International Journal of Systematic and Evolutionary Microbiology (2009), **59**, 234–247 **DOI** 10.1099/ijs.0.002741-0

Phylogenomics and protein signatures elucidating the evolutionary relationships among the *Gammaproteobacteria*

Beile Gao, Ritu Mohan and Radhey S. Gupta

Department of Biochemistry and Biomedical Science, McMaster University, Hamilton, ON L8N 3Z5, Canada

Correspondence
Radhey S. Gupta
gupta@mcmaster.ca

The class *Gammaproteobacteria*, which forms one of the largest groups within bacteria, is currently distinguished from other bacteria solely on the basis of its branching in phylogenetic trees. No molecular or biochemical characteristic is known that is unique to the class *Gammaproteobacteria* or its different subgroups (orders). The relationship among different orders of gammaproteobacteria is also not clear. In this study, we present detailed phylogenomic and comparative genomic analyses on gammaproteobacteria that clarify some of these issues. Phylogenetic trees based on concatenated sequences for 13 and 36 universally distributed proteins were constructed for 45 members of the class *Gammaproteobacteria* covering 13 of its 14 orders. In these trees, species from a number of the subgroups formed distinct clades and their relative branching order was indicated as follows (from the most recent to the earliest diverging): *Enterobacteriales* >*Pasteurellales* >*Vibrionales*, *Aeromonadales* >*Alteromonadales* >*Oceanospirillales*, *Pseudomonadales* >*Chromatiales*, *Legionellales*, *Methylococcales*, *Xanthomonadales*, *Cardiobacteriales*, *Thiotrichales*. Four conserved indels in four widely distributed proteins that are specific for gammaproteobacteria are also described. A 2 aa deletion in 5′-phosphoribosyl-5-aminoimidazole-4-carboxamide transformylase (AICAR transformylase; PurH) was a distinctive characteristic of all gammaproteobacteria (except *Francisella tularensis*). Two other conserved indels (a 4 aa deletion in RNA polymerase β-subunit and a 1 aa deletion in ribosomal protein L16) were found uniquely in various species of the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales*, *Aeromonadales* and *Alteromonadales*, but were not found in other gammaproteobacteria. Lastly, a 2 aa deletion in leucyl-tRNA synthetase was commonly present in the above orders of the class *Gammaproteobacteria* and also in some members of the order *Oceanospirillales.* The presence of the conserved indels in these gammaproteobacterial orders indicates that species from these orders shared a common ancestor that was separate from other bacteria, a suggestion that is supported by phylogenetic studies. Systematic BLASTP searches were also conducted on various open reading frames (ORFs) in the genome of *Escherichia coli* K-12. These analyses identified 75 proteins that were unique to most members of the class *Gammaproteobacteria* or were restricted to species from some of its main orders (*Enterobacteriales*; *Enterobacteriales* and *Pasteurellales; Enterobacteriales*, *Pasteurellales, Vibrionales*, *Aeromonadales* and *Alteromonadales*; and the *Enterobacteriales, Pasteurellales*, *Vibrionales*, *Aeromonadales*, *Alteromonadales*, *Oceanospirillales* and *Pseudomonadales* etc.). The genes for these proteins have evolved at various stages during the evolution of gammaproteobacteria and their species distribution pattern, in conjunction with other results presented here, provide valuable information regarding the evolutionary relationships among these bacteria.

Abbreviations: AICAR transformylase, 5′-phosphoribosyl-5-aminoimidazole-4-carboxamide transformylase; COG, conserved orthologous groups; LGT, lateral gene transfer; ML, maximum likelihood; MP, maximum parsimony; NJ, neighbour-joining; ORFans, orphan genes; RGCs, rare genomic changes.

A list of proteins used in the phylogenetic analysis, a list of the bacterial strains used to produce the concatenated alignments, the concatenated sequence alignment for the group of 36 proteins obtained with the GBlock program, a neighbour-joining phylogenetic tree based on the concatenated sequences of the 36 proteins, a maximum-likelihood/maximum-parsimony phylogenetic tree based on 13 proteins and a partial sequence alignment of ribosomal protein L6 are available as supplementary material with the online version of this paper.

INTRODUCTION

The class *Gammaproteobacteria* constitutes a very large and diverse group of bacteria that exhibits enormous variety in terms of their phenotype and metabolic capabilities (Woese *et al.*, 1985; Stackebrandt *et al.*, 1988; Brenner *et al.*, 2005; Kersters *et al.*, 2006). Although the majority of gammaproteobacteria are chemo-organotrophs, this group also includes several phototrophs and chemolithotrophs that derive their metabolic energy via hydrogen-, sulfur- or iron-oxidation (Stackebrandt *et al.*, 1988; Gupta, 2000; Brenner *et al.*, 2005; Kersters *et al.*, 2006). The class *Gammaproteobacteria* also includes enteric bacteria (including the thoroughly studied model organism *Escherichia coli*) and it is well known for harbouring large numbers of human, animal and plant pathogens such as members of the genera *Salmonella*, *Shigella*, *Vibrio*, *Yersinia, Pasteurella*, *Pseudomonas*, *Xanthomonas*, *Erwinia*, etc. (Brenner *et al.*, 2005; Kersters *et al.*, 2006). A number of species from this group (e.g. from the genera *Buchnera*, *Coxiella*, '*Candidatus* Blochmannia' etc.) are obligate intracellular parasites of mammalian, bird and arthropod species and live endosymbiotically within their host cells (Belda *et al.*, 2005; Brenner *et al.*, 2005; Kersters *et al.*, 2006). In the current taxonomic scheme based on 16S rRNA gene sequences, the *Gammaproteobacteria* are recognized as a class within the phylum *Proteobacteria* (Stackebrandt *et al.*, 1988; De Ley, 1992; Brenner *et al.*, 2005; Kersters *et al.*, 2006). In phylogenetic trees, the class *Gammaproteobacteria* shows a close relationship to the class *Betaproteobacteria* and the other three classes of proteobacteria (*Alphaproteobacteria*, *Deltaproteobacteria* and *Epsilonproteobacteria*) are more distantly related (Gupta, 2000; Ludwig & Klenk, 2005; Kersters *et al.*, 2006; Gupta & Sneath, 2007). Based on their branching in the 16S rRNA gene trees, the class *Gammaproteobacteria* has been divided into 14 main orders or subgroups: the *Enterobacteriales*, *Pseudomonadales*, *Alteromonadales*, *Vibrionales*, *Pasteurellales*, *Chromatiales, Xanthomonadales*, *Thiotrichales*, *Legionellales, Methylococcales*, *Oceanospirillales*, *Acidithiobacillales, Cardiobacteriales* and *Aeromonadales* (Garrity *et al.*, 2005; Brenner *et al.*, 2005; Kersters *et al.*, 2006). Although gammaproteobacteria are among the most extensively studied bacterial groups, they are presently defined solely on the basis of their clustering and branching pattern in phylogenetic trees (Woese *et al.*, 1985; De Ley, 1992; Ludwig & Klenk, 2005; Kersters *et al.*, 2006). No unique morphological, molecular or biochemical characteristic has been identified that can distinguish members of the class *Gammaproteobacteria* or its main orders from other bacteria.

Since the sequencing of the genome for *Haemophilus influenzae* in 1995 (Fleischmann *et al.*, 1995), sequence data for additional bacterial genomes have been accumulating at an increasingly accelerated pace. Of the present >550 completely sequenced bacterial genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi), more than half are from proteobacteria and of these about 25 % are from gammaproteobacteria, making them the most densely sequenced bacterial group. Comparative analyses of these genomes provide a huge and unprecedented resource for discovering novel molecular characteristics that are either unique to particular species or are shared by different gammaproteobacteria and they can also provide valuable tools for biochemical, diagnostic, taxonomic and evolutionary studies (Koonin & Galperin, 1997; Binnewies *et al.*, 2006). As the class *Gammaproteobacteria* includes many medically important groups of bacteria, such as the orders *Enterobacteriales*, *Vibrionales*, *Pasteurellales* and *Pseudomonadales*, a number of comparative genomic studies have been conducted to identify proteins that are unique to particular gammaproteobacterial species that could be responsible for disease causation or virulence (Van Sluys *et al.*, 2002; Edwards *et al.*, 2002; Whittam & Bumbaugh, 2002; Deng *et al.*, 2003; Howard *et al.*, 2006; Binnewies *et al.*, 2006). However, such studies have focused on closely related species, mainly at the species or genus level, and no studies have been conducted to search for proteins or molecular markers that are specific to either all or many of the orders of the class *Gammaproteobacteria*. In an earlier study, Daubin & Ochman (2004) analysed the *E. coli* genome to search for orphan genes (ORFans) that were restricted to gammaproteobacteria at different phylogenetic depths. Although their work suggested that >2000 genes were native to these bacteria (Daubin & Ochman, 2004), at that time very few gammaproteobacterial genomes were available and most of the identified ORFans were present in only very few (two or so) representatives from each 'clade'. Hence, based on earlier work, it is still not known if any proteins are uniquely shared by all or most of the sequenced gammaproteobacteria or by some of the main orders within these bacteria. The genomic data have also been used by some authors to examine the evolutionary relationships among gammaproteobacteria based on different sets of genes/protein sequences (Kunisawa, 2001; Lerat *et al.*, 2003; Brown & Volker, 2004; Belda *et al.*, 2005; Ciccarelli *et al.*, 2006; Mrazek *et al.*, 2006; Lee & Côté, 2006). However, most of these studies were again based on a limited number of species from a small number of orders of the class *Gammaproteobacteria*.

