Multigenetic characterization of ‘Candidatus Xenohaliotis californiensis’

Francesco Cicala,¹ James D. Moore,² Jorge Cáceres-Martínez,³ Miguel A. Del Río-Portilla,³ Mónica Hernández-Rodríguez,³ Rebeca Vásquez-Yeomans³ and Axayácatl Rocha-Olivares¹,*

Abstract
‘Candidatus Xenohaliotis californiensis’ (or Ca.Xc) is the aetiological agent of withering syndrome, a chronic wasting disease affecting most if not all North American species of abalone, and has been described as a Rickettsiales-like prokaryote. Genetic data regarding this species are limited to the 16S rRNA gene. The inability to grow it axenically has hindered its genetic and genomic characterization and, in consequence, a thorough analysis of its systematics. Here, we amplified and sequenced five genes (16S rRNA, 23S rRNA, ftsZ, virD4 and virB11) of Ca.Xc from infected abalone to analyse its phylogenic position. Phylogenies from concatenated DNA and amino acid sequences with representative genera of most Rickettsiales unequivocally place Ca.Xc in the family Anaplasmataceae. Furthermore, the family has two reciprocally monophyletic lineages: one leading to (Neorickettsia, Ca.Xc) and the other to ((Ehrlichia, Anaplasma), Wolbachia). A molecular-clock Bayesian reconstruction places Ca.Xc as the most basal lineage in Anaplasmataceae. These phylogenetic hypotheses shed light on patterns of host evolution and of ecological transitions. Specifically, Neorickettsia and Ca.Xc inhabit aquatic hosts whereas the remaining Anaplasmataceae are found in terrestrial hosts. Additionally, our evolutionary timeline places the directly transmitted marine Ca.Xc as the basal Anaplasmataceae, ancestral to both freshwater and terrestrial species with adaptations leading to more complex life cycles involving intermediate vectors or reservoir species; this supports the hypothesis of a marine origin for this bacterial family.

INTRODUCTION
Among Alphaproteobacteria, the order Rickettsiales comprises a diversity of obligate endocellular Gram-negative bacteria infecting a wide variety of metazoa. Best known are the pathogenic bacteria of vertebrates transmitted by ticks (Anaplasma, Ehrlichia and Rickettsia) or by arthropods and/or fishes (Neorickettsia) as well as arthropod endosymbionts (Wolbachia) [1–3]. Infection of blood cells in vertebrates, including humans, causes clinical diseases collectively called rickettsioses, ehrlichiosis and anaplasmosis; febrile illnesses often followed by lymphadenopathy and haematological abnormalities among other medical disorders [1–5].

Recently, technological advances have increased our ability to detect and describe Rickettsiales-like prokaryotes (RLPs) using an expanded set of molecular and morphological analyses. In this context, a multi-genetic understanding has been essential in shedding light on the phylogenetic relationships of different bacterial groups including Alphaproteobacteria and those in the order Rickettsiales. For instance, the taxonomy of this class has been recently revised into new subclasses according to genetic information from both the small (16S rRNA) and the large (23S rRNA) subunit rRNA genes [6], and the systematic position and boundaries of several Rickettsiales species have been resolved using DNA sequences from 16S rRNA, groESL and surface protein genes [7]. The order Rickettsiales includes three recognized families (Rickettsiaceae, Anaplasmataceae and Holosporaceae) and two more have been recently described (‘Candidatus Paracaeedibacteraceae’ and ‘Candidatus Midichloriaceae’) [8, 9]. Additionally, an increasing number of poorly characterized species are provisionally labelled with the generic name ‘Rickettsia’ [10] or RLP [11], and require additional efforts to determine their phylogenetic and systematic positions [5].

Author affiliations: ¹Molecular Ecology Laboratory, Department of Biological Oceanography, CICESE, Carretera Ensenada-Tijuana 3918, Ensenada Baja California 22860, Mexico; ²Bodega Marine Laboratory, University of California at Davis, PO Box 247, Bodega Bay, CA, USA; ³Department of Aquaculture, CICESE, Carretera Ensenada-Tijuana 3918, Ensenada Baja California 22860, Mexico.

*Correspondence: Axayácatl Rocha-Olivares, arocha@cicese.mx

Keywords: Candidatus Xenohaliotis californiensis; withering syndrome; multigenetic analysis; phylogenetic reconstruction.

