig. 4.  
 The stronger a matrix cell is coloured, the more similarity  
 was detected between the gene content of two genomes. As  
 can be seen in the lower right triangle, all *V. cholerae*  
 genomes are highly similar, with similarity ranging between  
 64% and 93% for any given pair of genomes. No statistical  
 difference was observed when comparing clinical isolates  
 to environmental isolates. The two *A. fischeri* and the two



**Figure 2** Pan-genome family clustering of the 32 *Vibrio* genome sequences. The two plots represent weighted values for genes present in at least 90% of the genomes (*stabilome*) or genes found in only a

few (two to four) genomes (*mobilome*). The colours highlighting the species are the same as in Fig. 1



**Figure 3** Pan- and core genome plot of the 32 *Vibrionaceae* genomes. The colours highlighting species are the same as in Fig. 1

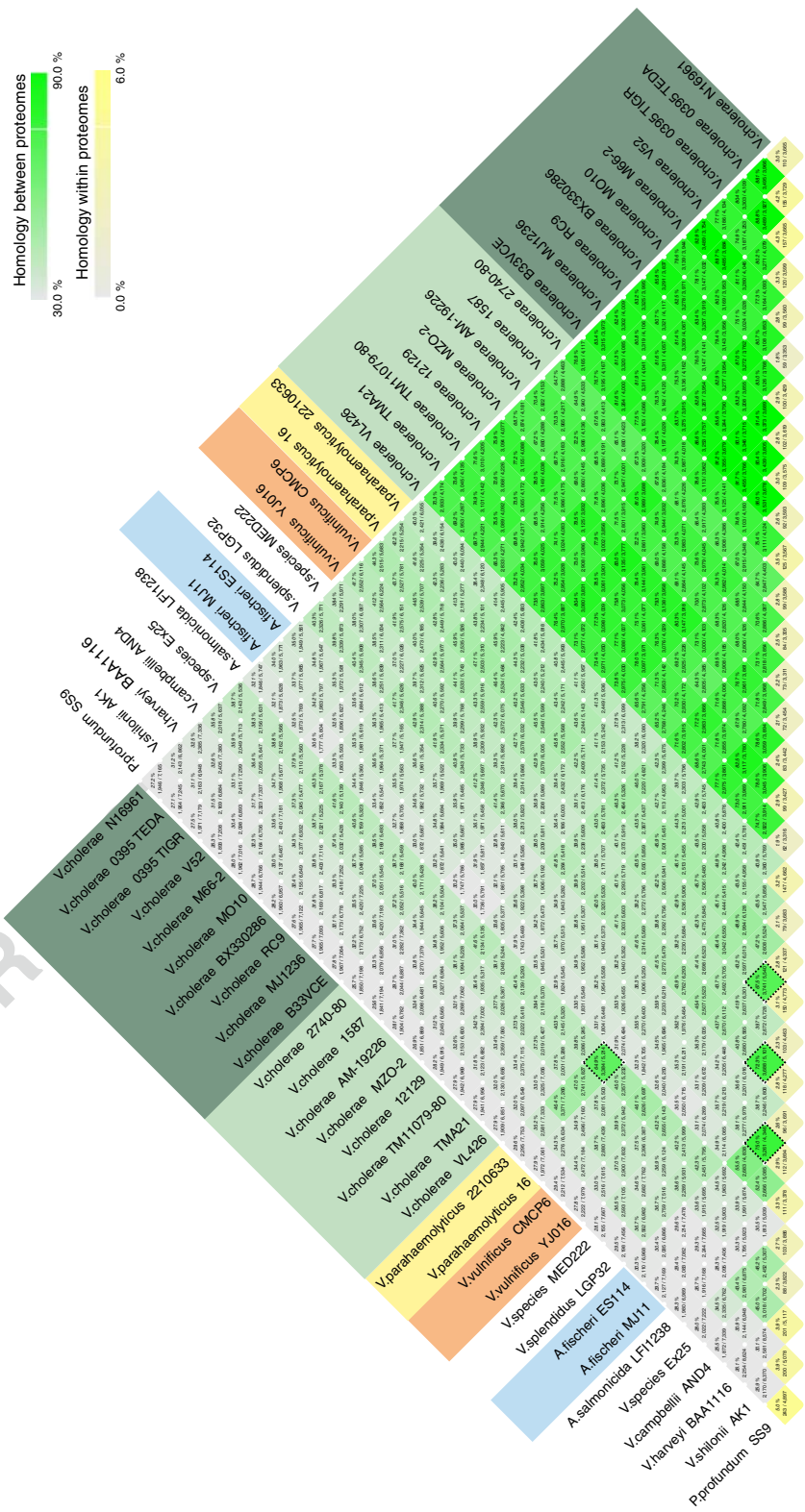
*V. vulnificus* genomes also share a high degree of identity within their species (75% and 67%, respectively), visible at the bottom of the matrix. In contrast, the two *V. parahaemolyticus* genomes only share 35% identity, which is not higher than the similarity detected between genomes of different species. With 72% similarity, isolate MED222 most closely matches *V. splendidus* and with 65% isolate EX25 again shares most similarity with *V. parahaemolyticus* 2210633.