To elucidate the evolutionary relationships amongst gammaproteobacteria, in the present study, a combination of phylogenomic and comparative genomic approaches was employed. This strategy has provided valuable insights into evolutionary relationships for a number of other groups/phyla of bacteria (for example, the *Alphaproteobacteria*, *Epsilonproteobacteria*, *Chlamydiae*, *Actinobacteria* and *Bacteroidetes–Chlorobi*) (Griffiths *et al.*, 2006; Gao *et al.*, 2006; Gupta, 2006; Gupta & Lorenzini, 2007; Gupta & Mok, 2007). In this work, we carried out detailed phylogenetic analyses on a broad range of gammaproteobacteria covering all the main orders of the class, based on concatenated sequences for 36 highly conserved and universally distributed proteins. In parallel, comparative

analyses were conducted on gammaproteobacterial genomes to identify molecular markers that were unique to this group of bacteria at different taxonomic levels. Of the two kinds of gammaproteobacterial-specific markers identified in this work, one type consisted of conserved inserts or deletions (i.e. indels) in widely distributed proteins that were restricted to either all or particular orders of these bacteria (Gupta, 2000). The other kind of molecular markers were whole proteins that were uniquely present in particular groups or orders of the class *Gammaproteobacteria*, but were not found elsewhere. The results obtained from all three of these approaches were concordant and provide valuable insights into the evolutionary relationships among gammaproteobacteria. The conserved indels and whole proteins that are specific for the class *Gammaproteobacteria* also provide valuable tools for genetic, biochemical and other studies on these bacteria which could lead to the identification of novel biochemical and/or physiological characteristics that are unique to them.

METHODS

Phylogenetic analyses and identification of conserved indels specific for gammaproteobacteria. Phylogenetic analyses were performed on a concatenated sequence alignment for 36 conserved and widely distributed proteins (set I). These proteins included 30 of the 31 (i.e. all except ribosomal protein S9, which was absent in one of the species) universally distributed proteins that were used by Ciccarelli *et al.* (2006) to construct a highly resolved tree of life. In addition, sequences for six other highly conserved proteins (50 ribosomal protein L2, DNA gyrase subunit A, DNA helicase II, DnaK, protein synthesis elongation factor-G and SecA translocase) were included in the dataset. The information regarding the lengths and clusters of orthologous groups (COG) for these proteins is provided in Supplementary Table S1 (available in IJSEM Online). For each of these proteins, sequences from 45 gammaproteobacterial species, along with a deep branching species *Caulobacter crescentus* (an alphaproteobacterium), were retrieved and multiple sequence alignments were created using the CLUSTAL_X 1.83 program (Jeanmougin *et al.*, 1998). The accession numbers for all of the sequenced genomes from which these sequences were retrieved, along with the information about which protein sequences were included in which concatenated set, is presented in Supplementary Table S2 (available in IJSEM Online). A concatenated sequence alignment for these proteins was imported into the Gblocks 0.91b program to remove poorly aligned regions (Castresana, 2000). The Gblocks program was used mainly with the default setting (namely, minimum number of sequences for a conserved position, 24; minimum number of sequences for a flank position, 39; maximum number of contiguous non-conserved positions, 8; minimum length of a block, 10; allowed gap positions, half). The original concatenated alignment contained a total of 14 309 aa positions, which after filtering with the Gblocks program was reduced to 10 993 aa positions (i.e. 78 % of the positions were retained). This filtered alignment, which was used for phylogenetic analyses, is presented as Supplementary Fig. S1 (available with the online version of this paper). A neighbour-joining (NJ) tree based on 1000 bootstrap replicates was constructed by the Kimura model (Kimura, 1983) using the TREECON 1.3b program (Van de Peer & De Wachter, 1994). The maximum-likelihood (ML) analysis was carried out using the WAG+F model with gamma distribution of evolutionary rates with four categories using the TREE-PUZZLE program with 10 000 puzzling steps (Schmidt *et al.*, 2002). A maximum-parsimony (MP) tree based on 1000 bootstrap replicates was computed using the MEGA 4.1 program (Tamura *et al.*, 2007).

In addition to the phylogenetic analyses on the above large dataset, phylogenetic trees for the same 45 gammaproteobacterial species were also constructed for many individual proteins (particularly those with lengths >400 aa) and for a smaller dataset of concatenated sequences for 13 large proteins [arginyl-tRNA synthetase, elongation factor-G, gyrase A, Hsp70, isoleucyl-tRNA synthetase, ribosomal L2 and S3 proteins, phenylalanyl-tRNA synthetase, RecA, RNA polymerase β-subunit (RpoB), SecA, SecY and UvrD] from the larger dataset. This dataset (set II) included '*Aquifex aeolicus*' as the outgroup species and the final alignment in this case (after removal of poorly aligned regions with Gblocks) consisted of 6501 positions.

The sequence alignments for these and a number of other proteins that have been previously constructed in our work were also inspected to identify any conserved indel that was restricted to particular subgroups of gammaproteobacteria (Gupta, 2000). Indels not flanked by conserved regions were not considered (Gupta, 1998). The group specificities of these and other indels were evaluated by carrying out detailed BLASTP searches on short sequence segments containing the indels and their flanking conserved regions. The sequence information for all indels was compiled into signature files presented in this study.

Identification of lineage-specific proteins. To identify proteins that were specific for gammaproteobacteria, BLASTP searches were performed on each individual protein or ORF in the genome of *E. coli* K-12, using the default parameters, without the low complexity filter, to identify different proteins where all the significant hits were from gammaproteobacteria (Altschul *et al.*, 1997). The results of BLAST searches were inspected for a sudden increase in the expected values (E-values) from the last gammaproteobacterial species in the search to the first non-gammaproteobacterial hit. The proteins that were of interest generally involved a large increase in E-values from the last gammaproteobacterial hit to the first hit from any other organism. Further, the E-values of these latter hits were generally higher than 10^{-3}, which indicates a weak level of similarity that could occur by chance (Gao *et al.*, 2006; Gupta, 2006). However, higher or lower E-values can sometimes be acceptable depending upon the length of the query sequence and that of the hit (Altschul *et al.*, 1997). All promising proteins were further analysed using the position-specific iterated (PSI) BLAST program (Schaffer *et al.*, 2001) to confirm their group specificity. In the present work, the focus was primarily on identifying those proteins that were distinctive characteristics of the higher taxonomic clades within the class *Gammaproteobacteria* (such as the order *Enterobacteriales*) or those that were uniquely present in the order *Enterobacteriales* and the other main orders of the class *Gammaproteobacteria*. The proteins that were unique to only *E. coli* K-12, or various *E. coli* strains, or were found in only a limited number of sequenced species of the order *Enterobacteriales*, are not reported here. Due to our focus on proteins that are broadly distributed in the gammaproteobacteria, the various proteins identified in this work were all present in different *E. coli* strains for which genome sequences were available. In addition to proteins that were specific for the indicated groups/orders of gammaproteobacteria, we also retained a few proteins where one or two isolated hits from other bacteria had acceptable E-values. We consider these proteins to be also specific for gammaproteobacteria and their presence in isolated unrelated species could be due to lateral gene transfer (LGT) (Doolittle, 1999; Gogarten *et al.*, 2002). For all proteins identified in this study, their protein identification numbers in the *E. coli* K-12 genome, accession numbers and information regarding COG numbers or any conserved domain are presented.

