Molecular Phylogenetics and Evolution 75 (2014) 103–117

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Ribosomal proteins: Toward a next generation standard for prokaryotic systematics?

Hemalatha Golaconda Ramulu [a,1], Mathieu Groussin [b,3], Emmanuel Talla [a], Remi Planel [a,2], Vincent Daubin [b], Céline Brochier-Armanet [b,*]

[a] Aix-Marseille Université, CNRS, UMR 7283, Laboratoire de Chimie Bactérienne, IMM, 31 chemin Joseph Aiguier, F-13402 Marseille, France
[b] Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France

ARTICLE INFO

Article history:
Received 15 October 2013
Revised 23 January 2014
Accepted 17 February 2014
Available online 28 February 2014

Keywords:
Non-homogeneous models
Horizontal gene transfer
Endosymbionts
Long branch attraction
SAR11
Compositional bias

ABSTRACT

The seminal work of Carl Woese and co-workers has contributed to promote the RNA component of the small subunit of the ribosome (SSU rRNA) as a "gold standard" of modern prokaryotic taxonomy and systematics, and an essential tool to explore microbial diversity. Yet, this marker has a limited resolving power, especially at deep phylogenetic depth and can lead to strongly biased trees. The ever-larger number of available complete genomes now calls for a novel standard dataset of robust protein markers that may complement SSU rRNA. In this respect, concatenation of ribosomal proteins (r-proteins) is being growingly used to reconstruct large-scale prokaryotic phylogenies, but their suitability for systematic and/or taxonomic purposes has not been specifically addressed. Using *Proteobacteria* as a case study, we show that amino acid and nucleic acid r-protein sequences contain a reliable phylogenetic signal at a wide range of taxonomic depths, which has not been totally blurred by mutational saturation or horizontal gene transfer. The use of accurate evolutionary models and reconstruction methods allows overcoming most tree reconstruction artefacts resulting from compositional biases and/or fast evolutionary rates. The inferred phylogenies allow clarifying the relationships among most proteobacterial orders and families, along with the position of several unclassified lineages, suggesting some possible revisions of the current classification. In addition, we investigate the root of the *Proteobacteria* by considering the time-variation of nucleic acid composition of r-protein sequences and the information carried by horizontal gene transfers, two approaches that do not require the use of an outgroup and limit tree reconstruction artefacts. Altogether, our analyses indicate that r-proteins may represent a promising standard for prokaryotic taxonomy and systematics.

1. Introduction

Reconstructing the evolutionary relationships among prokaryotic organisms is a major challenge in biology and key to many research fields, including evolution, ecology, and medicine (Gribaldo and Brochier, 2009). Since the seminal work of Carl Woese and colleagues at the end of the 70s (Woese and Fox, 1977), prokaryotic systematics and the exploration of microbial diversity have been relying mainly on phylogenetic analysis of the RNA component of the small subunit of the ribosome (SSU rRNA) (Lopez-Garcia and Moreira, 2008). However, single gene markers (including SSU rRNA) are not able to resolve all phylogenetic relationships with confidence, especially the most ancient ones (Gribaldo and Philippe, 2002). Taking the opportunity of the recent burst of complete genome sequencing projects, new phylogenetic approaches based on gene content, gene order, DNA-string comparison, shared rare genomic events, etc. have been developed (see (Delsuc et al., 2005) and references therein). Among them, special emphasis has been put on the simultaneous analysis of numerous protein coding genes through supertrees or supermatrices, which systematically provide better resolved trees than those based on single markers (including SSU rRNA) (Abby et al., 2012). In the case of prokaryotes, implementing such approaches may

* Corresponding author. Past address: Aix-Marseille Université, CNRS, UMR 7283, Laboratoire de Chimie Bactérienne, IMM, 31 chemin Joseph Aiguier, F-13402 Marseille, France.
E-mail address: celine.brochier-armanet@univ-lyon1.fr (C. Brochier-Armanet).
[1] Present address: Aix-Marseille Université, CNRS, UMR 7257, Laboratoire Architecture et Fonction des Macromolécules Biologiques, Campus de Luminy, 163 Avenue de Luminy, F-13288 Marseille CEDEX 09, France.
[2] Present address: Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France.
[3] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States.

http://dx.doi.org/10.1016/j.ympev.2014.02.013
1055-7903/

be complicated by horizontal gene transfer (HGT) which make the evolutionary history of genes different from that of organisms (Gribaldo and Brochier, 2009; Philippe and Douady, 2003). However, even if HGT is a very important evolutionary process, large scale phylogenetic analyses indicate that a congruent phylogenetic signal reflecting the phylogeny of organisms can be extracted from protein markers (Brochier et al., 2002; Lerat et al., 2003; Matte-Tailliez et al., 2002; Puigbo et al., 2010).