### BLAST Atlas

A BLAST atlas was constructed using *V. cholerae* N16961 (O1, El Tor) as the reference genome, shown in Fig. 5. The best blast hits identified in the query genomes are plotted in the lanes around the reference genome, with different colours for different species. In general, chromosome 1 is more strongly conserved than chromosome 2. A large part of chromosome 2 of N16961 displays very little conservation in the other genomes; this area represents a super integron [40] that contains the *V. cholerae*-specific repeat (VCR) sequences, as well as a high number of gene cassettes. The repeat sequences are visible as black boxes in the repeat lane of the reference genome (second inner lane). Although all *V. cholerae* genomes contain a superintegron, its genes are very diverse between isolates [34] which explains the lack of blast hits in this region.

Origins of *V. cholerae*

**Figure 4** BLAST matrix of the 32 *Vibrionaceae* genomes. The colours highlighting the species are the same as in Fig. 1. Since the reciprocal similarity (reported as percent) is not readable at this resolution, every matrix cell is coloured using the scales as indicated. The *bottom row* identifies hits (other than hits-to-self) found within a genome. Four matrix cells reporting high pairwise similarities are *outlined*; their numbers are specified in the text

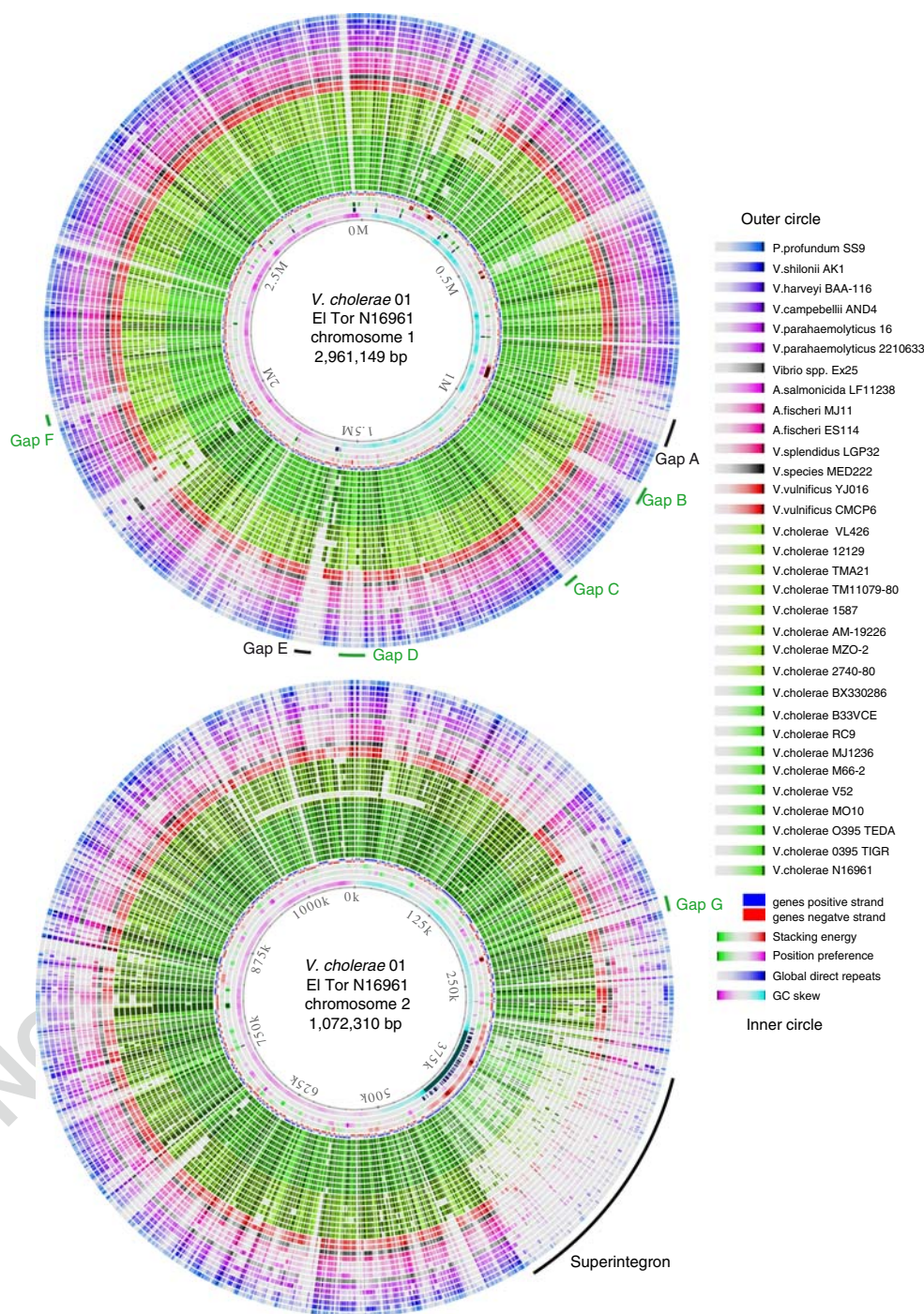


Q3

339 Several regions of the atlas have been highlighted. Gaps  
 340 B, C, D and F on chromosome 1 (indicated in green)  
 341 contain genes that are conserved in the represented  
 342 genomes of *V. cholerae* but not in the other *Vibrionaceae*.