Abbreviations: BI, Bayesian inference; Ca.Xc, ‘Candidatus Xenohaliotis californiensis’; DIG, digoxigenin-dUTP; ISH, in situ hybridization; ML, maximum-likelihood; MRCA, most recent common ancestor; mya, million years ago; RLP, Rickettsiales-like prokaryote; SH, Shimodaira–Hasegawa; T4SS, type IV secretion system.

Five supplementary tables and three supplementary figures are available with the online Supplementary Material.
In this context, ‘Candidatus Xenohaliotis californiensis’ (or Ca.Xc) has been described as a new obligate RLP endoparasite [11–13]. However, the study of its genetic organization has been hampered by the inability to grow it axenically. Even with the advent of new methods to study it outside of its host [14], the 16S rRNA gene remains the only known gene from this bacterium. Moreover, using 16S rRNA gene sequences, previous phylogenetic reconstructions placed Ca. Xc as an independent lineage inside Anaplasmataceae [13, 15–18]. However, the systematic position and evolutionary relationship of Ca.Xc remain uncertain due the low bootstrap values (typically around 70% bootstrap support) [11, 16–18].

Consequently, here we analyse the systematic position and evolutionary relationships of Ca.Xc within the order Rickettsiales, with special interest in its relationship to other members of the family Anaplasmataceae, using DNA sequences from five genes and the corresponding amino acid sequences from three of them.

**METHODS**

**Sample collection and Ca.Xc detection**

The source of Ca.Xc DNA included gastrointestinal tissue of wild-caught blue (Haliotis fulgens) and yellow (H. corrugata) abalone sampled from eight localities during commercial fishing operations on the Pacific coast of southern Baja California (Mexico). Approximately 30 mg of post-oesophageal tissue was collected and immediately transferred to sterile 1.5 ml microcentrifuge tubes containing molecular-grade ethanol, and DNA was extracted using DNeasy tissue kits (Qiagen). In addition, we analysed faecal material from farmed red abalone (H. rufescens) maintained in UC Davis Bodega Marine Laboratory, California, USA. Approximately 500 mg of faeces from infected specimens was collected using sterile pipettes and processed with the QiAamp DNA Stool Mini Kit (Qiagen).

To evaluate the presence of Ca.Xc, diagnostic PCRs were performed using 16S primers specific for Ca.Xc (RA5-1 and RA3-6, [19]). However, inconsistent results using the standard OIE protocol [13] required the design of a new set of diagnostic primers (ss16S-F: GCTCTAGTTTGGCTGGGTTCTTCA and ss16S-R: GAATGGCCATTITAAAGTAGTTGGACGG). For both primer sets, PCRs (15 µl final volume) contained: 90 ng of tissue DNA, 1× PCR buffer (Kapa Biosystems), 0.2 mM dNTPs (New England Biolabs), 0.3 µM of each primer and 1 U of Taq polymerase (Kapa Biosystems). Thermal cycling consisted of 4 min at 94°C, 40 cycles of 1 min at 94°C, 30 s at 62°C (RA-set; [19]) and 66°C (16Sss-set; new proposed diagnostic primers) and 30 s at 72°C, followed by 8 min at 72°C. Amplification was verified with 1.5% agarose gel electrophoresis. At least seven amplicons for each primer set were purified enzymatically (ExoSAP-IT; USB-Affymetrix) before cycle sequencing using an ABI 3730XL automatic sequencer (Macrogen). Resulting 16S rRNA gene sequences were trimmed and verified with Codon Code Aligner v3.7.1 (Codon Code Corporation) and aligned using CLUSTAL W within the program Geneious R9 [20]. Online BLAST searches [21] were used to confirm the taxonomic identity of the sequences.

**Gene selection and primer design**

A set of 16 genes was originally selected (products shown in parentheses): 16S rRNA (small ribosomal subunit RNA), 23S rRNA (large ribosomal subunit RNA), coxA (cytochrome oxidase A subunit), ftsZ (a cell division protein), virD4 (a type IV secretion system – T4SS – protein), virB11 (a T4SS protein), tpiA (triose phosphate isomerase), gap (glyceraldehyde-3-phosphate dehydrogenase), groEL (a chaperone protein), atpD (ATP synthase F1 beta subunit), tkt (transketolase, a pentose phosphate pathway protein), recA (recombinase A), omp (outer membrane protein), rpoB, rpoC and rpoD (RNA polymerase subunits). These loci were selected because (i) they have conserved regions allowing primer design across species of Rickettsiales, (ii) they are present in most Rickettsiales species and (iii) they have proven to be useful in molecular systematic analyses of this group [7, 22–25]. DNA sequences from the available target genes of Rickettsiales species were downloaded from GenBank and aligned using the Multiple Expectation-Maximization for Motif Elicitation (MEME) algorithm. This algorithm allows multiple alignment of DNA or amino acid sequences and facilitates searches for conserved uninterrupted motifs that can be used to design primers [26]. Primers were designed using the MEME suite with a maximum degeneracy of 25–35% of bases (http://memesuite.org/).

