# *Ca*. Endozoicomonas cretensis: A Novel Fish Pathogen Characterized by Genome Plasticity

Weihong Qi[1], Maria Chiara Cascarano[2], Ralph Schlapbach[1], Pantelis Katharios[2], Lloyd Vaughan[3,4], and Helena M.B. Seth-Smith[1,3,*]

[1]Functional Genomics Center Zurich, University of Zurich, Switzerland

[2]Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Crete, Greece

[3]Institute for Veterinary Pathology, Vetsuisse Faculty, University of Zurich, Switzerland

[4]Pathovet AG, Tagelswangen, Switzerland

*Corresponding author: E-mail helena.seth-smith@unibas.ch.

Accepted: May 2, 2018

Data deposition: ERZ494307, KP890196-KP890204

## Abstract

*Endozoicomonas* bacteria are generally beneficial symbionts of diverse marine invertebrates including reef-building corals, sponges, sea squirts, sea slugs, molluscs, and Bryozoans. In contrast, the recently reported *Ca*. Endozoicomonas cretensis was identified as a vertebrate pathogen, causing epitheliocystis in fish larvae resulting in massive mortality. Here, we described the *Ca*. E. cretensis draft genome, currently undergoing genome decay as evidenced by massive insertion sequence (IS element) expansion and pseudogene formation. Many of the insertion sequences are also predicted to carry outward-directed promoters, implying that they may be able to modulate the expression of neighbouring coding sequences (CDSs). Comparative genomic analysis has revealed many *Ca*. E. cretensis-specific CDSs, phage integration and novel gene families. Potential virulence related CDSs and machineries were identified in the genome, including secretion systems and related effector proteins, and systems related to biofilm formation and directed cell movement. Mucin degradation would be of importance to a fish pathogen, and many candidate CDSs associated with this pathway have been identified. The genome may reflect a bacterium in the process of changing niche from symbiont to pathogen, through expansion of virulence genes and some loss of metabolic capacity.

**Key words:** host–pathogen, resistance, virulence, genome degradation, genome decay, mobile elements.

## Introduction

*Endozoicomonas* bacteria are facultative intracellular or cell-associated symbiotic bacteria, found in close association with diverse marine invertebrate hosts including reef-building corals, sponges, Bryozoans, sea squirts, sea slugs, and molluscs (Jensen et al. 2010; Morrow et al. 2012; Forget and Juniper 2013; Fiore et al. 2015; Bourne et al. 2016; Miller et al. 2016; Neave et al. 2016; Schreiber et al. 2016a). Their reputation as beneficial symbionts stands in contrast to the recent discovery of the endozoicomonal pathogen, *Ca*. Endozoicomonas cretensis, causing epitheliocystis in sharpsnout seabream larvae, *Diplodus puntazzo*, and massive mortalities in aquaculture facilities (Katharios et al. 2015). Bacteria carrying an identical 16S rRNA gene sequence were also found to be present during an epitheliocystis outbreak in cobia larvae in Colombia (Mendoza et al. 2013), suggesting that this is indeed a novel pathogen of fish larvae.

Several whole genome sequences of symbiotic *Endozoicomonas* have been published, from cultured strains (Neave et al. 2014; Appolinario et al. 2016; Ding et al. 2016; Schreiber et al. 2016b) and directly from marine invertebrates using culture-independent and metagenomic binning methods (Miller et al. 2016; Neave et al. 2017). *Endozoicomonas* genomes in general are large (over 5 Mb) and contain many genes for the transport of molecules and secretion of proteins. An interpretation of this is that the bacteria have a free-living independent stage and can also exist associated with diverse marine hosts, either

symbiotically or pathogenically (Neave et al. 2016). *Endozicomonas montiporae* contains over 450 mobile elements, causing some gene disruption, leading to the idea it is a recently host-restricted symbiont (Ding et al. 2016). Host restriction and increased pathogenicity as a result of mobile element expansion and genome reduction has been observed in several pathogens (Cole et al. 1998; Parkhill et al. 2003; Holden et al. 2009).

We now present the refined, annotated genome of *Ca.* E. cretensis (Katharios et al. 2015). This genome has been derived from infected fish tissue, as *Ca.* E. cretensis infected sharpsnout seabream larvae have only been found during one experimental season, and the pathogen has not been successfully cultured. As the material available for sequencing was very limited and not of high quality, we were unable to capitalise on progresses in long read sequencing technologies, and have used metagenomic binning methods and manual curation to produce an improvement on our earlier draft (Katharios et al. 2015). This refined draft offers an opportunity to study how this *Ca*. Endozoicomonas species, related to species symbiotically associated with invertebrates, appears to have evolved to become a vertebrate pathogen. The findings suggest a genome encoding pathogenic potential, undergoing niche adaptation through IS element expansion and loss of functional genes.

## Materials and Methods

### Metagenome Sequencing, Assembly, and Binning

In the previous report (Katharios et al. 2015), genomic DNA isolated from a tail section of an infected larva (sample Dpd28tailN) was sequenced successfully using the low input Nextera protocol on an Illumina MiSeq sequencer (supplementary table S1, Supplementary Material online).

Raw reads were preprocessed using Trimmomatic-0.32 (Bolger et al. 2014) and quality controlled reads (average phred quality above 20 and longer than 36 nt) were assembled using SPAdes-3.1.0 (Bankevich et al. 2012) with both single-cell and multicell modes (Nurk et al. 2013). Quality controlled reads were mapped back to the assembled contigs using Bowtie2-2.2.3 (–all –no-mixed) (Langmead and Salzberg 2012) and genome assembly likelihoods were computed using CGAL-0.9.6 (Rahman and Pachter 2013). For each sample the metagenome assembly with higher genome assembly likelihood was selected for downstream analysis (supplementary table S1, Supplementary Material online). Due to the limited amount of input DNA, only one paired-end (PE) library prepared with the low input Nextera protocol was sequenced per sample. The material was not sufficient to generate further PE and mate-paired libraries with different insert sizes, which would have improved scaffolding of this genome (Ekblom and Wolf 2014), and the scaffold and contig statistics were almost identical (supplementary table S1, Supplementary Material online), as SPAdes was able to scaffold only a few contigs.

The metagenome assembly was scanned using ncbi-blast-2.2.29+ (Altschul et al. 1990) for the presence of *Ca.* E. cretensis 16S rRNA gene sequences (EMBL accession number LN626318) to confirm the presence of the target pathogen. The completeness and diversity of bacterial genomes harbored in the metagenome was estimated using the 40 marker genes universal for all bacteria and archaea (Wu et al. 2013), identified via hmmsearch in hmmer-3.1b2 (Mistry et al. 2013). Taxonomic content was analyzed using MEGAN5 (Huson et al. 2011) based on BLAST comparison against the NCBI nonredundant DNA database (nt) (supplementary table S1, Supplementary Material online). The metagenome was first binned using MaxBin-1.4.2 (Wu et al. 2014): CDSs from scaffolds longer than 1 Kb were predicted using FragGeneScan-1.18 (Rho et al. 2010). All predicted genes were scanned using hmmsearch in hmmer-3.1b2 (Mistry et al. 2013) for the 107 single-copy bacterial marker genes conserved in 95% of all sequenced bacteria (Wu et al. 2013), to estimate the number of bins and initialize the expectation–maximization process based on tetranucleotide frequencies and scaffold coverage levels (supplementary fig. S1, Supplementary Material online). For each bin, the presence of *Ca.* E. cretensis 16S rRNA gene sequences was checked as described above. For increased accuracy and resolution, completeness and copy number of bacterial genomes within each bin were estimated using the 107 single-copy marker genes (supplementary table S2, Supplementary Material online). The most complete, *Ca.* E. cretensis 16S rRNA gene-positive, bin was selected and host contamination was further removed based on MEGAN taxonomic assignments.