The gaps marked A, E and G indicate regions that are  
 specific to the toxigenic, clinical isolates only. Annotated,  
*V. cholerae*-specific genes present in all these regions are  
 listed in Table 2 (hypothetical genes are excluded). Genes

**Figure 5** BLAST atlas with *V. cholerae* strain N16961 as a reference strain, showing chromosomes 1 (top) and 2 (bottom). The best BLAST hits identified with genes from N16961 in the other *V. cholerae* genomes are represented in dark red, for the location as it appears in N16961. Blast hits in the other genomes are shown in various colours as indicated to the right. Major areas conserved in *V. cholerae* but not in other *Vibrionaceae* are identified as gap B, gap C, gap D and gap F in green; areas that are found in toxigenic *V. cholerae* only are marked black as gap A, gap E and gap G. The superintegron on chromosome 2 of *V. cholerae* is also indicated



Q3

347 specific for toxinogenic *V. cholerae* identified in gap A  
 348 include, amongst others, biosynthesis genes for the toxin  
 349 co-regulated pilus (which is required for transmission of the  
 350 prophage CTX $\Phi$  carrying the enterotoxin genes), as well as  
 351 genes encoding citrate lyase. Note that the genes in gap A  
 352 are also found in the environmental isolate *V. cholerae*  
 353 2740-80.

354 Gap B contains a number of outer membrane protein  
 355 genes involved in sugar modification that are found in all *V.*  
 356 *cholerae* genomes. Genes from gap C encoding a histidine

kinase two-component signal transduction regulatory sys- 357  
 tem are also conserved within the species, as genes in gaps 358  
 D and F, involved in chemotaxis and possible multidrug 359  
 resistance. 360

Gap E, containing genes conserved in toxigenic strains 361  
 only, holds the prophage CTX $\Phi$  that contains the genes 362  
 encoding cholera enterotoxin subunits A and B; this 363  
 enterotoxin is responsible for the excessive, watery diar- 364  
 rhoea typical for cholera. Upon binding to target cell GM1 365  
 gangliosides, enterotoxin enters the cell and stimulates 366

Origins of *V. cholerae*

**Table 2** A selection of genes located in the gaps marked in Fig. 5

t2.1	Gap A (850000–913000)	
t2.2	852903–851557	Citrate/sodium symporter
t2.3	853165–854235	Citrate (pro-3S)-lyase ligase
t2.4	854287–854583	Citrate lyase subunit gamma
t2.5	854565–855455	Citrate lyase, beta subunit
t2.6	855391–856995	Citrate lyase, alpha subunit
t2.7	856992–857528	citX protein
t2.8	857506–858447	citG protein
t2.9	869812–866873	Helicase-related protein
t2.10	870391–869813	Tellurite resistance protein-related
t2.11	871298–870819	Transcriptional regulator, putative
t2.12	873242–874225	Transposase, putative
t2.13	876974–880015	ToxR-activated gene A protein
t2.14	881390–884728	Inner membrane protein, putative
t2.15	885773–886267	tagD protein
t2.16	888405–886543	Toxin co-regulated pilus biosynthesis
t2.17	888846–889511	Toxin co-regulated pilus biosynthesis
t2.18	889496–889906	Toxin co-regulated pilus biosynthesis
t2.19	890449–891123	Toxin co-regulated pilin
t2.20	891203–892495	Toxin co-regulated pilus biosynthesis
t2.21	892495–892947	Toxin co-regulated pilus biosynthesis
t2.22	892950–894419	Toxin co-regulated pilus biosynthesis
t2.23	894412–894867	Toxin co-regulated pilus biosynthesis
t2.24	894855–895691	Toxin co-regulated pilus biosynthesis
t2.25	895707–896165	Toxin co-regulated pilus biosynthesis
t2.26	896155–897666	Toxin co-regulated pilus biosynthesis
t2.27	897641–898663	Toxin co-regulated pilus biosynthesis
t2.28	898673–899689	Toxin co-regulated pilus biosynthesis
t2.29	899896–900726	TCP pilus virulence regulatory protein
t2.30	900726–901487	Leader peptidase TcpJ
t2.31	901494–903374	Accessory colonization factor AcfB
t2.32	903380–904150	Accessory colonization factor AcfC
t2.33	904648–905556	tagE protein
t2.34	906206–905559	Accessory colonization factor AcfA
t2.35	914124–912856	Phage family integrase
t2.36	Gap B (975000–1010000)	
t2.37	978644–979144	Phosphotyrosine protein phosphatase
t2.38	981833–982387	Serine acetyltransferase-related protein
t2.39	982384–983532	Exopolysacch. biosynth protein EpsF
t2.40	983529–984938	Polysacch. export protein, putative (gfcE)
t2.41	986166–986597	Serine acetyltransferase-related protein
t2.42	986597–987937	capK protein, putative
t2.43	987913–989010	Polysaccharide biosynthesis protein, putative
t2.44	1001910–1002437	Polysaccharide export-related protein (gfcE)
t2.45	1002462–1004675	Putative exopolysacch. biosynth protein
t2.46	Gap C (1130000–1160000)	
t2.47	1139646–1142912	Chitinase, putative
t2.48	1147856–1148998	Response regulator
t2.49	1149033–1149398	Response regulator
t2.50	1149990–1151309	Sensory box sensor histidine kinase