In the post genomic era, the first step of most of studies aiming at investigating the phylogeny of a particular taxonomic group consists in identifying the best-suited set of genes (or proteins) to address the question. Among them, the analysis of the core genome (i.e. the set of orthologous genes present in a single copy per genome) is becoming growingly popular (Kelly et al., 2010; Lang et al., 2013; Touchon et al., 2009; Williams et al., 2010). This is because core genes are numerous, easy to identify in complete genomes and supposed to be less affected by HGT, gene duplications and losses. Such approaches have led to significant improvements on our knowledge of the evolutionary history of prokaryotes, in particular by allowing resolving difficult evolutionary issues such as the origin of obligate enterobacteriales endosymbionts of insects (Husnik et al., 2011). Accordingly, we are witnessing a bloom of methodological developments and databases allowing the selection of core genes at different taxonomic levels (see for instance (Kuzniar et al., 2008; DeLuca et al., 2012; Marthey et al., 2008; Wang and Wu, 2013; Wu et al., 2013)). However, the application of such strategies to systematics or taxonomic purposes deserves careful consideration. Indeed, the identification of core genes strongly depends on the taxonomic sampling of the lineage under study. For instance, the core genome of *Escherichia coli* varies according to the number and the type of strains considered (Touchon et al., 2009). Moreover, core genomes are not comparable from one lineage to another. For example, the core genome of *Escherichia* (Touchon et al., 2009) is different in size and gene content from that of *Mycobacterium* (Tettelin et al., 2005). Finally, the methods (e.g. sequence similarity, single gene phylogeny, etc.) and parameters used to identify orthologous genes/proteins can also strongly influence the delineation of core genomes (Kuzniar et al., 2008). Altogether, differences in gene sampling make the comparison among different studies and the discrepancies observed from one study to the other difficult to interpret from a systematic and/or taxonomic point of view. In addition, the identification of the core genome can be time-consuming, and because it depends on the taxonomic sampling of the lineages under study, it has to be recomputed prior to each analysis. Finally, the combination of core genes can lead to very large supermatrices of characters that can impose a very heavy burden in calculation time, precluding their use to address routinely taxonomic and systematic issues.

There is therefore an urgent need to define a stable and standardised set of molecular markers suitable to address systematic and taxonomic issues that overcomes these major issues. These markers should (i) be largely conserved across prokaryotic lineages, (ii) be easily identifiable in complete genome sequences, (iii) be rarely transferred, and (iv) harbour a robust and reliable phylogenetic signal at various taxonomic levels. Among protein markers, those involved in translation, and in particular ribosomal proteins (r-proteins), fulfil most of these criteria. Indeed, while a few cases of HGTs have been reported (Brochier et al., 2000; Chen et al., 2009), r-proteins carry a robust phylogenetic signal that can be used to reconstruct ancient phylogenies in the three domains of life (Brochier et al., 2002; Brown et al., 2001; Ciccarelli et al., 2006; Matte-Tailliez et al., 2002; Swithers et al., 2009). Moreover, a growing number of mass spectrometry studies showed that r-proteins can be used to discriminate among closely related species and to type/identify rapidly strains within species (Suarez et al., 2013; Tamura et al., 2013). Finally, it is worth noting that at very large evolutionary scales (e.g. domain or phylum levels), core genomes are mainly composed of r-proteins (Ciccarelli et al., 2006). However, the resolving power of these markers for systematic and/or taxonomic purposes has not been specifically tested. In this study, we evaluate the value of r-proteins for prokaryotic taxonomy and systematics by using *Proteobacteria* as a case study.