RESULTS AND DISCUSSION

Phylogenetic analysis of gammaproteobacteria

The availability of genomic sequences now makes it possible to examine evolutionary relationships based on concatenated sequences for large numbers of proteins. This approach is more reliable than analysis based on any single gene or protein (Rokas *et al.*, 2003; Brown & Volker, 2004; Belda *et al.*, 2005; Ciccarelli *et al.*, 2006). We have performed phylogenetic analyses for gammaproteobacteria based on the combined sequences for 36 conserved proteins from 45 gammaproteobacterial species covering 13 of its 14 orders (all except for the order *Acidithiobacillales*). The ML phylogenetic tree for the gammaproteobacterial species based on this large dataset is shown in Fig. 1. The proportion of puzzled quartets (ML analysis), or percentage bootstrap scores in MP analysis, which supported different nodes (only values >50 % are shown) are indicated. A NJ tree for this dataset is provided

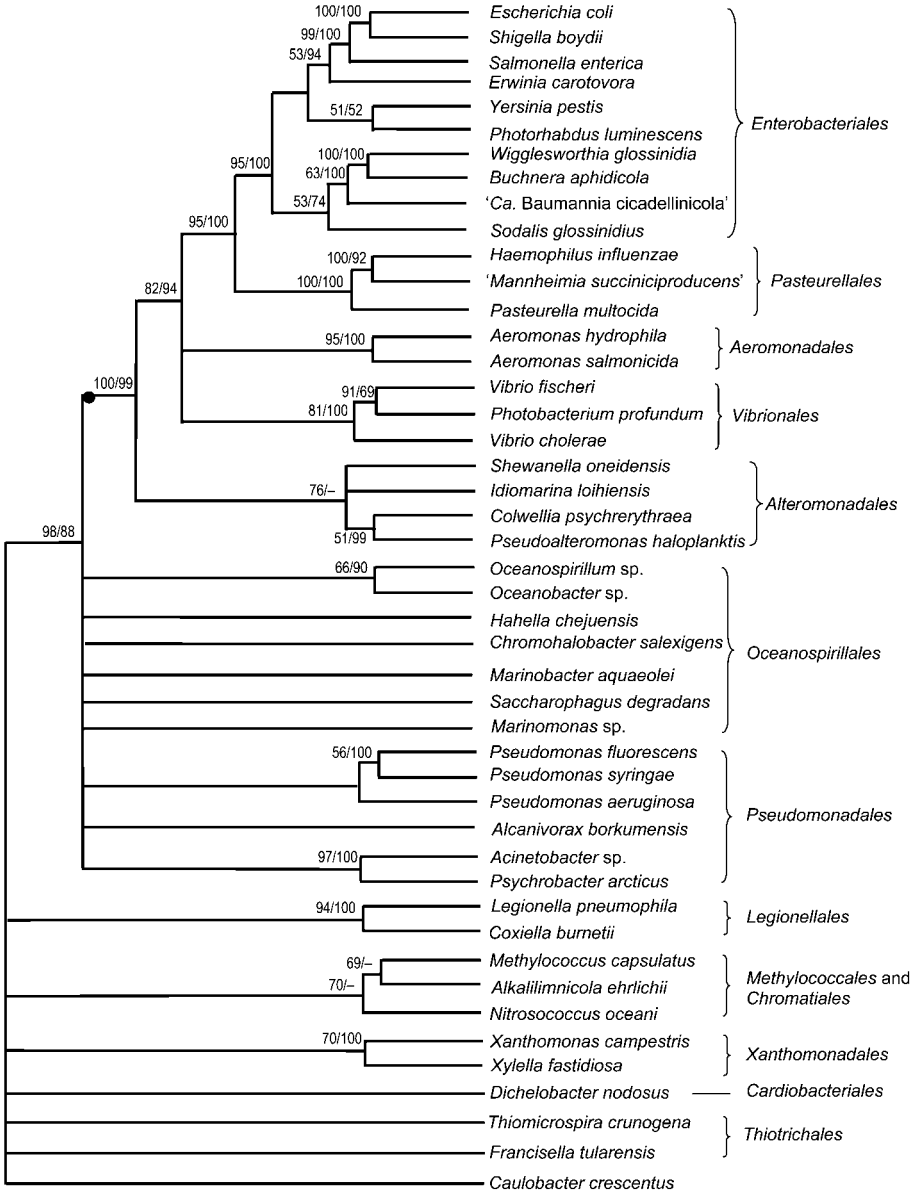

Fig. 1. A maximum-likelihood tree for gammaproteobacteria based on concatenated sequences for 36 proteins. The topology of this tree was very similar to that seen for the maximum-parsimony tree. The two numbers at the nodes (ML/MP) correspond to the proportion of the puzzling quartets (ML analysis) or % bootstrap scores (in the MP tree) that supported the indicated node. Only values above 50 % are shown. The filled circle on a node in this figure identifies the groups of species which uniquely share the RpoB indel shown in Fig. 3.

as Supplementary Fig. S2 (see IJSEM Online). The species from a number of orders of the class *Gammaproteobacteria* (e.g. the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales*, *Aeromonadales*, *Legionellales* and *Xanthomonadales*) formed distinct clades with good statistical support (i.e. relationships supported by >70 % bootstrap samples or puzzling quartets). Based on these trees, a clade consisting of the species of the order *Enterobacteriales* was found to be the most recently diverging lineage within the class *Gammaproteobacteria*. The late divergence of the order *Enterobacteriales* has also been observed in earlier studies (Olsen *et al.*, 1994; Lerat *et al.*, 2003; Brown & Volker, 2004; Ludwig & Klenk, 2005; Belda *et al.*, 2005; Mrazek *et al.*, 2006). Within this clade, various species corresponding to endosymbiotic bacteria (such as *Buchnera aphidicola*, *Wiggelsworthia glossinidia* and '*Candidatus* Baumannia cicadellinicola') formed a distinct deeper-branching cluster. It has been previously shown by Belda *et al.* (2005) that the deep branching of these bacteria is most probably due to their faster rate of evolution in comparison with the free-living enteric bacteria.

The phylogenetic trees also strongly supported a close relationship of the order *Enterobacteriales* to the order *Pasteurellales*. The combined clade of these two orders was linked at a higher level to the clades consisting of species of the orders *Vibrionales* and *Aeromonadales*. Although a clade consisting of these four orders was supported by different phylogenetic methods, the relative branching of the orders *Vibrionales* or *Aeromonadales* with respect to the order *Enterobacteriales–Pasteurellales* clade was not resolved. In the NJ tree (see Supplementary Fig. S2 in IJSEM Online), but not in the ML or MP trees (Fig. 1), these two orders were found to group together with strong bootstrap support. The phylogenetic trees also strongly indicated that the species of the order *Alteromonadales* formed an immediate outgroup of the above four orders. One additional clade that was reliably observed consisted of species from the above five orders as well as various species belonging to the orders *Oceanospirillales* and *Pseudomonadales*. It is noteworthy that species from the orders *Oceanospirillales* and *Pseudomonadales* did not form well-defined clades in the trees, indicating that these orders are phylogenetically heterogeneous. In comparison with these orders, the species from other gammaproteobacterial orders (such as the orders *Thiotrichales*, *Cardiobacteriales*, *Legionellales*, *Chromatiales*, *Methylococcales* and *Xanthomonadales*), consistently showed deeper branching in the trees and their relative branching positions were not resolved.

Phylogenetic trees were also constructed for many individual proteins (particularly those in our set with length >400 aa) and also on a smaller dataset of concatenated sequences for 13 large proteins from this set (see Methods). The relationships observed with this smaller dataset of concatenated protein sequences were identical to those shown here with the larger dataset and the results for the ML/MP tree for this dataset are provided as a supplementary figure (see Supplementary Fig. S3 in IJSEM Online). This smaller dataset of protein sequences was rooted using '*Aquifex aeolicus*' and this rooting did not affect the branching pattern or interrelationships among different gammaproteobacterial orders. The phylogenetic trees for most individual proteins (RpoB, SecA, DnaK, gyraseA, IleRS, SecY, PheRS, RpoA, ArgRS) supported similar relationships as seen here (Fig. 1 and Supplementary Fig. S2), but due to smaller number of positions in these alignments, the bootstrap scores for many nodes were low and not resolved (results not shown). However, in the phylogenetic trees for some proteins (for example, UvrD helicase, GTP binding protein, EF-G and O-sialylglycoprotein endopeptidase), the endosymbiotic bacteria (such as *Buchnera aphidicola*, *W. glossinidia* and '*Ca.* Baumannia cicadellinicola') did not group with other members of the order *Enterobacteriales* and instead they branched deeply in the tree (results not shown). In all of these cases, the branches for these species were very long, which can lead to artefactual deeper branching in the trees (Felsenstein, 1978; Gribaldo & Philippe, 2002; Belda *et al.*, 2005).

Conserved indels that are specific for gammaproteobacteria and their subgroups

Rare genomic changes (RGCs) such as conserved inserts and deletions in genes/proteins that are restricted to species from well-defined taxonomic groups provide a powerful means for inferring as well as confirming evolutionary relationships (Rivera & Lake, 1992; Gupta, 1998; Rokas & Holland, 2000). In many cases, these RGCs have been instrumental in elucidating relationships that were not resolved by phylogenetic trees (Rivera & Lake, 1992; Baldauf & Palmer, 1993; Gupta, 1998; Rokas & Holland, 2000; Kunisawa, 2001). We have identified a number of conserved indels in important housekeeping proteins that are helpful in clarifying the evolutionary relationships among gammaproteobacteria. We previously described two conserved indels in the proteins AICAR-transformylase (PurH) and ribosomal protein L16, which appeared to be restricted to gammaproteobacteria (Gupta, 2000). However, sequence information for these proteins at that time was available for a limited number of gammaproteobacterial species belonging to only certain orders. Hence, it was of importance to re-examine the species distribution of these indels.