**Optimization PCR, cloning and sequencing**

Amplification of all loci was carried out in 20 µl reactions containing 100 ng of faecal or 90 ng of tissue DNA, 1× PCR buffer, 1.5 mM MgCl2 (both Kapa Biosystems), 0.2 mM dNTPs (New England Biolabs), 0.5 µM of each primer, 0.4 mM BSA (New England Biolabs) and 1 U of Taq polymerase (Kapa Biosystems). Thermal cycling conditions were 94°C for 4 min, followed by 40 cycles of 94°C for 1 min, a gradient (40–60°C) as annealing temperature for 30 s and 72°C for 30 s, and a final extension for 8 min at 72°C. After verification using 1.5% agarose gel electrophoresis, amplicons of expected sizes were cleaned with a QIAquick PCR Purification kit (Qiagen) and cloned using a PCR Cloning Kit (New England Biolabs). Forty transformed colonies for each locus were used as DNA template in 15 µl colony PCRs consisting of 1× PCR buffer (Kapa Biosystems), 0.2 mM dNTPs (New England Biolabs), 0.3 µM of each vector primer (provided in the cloning kit) and 1 U of Taq polymerase (Kapa Biosystems). Thermal cycling conditions were as follows: one cycle of 4 min at 94°C, 30 cycles of 1 min at 94°C, 30 s at 57°C and 30 s at 72°C, followed by a final 8 min incubation at 72°C. Amplicon size and quality were verified by 1.5% agarose gel electrophoresis before enzymatic purification (ExoSAP-IT; USB-Affymetrix) and cycle-sequenced using an ABI 3730XL automatic DNA sequencer (Macrogen). DNA sequences were verified using Codon Code Aligner v3.7.1 (Codon Code Corporation).
DNA sequence control tests

DNA sequences obtained from colony PCR experiments were subject to several controls in order to maximize the probability of their being from Ca.Xc. Initially, an identity control was assessed by comparison with the GenBank database using BLAST [21]. For the confirmed rickettsial sequences, a second control consisted of *in situ* hybridization (ISH) assays to verify their physical association with bacterial inclusions in infected tissues. Digoxigenin-dUTP (DIG)-labelled probes were prepared using the PCR DIG Probe Synthesis Kit (Roche Applied Science). Each DIG probe was obtained by nested PCR using locus generic primers and purified amplicons from colony PCRs as templates. Subsequently, ISH was carried out on histological slides prepared with post-oesophagus, digestive gland, kidney and gill tissues from infected red abalone following Antonio et al. [27]. A third control consisted of comparing haplotype-specific PCR co-amplification patterns. For this, we designed specific primers for the DNA sequences (haplotypes) being tested as originating from Ca.Xc for each locus. These haplotype-specific primers were then used in conjunction with the improved diagnostic 16S rRNA primers (ss16S set) to assess co-amplification patterns in infected and uninfected abalone (*n*=16 yellow, *n*=15 blue and *n*=8 red abalone). Amplification of both markers in the same organism (i.e. co-amplification) was interpreted as corroborating evidence that the source of the sequence was Ca.Xc, whereas lack of co-amplification was interpreted as evidence that the source was another RLP. For final corroboration, all PCR products from these experiments were sequenced as described above.