Proteobacteria (from the Greek god "Proteus", who was capable of assuming many different shapes (Stackebrandt et al., 1988)) represent the largest and phenotypically most diverse bacterial lineage. It encompasses the majority of known Gram-negative bacteria (Kersters et al., 2006), shows a wide diversity of metabolisms and morphologies, and includes a large number of human, animal, and plant symbionts/pathogens of ecological, medical, industrial, and agricultural interest. *Proteobacteria* are also highly relevant from an evolutionary point of view, as the endosymbiosis of the alphaproteobacterial ancestor of mitochondria represents a key step in eukaryogenesis (Embley et al., 2003; Lang et al., 1999). The importance of this phylum is exemplified by the huge number of complete genome sequences in public databases: they represented 40% of bacterial complete genome sequences available at the NCBI in August 2013 (http://www.ncbi.nlm.nih.gov/). Based on SSU rRNA and protein analyses, *Proteobacteria* have been divided into five classes, which were arbitrarily designated as Alpha, Beta, Gamma, Delta and Epsilon (see (Kersters et al., 2006). Recently, a new candidate division, the *Zetaproteobacteria*, has been proposed for *Mariprofundus ferrooxydans* PV-1 and JV-1 strains, the first cultivated neutrophilic Fe-oxidising bacteria, and their uncultivated relatives (Emerson et al., 2007). While proteobacterial classes are well defined, the location of the root of *Proteobacteria* and the relationships among the orders and families composing each class have not been fully clarified. In particular, the phylogenetic position of a number of newly described lineages and of fast evolving species has remained elusive or is hotly debated. Due to their diversity and abundance, *Proteobacteria* represent an interesting case study to assess the value of r-proteins for prokaryotic taxonomic and systematic studies.

2. Materials and methods

2.1. Data set construction

A subset of 472 proteomes representative of proteobacterial diversity was downloaded at the NCBI (http://www.ncbi.nlm.nih.gov/) (Supplementary Table S1). The sequences were gathered in a local database. The sequences from the recently published genome of *Magnetospira* sp. QH-2 (Ji et al., 2013) and from the magnetoovoid strain MO-1 (a second representative of *Magnetococcales*, *Alphaproteobacteria*) ongoing project were kindly provided by Dr. Long-Fey Wu (personal communication) and included in the database. The database was screened with BLASTP (Altschul et al., 1997) to identify the homologues of the 55 proteobacterial r-proteins (33 LSU and 22 SSU r-proteins) using *Escherichia coli* sequences as seeds. The absence of any r-protein in a given proteome was verified by screening the nucleic acid sequence of the corresponding genome with TBLASTN. Accession numbers of retrieved r-proteins sequences are given in Supplementary Table S1. The nucleotide sequences corresponding to these r-proteins were also retrieved from the NCBI.

The 55 resulting datasets were aligned using MUSCLE 3.6 (Edgar, 2004). The use of other programs (i.e. CLUSTALW (Larkin et al., 2007) and MAFFT (Katoh and Toh, 2008)) provided very similar multiple alignments (not shown). The resulting alignments were used as template to align the corresponding nucleic acid sequences. At this step, the r-protein S1 was discarded due to the

presence of numerous repeats preventing the construction of an accurate alignment. The 54 amino acid and nucleic acid alignments were visually inspected and adjusted when necessary using ED from the MUST package (Philippe, 1993). Protein and nucleic acid alignments were trimmed using the NET application from the MUST package (Philippe, 1993).

2.2. Supermatrix construction

The trimmed alignments of individual r-proteins from the 474 proteomes were combined to build supermatrices. When a proteome harboured several copies of a given r-protein, the less divergent homologue was retained. To analyse the impact of data sampling on supermatrix reconstruction, we constructed 100 supermatrices using the 33 LSU r-proteins and 100 supermatrices using the 21 SSU r-proteins by gathering alignments containing an increasing number of missing homologues (from 0 up to 99). Examination of the resulting supermatrices suggested that for LSU and SSU r-proteins a maximum of ten missing species represented a good compromise between the amount of missing data and the length of the resulting alignments (upper graph, Supplementary Fig. S1). They corresponded to the concatenation of 28 LSU and 20 SSU r-proteins. As expected, the Maximum Likelihood (ML) trees inferred with these two supermatrices showed similar topologies (not shown), confirming that LSU and SSU r-proteins carry a consistent phylogenetic signal. The two supermatrices were therefore combined into a single alignment (FAA-474) representing 5,228 amino acid positions. The maximum likelihood phylogeny inferred with the FAA-474 is shown in Supplementary Fig. S2.