Fig. 2 shows the partial sequence alignment of the PurH protein showing the 2 aa deletion that is common to various gammaproteobacteria. As can be seen, this 2 aa deletion, located in a conserved region, is uniquely shared by different gammaproteobacteria, but it is not found in other classes of the phylum *Proteobacteria* or other bacterial phyla. The only gammaproteobacterium in which this indel is absent is *Francisella tularensis*, which corresponds to one of the deepest branches in the phylogenetic tree (Fig. 1 and Supplementary Fig. S2). Although, *F. tularensis* is currently in the order *Thiotrichales* within the class

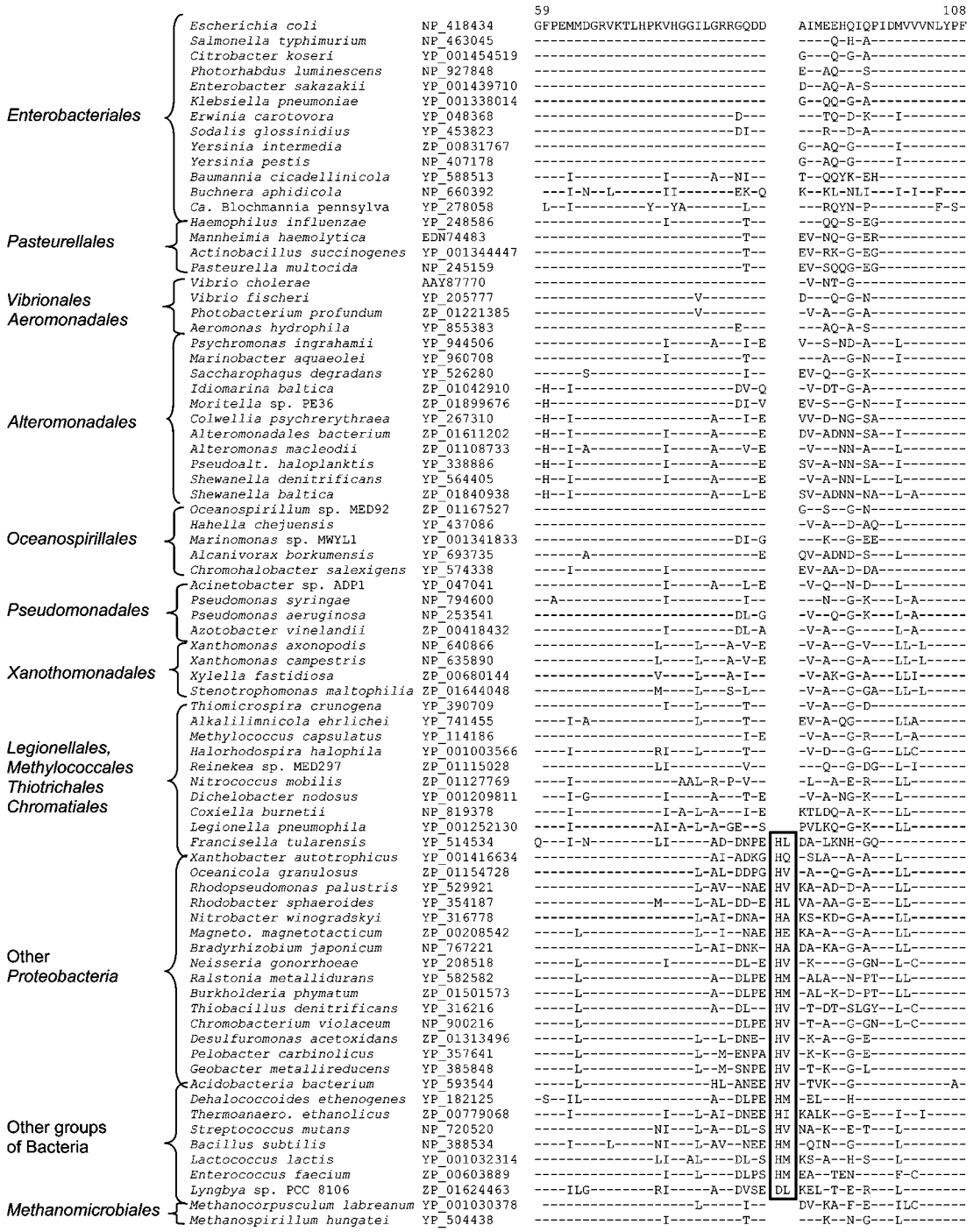

Fig. 2. Partial sequence alignments of AIACR-transformylase (PurH) showing a 2 aa deletion (the corresponding region in other species is boxed) that is uniquely found in various gammaproteobacteria, but is absent in all other bacteria. The dashes (–) in this and other alignments denote identity with the amino acid on the top line. The position of this sequence in *E. coli* protein is marked on the top. A 2 aa deletion in this position is also present in some methanogenic *Archaea* (*Methanomicrobiales*), which is probably of independent origin (see results). Sequence information for only representative species is presented. All other available species from these groups behaved similarly.

Gammaproteobacteria, in phylogenetic trees where members of the class *Betaproteobacteria* are also included, this species forms an outgroup from all of the gamma- and betaproteobacterial species (results not shown). These results indicate that the placement of this species within the class *Gammaproteobacteria* is probably incorrect and

that the absence of the PurH indel in this species may not constitute an exception. However, based upon these results, other possibilities (e.g. this indel occurred after the branching of *F. tularensis* or the *purH* gene was acquired by this species by LGT) cannot be excluded. Nevertheless, the shared presence of this deletion in all gammaproteobacteria except *F. tularensis* (sequence information for >200 gammaproteobacteria is currently available) and its absence in all other bacteria, indicates that the RGC responsible for this deletion probably occurred in a common ancestor of all or most gammaproteobacteria. This RGC thus provides a good molecular marker for this large and important class of proteobacteria. It is interesting to note that besides the class *Gammaproteobacteria*, a 2 aa deletion in this position is also present in three archaeal species belonging to the order *Methanomicrobiales*. The sequences for two of the members of the order *Methanomicrobiales* are shown in the sequence alignment in Fig. 2. In a phylogenetic tree for the PurH sequences, the class *Gammaproteobacteria* and order *Methanomicrobiales* do not group together (results not shown), indicating that the shared absence of this indel in these two groups is not due to LGT, but is very probably due to independent genetic events.

The partial sequence alignment for the ribosomal protein L16 is presented in Supplementary Fig. S4 (see IJSEM Online). Unlike the indel in PAC formyltransferase, the 1 aa deletion in this protein is specifically present in various species from the orders *Enterobacteriales*, *Pasteurellales* and *Vibrionales*, and also several species of the order *Alteromonadales*, but it is not found in other members of the class *Gammaproteobacteria* or other bacterial phyla. This indel supports a close relationship between the species belonging to these orders. The presence of this indel in some species of the order *Alteromonadales* but not others suggests that the species from this order are not phylogenetically homogeneous, a feature also observed in our phylogenetic analysis. In the ML/MP tree shown in Fig. 1, the clade corresponding to the order *Alteromonadales* is weakly supported only by ML analysis and it is not supported by MP analysis. In the NJ tree (see Supplementary Fig. S2), these species do not group together, with *Idiomarina loihiensis* branching deeper than other species of the order *Alteromonadales*.

Two other novel conserved indels that are specific for certain orders or subgroups of gammaproteobacteria were identified in the present study. In the β-subunit of RNA polymerase (RpoB), a 4 aa deletion was uniquely present in various species from the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales*, *Aeromonadales* and *Alteromonadales* (Fig. 3), but it was not found in any other gammaproteobacteria or other groups of bacteria. The genetic change responsible for this indel most probably occurred in a common ancestor of these particular orders after the divergence of other gammaproteobacteria at a stage marked by the filled circle in Fig. 1. Interestingly, this deletion in RpoB was not present in species of the genus *Marinobacter*, which are indicated to belong to the order *Alteromonadales*. However, in the phylogenetic trees shown in Fig. 1 and Supplementary Fig. S2, *Marinobacter aquaeolei* did not group with other species of the order *Alteromonadale*s, but branched outside of the clade comprising these *Alteromonadales* species as well as various species from the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales* and *Aeromonadales*. Both these observations indicate that the genus *Marinobacter* is a deeper branching genus when compared with other genera of the order *Alteromonadales*. Another useful indel for the gammaproteobacteria is present in the protein leucyl-tRNA synthetase (Fig. 4). In this case, a 2 aa deletion is present in various species belonging to the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales*, *Aeromonadales*, *Alteromonadales* and *Oceanospirillales*, but is not found in other gammaproteobacterial orders or in other bacteria. This indel suggests that the species from the order *Oceanospirillales* are more closely related to the above orders in comparison with the order *Pseudomonadales*.