### Genome Draft Reconstruction, Annotation and Analysis

The previously reported draft (Katharios et al. 2015) was compared against the selected bin from the same sample and discrepancies were manually inspected. The final set of scaffolds was ordered against the genome of *Endozoicomonas elysicola* DSM 22380 (Neave et al. 2014) using ABACAS-1.3.1 (Assefa et al. 2009), and unordered scaffolds were appended. The refined draft was automatically annotated using PROKKA-1.10 (Seeman 2014), with annotation manually curated through comparison to the genome of *E. elysicola* DSM 22380 and checking of blastp hits, focusing on pseudogenes and ISs. To confirm the accuracy of parts of the assembly, reads were mapped using BWA v0.7.13 (Li and Durbin 2009).

Full length ISs (table 1) were generated by PCR using the primers and conditions in supplementary table S3, Supplementary Material online, and capillary sequencing, and assigned to families using ISfinder (Siguier et al. 2006; table 1). Promoter sequences within full length ISs were predicted using BPROM (https://omictools.com/bprom-tool; last accessed October 2017; Linear discriminant function (LDF)

**Table 1**
IS Element Families Identified within the Draft Genome of *Ca*. Endozoicomonas cretensis Sample Dpd28tailN

| IS Name | IS Family | # CDSs | Inverted Repeat Sequence | Duplicated Insertion Site | Approx. # in Dpd28tailN Genome | # Dpd28tailN Genes Putatively Disrupted | GenBank Accession Number of Full Length IS Elements | Length (bp) |
|---|---|---|---|---|---|---|---|---|
| ISEcret1 | IS1634 | 1 | CTGTCTTTCACCAC | 6 bp (5–7bp) | >65–80 | 12 | KP890196.1 | 1,731 |
| ISEcret2 | ND | 1 | CTCWGCTTTAGAGCWT | 7–11 bp | >60–75 | 20 | KP890197.1 | 1,535 |
| ISEcret3 | ISL3 | 1 | GGYTCTTTTKAA | 8 bp | >35–46 | 7 | KP890204.1 | 1,332 |
| ISEcret4 | IS1 | 2 | GGTGATGTRTCA | 8 bp | >64 (21 truncated) | 7 | KP890198.1 | 766 |
| ISEcret5 | IS630 | 1 | ATRCCAATYGCYTTTTC | 2 bp (TA) | >49 | 12 | KP890199.1 | 1,149 |
| ISEcret6 | ISNCY | 1 | CAGCRRTTCCCRCT | 9 bp | >22 | 5 | KP890200.1 | 1,603 |
| ISEcret7 | IS5 | 1 | GGAMCCTCTGAAAAA | 4 bp | >12–14 | 4 | KP890201.1 | 1,143 |
| ISEcret8 | IS481 | 1 | TVKAGWAGTTCAGAC | 7 bp | >66 | 6 | KP890202.1 | 1,206 |
| ISEcret9 | IS1 | 2 | GRTRRRRGTTCARA | 8 bp | >34 (9 truncated) | 3 | KP890203.1 | 791 |

NOTE.—All have been submitted to ISfinder under the given names. Accession numbers are provided.

value > 3.0) (Solovyev and Salamov 2011). IS associated terminal inverted repeats were determined through manual curation of ISs and comparison with feature within ISfinder. Sequence alignments were used to build IS-family specific hidden markov models (HMMs) using hmmbuild in hmmer-3.1b2, which were used by hmmsearch to search the refined draft.

Multiple metrics for bacterial species and genus classification were applied to the *Ca*. E. cretensis genome draft and related genomes: percentage of conserved proteins (POCP) (Qin et al. 2014), average nucleotide identity (ANI) (Goris et al. 2007), and digital DNA–DNA hybridisation (dDDH) (ggdc.dsmz.de/distalcalc2.php) (Auch et al. 2010). HMMs of 118 marker genes specific to *Gammaproteobacteria* (Wu et al. 2013) were downloaded (https://figshare.com/articles/Systematically_identify_phylogenetic_markers_at_different_taxonomic_levels_for_bacteria_and_archaea/722713; last accessed September 2017) and used by hmmsearch to scan the *Ca*. E. cretensis genome draft, the other *Endozoicomonas* genomes, and *Pseudomonas aeruginosa* PA01 genome (GCF_000006765.1). Protein sequences of single copy marker genes present in all 13 genomes were aligned using clustalw2 in clustalw-2.1 (Larkin et al. 2007). Concatenated protein sequence alignment was edited using Gblocks-0.91b to remove poorly aligned and divergent regions, before it was used in MEGA7 (Kumar et al. 2016) to generate a maximum-likelihood tree.

For CDSs within the *Ca*. E. cretensis genome draft, KEGG orthology (Kanehisa et al. 2010, 2012) and COG annotation (Galperin et al. 2015) were performed using KEGG Automatic Annotation Server (KAAS) (Moriya et al. 2007) and COGNITOR within the COGsoft.04.19.2012 package (Kristensen et al. 2010). Species-specific genes and paralogous genes were identified used Roary-f299b01 (95% protein sequence identity, MCL inflation value = 1.5) (Page et al. 2015). For estimating genetic redundancy, genes assigned with the same COG or KEGG IDs, as well as genes with paralogs, were counted as functionally redundant. Secreted, pathogenic, phage and antibiotic resistance encoding genes were predicted using EffectiveDB (Jehl et al. 2011), MP3 (Gupta et al. 2014), PHAST (http://phast.wishartlab.com/; last accessed November 2017) (Zhou et al. 2011), PhageFinder-v2.1 (Fouts 2006) and ResFinder (Zankari et al. 2012), respectively. To predict plasmid sequences in the genome draft, scaffolds longer than 1 Kb were analysed using PlasFlow-1.0.7 (probability threshold = 0.99) (Krawczyk et al. 2018).

The final *Ca*. E. cretensis genome draft (supplementary file S1, Supplementary Material online) has been deposited at ENA under the accession ERZ494307, and IS elements under KP890196-KP890204.

## Results and Discussion

### The *Ca*. E. cretensis Genome Draft

A refined *Ca*. E. cretensis genome draft from sample Dpd28tailN was produced (Katharios et al. 2015). DNA isolated from a microdissected tail section of an infected larva (Dpd28tailN) was sequenced and assembled (table 2 and supplementary table S1, Supplementary Material online). The presence of *Ca*. E. cretensis in the sample was confirmed by scaffolds matching the representative 16S rRNA gene sequence. All the 40 single-copy marker genes universal to all bacteria and archaea were identified in the assembly, giving an initial indication that this metagenome could harbour a complete *Ca*. E. cretensis genome. The assembly was classified using MaxBin, which binned sequences using an expectation–maximization algorithm based on both tetranucleotide frequencies and scaffold coverage (supplementary table S2 and fig. S1, Supplementary Material online). To increase accuracy, the binning was initiated with a broader set of bacterial marker genes, the 107 single copy marker genes that are conserved in 95% of all sequenced bacteria (Wu et al. 2013). One *Ca*. E. cretensis 16S rRNA gene positive bin was produced, where the set of the bacterial marker genes was almost complete (99.1%). The scaffolds from this bin had an