**Table 2** (continued)

1151321–1152625	Sensor histidine kinase	t2.52
1152625–1154235	Response regulator	t2.53
1154252–1155595	Response regulator	t2.54
1157228–1155624	Sensor histidine kinase	t2.55
1158044–1157232	Periplasmic binding protein-related	t2.56
Gap D (1478000–1520000)		t2.57
2086826–2087584	CDP-diacylglycerol-glyc.-3-phosph-3-phosphatidyltransferase	t2.58
2087587–2088519	Phosphatidate cytidylyltransferase	t2.59
2094741–2095604	PvcB protein	t2.60
2098112–2097183	LysR family transcriptional regulator	t2.61
2098432–2100258	pvcA protein	t2.62
2117923–2119977	Methyl-accepting chemotaxis protein	t2.63
2120575–2120030	Transcriptional regulator	t2.64
2120663–2121826	Benzoate transport protein	t2.65
Gap E (1537000–1587500)		t2.66
1541452–1543170	Sensor histidine kinase/response regulator	t2.67
1545396–1543231	Toxin secretion transporter, putative	t2.68
1546802–1545399	RTX toxin transporter	t2.69
1548919–1546757	RTX toxin transporter	t2.70
1549662–1550123	RTX toxin activating protein	t2.71
1550108–1563784	RTX toxin RtxA	t2.72
1564376–1564152	RstC protein	t2.73
1564844–1564470	RstB1 protein	t2.74
1565901–1564822	RstA1 protein	t2.75
1566027–1566365	Transcriptional repressor RstR	t2.76
1567341–1566967	Cholera enterotoxin, B subunit	t2.77
1568114–1567338	Cholera enterotoxin, A subunit	t2.78
1569412–1568213	Zona occludens toxin	t2.79
1569702–1569409	Accessory cholera enterotoxin	t2.80
1571241–1570993	Colonization factor	t2.81
1571760–1571377	RstB2 protein	t2.82
1572817–1571738	RstA1 protein	t2.83
1572943–1573281	Transcriptional repressor RstR	t2.84
1577272–1575704	Phage replication protein Cri	t2.85
1582123–1580555	Phage replication protein Cri	t2.86
1583160–1583513	Transposase OrfAB, subunit A	t2.87
1583510–1584382	Transposase OrfAB, subunit B	t2.88
Gap F (1896000–1956000)		t2.89
1896092–1897327	Phage family integrase	t2.90
1900831–1898009	Helicase, putative	t2.91
1903632–1902898	Chemotaxis protein MotB-related	t2.92
1908858–1905790	Type I restriction enzyme HsdR	t2.93
1916009–1913628	DNA methylase HsdM, putative	t2.94
1933231–1935654	Neuraminidase	t2.95
1936007–1935801	Transcriptional regulator	t2.96
1936121–1936597	DNA repair protein RadC, putative	t2.97
1938391–1937519	Transposase OrfAB, subunit B	t2.98
1938732–1938388	Transposase OrfAB, subunit A	t2.99
1941671–1941351	Transcriptional regulator, putative	t2.100

**Table 2** (continued)

t2.101	1942032–1941658	Middle operon regulator-related
t2.102	1944457–1943306	eha protein
t2.103	Gap G (chromosome II, 21300–223000)	
t2.104	213207–214250	GMP reductase
t2.105	214574–215725	DNA methyltransferase
t2.106	220262–219825	IS1004 transposase

All gene annotations are taken from the reference genome *V. cholerae* strain N16961. Hypothetical proteins were excluded. Gaps A, E and G are conserved in pathogenic strains, whereas gaps B, C, D and F are conserved in all *V. cholerae* genomes analysed (Figure 1)

367 adenylate cyclase by ADP ribosylation. The resultant  
 368 increased cyclic AMP levels induce excessive electrolyte  
 369 movement and sodium plus water secretion [43]. Strain  
 370 M66-2 is believed to be a precursor of the seventh  
 371 pandemic *V. cholerae* that lacks the prophage CTXΦ and  
 372 the enterotoxin genes [11]. Gap E bears the RTX toxin  
 373 operon, which encodes a pore-forming cytotoxin [22]. An  
 374 RTX toxin is also present in environmental isolate 2740-80  
 375 and in *V. vulnificus*.