We constructed a second set of supermatrices using a more restricted taxonomic sampling (Supplementary Table S1) representing at best the taxonomic diversity of each proteobacterial class and its genetic diversity according to the ML phylogeny inferred with the FAA-474 supermatrix (Supplementary Fig. S2), and avoiding the over representation of a few taxa. To do so we selected at least one representative of each proteobacterial family, but no more than two representatives per genus, excepted in the case of Buchnera for which we selected three representatives due to the extreme divergence of strains within this lineage. Altogether this procedure led us to select 137 organisms. The supermatrices were built by allowing from 0 up to ten missing species per r-protein family. The examination of the resulting supermatrices showed that a maximum of three missing species represented the best compromise between the amount of missing data and the length of the resulting alignments (lower graph, Supplementary Fig. S1). They correspond to the concatenation of 27 LSU r-proteins and 19 SSU r-proteins. The ML trees inferred with these two supermatrices were consistent and in agreement with those inferred with the 474 species (Supplementary Figs. S3-S4 and S2, respectively). This confirmed that LSU and SSU r-proteins carried a consistent phylogenetic signal and indicated that the reduction of the taxonomic sampling did not bias the phylogenetic signal contained in these markers. The two supermatrices were combined into a single alignment (FAA-137) containing 5,124 amino acid positions. The nucleic acid version of this supermatrix will be referred to as FNT-137. The supermatrices are available upon request to CB-A or on the treeBASE repository (http://treebase.org/treebase-web/home.html).

2.3. Phylogenetic analyses

ML phylogenies of individual r-proteins were inferred using PhyML v. 3.1 (Guindon et al., 2010) with the Le and Gascuel (LG) model (Le and Gascuel, 2008). In order to take into account the heterogeneity of evolutionary rates across sites, we used a gamma distribution with four discrete classes of sites (Γ_4) and an estimated alpha parameter. The branch robustness of the ML trees was estimated with the non-parametric bootstrap procedure implemented in PhyML v.3.1 (100 replicates of the original dataset).

ML trees of the supermatrices were inferred with PhyML v.3.1 (Guindon et al., 2010). The best fitted evolutionary models were selected with ProtTest v2.4 (Abascal et al., 2005) for the amino acid supermatrices (FAA-474 and FAA-137) and with TreeFinder v.2011 (AICc criterion) (Jobb et al., 2004) for the nucleic acid supermatrix (FNT-137). The robustness of the FAA-137 and FNT-137 ML trees was estimated by the non-parametric bootstrap procedure implemented in PhyML (100 replicates of the original dataset), whereas the SH-like support was used for FAA-474 ML trees.

Additional ML trees of the FNT-137 supermatrix were inferred with the Galtier and Gouy (GG) (Galtier and Gouy, 1998) non-homogeneous model that was recently implemented in nhPhyML (Boussau and Gouy, 2006) in combination with a gamma distribution (Γ_5). The discrete version of the model was considered in all analyses, with three values of G + C equilibrium content (0.25, 0.5 and 0.75). Thus, for each branch, the best out of the three possible values was determined by maximum likelihood. nhPhyML requires a rooted tree as a starting point but does not allow topology exploration around the root so that no Nearest Neighbour Interchange (NNI) can be tested between any two lineages present on each side of the root. However, nhPhyML allows topology exploration on each sub-tree surrounding the root. According to the results from this study (see Results), we fixed the root position on the branch separating the *Epsilonproteobacteria* from the other proteobacterial classes to compute ML trees.

Bayesian analyses were performed with PhyloBayes 3.3b to investigate the relationships among species within each class (Lartillot et al., 2009). We used the CAT model in order to take into account across-site heterogeneities in the amino-acid replacement process (Lartillot and Philippe, 2004). For each class (i.e. *Alphaproteobacteria*, *Betaproteobacteria*, *Deltaproteobacteria*, *Gammaproteobacteria*, and *Epsilonproteobacteria*), we ran two MCMC chains in parallel with the CAT + Γ_4 model. The initial 500 trees were discarded as "burn-in". The remaining trees from each chain were used to test for convergence, compute the 50% majority rule consensus tree and the posterior probabilities by sampling one every ten trees. The chains were stopped when the maxdiff and the effective size became lower than 0.3 and greater than 100, respectively. Similar analyses were performed using the Dayhoff4 recoding option (CAT + REC4 + Γ_4). The four Dayhoff's amino acid families corresponded to [(A, G, P, S, T) (D, E, N, Q) (H, K, R) (F, Y, W, I, L, M, V)] plus cysteines treated as missing data (C=?).