**Table 2**
Properties and Genome Features of Metagenome Assembly and Genome Draft

| Draft | Dpd28tailN Metagenome Assembly | Dpd28tailN Genome Draft |
|---|---|---|
| # Scaffolds | 62,776 (≥0 bp) 4734 (≥1000 bp) | 648 |
| Total scaffold length (bp) | 39,315,042 (≥0 bp) 12699876 (≥1000 bp) | 5,898,394 |
| Largest scaffold (bp) | 91,550 | 91,550 |
| Scaffold N50 | 1,085 | 19,571 |
| % G+C | 46.69 | 46.85 |
| Completeness, 40 bacterial and archaeal markers | 100% | 100% |
| Completeness, 107 bacterial markers | — | 99.10% |
| Completeness, 118 gammaproteobacterial markers | — | 100% |
| Diversity | 3 | 1 |
| # Predicted genes | — | 5,858 |
| # KEGG annotated genes (%) | — | 2,447 (42) |
| # COG annotated genes (%) | — | 4,620 (79) |
| Coding density | — | 78.90% |
| Average gene length | — | 849 |
| rRNA operons | — | 7 |
| tRNAs | — | 77 |
| Pseudogenes | — | 477 |
| Transposases (incl. pseudogenes and partial) | — | 783 |
| ENA accession | Reads: ERR662023 | Analysis: ERZ494307 |

average coverage of 104×, whereas the other three bins had under 5× (supplementary table S2 and fig. S1, Supplementary Material online). This bin was selected for further study, and eukaryotic scaffolds identified through blastn ($n = 11$) were removed (the genome sequence of the sharpsnout seabream host is not yet available for automated removal of matching sequences).

This bin was manually curated through comparison against the original genome draft, and the genome of *E. elysicola* DSM 22380 (Neave et al. 2014). The genome draft was ordered against this reference where possible, and nonmatching scaffolds were appended. The final draft comprises 5.9 Mb over 638 scaffolds, (fig. 1). Genome features are given in table 2. This draft is estimated to be >99% complete, using three different gene sets, including the 118 markers genes specific to *Gammaproteobacteria* (Wu et al. 2013) (table 2). Pairwise ANI, POCP, and dDDH comparisons with other *Endozoicomonas* species confirm that *Ca.* E. cretensis represents a new *Endozoicomonas* species (table 3).

## Phylogeny and Genomic Features

To study the phylogenetic context of *Ca.* E. cretensis, we constructed a phylogenetic tree using protein sequences of *Gammaproteobacteria* markers genes found in all *Endozoicomonas* genomes in the databases (fig. 2). Although *Endozoicomonas* bacteria are found in diverse marine hosts, it is not clear if they and their hosts speciate in parallel. Some correlation between *Endozoicomonas* species phylogeny and host phylogeny can be seen (fig. 2), although they do not completely reflect each other (Neave et al. 2017), and further resolution of this relationship will come from further isolation of species. *E. elysicola* DSM22380, isolated from sea slug, appears as the most closely related species to *Ca.* E. cretensis; although fish is phylogenetically closer to sea squirt (host of *Endozoicomonas atrinae* WP70) than sea slug. The isolate S-B4-1U has been recently reclassified from *Endozoicomonas* to *Parendozoicomonas haliconae* (in press, personal communication), justifying its more distant phylogenetic position in the tree.

A major characteristic of the genome of *E. cretensis* Dpd28tailN is the presence of 773 (13.2% of CDSs) identified transposases, complete or partial. This compares against 145 (3.5%) identified in the genome of *E. montiporae* (Ding et al. 2016), and two annotated in *E. elysicola*. The insertion sequences which carry these transposases are largely responsible for the fragmented assembly, with the vast majority of scaffolds carrying an IS associated inverted repeat (ISIR) ($n = 442$, 69% of scaffolds with ≥1 ISIR). A search for putative plasmid-related sequences identified a set of 22 contigs; however, closer analysis of the encoded CDSs did not provide convincing evidence of the presence of a plasmid, with many carrying CDSs encoding putative phage proteins, hypothetical proteins or transposases. No *Endozoicomonas* genome to date has been reported to carry a plasmid (Neave et al. 2014; Appolinario et al. 2016; Ding et al. 2016; Schreiber et al. 2016b).

## *Ca.* E. cretensis Species-Specific and Virulence Related CDSs

*Ca.* Endozoicomonas cretensis Dpd28tailN genome draft is the first from an *Endozoicomonas* species associated with a

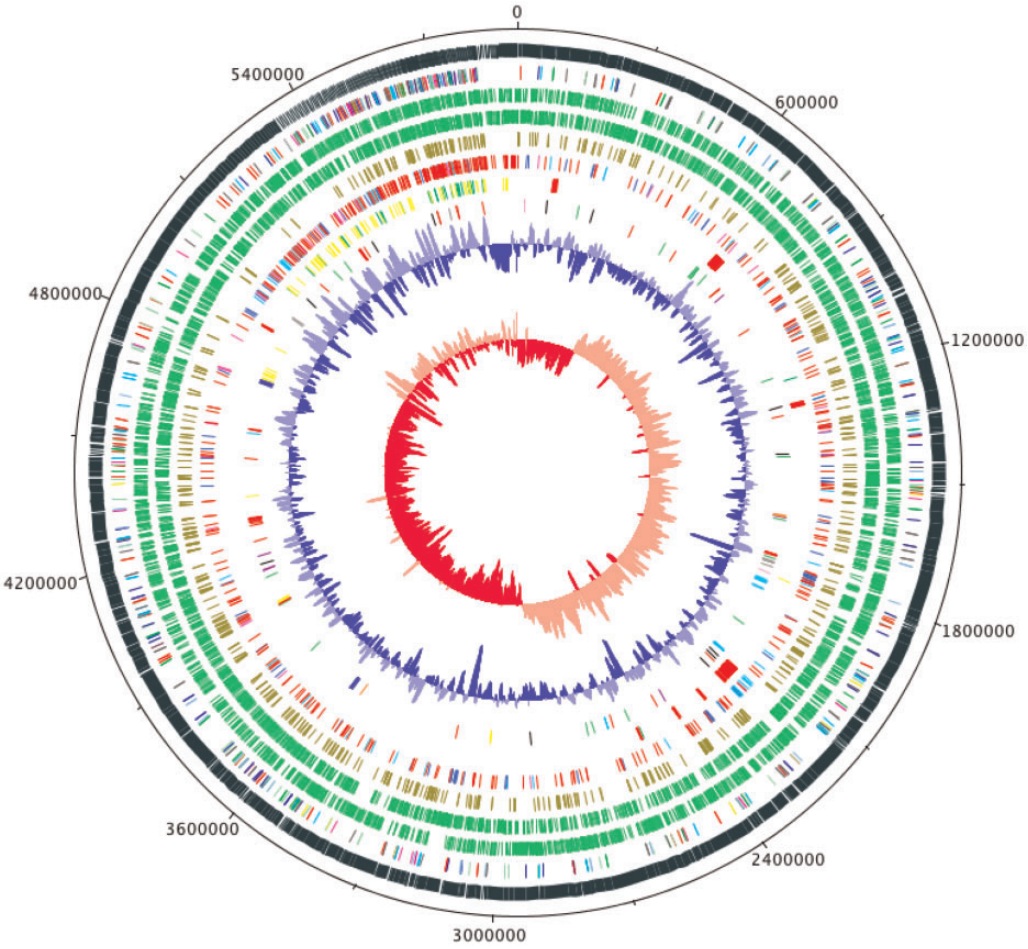


**Fig. 1.**—Circular representation of the genome of *Ca.* Endozoicomonas cretensis. Scaffolds were ordered against the genome of *E. elysicola* DSM 22380 (Neave et al. 2014). Scaffolds not aligned to *Endozoicomonas elysicola* were appended to the ordered scaffolds after 495520 bp. The tracks from the outside in represent: (1) the scaffolds ($n = 638$); (2) ISIR located at the ends of the scaffolds, colored by IS families; (3) forward CDSs; (4) reverse CDSs; (5) pseudogenes; (6) species-specific genes in enriched COG categories: replication, recombination and repair (cyan), cell wall/membrane/envelope biogenesis (blue), cell motility (magenta) and Mobilome: prophages, transposons (red); (7) phages (red), unordered phage genes (yellow) and newly expanded families of pathogenic genes (green); (8) virulence factors including T3SS (cyan), flagella (purple), chemotaxis (green), Tfp (red), T2SS (blue), mucin degradation genes (yellow), invasion biofilm formation genes (grey), invasin (orange), and effectors nucleomodulin (black) and E3 ubiquitin ligases (pink).