376 Gap G on chromosome 2 consists of a set of five genes,  
 377 all in the same orientation, in a putative operon, flanked by  
 378 genes on the complimentary strand. This appears to be a  
 379 remnant of a mobile element, as these genes are flanked by  
 380 a transposase gene on the 3' end, and there is a small global  
 381 repeat on the 5' end. Only the first two of the five genes have  
 382 an assigned function, with the first gene being a GMP  
 383 reductase, and the second a putative DNA methyltransferase.  
 384 The remaining three genes are hypothetical, but their  
 385 strikingly strong conservation in all pathogenic strains and  
 386 complete absence of homologues in the other *Vibrio* genomes  
 387 strongly point towards a potential biological significance.

388 **Discussion**

389 The recent availability of many *Vibrionaceae* genomes,  
 390 including a substantial number of *V. cholerae* genomes,  
 391 allows the possibility to take a closer look at the similarities  
 392 and differences of species within the genus *Vibrio*. This can  
 393 examine, on a genome scale, what distinguishes *V. cholerae*  
 394 from the other *Vibrio* species. Since not all *V. cholerae*  
 395 isolates are pathogenic, the presence of the prophage-  
 396 bearing cholera enterotoxin, the main virulence factor for  
 397 cholera, is not a suitable marker for this species. We  
 398 attempted to identify a set of *V. cholerae*-specific genes,  
 399 and also explored the internal diversity within the *V.*  
 400 *cholerae* genomes that have been sequenced to date.

401 On a phylogenetic tree based on the 16S ribosomal RNA  
 402 gene, those isolates that do not belong to the genus *Vibrio*  
 403 were positioned as outliers, as expected. This tree further

indicated the closest resembling 16S rRNA sequence for 404  
 the two sequenced *Vibrio* strains that are currently not 405  
 assigned to a species. It was observed that the two 406  
 sequenced *V. parahaemolyticus* strains were not placed 407  
 together. The complete gene content of each genome was 408  
 next compared by BLAST and the results were pooled into 409  
 gene families which were subjected to cluster analysis. This 410  
 provided evidence that the 18 *V. cholerae* genomes fall into 411  
 two subclusters, one mainly containing clinical isolates and 412  
 the other environmental isolates. 413

414 The gene family clustering, subsequent pan-genome  
 415 analysis and the pairwise BLAST results, as summarised  
 416 in the BLAST matrix, all supported the relatedness of  
 417 *Vibrio* species Ex25 to *V. parahaemolyticus* 2210633 but  
 418 not to *V. parahaemolyticus* 16. This latter genome was quite  
 419 different from *V. parahaemolyticus* 2210633 in all analyses.  
 420 Although it is possible that the species *V. parahaemolyticus*  
 421 is far more genetically diverse than *V. cholerae*, *A. fischeri*  
 422 or *V. vulnificus*, an alternative explanation is that one of the  
 423 sequenced isolates is perhaps incorrectly named as *V.*  
 424 *parahaemolyticus*. The similarity between *Vibrio* species  
 425 MED222 and *V. splendidus* based on gene families is in  
 426 agreement with their related 16S rRNA genes and pub-  
 427 lished data [21]. However, in contrast to what the ribosomal  
 428 gene suggests, our whole-genome comparison indicates that  
 429 the three *Aliivibrio* genomes (*A. salmonicida* and two *A.*  
 430 *fischeri*) are not so different from *Vibrio* after all. Their  
 431 recent placement in the genus *Aliivibrio*, a decision based  
 432 on five genes (the 16S rRNA gene and four housekeeping  
 433 genes) and phenotypical characteristics [48], appears not to  
 434 be reflective of the whole genome picture presented here.

435 The BLAST results were graphically summarised in a  
 436 BLAST atlas, which visualised *V. cholerae*-specific gene  
 437 clusters. These coded for polysaccharide biosynthesis  
 438 enzymes, response regulators and chemotaxis proteins,  
 439 amongst others. In addition, a *V. cholerae*-specific, histidine  
 440 kinase two-component signal transduction regulatory sys-  
 441 tem was identified. The two-component signal transduction  
 442 pathway is a powerful regulating system for bacteria to  
 443 adapt to a particular ecological niche. There is a precedent  
 444 for this claim, as the introduction of a single regulatory  
 445 protein in *Vibrio fischeri* strain MJ11 has been shown to  
 446 specifically enable colonization of the squid *Euprymna*  
 447 *scolopes* [26].

448 As expected, the main differences observed between *V.*  
 449 *cholerae* clinical isolates and the environmental strains are  
 450 due to genes related to virulence. Two exceptions are the  
 451 presence of a number of virulence genes in the environ-  
 452 mental strain *V. cholerae* 2740-80 and the absence of  
 453 enterotoxin genes in clinical isolate M66-2. It has already  
 454 been suggested that M66-2 might be a predecessor of  
 455 pandemic, enterotoxic *V. cholerae* [11]. From sequence  
 456 comparison of four housekeeping genes, it was concluded

457 that *V. cholerae* 2740-80 is intermediary between toxigenic  
458 and non-toxigenic isolates [30]. This view is confirmed by  
459 the data presented here, although we propose to consider  
460 the possibility that the isolate arose from a pandemic clone  
461 that has lost the CTX $\Phi$  prophage, rather than being a  
462 precursor of a pathogen.