Phylogenetic analyses

Phylogenetic analyses were conducted separately for each gene and using concatenated DNA and amino acid sequences (for protein-coding genes), the latter to improve the resolution of deeper nodes. To assess the phylogenetic position of Ca.Xc in the order Rickettsiales, homologous sequences were obtained from representative species of as many lineages as possible for which genomic data were available; these included the two well-characterized families Anaplasmataceae (Ehrlichia, Anaplasma, Wolbachia and Neorickettsia) and Rickettsiaceae (Rickettsia and Orientia), also including Holosporaceae (Holospora), ‘Ca. Paracaeceibacteraceae’ (Paracaeceibacter and Odyssella), ‘Ca. Midichloriaceae’ (‘Ca. Midichloria’ and an undescribed endosymbiont of *Acanthamoeba* str. UWCh8), as well as unclassified RLP Caedibacter and ‘Ca. Hepatobacter’ (Table S1, available in the online Supplementary Material). Phylogenetic reconstructions were outgroup-rooted with sequences from Alphaproteobacteria (SAR11 clade) for the 16S rRNA, 23S rRNA and *fzs* genes (Table S1). Given that the *rvh* T4SS evolved after the split of *Candidatus Pelagibacter* from the *Rickettsiales* ancestor [25], *virD4* and *virB11* phylogenetic reconstructions were rooted with other outgroup taxa. Furthermore, because the *rvh* P-T4SSs of *Rickettsiales* evolved via horizontal gene transfer from Gammaproteobacteria [25], we rooted these reconstructions with their closest *rvh* xenologue: *trw* P-T4SS sequences from *Xanthomonas* and *Lyso bacter* (Table S1). *Caedibacter varicaedens*, ‘*Ca. Caedibacter acanthamoebae*’, ‘*Ca. Hepatobacter penaei*’, ‘*Ca. Odysella thessalonicensis*’, ‘*Ca. Paracaeceibacter acanthamoebae*’, *Holospora obtuse* and *Holospora undulata* lack T4SS genes and hence they were excluded from *virD4* and *virB11* as well as from amino-acid-based reconstructions (Table S1). In concatenated DNA and amino acid analyses, outgroup sequences consisted of assembled chimeras from alpha- (16S rRNA, 23S rRNA and *fzs*) and gamma- (*virD4* and *virB11*) proteobacteria.

Alignments were obtained with **Clustal W** as implemented in Geneious R9 [20]. We used maximum-likelihood (ML) and Bayesian inference (BI) as methods of phylogenetic reconstruction. The best fit model of sequence evolution (Table S2) was estimated with ModelTest v2.1.3 for nucleotide sequences and ProtTest v2.4 for amino acid sequences [28, 29]. ML heuristic searches were conducted using 100 random taxon addition replicates with tree bisection and reconnection (TBR) branch swapping with RAxML v7.2.8 [30], whereas BI was conducted using Mr Bayes version 3.2 [31]. For each data set, the default number of generations (20 000) was increased until achieving a standard deviation smaller than 0.01. Topological congruence was assessed with the Shimodaira–Hasegawa test (SH-test; [32]) as implemented in **PAUP** version 4 [33]. The SH-test null distribution was estimated by non-parametric resampling generated by a log-likelihood RELL method in **PAUP** version 4 [33].

To investigate the chronology of evolutionary changes in hosts in the order *Rickettsiales* leading to Ca.Xc, we performed a Bayesian molecular clock constraint phylogeny with 16S rRNA gene sequences using the program BEAST v1.8 [34]. The Monte Carlo Markov chain consisted of 1 000 000 generations with sampling every 200 generations. We used the GTR+Γ+I substitution model and a rate of nucleotide substitution estimated for endosymbiont bacteria [35]. In this analysis we included data from additional *Rickettsiales* in order to widen the taxonomic scope of the reconstruction.

RESULTS

**Ca.Xc detection**

Inconsistent results and a high frequency of false negative assays with the World Organization for Animal Health (OIE) detection primers [19] led us to design a new set of diagnostic primers for Ca.Xc. DNA sequences of diagnostic amplicons confirmed the presence of Ca.Xc. Consequently, all diagnostic tests were performed with the proposed ss16S primer set (Table 1).