vertebrate host, and with pathogenicity. This allows us to interpret the genome in the context of adaption to, and virulence towards, a vertebrate host. Comparison of the Dpd28tailN genome draft with available *Endozoicomonas* genomes finds 2,462 species-specific CDSs (supplementary table S4, Supplementary Material online), which are either absent from, or divergent from (<95% amino acid similarity) CDSs within the genomes of the other species. To identify genes in particular classes and with functions that are over-represented in this set of species-specific genes, and may have an association with the disease phenotype, we annotated all 5,858 manually curated CDSs with multiple systems: KEGG, COG, MP3, and EffectiveDB.

Secretion of proteins across phospholipid membranes, bacterial motility and chemotaxis systems are essential strategies for many bacteria, often associated with virulence. Components of a Type Two Secretion System (T2SS), Type 4 pili (Tfp), Type Three Secretion System (T3SS), and flagella were identified within the genome (fig. 1) through KEGG and COG analysis. Additionally, 760 T3SS effectors, 106 T3SS effector chaperones, and 2,807 proteins with eukaryotic-like domains (ELD) were predicted by EffectiveDB, of which 991, 402, 49, and 776, respectively, are specific to *Ca.* E. cretensis (supplementary table S4, Supplementary Material online). Previous electron microscopy images indicate the presence of pili or flagella within *Ca.* E. cretensis epitheliocysts (Katharios et al. 2015). No Type IV secretion system (T4SS) components or effectors were identified.

The 2,462 species-specific CDSs occur in almost all major COG functional categories (fig. 3), with the "Mobilome" category being most highly enriched, followed by "Replication, recombination and repair," "Cell wall/membrane/envelope

**Table 3**
ANI, POCP, and dDDH Analysis of *Ca.* Endozoicomonas cretensis against Other *Endozoicomonas* Species

| Comparator Species | Strain | Accession Number | Genome Size (Mbp) | *Ca.* E. cretensis Dpd28tailN | | | |
|---|---|---|---|---|---|---|---|
| | | | | ANI | POCP | dDDH | %G + C Difference |
| *Endozoicomonas elysicola* | DSM 22380 | GCF_000710775.1 | 5.61 | 34.83 | 69.09 | 51.6 | 0.09 |
| *Endozoicomonas atrinae* | WP70 | GCF_001647025.1 | 6.69 | 7.48 | 51.58 | 31.6 | 1.09 |
| *Endozoicomonas montiporae* | CL-33(T) | GCF_000722565.1 | 5.43 | 0.08 | 51.32 | 24.8 | 1.62 |
| *Endozoicomonas arenosclerae* | E_MC227 | GCA_001562005.1 | 6.22 | 0.04 | 43.36 | 24.4 | 0.31 |
| *Endozoicomonas numazuensis* | DSM 25634 | GCF_000722635.1 | 6.34 | 0.05 | 50.19 | 24.1 | 0.17 |
| *Endozoicomonas montiporae* | LMG 24815 | GCF_000722565.1 | 5.6 | 0.08 | 51.32 | 23.6 | 1.62 |
| *Endozoicomonas arenosclerae* | Ab112 | GCF_001562015.1 | 6.45 | 0.09 | 49.78 | 23.3 | 0.81 |
| *Endozoicomonas* sp. | S-B4-1U | GCF_900174585.1 | 5.467 | 0.02 | 42.35 | 22.6 | 4.65 |
| *Endozoicomonas ascidiicola* | AVMART05 | GCF_001646945.1 | 6.13 | 0.14 | 59.62 | 22.3 | 0.14 |
| *Endozoicomonas ascidiicola* | KASP37 | GCF_001646955.1 | 6.51 | 0.1 | 61.54 | 22.2 | 0.2 |
| *Endozoicomonas* sp. | AB1-5 | GCA_001729985.1 | 4.049 | 0.01 | 51.98 | 20.2 | 1.57 |

NOTE.—Comparing *Ca.* E. cretensis Dpd28tailN genome draft against other published drafts. For POCP 69% is proposed as species cutoff (Goris et al. 2007) and 50% as genus cutoff (Qin et al. 2014). For ANI analysis, the species cutoff is 95%, and for dDDH (formula 2 used) 70% (Auch et al. 2010).

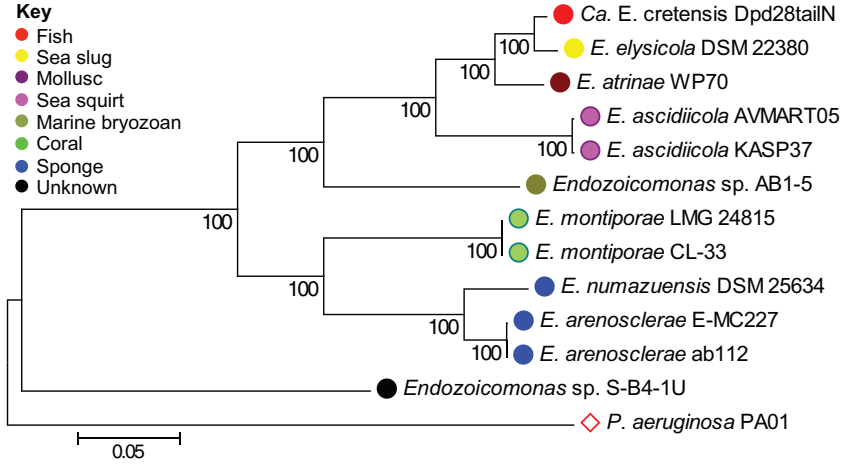


**FIG. 2.**—Phylogenetic relationship of *Ca.* Endozoicomonas cretensis to other *Endozoicomonas* species. Maximum-likelihood tree based on concatenated aligned protein sequences of 43 conserved single-copy marker genes, extracted from the *Ca.* E. cretensis genome draft, and 11 publically available *Endozoicomonas* genomes (table 4). The tree was rooted using the *Gammaproteobacterium Pseudomonas aeruginosa* PA01 (GCF_000006765.1). In total 6,120 sites were used, which were extracted from the 12,791 sites in the original protein alignment by Gblocks after eliminating poorly aligned and divergent regions. The scale bar indicates the number of substitutions per site.

biogenesis," and "Cell motility," The majority of the "Mobilome" category are transposase CDSs, whereas 65 (16%) are putative phage CDSs. All putative phages span multiple scaffolds, and we estimate that up to ten are present in the genome (supplementary table S5, Supplementary Material online), suggesting that phages are a source of species-specific CDSs.