463 In conclusion, several different methods of genome  
464 comparisons have yielded a picture of *V. cholerae* genomes  
465 as forming a distinct cluster, compared to related species,  
466 and a relatively small number of genes might be responsible  
467 for environmental niche adaptation and hence for genera-  
468 tion of this distinct species. Likely candidates include  
469 multiple two-component signal transduction regulatory  
470 proteins as well as chemotaxis proteins.

471  
472 **Acknowledgements** We would like to thank Tim Binnewies for  
473 early work on this project, and also to the Danish Research Councils  
474 and the DTU Globalization funds for financial support.

475 **Open Access** This article is distributed under the terms of the  
476 Creative Commons Attribution Noncommercial License which per-  
477 mits any noncommercial use, distribution, and reproduction in any  
478 medium, provided the original author(s) and source are credited.

## 479 References

- 480 1. Bassler B et al. (2007) CP000789.1: Direct submission to  
481 GenBank
- 482 2. Binnewies TT, Hallin PF, Staerfeldt HH, Ussery DW (2005)  
483 Genome update: proteome comparisons. *Microbiol* 151:1–4
- 484 3. Chen CY, Wu KM, Chang YC, Chang CH, Tsai HC, Liao TL, Liu  
485 YM, Chen HJ, Shen AB, Li JC, Su TL, Shao CP, Lee CT, Hor LI,  
486 Tsai SF (2003) Comparative genome analysis of *Vibrio vulnificus*,  
487 a marine pathogen. *Genome Res* 13:2577–2587
- 488 4. Clayton RA, Sutton G, Hinkle PS, Bult C, Fields C (1995)  
489 Intraspecific variation in small-subunit rRNA sequences in  
490 GenBank: why single sequences may not adequately represent  
491 prokaryotic taxa. *Int J Syst Bacteriol* 45:595–599
- 492 5. Colwell R, Grim CJ, Young S, Jaffe D, Gnerre S, Berlin A,  
493 Heiman D, Hepburn T, Shea T, Sykes S, Alvarado L, Kodira C,  
494 Heidelberg J, Lander E, Galagan J, Nusbaum C, Birren B (2008)  
495 NZ\_AAKF000000000: Direct submission to GenBank
- 496 6. Doolittle WF (1995) Phylogenetic classification and the universal  
497 tree. *Science* 284:2124–2129
- 498 7. Doolittle WF, Papke RT (2006) Genomics and the bacterial  
499 species problem. *Genome Biol* 7:116
- 500 8. Doolittle WF, Zhaxybayeva O (2009) On the origin of prokaryotic  
501 species. *Genome Res* 19:744–756
- 502 9. Edwards R, Ferreira S, Johnson J, Kravitz S, Beeson K, Sutton G,  
503 Rogers Y-H, Friedman R, Frazier M, Venter JC (2008)  
504 NZ\_ACCV000000000: Direct submission to GenBank
- 505 10. Farmer JJ, Janda JM (2005) Vibrionaceae. In: *Bergey's manual of*  
506 *systematic bacteriology*, 2nd edn, vol 2 part B, pp. 491–546
- 507 11. Feng L, Reeves PR, Lan R, Ren Y, Gao C, Zhou Z, Ren Y, Cheng  
508 J, Wang W, Wang J, Qian W, Li D, Wang L (2008) A recalibrated  
509 molecular clock and independent origins for the cholera pandemic  
510 clones. *PLoS ONE* 3:e4053
- 511 12. Gevers D, Cohan FM, Lawrence JG, Sprat BG, Coeyne T, Feil EJ,  
512 Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL,