2.4. Detection of HGT

We used Prunier (Abby et al., 2010) to search for possible HGT events in individual r-protein trees built with the subset of 137 species. According to a reference phylogeny, Prunier explores HGT scenarios on single marker trees and finds the most parsimonious one (i.e. the one involving the least number of HGT events). We used the ML phylogeny of FAA-137 as a reference tree, because it was previously shown that supermatrix approaches provide good frames to detect HGTs in single marker trees (Abby et al., 2012). The branch robustness is taken into account by Prunier in order to minimise the impact of phylogenetic reconstruction errors and the lack of phylogenetic signal. Here, we considered a threshold of 80% bootstrap values in individual r-protein ML trees and set the "forward" parameter to 2. Because HGT scenarios depend on the position of the root in the reference tree, we tested the 271 possible roots of the reference phylogeny in order to find the most parsimonious HGT scenario over all protein families. Similar analyses were performed using a more restricted taxonomic sampling (52 species).

2.5. Rooting the proteobacterial phylogeny

The non-homogeneous models, such as the GG model implemented in nhPhyML, render the final likelihood of a tree dependent on its root position, contrary to standard homogeneous models. In fact, the GG model allows base compositions in terms of G + C frequencies to vary between lineages and the root is assigned its own G + C composition (which is estimated by Maximum Likelihood), so that the placement of the root changes the likelihood of the tree. Accordingly, it is possible to use these models to identify the most likely location of the root of an unrooted phylogenetic tree (Yang and Roberts, 1995), provided the data has retained enough compositional signal to discriminate between alternative roots. We applied this approach to determine the most likely position of the root of *Proteobacteria*. To do so, we used the ML phylogeny inferred with the FAA-137 supermatrix with the LG + Γ_4 + I model and the first two codon positions of the FNT-137 supermatrix using the GTR + Γ_4 or the GG + Γ_5 model. Three species were removed from the analysis because their position in the ML trees was unresolved (*Mariprofundus ferrooxydans* PV-1) or because of huge evolutionary rates ('*Candidatus* (*Ca.*) Hodgkinia cicadicola' and '*Ca.* Carsonella ruddii PV'). Nine putative root positions were tested. The likelihoods of the resulting trees were further compared for statistical significance with the AU test (Shimodaira, 2002) implemented in the CONSEL program (Shimodaira and Hasegawa, 2001). An independent approach based on the pattern of HGTs inferred by Prunier was used to determine the location of the root of *Proteobacteria* (see below).

2.6. Mutational saturation level

The mutational saturation level of FAA-137 was estimated by comparing the evolutionary distance deduced from ML trees inferred with PhyML to the p-distance (i.e. observed divergence) deduced from the multiple alignment between each pair of sequences (Chiari et al., 2012; Philippe and Forterre, 1999; Philippe et al., 1994). A similar analysis was performed for FNT-137 but by considering each of the three codon position separately.

3. Results and discussion

3.1. Taxonomic distribution of r-proteins in Proteobacteria and supermatrix construction

The survey of the proteobacterial proteomes with BLASTP highlighted missing r-proteins in many lineages, yet most of these absences corresponded to annotation errors because the corresponding genes could be easily identified in the corresponding genomic sequences with TBLASTN (Supplementary Table S1). More precisely, annotation errors were detected in half of the analysed proteomes (240 out of 474), and a few genomes presented more than 10 unannotated r-protein genes. This observation was in agreement with a recent survey of r-proteins in complete prokaryotic genomes (Yutin et al., 2012) and underlined the poor quality of the annotation of some genome sequences. Beside annotation errors, a few r-proteins were truly missing in some proteobacterial lineages (Supplementary Table S1). For instance, L30 and L32 were absent from the genomes of all *Epsilonproteobacteria*; L34 and L36 were missing in *Mariprofundus ferrooxydans* PV-1 (the only representative of *Zetaproteobacteria*), whereas L32 was missing in *Magnetococcus marinus* MC-I and its close relative, the magnetovoid strain MO-1. Regarding S22, an extremely restricted taxonomic distribution was observed, the protein being present in only 50 closely related species belonging to *Enterobacteriales* (a gammaproteobacterial order). This indicated that S22, which is associated to stationary phase ribosomes (see (Maki et al., 2000) and references therein), appeared late during the evolution of *Proteobacteria*. Conversely, a few r-proteins were present in multiple copies in a few genomes (Supplementary Table S1). For instance, this is the case for L31 and S21, for which four copies are found in *Escherichia coli* O157:H7 EDL933 and *Burkholderia* (*Betaproteobacteria*), respectively. The loss or, alternatively, the presence of multiple copies of some r-proteins in some taxa is puzzling and should be further investigated from a functional point of view. However, to a few exceptions, our results indicated that the majority of the 55 r-proteins were present in a single copy in *Proteobacteria* and therefore that the set of r-proteins (and thus the ribosome) has not significantly changed during the diversification of this phylum.