Of the species-specific CDSs belonging to COG category "Replication, recombination and repair," 47 are predicted to have ELD. Proteins containing ELD domains have been proposed to act as bacterial effectors with biological roles in host cells, because they are found to have a higher frequency in genomes of host-associated bacteria compared with non host-associated bacteria (Jehl et al. 2011). These CDSs are predicted to coordinate chromatin modification (*Ecret_2685*), replication (*Ecret_2665, Ecret_1488, Ecret_5076*), recombination (*Ecret_7122, Ecret_1716, Ecret_0306, Ecret_6327, Ecret_6729, Ecret_2726*) and repair (*Ecret_3562, Ecret_2665, Ecret_0113, Ecret_2665, Ecret_1488, Ecret_5076*). This raises the possibility that *Ca.* E. cretensis can deliver effectors into the host cell nucleus, and subvert host defences by directly interfering with transcription, DNA replication and repair through chromatin-remodeling. This ability is implicated in the virulence of several intracellular bacterial pathogens (Bierne et al. 2009; Pennini et al. 2010; Bierne and Cossart 2012). Some of the *Ca.* E. cretensis nucleomodulin genes are associated with phages, suggesting horizontal gene transfer as an important

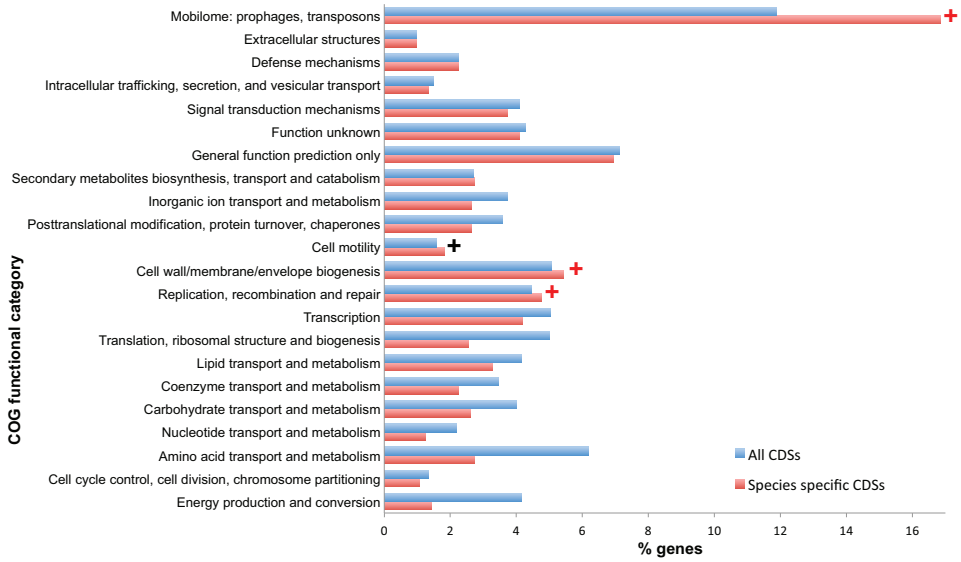


**Fig. 3.**—Functional categories enriched with *Ca.* Endozoicomonas cretensis Dpd28tailN specific genes. COG (Clusters of Orthologous Groups) functional categories where the numbers of species-specific genes (pink horizontal bars) are more (+) than expected by Fisher's exactly test (red: $P$ value $< 0.01$; black: $0.01 < P$ value $< 0.05$) are marked. The numbers of "all genes" in each category are shown as the neighboring blue horizontal bars. For visualization purpose, only categories with more than two genes are shown.

mechanism of virulence evolution in *Ca.* E. cretensis (supplementary table 5 and fig. S1, Supplementary Material online).

Among the species-specific CDSs within the enriched "Cell wall/membrane/envelope biogenesis" COG category, we found one CDS predicted to encode an E3 ubiquitin ligase (*Ecret_0032*) and one (*Ecret_2214*) to encode the adaptor and target recognizing subunit of E3 ubiquitin ligases. Bacterial E3 ubiquitin ligases are T3SS effectors involved in impairing host inflammatory responses and host detection of the pathogen, with demonstrated virulence effects in *Salmonella*, *Shigella*, and *Yersinia* (Maculins et al. 2016).

Species-specific CDSs within the enriched "Cell motility" COG category include CDSs involved in bacterial chemotaxis and Tfp assembly, which could together lead to directed cell movement. More importantly, it has been demonstrated that Tfp are used by the pathogen *P. aeruginosa* to sense initial contact with surfaces, which in turn regulates the transcription of hundreds of genes associated with pathogenicity and surface-specific twitching motility (Persat et al. 2015).

To investigate systems which may contribute to virulence, KEGG pathways were analyzed in further detail. It has been demonstrated that bacteria can have a chemotactic response to specific fish mucins (O'Toole et al. 1999). Pathogenic strains, such as *Ca.* E. cretensis, need to be able to degrade the mucous layer of the host to access epithelial cells (McGuckin et al. 2011), and may use mucins as a carbon source (Schreiber et al. 2016a; Ottman et al. 2017). *Ecret_5397* is predicted to encode a secreted beta-N-acetylglucosaminidase. This putatively pathogenic enzyme, also present in *E. montiporae* (Ding et al. 2016), can hydrolyze the glycosidic bond in N-linked sugar chains in glycoproteins (Yin et al. 2009) and has been suggested to play a role in the dissociation of mucin, allowing microorganisms to penetrate the coral mucous layer. In addition, glycosulphatases, sialidases, sialate O-acetyl esterase, metalloproteinase, $\alpha$-glycosidases, $\beta$-glycosidases, $\alpha$-2, 3/2, 8-N-acetylneuraminidase, and mucin-depolymerizing enzymes are required (McGuckin et al. 2011), as sialic acids and sulfated polar groups of the mucins can inhibit the digestion (Macfarlane et al. 2005). In the genome of *Ca.* E. cretensis, many of these mucin-degradation associated CDSs are organized in clusters, whereas others appear to be scattered throughout the genome (fig. 1). One cluster comprises: two versions of *betC* encoding a choline-sulfatase (*Ecret_5999 and Ecret_6016)*; two intact (*Ecret_6000, Ecret_6027*) and one disrupted (*Ecret_6009*) CDSs encoding chondroitin sulfate lyase, which is a major adhesion-related virulence factor in *Flavobacterium psychorphilum* (Suomalainen et al. 2006); *Ecret_6008* encoding a putative exported protein; *Ecret_6015* encoding an arylsulfatase; and *Ecret_6033* encoding hyaluronate lyase precursor. Two CDSs in the *Ca.* E. cretensis genome are predicted to encode sialidases and predicted as pathogenic: *Ecret_3702*, which is unique in *Ca.* E. cretensis, and *Ecret_4889*. Neuraminidases or sialidases are particularly important because, apart from cleaving sialic acids, their action has been connected with the adhesion mediated by Tfp (Soong et al. 2006). Both have neighbouring CDSs encoding branched chain amino acid transporters (*Ecret_3710/3711* and *Ecret_4891/4879*), suggesting that the extracellular digestion of mucin (secreted enzyme mediated) is linked to the transport of the degraded oligosaccharides. Finally, an excreted

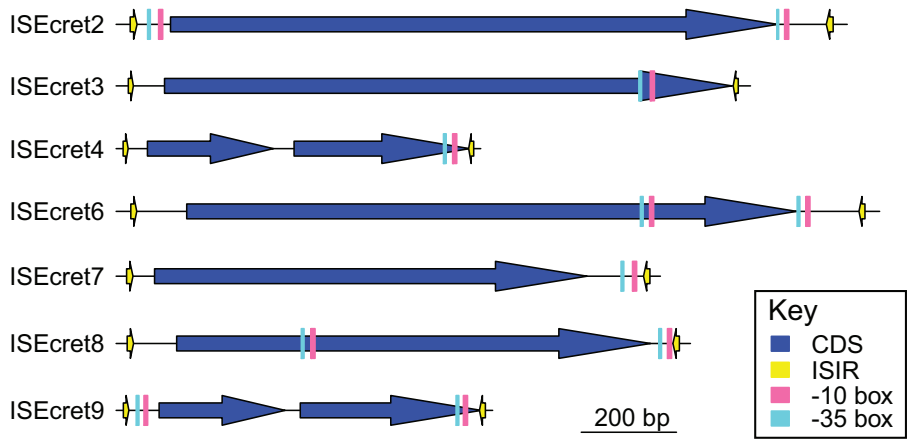


**Fig. 4.**—Location of outward-directed promoters predicted within *Ca.* Endozoicomonas cretensis IS elements. "-10 box" and "-35 box" represent the two short conserved sequence elements in the bacterial promoters, which are, respectively, approximately 10 and 35 nucleotides upstream of the transcription start site (TSS). Each CDS is shown as a blue arrow, with the flanking yellow arrows representing the ISIRs.

glucosidase (*Ecret_6105*) may be further related to the process of mucin degradation.