**Gene amplification, sequencing and control tests**

Fifteen pairs of generic rickettsial primers were designed using the MEME algorithm, and most of them amplified products of the expected size (Table S3). Only those designed for *recA*, *rpoB* and *rpoC* failed to yield positive amplifications. Cloning of PCR products revealed sequences from more than one bacterial species for all except the *virB11* gene, for which a single haplotype was obtained. The
BLAST analyses revealed that only 23S rRNA, ftsZ, virD4, virB11, tipA, coxA, gap and rkt genes contained at least one rickettsial haplotype (72–85% similar to the closest rickettsial sequence from GenBank, Table S4). Successive analyses were carried out on these rickettsial haplotypes. Physical confirmation of Ca.Xc origin through ISH was observed for the ribosomal gene probes (16S rRNA and 23S rRNA) in the post-oesophagus of infected red abalone (Fig. S1). Other probes did not produce detectable signal in the histological slides. However, co-amplification patterns provided reliable evidence that the source of rickettsial gene sequences was Ca.Xc. In total, 32 of the 39 tested abalone were carriers of Ca.Xc, and positive co-amplification of the 16S rRNA gene (Table S5) with the haplotype-specific primers was confirmed for the genes 23S rRNA, ftsZ, virD4 and virB11 (Table 1). Co-amplification mismatches with the genes coxA, tipA, gap and tkt suggest a different RLP source for these gene sequences. Accordingly, phylogenetic analyses were conducted on 16S rRNA, 23S rRNA, ftsZ, virD4 and virB11 gene sequences.

**Nucleotide and amino acid phylogenetic analyses**

Phylogenetic reconstructions were based on 1304 bp for 16S rRNA, 312 bp for 23S rRNA, 429 bp for ftsZ, 800 bp for virB11 and 1111 bp for virD4, for a total of 3956 bp for the concatenated analysis. For each gene ML and BI reconstructions were identical or not significantly different (SH-tests: 16S rRNA P=0.052, 23S rRNA P=0.076, ftsZ P=0.081, virB11 P=0.065, virD4 P=1.0), and hence we focus on the BI reconstruction. Furthermore, no significant differences were found between the BI individual-gene trees and the concatenated phylogeny in three of the five genes (SH-tests: 16S rRNA P=0.36, 23S rRNA P=0.44, ftsZ P=0.25, virB11 P=0.002, virD4 P=0.015). Topological differences between the concatenated tree and virB11 and virD4 trees, computed on the subset of taxa possessing these genes, were related to insufficient resolution of deep nodes (Figs S2 and S3), so we focus on the results of the concatenated tree as the most reliable reconstruction. Salient features of this tree included the monophyly of three well-supported monophyletic families *Anaplasmataceae*, 'Ca. Midichloriaeaceae' and *Rickettsiaceae*, with the last named as the most ancestral. This lineage also included the unclassified *Rickettsiales* bacterium strain Ac37b (GenBank accession no. CP009217) as sister to *Rickettsiaceae*. Of the remaining *Rickettsiales* bacteria, a second well-supported group, sister to the one grouping *Anaplasmataceae*, 'Ca. Midichloriaeaceae' and *Rickettsiaceae*, included the family *Holosporaceae*, 'Ca. Hepatobacter' and *Caedibacter*. 'Ca. Paracaedibacteraceae' was the most ancestral lineage among *Rickettsiales*. Ca.Xc was included in the family *Anaplasmataceae* (posterior probability =1.0), appearing as sister to *Neorickettsia* and with it in reciprocal monophyly to the monophyletic lineage grouping *Wolbachia, Ehrlichia* and *Anaplasma* (topology A, Fig. 1).

In order to improve deep resolution, we carried out phylogenetic reconstructions based on amino acid sequences of the protein-coding genes (136 codons for ftsZ, 273 codons for virB11 and 383 codons for virD4) separately and concatenated (792 codons). Protein sequences satisfactorily resolved most of the shallow branches of the phylogeny encompassing the monophyly of several genera and the clade *Ehrlichia, Anaplasma* and *Wolbachia*. The concatenated amino acid phylogeny was well resolved and placed Ca.Xc as sister to *Neorickettsia*, in the same topology as the DNA-based tree (topology A, Fig. 2).

The clock constraint tree places the transition to rickettsial intracellular habituation between 1070 and 1030 million years ago (mya). The most recent common ancestor (MRCA) of the families *Anaplasmataceae*, 'Ca. Midichloriaeaceae' and *Rickettsiaceae* lineages lived ca. 870 mya. In *Anaplasmataceae*, the MRCA of bacteria infecting terrestrial hosts (*Ehrlichia, Anaplasma* and *Wolbachia*) lived 455 mya, which diverged from bacteria infecting aquatic hosts (*Neorickettsia and Ca.Xc*) after 568 mya. Notably, in this reconstruction the marine Ca.Xc RLP appears as the most basal lineage of the family *Anaplasmataceae* (topology B, Fig. 3).