It should be acknowledged that in the present study we have not carried out a comprehensive sequence alignment of all gammaproteobacterial proteins to identify different conserved indels that might be specific for this class of bacteria or for its different subgroups. Hence, it is likely that in future many other conserved indels that are specific for different gammaproteobacterial subgroups will be identified, providing additional molecular markers and further insights into the evolution of these bacteria.

Comparative genomic studies to identify proteins that are specific for gammaproteobacteria

We have also performed systematic BLASTP searches on various ORFs in the *E. coli* K-12 genome to identify proteins that are unique to the gammaproteobacteria at a higher taxonomic level. This genome was chosen as the query because *E. coli* belongs to the order *Enterobacteriales*, which, based upon our phylogenetic analysis (Fig. 1 and Supplementary Figs S2 and S3), is the most recently diverged group/order within the class *Gammaproteobacteria*. Hence, by using probes from this genome, which lies at the 'tip' of the phylogenetic tree, it should be possible to identify proteins that are specific for gammaproteobacteria at different phylogenetic depths. The genome of *E. coli* is also well annotated and extensive functional and gene mutation studies have been conducted on this organism (Blattner *et al.*, 1997; Gerdes *et al.*, 2003; Kang *et al.*, 2004; Chen *et al.*, 2006). The objective of our comparative genomic studies in this work was to identify proteins that were distinctive characteristics for either most species of the order *Enterobacteriales* or were uniquely present in this order as well as other orders of the class *Gammaproteobacteria* (see Methods). Because our query sequences were from *E. coli* K-12, these studies will not have detected certain proteins that might be present in other gammaproteobacteria, but absent in *E. coli* K-12.

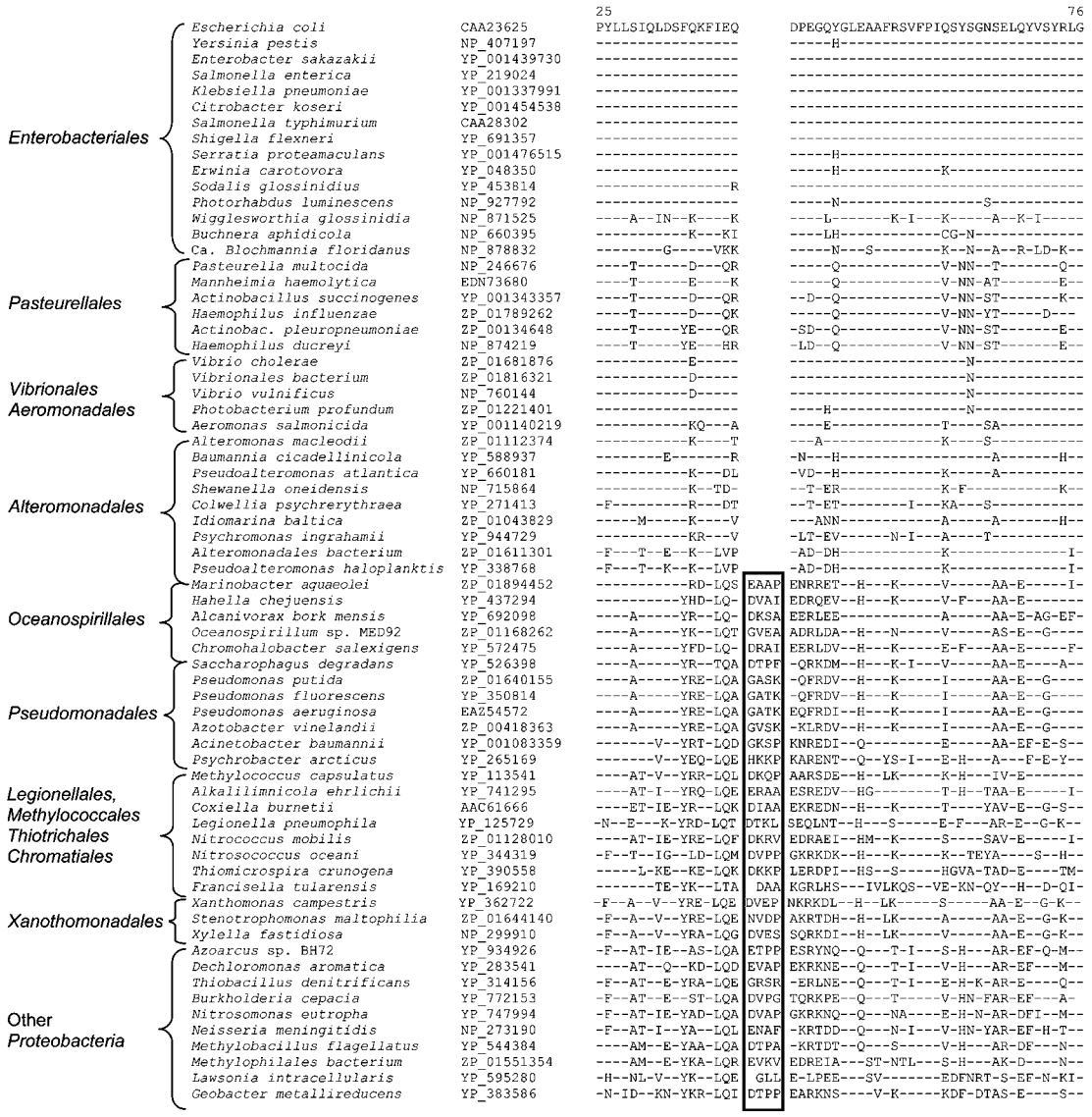

Fig. 3. Partial sequence alignments of RNA polymerase β-subunit (RpoB) showing a 4 aa deletion (corresponding region in other species boxed) that is uniquely found in various species from the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales*, *Aeromonadales* and *Alteromonadales*, but absent in all other gammaproteobacteria or other groups of bacteria. The dashes (–) denote identity with the amino acid on the top line. Sequence information for only representative species is presented.

Likewise, these studies will also not have detected proteins that are specific for other orders of the class *Gammaproteobacteria,* but which are not found in *E. coli* K-12.

Our analyses identified 75 gammaproteobacteria-specific proteins that met these criteria and a brief account of their species distribution and other relevant information is provided. The first five proteins in Table 1(a) are largely specific for the order *Enterobacteriales*. Except for one or two hits mainly from other gammaproteobacteria, all other hits for these proteins are for species of the order *Enterobacteriales*. The next three proteins in this Table are mainly found in various sequenced species of the orders *Enterobacteriales* and *Pasteurellales*. Of these, all significant BLAST hits for proteins b2343 and b3793 are from these two orders, whereas for protein b4481, two hits are also seen for species of the order *Oceanospirillales*. Of these three proteins, b3793, which is annotated as putative ECA polymerase, is essential for *E. coli* cells (Gerdes *et al.*, 2003).

Table 1(b) lists 24 proteins that are mainly restricted to species from the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales* and *Aeromonadales*. Two of these proteins (b0919 and b4311) are only found in various members of the orders *Enterobacteriales* and *Vibrionales*, whereas protein b3790 is present only in the orders *Enterobacteriales*, *Pasteurellales* and *Vibrionales*. Some of