Antibiotics are rarely used in the facility where the samples were collected, meaning that it is unlikely that *Ca.* E. cretensis has been exposed to antimicrobials in the location that it was found. We did not identify any specific antimicrobial-resistant CDSs, using Resfinder, although KEGG analysis revealed several efflux pumps (*Ecret_1073, Ecret_1695, Ecret_1696, Ecret_1939, Ecret_2069, Ecret_2555, Ecret_3672, Ecret_5349, Ecret_5351*).

KEGG pathway analysis also helped to identify other putative virulence related genes, such as those encoding adhesin/invasin (*Ecret_1991* and *Ecret_4299*), and involved in biofilm formation (*Ecret_2152, Ecret_6208, Ecret_2155, Ecret_2156, Ecret_2170, Ecret_7277, Ecret_6213*).

## Expansion of IS Elements, Genome Degradation, and Genetic Redundancy

Nine novel IS elements (ISEcret1-9) have been identified in the genome, and representative full length ISs were determined (table 1), ranging from 700 bp to 1,800 bp. Analysis of ISIRs estimate that several of the families have copy numbers over 60, giving a genomic total in excess of 400 IS elements (table 1). As these elements are in such a high copy number, and are longer than Illumina reads and library fragments, the IS elements remain unassembled in the genome draft and are responsible for the high number of scaffolds. The IS elements within this *Ca.* E. cretensis genome draft are different from those found disrupting the genome of *E. montiporae* (Ding et al. 2016), indicating an independent and parallel process occurring in the two species.

Several families of IS elements are reported to carry promoters which regulate expression of CDSs adjacent to the insertion site (Tolmasky and Crosa 1995; Han et al. 2011). We found at least one complete outward-directed promoter in seven of the full length ISEcrets, in most cases downstream of the transposase CDS (fig. 4). ISEcret2 and ISEcret9 are predicted to harbour two complete promoters, one at each end. We analysed all CDSs downstream of the IS promoters (supplementary table S6, Supplementary Material online), identifying 58 CDSs whose regulation could potentially be altered; of these *Ca.* E. cretensis specific ($n = 39$) and putatively pathogenic CDSs ($n = 31$) make up the majority. Promoters associated with ISEcret2 are associated with the highest number of candidate targets ($n = 20$). Without gene expression data, it is not possible to predict the effect of these insertions: functional studies would be needed to confirm any IS-mediated gene upregulation in *Ca.* E. cretensis.

The *Ca.* E. cretensis draft genome also shows the disruptive effect of IS elements on a genome, where a gene is truncated by IS element insertion. Using ISEcret1 as an example, insertional disruption of the following CDSs is apparent: *Ecret_0863, Ecret_1768, Ecret_1982, Ecret_3674, Ecret_4929, Ecret_5292, Ecret_5679*. All the identified IS elements are responsible for several disruptions of CDSs (table 1). Approximately 200 CDSs have been insertionally disrupted; this compares to 56 in *E. montiporae* (Ding et al. 2016).

In total, 475 CDSs (8.1%) have been identified as being disrupted, through IS element insertions, frameshifts or mutations to create a premature stop codon. Many of the pseudogenes are actually predicted to encode transposases ($n = 210$; 26.8% of all transposase CDSs). Almost all (385, >80%) of the pseudogenes have been created from redundant genes, where a CDS with equivalent function exists in the genome, making one copy nonessential, and therefore not subject to selective pressure for maintaining function. For example, CDSs encoding lipid synthesis proteins are significantly enriched with pseudogenes, but detailed analysis of the fatty acid biosynthesis pathway reveals that the pathway is still intact.

**Table 4**
Newly Expanded Family of Pathogenic Genes in *Ca.* Endozoicomonas cretensis Dpd28tailN Genome and the Predicted Properties of Family Members

| Locus_tag | ID | Predicted Product | Pseudogene | Virulence Related (as predicted by MP3) | Type III Secreted Proteins (predicted by EffectiveT3) | Species-Specific | Novel Gene Family ID |
|---|---|---|---|---|---|---|---|
| *Ecret_1399* | Dp_catedit6.1569 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_11 |
| *Ecret_6420* | Dp_catedit6.7334 | Hypothetical protein | No | Yes | No | Yes | group_11 |
| *Ecret_6838* | Dp_catedit6.7799 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_11 |
| *Ecret_7145* | Dp_catedit6.8177 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_11 |
| *Ecret_7201* | Dp_catedit6.8244 | Putative exported protein | No | Yes | No | Yes | group_11 |
| *Ecret_7267* | Dp_catedit6.8330 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_11 |
| *Ecret_7540* | Dp_catedit6.8679 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_11 |
| *Ecret_6703* | Dp_catedit6.7645 | Conserved hypothetical protein (partial) | No | Yes | Yes | No | group_11 |
| *Ecret_7156* | Dp_catedit6.8191 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_19 |
| *Ecret_7409* | Dp_catedit6.8509 | Hypothetical protein (partial) | No | No | No | No | group_19 |
| *Ecret_7514* | Dp_catedit6.8649 | Conserved hypothetical protein (partial) | No | No | No | No | group_19 |
| *Ecret_7369* | Dp_catedit6.8458 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_41 |
| *Ecret_7467* | Dp_catedit6.8587 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_41 |
| *Ecret_7552* | Dp_catedit6.8693 | Conserved hypothetical protein | No | Yes | No | Yes | group_5667 |
| *Ecret_3116* | Dp_catedit6.3531 | Protein of unknown function (DUF2523) | No | Yes | No | No | group_5667 |
| *Ecret_7436* | Dp_catedit6.8545 | Conserved hypothetical protein (partial) | No | Yes | Yes | Yes | group_5668 |
| *Ecret_7200* | Dp_catedit6.8242 | Conserved hypothetical protein (partial) | No | Yes | Yes | Yes | group_5668 |
| *Ecret_6312* | Dp_catedit6.7219 | Bacteriophage replication gene A protein (GPA) | No | No | No | Yes | group_69 |
| *Ecret_7264* | Dp_catedit6.8324 | Phage replication protein A (partial) | No | N0 | No | Yes | group_69 |
| *Ecret_7618* | Dp_catedit6.8776 | Conserved hypothetical protein (partial) | No | Yes | Yes | Yes | group_7 |
| *Ecret_6493* | Dp_catedit6.7413 | Hypothetical protein | No | Yes | Yes | Yes | group_7 |
| *Ecret_6829* | Dp_catedit6.7789 | Conserved hypothetical protein (pseudogene) | Yes | Yes | Yes | Yes | group_7 |
| *Ecret_6626* | Dp_catedit6.7560 | Conserved hypothetical protein | No | Yes | No | Yes | group_7 |
| *Ecret_6814* | Dp_catedit6.7772 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_7 |
| *Ecret_6966* | Dp_catedit6.7953 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_7 |
| *Ecret_7287* | Dp_catedit6.8358 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_7 |
| *Ecret_7427* | Dp_catedit6.8533 | Conserved hypothetical protein (partial) | No | Yes | No | Yes | group_7 |
| *Ecret_6492* | Dp_catedit6.7412 | Conserved hypothetical protein (partial) | No | No | No | Yes | group_7 |
| *Ecret_6621* | Dp_catedit6.7554 | Conserved hypothetical protein (partial) | No | Yes | No | No | group_7 |
| *Ecret_6625* | Dp_catedit6.7559 | Conserved hypothetical protein | No | Yes | No | No | group_7 |
| *Ecret_7508* | Dp_catedit6.8642 | Hypothetical protein | No | Yes | No | Yes | group_70 |
| *Ecret_7506* | Dp_catedit6.8640 | Hypothetical protein | No | No | No | Yes | group_70 |