- Swings J (2005) Re-evaluating prokaryotic species. *Nat Rev* 513  
*Microbiol* 3:733–739 514
13. Hagstrom A, Ferreira S, Johnson J, Kravitz S, Beeson K, Sutton  
515 G, Rogers Y-H, Friedman R, Frazier M, Venter JC (2007)  
516 NZ\_ABGR000000000: Direct submission to GenBank 517
14. Hallin PF, Binnewies TT, Ussery DW (2008) The genome 518  
BLASTAtlas—a GeneWiz extension for visualization of whole-  
519 genome homology. *Mol Biosyst* 4:363–371 520
15. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML,  
521 Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, Gill  
522 SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva  
523 MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald  
524 L, Utterback T, Fleishmann RD, Nierman WC, White O, Salzberg  
525 SL, Smith HO, Colwell RR, Mekalanos JJ, Venter JC, Fraser CM  
526 (2000) DNA sequence of both chromosomes of the cholera  
527 pathogen *Vibrio cholerae*. *Nature* 406:477–483 528
16. Heidelberg J, Sebastian Y. NZ\_AAKJ000000000, NZ\_AAUT000000000,  
529 NZ\_AAKK000000000, NZ\_AAUR000000000, NZ\_AAWF000000000:  
530 Direct submission to GenBank 531
17. Hjerde E, Lorentzen MS, Holden MT, Seeger K, Paulsen S, Bason  
532 N, Churcher C, Harris D, Norbertczak H, Quail MA, Sanders S,  
533 Thurston S, Parkhill J, Willassen NP, Thomson NR (2008) The  
534 genome sequence of the fish pathogen *Aliivibrio salmonicida*  
535 strain LFH1238 shows extensive evidence of gene decay. *BMC*  
536 *Genomics* 9:616 537
18. Konstantinidis T, Ramette A, Tiedje JA (2006) The bacterial  
538 species definition in the genomic era. *Phil Trans R Soc B*  
539 361:1929–1940 540
19. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T,  
541 Ussery DW (2007) RNAmmer: consistent and rapid annotation of  
542 ribosomal RNA genes. *Nucleic Acids Res* 35:3100–3108 543
20. Larsen TS, Krogh A (2003) EasyGene—a prokaryotic gene finder  
544 that ranks ORFs by statistical significance. *BMC Bioinformatics* 4:29  
545 546
21. Le Roux F, Zouine M, Chakroun N, Binesse J, Saulnier D,  
547 Bouchier C, Zidane N, Ma L, Rusniok C, Lajus A, Buchrieser C,  
548 Médigue C, Polz MF, Mazel D (2009) Genome sequence of *Vibrio*  
549 *splendidus*: an abundant planktonic marine species with a large  
550 genotypic diversity. *Environ Microbiol* 11:1959–1970 551
22. Lin W, Fullner KJ, Clayton R, Sexton JA, Rogers MB, Calia KE,  
552 Calderwood SB, Fraser C, Mekalanos JJ (1999) Identification of  
553 a *Vibrio cholerae* RTX toxin gene cluster that is tightly linked to  
554 the cholera toxin prophage. *Proc Natl Acad Sci U S A* 96:1071–  
555 1076 556
23. Loytynoja A, Goldman N (2005) An algorithm for progressive  
556 multiple alignment of sequences with insertions. *Proc Natl Acad*  
557 *Sci U S A* 102:10557–10562 558
24. Loytynoja A, Goldman N (2008) Phylogeny-aware gap placement  
559 prevents errors in sequence alignment and evolutionary analysis.  
560 *Science* 320:1632–1635 561
25. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T,  
562 Tagomori K, Iijima Y, Najima M, Nakano M, Yamashita A,  
563 Kubota Y, Kimura S, Yasunaga T, Honda T, Shinagawa H, Hattori  
564 M, Iida T (2003) Genome sequence of *Vibrio parahaemolyticus*: a  
565 pathogenic mechanism distinct from that of *V. cholerae*. *Lancet*  
566 361:743–749 567
26. Mandel MJ, Wollenberg MS, Stabb EV, Visick KL, Ruby EG  
568 (2009) A single regulatory gene is sufficient to alter bacterial host  
569 range. *Nature* 458:215–218 570
27. Mazel D, Le Roux F (2008) FM954973.1: Direct submission to  
571 GenBank 572
28. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005)  
573 The microbial pan-genome. *Curr Opin Genet Dev* 15:589–594 574
29. Medrano-Soto A, Moreno-Hagelsieb G, Vinuesa P, Christen JA,  
575 Collado-Vides J (2001) Successful lateral transfer requires codon  
576 usage compatibility between foreign genes and recipient genomes.  
577 *Mol Biol Evol* 21:1884–1894 578