### Table 1. Haplotype-specific primers for Ca.Xc used in control tests

<table>
<thead>
<tr>
<th>Locus</th>
<th>ID primers</th>
<th>Sequence (5'-3')</th>
<th>Length</th>
<th>Annealing temperature (°C)</th>
<th>Expected fragment size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S rRNA</td>
<td>ss16S.F</td>
<td>GCCTCAGTTTGCTGGTCTTCCA</td>
<td>24</td>
<td>66</td>
<td>426</td>
</tr>
<tr>
<td></td>
<td>ss16S.R</td>
<td>GAATGGCCACCTTTAAAGTTAGCGG</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S rRNA</td>
<td>ss16SV1.F</td>
<td>GCCTTACCTTTCCGGCAGCAAA</td>
<td>26</td>
<td>63</td>
<td>1412</td>
</tr>
<tr>
<td></td>
<td>ss16SV2.R</td>
<td>GGTGGTTCACCTCCGTATCAA</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23S rRNA</td>
<td>23S.s1F</td>
<td>CAGTCTCAGTTTGCTTCTCTC</td>
<td>24</td>
<td>66</td>
<td>376</td>
</tr>
<tr>
<td></td>
<td>55.s1R</td>
<td>GGAGTGTTCACCTCCGTATCAA</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>virD4</td>
<td>VirD4.s1F</td>
<td>CCTCACTTTTGGCTGTTGTT</td>
<td>23</td>
<td>66</td>
<td>289</td>
</tr>
<tr>
<td></td>
<td>VirD4.s2R</td>
<td>CAACTAAGCTGCAAAATGTAGGC</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>virB11</td>
<td>VirB11.ss2F</td>
<td>GCTCAACTAAGCTAGCGAAGAT</td>
<td>26</td>
<td>60</td>
<td>508</td>
</tr>
<tr>
<td></td>
<td>VirB11.ss1R</td>
<td>CGCAGCAAGGTCACTCTCCTTG</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftsZ</td>
<td>ftsZ.ss1F</td>
<td>CCGTCAATTGCAAAGATTTTCCT</td>
<td>25</td>
<td>61</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>ftsZ.ss1R</td>
<td>CTCTACCTTTGGCATTGCATAACGAT</td>
<td>26</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 1.* Haplotype-specific primers for Ca.Xc used in control tests.
DISCUSSION

Sequencing of Ca.Xc DNA

Our experimental approach successfully allowed us to sequence the target Ca.Xc DNA without the need for isolation or axenic culture. The physical association of the amplified DNA with bacterial inclusions in infected tissues was confirmed through ISH for the two ribosomal genes. We speculate that the lack of detectable ISH signal in the non-target genes used in this study are single copy in rickettsial genomes, using the same experimental approach, is probably due to a number of target gene sequences below the detection level of ISH [36]. To our knowledge, the structural genes used in this study are single copy in rickettsial bacteria and their transcription levels may provide an insufficient number of mRNA copies for ISH detection. In contrast, rRNA can accumulate in the cytoplasm reaching millions of copies and up 33% of the total dry cell weight [37]. Nevertheless, the adoption of a third verification protocol based on co-amplification of target genes with Ca.Xc-diagnostic 16S rRNA allowed us to differentiate target and non-target sequences.

Ca.Xc phylogenetic affinities

Although we found topological differences in individual gene trees, they were only significantly different from the concatenated analysis in the T4SS gene trees. These differences may relate to the distinct evolutionary dynamics of these genes, which are well documented [25]. Nevertheless, the reconstruction carried out with the five concatenated genes (3956 bp) represents the most extensive analysis of Ca.Xc evolutionary relationships and provided improved resolution surrounding its phylogenetic position among Rickettsiales (most posterior probabilities >95% in Anaplasmataceae) compared with published reconstructions based on the 16S rRNA gene [11, 16–18].

The use of DNA sequences may sometimes lead to alignments that do not reflect the appropriate mutational history and thus to incorrect phylogenetic trees [38]. The difficulties in reconstructing accurate phylogenetic trees may relate to speciation events closely spaced in time that do not allow for the evolution of synapomorphic characters or, in the case of ancient nodes and relationships, to long terminal branches abounding with multiple substitutions at the same sequence position, erasing the original phylogenetic signal and producing homoplasies. In consequence, resolving some deep divergences may be challenging even with the use of long DNA sequences [38, 39]. The use of proteins or expressed sequence tags may help to resolve phylogenetic incongruences and is becoming a standard approach to
investigate evolutionary relationships [38, 40]. Hence, we investigated phylogenetic reconstructions based on amino acid sequences.