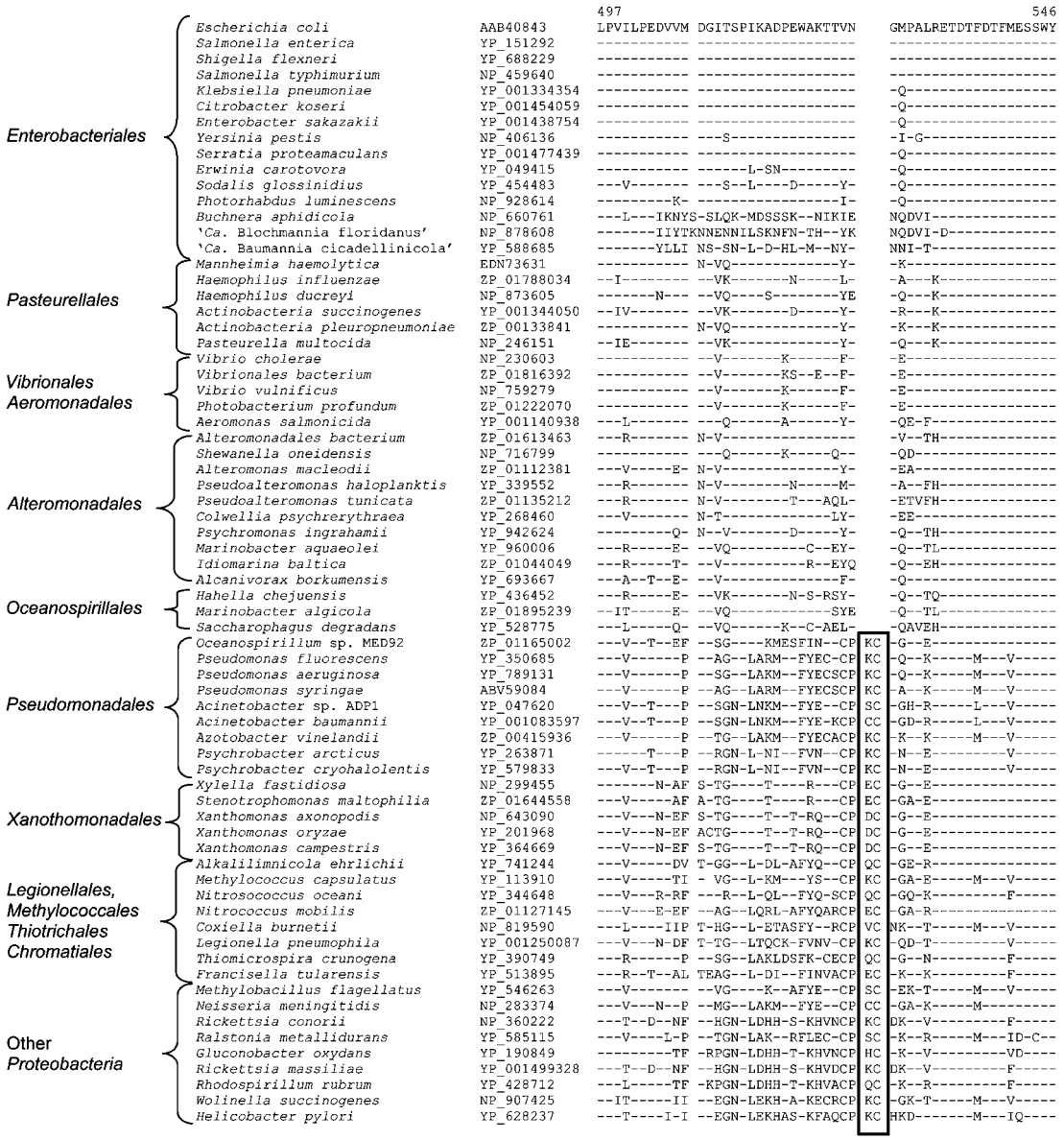

Fig. 4. Partial sequence alignments of leucy-tRNA synthetase (LeuRS) showing a 2 aa deletion (corresponding region in other species boxed) that is uniquely found in various species from the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales*, *Aeromonadales*, *Alteromonadales* and also some *Oceanospirillales*, but absent in all other gammaproteobacteria or other groups of bacteria. The dashes (–) denote identity with the amino acid on the top line.

the proteins listed in Table 1(b) are missing in the order *Aeromonadales* (for example, b0956, b1811, b2510 and b4372) or from both the orders *Aeromonadales* and *Pasteurellales*. The absence of some of these proteins in species of the order *Pasteurellales*, which are obligate parasites, is probably due to gene loss. The species distribution profile of these proteins suggests that species from the orders *Enterobacteriales* and *Pasteurellales* are more closely related to the orders *Vibrionales* and *Aeromonadales* when compared with members of the order *Alteromonadales* or other orders of gammaproteobacteria; this is supported by phylogenetic studies (Fig. 1 and Supplementary Figs S2 and S3). Of the proteins listed in Table 1(b), three proteins, b0922 (MukF), b0923 (MukE) and b0924 (MukB), which are encoded by neighbouring genes, form a complex, MukBEF, which is involved in chromosome partition and DNA repair (Gloyd *et al.*, 2007). Another protein SeqA (b0687), which shows similar species distribution to these three proteins [except that it is also present in some members of the order *Alteromonadales*: Table 1(c)], also interacts with the MukBEF complex in the cell division process (Yamazoe

Table 1. Gammaproteobacteria-specific proteins that are limited to particular orders

The proteins listed in this Table are largely specific for the indicated groups/orders of gammaproteobacteria, as indicated by the BLASTP and PSI-BLAST searches. All of these proteins may not be present in all species from these groups and in some cases they may be entirely missing from certain orders of bacteria (see text). For some of these proteins (marked by superscripts), one or two isolated hits from other bacteria or organisms that are deemed significant are also observed (noted below). For a number of proteins in Table 1(c, d) and Table 2, significant hits are also observed for a single alphaproteobacterial sp. HTCC 2255. This particular species also lacks the gammaproteobacteria-specific indel in the PurH protein, indicating that it is probably a gammaproteobacterium that is incorrectly classified as an alphaproteobacterium.

Gammaproteobacteria-specific proteins			
(a) Proteins specific to the orders *Enterobacteriales* or *Enterobacteriales* and *Pasteurellales*			
b0193 [NP_414735] YaeF		Specific for *Enterobacteriales*	
b0246 [NP_414780] YafW		Specific for *Enterobacteriales* (except *Aeromonas* + *Shewanella*)	
b1726 [NP_416240] YniB		Specific for *Enterobacteriales* (except *Shewanella* + *Pseudomonas*)	
b2004 [NP_416508] YeeU		Specific for *Enterobacteriales* (except *Aeromonas* + *Shewanella*)	
b2586 [NP_417081] YfiM	COG5544	Specific for *Enterobacteriales* (except *Pseudomonas* + 1 betaproteobacterium)	
b2343 [NP_416845] YfcZ		Specific for *Enterobacteriales* + *Pasteurellales*	
b3793 [NP_418241] WecF*		Specific for *Enterobacteriales* + *Pasteurellales*	
b4481 [YP_026257] RffT	CDD70879	Specific for *Enterobacteriales* + *Pasteurellales* (exceptions 2 *Oceanospirillales* species)	
(b) Proteins specific to the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales* and *Aeromonadales*			
b0163 [NP_414705] YaeH		b1811 [NP_416325] YoaH	
b0240 [NP_414775] Crl		b2510 [NP_417005] YfgJ	
b0466 [NP_414999] YbaM		b3790 [YP_026256] RffC	
b0467 [NP_415000] PriC*	COG3923	b3858 [NP_418295] YihD†	
b0735 [NP_415263] YbgE		b3922 [NP_418357] YiiS	COG3691
b0919 [NP_415439] YcbJ		b3964 [NP_418399] YijD	
b0922 [NP_415442] MukF*‡	COG3006	b4151 [NP_418575] FrdD*†	CDD48052
b0923 [NP_415443] MukE*	COG3095	b4372 [NP_418789] Hold	COG3050
b0924 [NP_415444] MukB*	COG3096	b4216 [NP_418637] YtfJ§	COG3054
b0956 [NP_415476] YcbG	COG3120	b2929 [NP_417404] YggD†	COG3722
b1248 [NP_415764] YciU	COG3099	b3601 [NP_418058] MtlR	
b1273 [NP_415789] YciN		b4311 [NP_418731] YjhA	COG1629
(c) Proteins specific to the orders *Enterobacteriales*, *Vibrionales*, *Aeromonadales*, *Pasteurellales* and *Alteromonadales*			
b0119 [NP_414661] YacL	COG3112	b2793 [NP_417273] Syd	CDD70800
b0196 [NP_414738] RcsF		b2831 [NP_417308] MutH	CDD29958
b0685 [NP_415211] YbfE		b2900 [NP_417376] YqfB*	COG3097
b0687 [NP_415213] SeqA*	COG3057	b3466 [NP_417923] YhhL*	
b0946 [NP_415466] YcbW		b3739 [NP_418195] AtpI	COG3312
b0964 [NP_415484] YccT	COG3110	b3764 [NP_418213] YifE	COG3085
b1610 [NP_416127] Tus*	CDD69020	b3938 [NP_418373] MetJ	COG3060
b2187 [NP_416692] YejL	COG3082	b3999 [NP_418427] YjaG	COG3068
b2295 [NP_416798] YfbV		b4217 [NP_418638] YtfK	
b2325 [NP_416828] YfcL		b4255 [NP_418676] YjgD	COG3076
(d) Proteins specific to the orders *Enterobacteriales*, *Pasteurellales*, *Aeromonadales*, *Vibrionales*, *Oceanospirillales*, *Alteromonadales* and *Pseudomonadales*			
b0411 [NP_414945] Tsx	COG3248	b3995 [NP_418423] Rsd*	COG3160
b0953 [NP_415473] Rmf		b4550 [YP_588467] YhdL‖	COG3036
b2792 [NP_417272] YqcC		b4551 [YP_588468] YheV‖¶	
b2944 [NP_417419] SprT	COG3091		
(e) Proteins specific to gammaproteobacteria with sporadic distribution			
b0985 [NP_415505] YmcB	CDD69756	b2626 [NP_417115] YfjJ	CP4-57 prophage
b0986 [NP_415506] YmcC		b2969 [NP_417443] GspL	COG3297
b1036 [NP_415554] YcdZ*#	CDD69987	b3369 [NP_417828] YhfL	
b2252 [NP_416755] Ais		b3382 [NP_417841] YhfY	
b2419 [NP_416914] YfeK		b4028 [NP_418452] YjbG	
b2602 [NP_417093] YfiL		b4181 [NP_418602] YjfI	COG3789