- 579 30. Mohapatra SS, Ramachandran D, Mantri CK, Colwell RR, Singh 616  
 580 DV (2009) Determination of relationships among non-toxicogenic 617  
 581 *Vibrio cholerae* O1 biotype El Tor strains from housekeeping gene 618  
 582 sequences and ribotype patterns. *Res Microbiol* 160:57–62 619  
 583 31. Munk A, Tapia R, Green L, Rogers Y, Detter JC, Bruce D, Brettin TS, 620  
 584 Colwell R, Grim C, Vonstein V, Bartels D. CP001485.1, 621  
 585 NZ\_ACHV000000000, NZ\_ACHY000000000, NZ\_ACHW000000000, 622  
 586 NZ\_ACHX000000000, NZ\_ACHZ000000000, NZ\_ACIA000000000, 623  
 587 NZ\_ACFQ000000000: Direct submission to GenBank 624  
 588 32. Murray RG, Stackebrandt E (1995) Taxonomic note: implementa- 625 **Q5**  
 589 tion of the provisional status Candidatus for incompletely 626  
 590 described prokaryotes. *Int J Syst Bacteriol* 45:186–187 627  
 591 33. Nierman WC (2006) NZ\_AATY000000000: Direct submission to 628  
 592 GenBank 629  
 593 34. Pang B, Yan M, Cui Z, Ye X, Diao B, Ren Y, Gao S, Zhang L, 630  
 594 Kan B (2007) Genetic diversity of toxigenic and nontoxigenic 631  
 595 *Vibrio cholerae* serogroups O1 and O139 revealed by array-based 632  
 596 comparative genomic hybridization. *J Bacteriol* 189:4837–4879 633  
 597 35. Philippe H, Douady CJ (2003) Horizontal gene transfer and 634  
 598 phylogenetics. *Curr Opin Microbiol* 6:498–505 635  
 599 36. Pinhassi J, Pedros-Alio C, Ferreira S, Johnson J, Kravitz S, 636  
 600 Halpern A, Remington K, Beeson K, Tran B, Rogers Y-H, 637  
 601 Friedman R, Venter JC (2006) NZ\_AAND000000000: Direct 638  
 602 submission to GenBank 639  
 603 37. Pupo GM, Lan R, Reeves PR (2000) Multiple independent origins 640  
 604 of *Shigella* clones of *Escherichia coli* and convergent evolution of 641  
 605 many of their characteristics. *Proc Natl Acad Sci U S A* 642  
 606 97:10567–10572 643  
 607 38. Rhee JH, Kim SY, Chung SS, Lee SE, Choy HE (2002) 644  
 608 AE016795.2: Direct submission to GenBank 645  
 609 39. Riley MA, Lizotte-Waniewski M (2009) Population genomics and 646  
 610 the bacterial species concept. *Methods Mol Biol* 532:367–377 647  
 611 40. Rowe-Magnus DA, Guérout AM, Mazel D (1999) Super- 648  
 612 integrons. *Res Microbiol* 150:641–651 649  
 613 41. Rosenberg E, Ferreira S, Johnson J, Kravitz S, Beeson K, Sutton 650  
 614 G, Rogers Y-H, Friedman R, Frazier M, Venter JC (2006) 651  
 615 NZ\_ABCH000000000: Direct submission to GenBank 652  
 42. Ruby EG, Urbanowski M, Campbell J, Dunn A, Faini M, 616  
 Gunsalus R, Lostroh P, Lupp C, McCann J, Millikan D, Schaefer 617  
 A, Stabb E, Stevens A, Visick K, Whistler C, Greenberg EP 618  
 (2005) Complete genome sequence of *Vibrio fischeri*: a symbiotic 619  
 bacterium with pathogenic congeners. *Proc Natl Acad Sci U S A* 620  
 102:3004–3009 621  
 43. Sánchez J, Holmgren J (2005) Virulence factors, pathogenesis and 622  
 vaccine protection in cholera and ETEC diarrhoea. *Curr Opin* 623  
*Immunol* 17:388–398 624  
 44. Snippen L, Ussery D (2009) Pan-genome family trees. *SIGS* 625  
*Journal* (in press) 626  
 45. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, 627  
 Kämpfer P, Maiden MC, Nesme X, Rosselló-Mora R, Swings J, 628  
 Trüper HG, Vauterin L, Ward AC, Whitman WB (2002) Report of 629  
 the ad hoc committee for the re-evaluation of the species 630  
 definition in bacteriology. *Int J Syst Evol Microbiol* 52:1043– 631  
 1047 632  
 46. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular 633  
 Evolutionary Genetics Analysis (MEGA) software version 4.0. 634  
*Mol Biol Evol* 24:1596–1599 635  
 47. Thompson FL, Iida T, Swings J (2004) Biodiversity of vibrios. 636  
*Microbiol Mol Biol Rev* 68:403–431 637  
 48. Urbanczyk H, Ast JC, Higgins MJ, Carson J, Dunlap PV (2007) 638  
 Reclassification of *Vibrio fischeri*, *Vibrio logei*, *Vibrio salmonicida* 639  
 and *Vibrio wodanis* as *Aliivibrio fischeri* gen. nov., comb. 640  
 nov., *Aliivibrio logei* comb. nov., *Aliivibrio salmonicida* comb. 641  
 nov. and *Aliivibrio wodanis* comb. nov. *Int J Syst Evol Microbiol* 642  
 57:2823–2829 643  
 49. Vezzi A, Campanaro S, D'Angelo M, Simonato F, Vitulo N, Lauro 644  
 FM, Cestaro A, Malacrida G, Simionati B, Cannata N, Romualdi 645  
 C, Bartlett DH, Valle G (2005) Life at depth: *Photobacterium* 646  
*profundum* genome sequence and expression analysis. *Science* 647  
 30:1459–1461 648  
 50. Wang L, Feng L, Reeves P, Lan R, Ren Y, Gao C, Zhou Z, Ren Y, 649  
 Wang W (2008) CP001233.1. CP001235.1: Direct submission to 650  
 GenBank 651  
 51. Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51:221–271 652

## AUTHOR QUERY

### AUTHOR PLEASE ANSWER QUERY.

- Q1. Please check authors' affiliations if these were correctly presented.
- Q2. All occurrences of "phylogenic" in the article have been changed to "phylogenetic". Please check.
- Q3. Figures 2-5 has poor quality with data smaller than 6 pts. Please provide better quality of the said figures.
- Q4. Farmer and Janda (2005) Please provide complete bibliographic details for this reference item.
- Q5. Snippen and Ussery 2009 (in press) Please update the publication status of this reference.

UNCORRECTED PROOF