In contrast to recent studies that use only 16S rRNA gene sequences [11, 16–18], our most reliable reconstructions (DNA and amino acid concatenated phylogenetic trees) placed Ca.Xc unequivocally as a member of the family Anaplasmataceae and as sister to Neorickettsia, in reciprocal monophyly to the remaining family members (topology A). Additionally, our molecular clock phylogeny suggests a basal position of Ca.Xc within the family Anaplasmataceae (topology B). Given the aquatic habits of the Ca.Xc–Neorickettsia clade, our finding supports the hypothesis of an aquatic origin of Anaplasmataceae and the subsequent transition to the terrestrial environment by the lineages giving rise to Wolbachia, Anaplasma and Ehrlichia [5].

**Evolutionary timeline and habitat/host transitions**

In agreement with previous publications, the origin of species belonging to the families Rickettsiaceae, Anaplasmataceae and ‘Ca. Midichloriaeae’ evolved around 870 mya [41]. At that time, genome changes may have involved genetic reorganization, which in turn may have modified the metabolic pathways used by these bacterial families [10, 41, 42]. The position of Ca.Xc as the most basal Anaplasmataceae lineage (topology B) in the molecular clock constraint tree is noteworthy because in the concatenated analysis it is not (topology A). Given the wider taxonomic scope of the molecular clock tree, this could result from improved character polarization provided by increased taxonomic sampling. Also, the basal position of Ca.Xc may have resulted from enforcing a uniform substitution rate in the ultrametric tree, placing the longest branch among Anaplasmataceae sequences (i.e. the one leading to Ca.Xc) at the most basal position in the family.

In light of the above and given that Ca.Xc has been detected only in marine invertebrate hosts, it is possible to re-evaluate the evolution of some biological characteristics within the family Anaplasmataceae consistent with our phylogenetic insights: (i) the ability to live in the marine environment should be considered an ancestral characteristic of this bacterial family; (ii) the ability to infect invertebrate hosts is a plesiomorphic character; and (iii) because transmission of Ca.Xc is direct and presumed to be faecal–oral [43], the absence of known intermediate vectors or reservoir organisms mediating transmission is an ancestral character [14]. Consequently, vector-mediated infection
research is needed to determine if the detected sequences
previously proposed by Crosson et al. [45] may harbour still uncharacterized RLP organisms, as previ-
ous RLP organisms, as previously proposed by Crosson et al. [45]. However, further
research is needed to determine if the detected sequences belong to potentially pathogenic species.

Diversity of rickettsial endosymbionts in abalone
gastrointestinal tract

Given the approach of subsequent refinement of the primers used to target Ca.Xc DNA, we detected new bacterial DNA sequences assignable to the order Rickettsiales but not to the newly described Ca.Xc. This suggests that abalone species may harbour still uncharacterized RLP organisms, as previously proposed by Crosson et al. [45]. However, further research is needed to determine if the detected sequences belong to potentially pathogenic species.

Acknowledgements

We are grateful to Nallely Cano Domínguez for assistance in the labora-

atory, and Ashley Vater, Andrea Lievana-MacTavish and Roberto

Cruz-Flores for a critical assessment of an earlier version of the

manuscript.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Rikihisa Y. New findings on members of the family Anaplasmata-
445.
2. Weinert LA, Warren JH, Aeby A. Evolution and diversity of rickettsia

3. Darby AC, Cho NH, Fuxelius HH, Westberg J, Andersson SG. Intracellular pathogens go extreme: genome evolution in the Rick-

diversity of Rickettsiales bacteria in two species of ticks from

5. Ferla MP, Thrash JC, Giovanni S, Patrick WM. New rRNA gene-
based phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylo-

Fig. 3. Bayesian molecular clock constraint phylogeny of the order Rickettsiales based on 16S rRNA gene sequences. Numbers at nodes indicate divergence time in millions of years. "The collapsed node labelled as ‘Uncultured marine Alphaproteobacteria’ includes: Alphaproteobacterium strain HMB59 (CP003801); Uncultured marine micro-organism clone NB062806_218 (KC425551) and

Uncultured Alphaproteobacterium clone ARTE1_103 (GU230260).