*These proteins are indicated to be essential for the survival of *E. coli* cells (Gerdes *et al.*, 2003).
†Exception: *Psychromonas* sp. CNPT3.
‡Exception: *Sorangium cellulosum*.
§Exceptions: *Nitratiruptor* sp. SB155-2 and *Sulurovum* sp. NBC37-1.
‖Exception: *Neisseria meningitidis* MC58.
¶Exception: *Anopheles gambiae*.
#Exception: *Clostridia* sp.

et al., 2005). All four of these proteins, as well as two other proteins, b0467 and b4151, which are annotated as primosomal replication protein N and the fumarate reductase subunit D, respectively, are essential for the growth of *E. coli* cells (Gerdes *et al.*, 2003). The species distribution profiles of the Muk and SeqA proteins indicate that this novel mechanism for chromosome partition, which is limited to only certain orders of gammaproteobacteria, evolved very late in evolution.

Table 1(c) lists 20 proteins that we consider to be mainly specific for members of the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales*, *Aeromonadales* and *Alteromonadales*. Several of these proteins, such as b0119, b0687, b0964, b2187, b2295, b2900 and b3938, are entirely specific for these orders, whereas a number of others are also found in one or two species from other groups. The absence of some of these proteins in the order *Pasteurellales* is again presumably due to gene loss. The species distribution profiles of these proteins suggest that species from these orders shared a common ancestor exclusive of other gammaproteobacteria. This conclusion is strongly supported by phylogenetic analyses (Fig. 1 and Supplementary Figs S2 and S3) and by the conserved indel in the RpoB protein (Fig. 3). Five of these proteins (b0163, b0466, b3764, b3999 and b4255) are also present in some of the species from the order *Oceanospirillales*, indicating that species from this order show a close relationship to these other orders, a conclusion also supported by the signature indel in the LeuRS protein (Fig. 4). Of the proteins in Table 1(c), in addition to SeqA (b0687), three further proteins, b1610, b2900 and b3466, are essential for the survival of *E. coli* cells (Gerdes *et al.*, 2003). Of these proteins, b1610 (Tus) is annotated as a DNA replication terminus site binding protein, whereas the functions of the other two are unknown.

Seven additional proteins in Table 1(d) (b0411, b0953, b2792, b2944, b3995, b4550 and b4551) are commonly shared by most of the species from the orders *Enterobacteriales*, *Pasteurellales*, *Vibrionales*, *Aeromonadales*, *Alteromonadales*, *Oceanospirillales* and *Pseudomonadales*. The presence of these proteins suggests that species from these orders shared a common ancestor exclusive of other gammaproteobacteria, which is in accordance with our phylogenetic analyses (Fig. 1 and Supplementary Figs S2 and S3). It is of interest to note that for a number of proteins in Table 1(c, d) and Table 2, significant BLAST hits are also observed for a single alphaproteobacterial strain, sp. HTCC 2255. This species also contains a 2 aa deletion in the PurH protein, which is specific for the class *Gammaproteobacteria*, thus making a strong case for its grouping with the class *Gammaproteobacteria* rather than the class *Alphaproteobacteria*. Twelve additional proteins listed in Table 1(e) are also specific for gammaproteobacteria, but they are present sporadically in species from a number of different orders. The species distributions of these proteins can be accounted for by their evolution at various stages in

Table 2. Proteins specific for most gammaproteobacteria

The four proteins listed in this Table are uniquely found in the broadest range of gammaproteobacteria. All significant BLAST hits for these proteins were from gammaproteobacteria. The first column indicates the number of sequenced genomes from different orders of the class *Gammaproteobacteria*. Many of these entries are for different strains of the same species (e.g. of the eight genomes of the order *Thiotrichales*, seven are for *F. tularensis*). None of these proteins are found in *F. tularensis* and the grouping of this species with the class *Gammaproteobacteria* is questionable (see text). The numbers in different columns under various proteins indicate the number of genomes from different orders where significant BLAST hit to the query protein was observed. The header row indicates the ID number of the protein from the *E. coli* K-12 genome. The accession numbers and the COG numbers for these proteins are given in the first and second rows. The cellular functions of these proteins are not known. However, the proteins marked with * are essential for the survival of *E. coli* cells (Gerdes *et al.*, 2003).

	b0354	**b1132***	**b1179**	**b3033***
Accession no.	[NP_414888]	[NP_415650]	[NP_415697]	[NP_417505]
COG or CDD no.	COG3122	COG2915	COG3100	COG3251
Gene name	*yaiL*	*ycfC*	*ycgL*	*yqiB*
Enterobacteriales (48)	38	42	40	39
Pasteurellales (10)	0	10	10	0
Vibrionales (8)	4	8	8	8
Aeromonadales (2)	0	2	2	2
Alteromonadales (22)	21	22	22	21
Oceanospirillales (3)	2	3	3	3
Pseudomonadales (19)	19	15	19	14
Legionellales (6)	0	0	0	0
Methylococcales (1)	1	1	0	1
Chromatiales (3)	2	3	1	2
Thiotrichales (8)	1	1	1	0
Xanthomonadales (8)	8	8	6	8

the divergence of gammaproteobacteria (as noted above for other proteins) followed by gene losses in specific species or lineages.

Lastly, we describe four proteins (b0354, b1132, b1179 and b3033) that are present in most of the gammaproteobacteria, but which are not found in any other bacteria (see Table 2). Except for a few orders that contain either intracellular or parasitic bacteria, these proteins are present in the majority of the sequenced genomes from other orders of gammaproteobacteria. These proteins are also not found in different strains of *F. tularensis*, again supporting our contention that the grouping of this species with gammaproteobacteria is incorrect. Of all the gammaproteobacteria-specific proteins identified in our analyses, these four proteins show the broadest species distribution and we suggest that their genes first evolved in a common ancestor of all of the gammaproteobacteria, followed by gene losses in certain groups, where their cellular functions were not required. Of these four proteins, b1132 and b3033 are essential for the survival of *E. coli* cells (Gerdes *et al.*, 2003). Although some of these proteins have been assigned to specific COG groups (Tatusov *et al.*, 2000), their cellular functions are not known at present.

Main inferences from phylogenomic and comparative genomic analyses

The results of our analyses indicate that the main orders within the class *Gammaproteobacteria* have branched or diverged in the following order (from earliest to most recent): *Thiotrichales*, *Cardiobacteriales*, *Xanthomonadales*, *Chromatiales*, *Legionellales*, *Methylococcales* >*Pseudomonadales*, *Oceanospirillales* >*Alteromonadales* >*Aeromonadales*, *Vibrionales* >*Pasteurellales* >*Enterobacteriales*. While the positions of the late branching orders are clearly resolved, the relationships amongst the early branching groups remain unclear. This branching order is supported not only by phylogenetic trees based on a large number of proteins, but it is also independently supported by the identification of a number of conserved indels and many proteins for which the RGCs or genes were introduced after some of the major branch points in this scheme.

The gammaproteobacteria have previously been characterized solely on the basis of their branching pattern in trees based on 16S rRNA gene sequences. However, our results show that the 2 aa deletion in the PurH protein is a distinctive characteristic of all gammaproteobacteria (>240 entries currently in the database), with the sole exception of *F. tularensis*, whose grouping with other gammaproteobacteria is questionable. The indel in the PurH protein thus provides the first known molecular marker that can be used to define and circumscribe the class *Gammaproteobacteria*. We have also identified four proteins (b0354, b1132, b1179 and b3033) that are uniquely found in most gammaproteobacterial species; the main exceptions being in the endosymbiotic or parasitic bacteria. Although these proteins are not present in all gammaproteobacteria, they also provide novel and useful molecular markers for this large and diverse group.