Among the 90 pseudogenes without a functional counterpart, most are hypothetical proteins thus the functional impacts are unknown. A few are predicted (when intact) to encode key enzymes in multiple metabolism pathways. For example: *arcC* (*Ecret_6060*) is used in arginine biosynthesis, and the metabolism of purine, nitrogen, and carbon; *ushA* (*Ecret_0258, Ecret_5722*) is involved in metabolism of purine, pyrimidine, nicotinate and nicotinamide; *mhpD* (*Ecret_2614*) is important for degradation of benzoate, dioxin, xylene, and aromatic compounds; *pldA* (*Ecret_2372*) is important for the metabolism of glycerophospholipid, ether lipid, arachidonic acid, linoleic acid, and alpha-linolenic acid. That *Ca.* E. cretensis appears not to be culturable on marine agar, in contrast to other *Endozoicomonas* species, might be the result of functional gene loss due to the creation of pseudogenes. This may also imply that this species is no longer capable of free-living, and may be more reliant on a host.

Genetic redundancy appears to be a feature of *Endozoicomonas* genomes. KEGG analysis identified 313 gene families comprising 867 CDSs, whereas COG analysis found 739 gene families with 3,202 CDSs. Over 75% of the gene families were also present in the *E. elysicola* DSM22380 genome, with comparable family sizes (supplementary table S7, Supplementary Material online), representing more ancestral duplications. The remaining gene families were clustered de novo, identifying eight novel families within the genome of

*Ca.* E. cretensis, comprising 32 CDSs, where family members shared significant sequence similarities (amino acid similarities > 95%; table 4). Of these 32 CDSs, 26 are predicted to be species-specific, and of these, 22 virulence related, suggesting that these families have expanded within *Ca.* E. cretensis and may contribute to the pathogenicity. Due to the origin of the genome draft, it is possible that some of this redundancy is technical rather than biological (Neave et al. 2017), but we estimate this to be low and conclude that the observed redundancy is largely biological. Genetic redundancy has been recognized as a strategy that facilitates adaptation in bacterial populations (Stover et al. 2000; Toll-Riera et al. 2016), allowing them to evolve new metabolic functions without compromising existing functions, thus potentiating innovation by minimizing the associated cost.

## Conclusions

Here, we present a thorough study on the first genome draft of a pathogenic *Endozoicomonas* species, *Ca.* E. cretensis. A reference genome draft has been constructed from infected host material, in the absence of a cultured strain, and reveals potential mechanisms for bacterial adaptation, such as IS mediated gene regulation, gene disruption and the presence of species-specific virulence related CDSs. An interesting attribute of *Ca.* E. cretensis is its arsenal of mucin-degrading enzymes. CDSs implicated in virulence include nucleomodulins that manipulate the host nucleus, and E3 ubiquitin ligases, which impair host inflammatory responses. The T3SS, Tfp, flagella, bacterial motility and chemotaxis systems found in the genome draft can also play important roles in promoting virulence, from enhancing attachment to host cells, to directly intoxicating them thus disrupting their functions.

In a genome draft comprising so many scaffolds, it is challenging to define the full evolutionary path and pathogenic potential. Successfully culturing this bacterium would be of great value, providing sufficient high molecular weight DNA without contamination, thus allowing chromosome-level assembly through long read sequencing technologies. Clonal cultures would also permit further phenotypic, genetic and functional studies, such as the confirmation of IS mediated gene regulation suggested by the genome draft. The evolution of *Endozoicomonas* species and emergence of their diverse lifestyles are fascinating, and deserve further in depth investigation. Different *Endozoicomonas* species associated with the same host, or collected from a broader set of related hosts would aid studies into the relationship between *Endozoicomonas* and host evolution.

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

This work was supported by the European Union through Marie Curie Intra-European Fellowship grant number 332058 to H.M.B.S.S. and an FP7 Aquaexcel-TNA project 01-05-15-0004-B to L.V. and P.K.

## Literature Cited

Altschul SF, Gish W, Miller W, Myers E, Lipman D, 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Appolinario L, et al. 2016. Description of *Endozoicomonas arenosclerae* sp. nov. using a genomic taxonomy approach. Antonie Van Leeuwenhoek 109(3):431–438.

Assefa S, Keane T, Otto T, Newbold C, Berriman M, 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25(15):1968–1969.

Auch A, Von Jan M, Klenk H-P, Göker M, 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci. 2(1):117–134. doi: 10.4056/sigs.531120

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bierne H, Cossart P, 2012. When bacteria target the nucleus: the emerging family of nucleomodulins. Cell Microbiol. 14(5):622–633. doi: 10.1111/j.1462-5822.2012.01758.x

Bierne H, et al. 2009. Human BAHD1 promotes heterochromatic gene silencing. Proc Natl Acad Sci U S A. 106(33):13826–13831. doi: 10.1073/pnas.0901259106

Bolger AM, Lohse M, Usadel B, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. doi: 10.1093/bioinformatics/btu170

Bourne DG, Morrow KM, Webster NS, 2016. Insights into the coral microbiome: underpinning the health and resilience of reef ecosystems. Ann Rev Microbiol. 70:317–340.

Cole ST, et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393(6685):537–544. doi: 10.1038/31159

Ding JY, Shiu JH, Chen WM, Chiang YR, Tang SL, 2016. Genomic insight into the host-endosymbiont relationship of *Endozoicomonas montiporae* CL-33(T) with its coral host. Front Microbiol. 7: 251. doi: 10.3389/fmicb.2016.00251

Ekblom R, Wolf JB, 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 7(9):1026–1042. doi: 10.1111/eva.12178

Fiore C, Labrie M, Jarett J, Lesser M, 2015. Transcriptional activity of the giant barrel sponge, *Xestospongia muta* holobiont: molecular evidence for metabolic interchange. Front Microbiol. 6:364.

Forget N, Juniper K, 2013. Free-living bacterial communities associated with tubeworm (*Ridgeia piscesae*) aggregations in contrasting diffuse flow hydrothermal vent habitats at the Main Endeavour Field, Juan de Fuca Ridge. MicrobiologyOpen 2(2):259–275.

Fouts D, 2006. *Phage_Finder*: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. 34(20):5839–5851.

Galperin M, Makarova K, Wolf Y, Koonin E, 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 43(Database issue):D261–D269.

Goris J, et al. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 57(1):81–91. doi: 10.1099/ijs.0.64483-0

Gupta A, Kapil R, Dhakan D, Sharma V, 2014. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. PLoS One 9(4):e93907.

Han HJ, Kuwae A, Abe A, Arakawa Y, Kamachi K, 2011. Differential expression of type III effector BteA protein due to IS481 insertion in *Bordetella pertussis*. PLoS One 6(3):e17797. doi: 10.1371/journal.pone.0017797

Holden MT, et al. 2009. The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol 191(1):261–277. doi: 10.1128/JB.01230-08

Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC, 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 21(9):1552–1560. doi: 10.1101/gr.120618.111

Jehl M, Arnold R, Rattei T, 2011. Effective: a database of predicted secreted bacterial proteins. Nucleic Acids Res. 39(Database issue):D591–D595.