Of the 75 proteins that are specific for either the order *Enterobacteriales* or higher clades within the class *Gammaproteobacteria*, most are of unknown functions with a few exceptions. A number of these proteins are essential for the growth of *E. coli* cells (see Tables 1 and 2) (Gerdes *et al.*, 2003). The remainder of these proteins, although they are not required for growth under laboratory conditions, are also expected to be important for these bacteria in their natural environments based on their high degree of conservation and persistence (Fang *et al.*, 2005). It is thus of great importance to understand the cellular functions of these unique and broadly distributed proteins. In addition to providing significant insights in to possible novel biochemical or physiological characteristics that are common to many or all gammaproteobacteria, they may also provide potential drug targets for a large group of disease-causing bacteria which belong to this class.

ACKNOWLEDGEMENTS

This work was supported by a research grant from the Canadian Institute of Health Research. R. M. was a visiting student from the University of Sydney, Australia.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein databases search programs. *Nucleic Acids Res* **25**, 3389–3402.

Baldauf, S. L. & Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* **90**, 11558–11562.

Belda, E., Moya, A. & Silva, F. J. (2005). Genome rearrangement distances and gene order phylogeny in γ-proteobacteria. *Mol Biol Evol* **22**, 1456–1467.

Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., Hampson, D. J., Bellgard, M., Wassenaar, T. M. & Ussery, D. W. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics* **6**, 165–185.

Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K. & other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462.

Brenner, D. J., Krieg, N. R., Staley, J. T. & Garrity, G. M. (2005). *Bergey's Manual of Systematic Bacteriology*. New York: Springer.

Brown, J. R. & Volker, C. (2004). Phylogeny of γ-proteobacteria: resolution of one branch of the universal tree? *Bioessays* **26**, 463–468.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540–552.

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R. R. & other authors (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc Natl Acad Sci U S A* **103**, 5977–5982.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287.

Daubin, V. & Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res* **14**, 1036–1042.

De Ley, J. (1992). The Proteobacteria: ribosomal RNA cistron similarities and bacterial taxonomy. In *The Prokaryotes*, pp. 2111–2140. Edited by A. Balows, H. G. Trüper, M. Dworkin, W. Harder & K. H. Schleifer. New York: Springer-Verlag.

Deng, W., Liou, S. R., Plunkett, G., III, Mayhew, G. F., Rose, D. J., Burland, V., Kodoyianni, V., Schwartz, D. C. & Blattner, F. R. (2003). Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J Bacteriol* **185**, 2330–2337.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128.

Edwards, R. A., Olsen, G. J. & Maloy, S. R. (2002). Comparative genomics of closely related salmonellae. *Trends Microbiol* **10**, 94–99.

Fang, G., Rocha, E. & Danchin, A. (2005). How essential are nonessential genes? *Mol Biol Evol* **22**, 2147–2156.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* **27**, 401–410.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B A. & other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Gao, B., Parmanathan, R. & Gupta, R. S. (2006). Signature proteins that are distinctive characteristics of actinobacteria and their subgroups. *Antonie van Leeuwenhoek* **90**, 69–91.

Garrity, G. M., Bell, J. A. & Lilburn, T. G. (2005). The revised road map to the manual. In *Bergey's Manual of Systematic Bacteriology, Volume 2, Part A, Introductory Essays*, pp. 159–220. Edited by D. J. Brenner, N. R. Krieg & J. T. Staley. New York: Springer.

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I. & other authors (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol* **185**, 5673–5684.

Gloyd, M., Ghirlando, R., Matthews, L. A. & Guarne, A. (2007). MukE and MukF form two distinct high affinity complexes. *J Biol Chem* **282**, 14373–14378.

Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226–2238.

Gribaldo, S. & Philippe, H. (2002). Ancient phylogenetic relationships. *Theor Popul Biol* **61**, 391–408.

Griffiths, E., Ventresca, M. S. & Gupta, R. S. (2006). BLAST screening of chlamydial genomes to identify signature proteins that are unique for the *Chlamydiales*, *Chlamydiaceae*, *Chlamydophila* and *Chlamydia* groups of species. *BMC Genomics* **7**, 14.

Gupta, R. S. (1998). Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* **62**, 1435–1491.

Gupta, R. S. (2000). The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol Rev* **24**, 367–402.

Gupta, R. S. (2006). Molecular signatures (unique proteins and conserved indels) that are specific for the epsilon proteobacteria (*Campylobacterales*). *BMC Genomics* **7**, 167.

Gupta, R. S. & Lorenzini, E. (2007). Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the bacteroidetes and chlorobi species. *BMC Evol Biol* **7**, 71.

Gupta, R. S. & Mok, A. (2007). Phylogenomics and signature proteins for the alpha proteobacteria and its main groups. *BMC Microbiol* **7**, 106.

Gupta, R. S. & Sneath, P. H. A. (2007). Application of the character compatibility approach to generalized molecular sequence data: branching order of the proteobacterial subdivisions. *J Mol Evol* **64**, 90–100.

Howard, S. L., Gaunt, M. W., Hinds, J., Witney, A. A., Stabler, R. & Wren, B. W. (2006). Application of comparative phylogenomics to study the evolution of *Yersinia enterocolitica* and to identify genetic differences relating to pathogenicity. *J Bacteriol* **188**, 3645–3653.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998). Multiple sequence alignment with CLUSTAL_X. *Trends Biochem Sci* **23**, 403–405.

Kang, Y., Durfee, T., Glasner, J. D., Qiu, Y., Frisch, D., Winterberg, K. M. & Blattner, F. R. (2004). Systematic mutagenesis of the *Escherichia coli* genome. *J Bacteriol* **186**, 4921–4930.

Kersters, K., Devos, P., Gillis, M., Swings, J., Vandamme, P. & Stackebrandt, E. (2006). Introduction to the Proteobacteria. In *The Prokaryotes: A Handbook on the Biology of Bacteria*, pp. 3–37. Edited by M. Dworkin, S. Falkow, E. Rosenberg, K. H. Schleifer & E. Stackebrandt. New York: Springer.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Koonin, E. V. & Galperin, M. Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev* **7**, 757–763.

Kunisawa, T. (2001). Gene arrangements and phylogeny in the class *Proteobacteria*. *J Theor Biol* **213**, 9–19.

Lee, H.-Y. & Côté, J. C. (2006). Phylogenetic analysis of γ-proteobacteria inferred from nucleotide sequence comparisons of the house-keeping genes *adk*, *aroE* and *gdh*: comparisons with phylogeny inferred from 16S rRNA gene sequences. *J Gen Appl Microbiol* **52**, 147–158.

Lerat, E., Daubin, V. & Moran, N. A. (2003). From gene trees to organismal phylogeny in prokaryotes: the case of the γ-proteobacteria. *PLoS Biol* **1**, E19.

Ludwig, W. & Klenk, H.-P. (2005). Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In *Bergey's Manual of Systematic Bacteriology*, pp. 49–65. Edited by D. J. Brenner, N. R. Krieg, J. T. Staley & G. M. Garrity. Berlin: Springer-Verlag.

Mrazek, J., Spormann, A. M. & Karlin, S. (2006). Genomic comparisons among γ-proteobacteria. *Environ Microbiol* **8**, 273–288.

Olsen, G. J., Woese, C. R. & Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* **176**, 1–6.

Rivera, M. C. & Lake, J. A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257**, 74–76.

Rokas, A. & Holland, P. W. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol* **15**, 454–459.

Rokas, A., Williams, B. L., King, N. & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804.

Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y.I., Koonin, E.V. & Altschul, S. F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* **29**, 2994–3005.

Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.

Stackebrandt, E., Murray, R. G. E. & Trüper, H. G. (1988). *Proteobacteria* classis nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives". *Int J Syst Bacteriol* **38**, 321–325.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596–1599.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33–36.

Van de Peer, Y. & De Wachter, R. (1994). TREECON for windows: a software package for the construction and drawing of evolutionary trees for the Microsoft windows environment. *Comput Appl Biosci* **10**, 569–570.

Van Sluys, M. A., Monteiro-Vitorello, C. B., Camargo, L. E., Menck, C. F., Da Silva, A. C., Ferro, J. A., Oliveira, M. C., Setubal, J. C., Kitajima, J. P. & Simpson, A. J. (2002). Comparative genomic analysis of plant-associated bacteria. *Annu Rev Phytopathol* **40**, 169–189.

Whittam, T. S. & Bumbaugh, A. C. (2002). Inferences from whole-genome sequences of bacterial pathogens. *Curr Opin Genet Dev* **12**, 719–725.

Woese, C. R., Weisburg, W. G., Hahn, C. M., Paster, B. J., Zablen, L. B., Lewis, B. J., Macke, T. J., Ludwig, W. & Stackebrandt, E. (1985). The phylogeny of purple bacteria: the gamma subdivision. *Syst Appl Microbiol* **6**, 25–33.

Yamazoe, M., Adachi, S., Kanaya, S., Ohsumi, K. & Hiraga, S. (2005). Sequential binding of SeqA protein to nascent DNA segments at replication forks in synchronized cultures of *Escherichia coli*. *Mol Microbiol* **55**, 289–298.