Jensen S, Duperron S, Birkeland N-K, Hovland M, 2010. Intracellular Oceanospirillales bacteria inhabit gills of *Acesta* bivalves. FEMS Microbiol Ecol. 74(3):523–533.

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M, 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38(Database issue):D355–D360.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M, 2012. KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40(Database issue):D109–D114.

Katharios P, et al. 2015. Environmental marine pathogen isolation using mesocosm culture of sharpsnout seabream: striking genomic and morphological features of novel *Endozoicomonas* sp. Sci Rep. 5(1):17609. doi: 10.1038/srep17609

Krawczyk PS, Lipinski L, Dziembowski A, 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res. 46(6):e35. doi: 10.1093/nar/gkx1321

Kristensen DM, et al. 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics 26(12):1481–1487. doi: 10.1093/bioinformatics/btq229

Kumar S, Stecher G, Tamura K, 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 33(7):1870–1874.

Langmead B, Salzberg SL, 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359. doi: 10.1038/nmeth.1923

Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948. doi: 10.1093/bioinformatics/btm404

Li H, Durbin R, 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.

Macfarlane S, Woodmansey EJ, Macfarlane GT, 2005. Colonization of mucin by human intestinal bacteria and establishment of biofilm communities in a two-stage continuous culture system. Appl Environ Microbiol. 71(11):7483–7492.

Maculins T, Fiskin E, Bhogaraju S, Dikic I, 2016. Bacteria-host relationship: ubiquitin ligases as weapons of invasion. Cell Res. 26(4):499–510.

McGuckin MA, Lindén SK, Sutton P, Florin TH, 2011. Mucin dynamics and enteric pathogens. Nat Rev Microbiol. 9(4):265–278. doi: 10.1038/nrmicro2538

Mendoza M, et al. 2013. A novel agent (*Endozoicomonas elysicola*) responsible for epitheliocystis in cobia *Rachycentrum canadum* larvae. Dis Aquat Org. 106(1):31–37. doi: 10.3354/dao02636

Miller IJ, Weyna TR, Fong SS, Lim-Fong GE, Kwan JC, 2016. Single sample resolution of rare microbial dark matter in a marine invertebrate metagenome. Sci Rep. 6(1):34362. doi: 10.1038/srep34362

Mistry J, Finn R, Eddy SR, Bateman A, Punta M, 2013. Challenges in homology search: hMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41(12):e121.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M, 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35(Web Server):W182–W185. doi: 10.1093/nar/gkm321

Morrow KM, Moss AG, Chadwick NE, Liles MR, 2012. Bacterial associates of two caribbean coral species reveal species-specific distribution and geographic variability. Appl Environ Microbiol 78(18):6438–6449. doi: 10.1128/AEM.01162-12

Neave MJ, Apprill A, Ferrier-Pagès C, Voolstra CR, 2016. Diversity and function of prevalent symbiotic marine bacteria in the genus *Endozoicomonas*. Appl Microbiol Biotechnol. 100(19):8315–8324. doi: 10.1007/s00253-016-7777-0

Neave MJ, Michell CT, Apprill A, Voolstra CR, 2017. *Endozoicomonas* genomes reveal functional adaptation and plasticity in bacterial strains symbiotically associated with diverse marine hosts. Sci Rep. 7:40579. doi: 10.1038/srep40579

Neave MJ, Michell CT, Apprill A, Voolstra CR, 2014. Whole-genome sequences of three symbiotic *Endozoicomonas* strains. Genome Announc. 2(4):e00802-14–e00814. doi: 10.1128/genomeA.00802-14

Nurk S, et al. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol. 20(10):714–737. doi: 10.1089/cmb.2013.0084

O'Toole R, et al. 1999. The chemotactic response of Vibrio anguillarum to fish intestinal mucus is mediated by a combination of multiple mucus components. J Bacteriol. 181:4308–4317.

Ottman N, et al. 2017. Genome-scale model and omics analysis of metabolic capacities of *Akkermansia muciniphila* reveal a preferential mucin-degrading lifestyle. Appl Environ Microbiol. 83:e01014–e01017.

Page AJ, et al. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22):3691–3693.

Parkhill J, et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet .35(1):32–40. doi: 10.1038/ng1227

Pennini ME, Perrinet S, Dautry-Varsat A, Subtil A, 2010. Histone methylation by NUE, a novel nuclear effector of the intracellular pathogen *Chlamydia trachomatis*. PLoS Pathog. 6(7):e1000995. doi: 10.1371/journal.ppat.1000995

Persat A, Inclan YF, Engel JN, Stone HA, Gitai Z, 2015. Type IV pili mechanochemically regulate virulence factors in *Pseudomonas aeruginosa*. Proc Natl Acad Sci U S A. 112(24):7563–7568.

Qin Q-L, et al. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. J Bacteriol. 196(12):2210–2215.

Rahman A, Pachter L, 2013. CGAL: computing genome assembly likelihoods. Genome Biol. 14(1):R8. doi: 10.1186/gb-2013-14-1-r8

Rho M, Tang H, Ye Y, 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 38(20):e191. doi: 10.1093/nar/gkq747

Schreiber L, et al. 2016. *Endozoicomonas* are specific, facultative symbionts of sea squirts. Front Microbiol. 7:1–15. doi: 10.3389/fmicb.2016.01042

Schreiber L, Kjeldsen KU, Obst M, Funch P, Schramm A, 2016. Description of *Endozoicomonas ascidiicola* sp. nov., isolated from *Scandinavian ascidians*. Syst Appl Microbiol. 39(5):313–318. doi: 10.1016/j.syapm.2016.05.008

Seeman T, 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M, 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34(Database issue):D32–D36.

Solovyev V, Salamov A. 2011. Automatic annotation of microbial genomes and metagenomic sequences. In: Li R, editor.

Metagenomics and its applications in agriculture, biomedicine and environmental studies. Nova Science Publishers. p. 61–78.
Soong G, et al. 2006. Bacterial neuraminidase facilitates mucosal infection by participating in biofilm production. J Clin Invest. 116(8):2297.
Stover C, et al. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature 406(6799): 959–964.
Suomalainen LR, Tiirola M, Valtonen ET, 2006. Chondroitin AC lyase activity is related to virulence of fish pathogenic *Flavobacterium columnare*. J Fish Dis. 29(12):757–763. doi: 10.1111/j.1365-2761.2006.00771.x
Toll-Riera M, Millan A, Wagner A, MacLean R, 2016. The genomic basis of evolutionary innovation in *Pseudomonas aeruginosa*. PLoS Genet. 12(5):e1006005.
Tolmasky ME, Crosa JH, 1995. Iron transport genes of the pJM1-mediated iron uptake system of *Vibrio anguillarum* are included in a transposon-like structure. Plasmid 33(3):180–190. doi: 10.1006/plas.1995.1019
Wu D, Jospin G, Eisen J, 2013. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS One 8(10):e77033.
Wu Y, Tang Y-H, Tringe S, Simmons B, Singer S, 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2(1):26.
Yin J, et al. 2009. Structural basis and catalytic mechanism for the dual functional endo-$\beta$-N-acetylglucosaminidase A. PLoS One 4(3):e4658.
Zankari E, et al. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 67(11):2640–2644.
Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart D, 2011. PHAST: a fast phage search tool. Nucleic Acids Res. 39(suppl):W347–W352.

**Associate editor**: Dan Graur