The burst of genome sequencing projects allows combining protein markers to investigate the phylogeny of organisms (Delsuc et al., 2005). In the case of r-proteins, their relative short size hinders the use of supertree approaches, especially when large taxonomic samplings are considered. In contrast, supermatrices have been shown to be particularly well-suited for this type of data (Brochier et al., 2002; Matte-Tailliez et al., 2002). Briefly, this approach consists in combining the alignments of single phylogenetic markers into a single large alignment (called a supermatrix), which is then used for phylogenetic reconstruction (Delsuc et al., 2005). We applied this strategy to build the FAA-474 supermatrix that gathered the 28 LSU and 20 SSU r-proteins presenting a sufficient taxonomic sampling (see methods). The ML tree inferred with this supermatrix recovered the monophyly of most proteobacterial taxa (Supplementary Fig. S2), confirming that r-proteins and SSU rRNA carry an overall consistent phylogenetic signal. Due to biases in the taxonomic distribution of genome projects, some taxa were overrepresented (Supplementary Table S1). To limit taxonomic sampling biases and to reduce computation time, we selected a subset of 137 organisms encompassing most of the taxonomic and genetic diversity of each proteobacterial class (see Material and Methods) to investigate in more detail the phylogenetic signal contained in r-proteins and the phylogeny of this bacterial phylum (species in bold in Supplementary Fig. S2 and Table S1). To do so, we built two supermatrices, FAA-137 and FNT-137, which gathered, respectively, the amino acid and the nucleic acids alignments of 46 (27 LSU and 19 SSU) r-proteins.

3.2. Protein and nucleic acid sequences of r-proteins contain a reliable phylogenetic signal

The decay of the ancient phylogenetic signal contained in molecular data by successive substitutions occurring at the same position is a frequent problem encountered in phylogeny. This phenomenon is called mutational saturation. Beside the loss of information, mutational saturation may generate tree reconstruction artefacts such as Long Branch Attraction (LBA) which tends to group together sequences associated to long branches (Bergsten, 2005; Felsenstein, 1978; Philippe and Laurent, 1998). This has been extremely well documented in the case of ancient phylogenies (Gribaldo and Philippe, 2002). Because *Proteobacteria* are an ancient phylum, a certain level of mutational saturation is expected in their r-protein sequences, and thus in the FAA-137 and FNT-137 supermatrices.

The level of mutational saturation can be revealed by comparing the p-distances (i.e. the observed substitutions) between each pair of sequences to the corresponding ML-estimated distances (Chiari et al., 2012; Philippe and Forterre, 1999; Philippe et al., 1994). The stronger the correlation between the two measures is, the lower the level of mutational saturation is. This can be visualised graphically by plotting the two measures. The slope of the linear regression indicates the level of mutational saturation contained in the data. In theory, when the level of saturation is

close to zero, meaning that no multiple substitutions occurred, both distances should be equal. A relatively good correlation between the two distances was observed in the case of FAA-137 (slope of the linear regression = 0.1296 and R^2 = 0.847, Fig. 1a). Expectedly, the ML and p-distances were strongly correlated among closely related species and/or slowly evolving sequences, whereas the highest discrepancies were observed for pair of sequences involving the two very fast evolving endosymbionts '*Ca.* Hodgkinia cicadicola' and '*Ca.* Carsonella ruddii' (surrounded by a dot line, Fig. 1a), for which more than 3 substitutions per site were inferred with ML, whereas only 0.6 to 0.7 substitutions per site were observed at the sequence level. This indicates that many substitutions have occurred in these sequences (large ML-distances), but are hidden due to mutational saturation (moderate p-distances). Indeed, the removal of these two organisms improved the correlation between the p- and the ML distances (slope of the linear regression = 0.1669 and R^2 = 0.9204, not shown), confirming their strong impact on the global saturation level of the data. Altogether, this suggests that the level of mutational saturation in the FAA-137 supermatrix is moderate and that most of the ancient phylogenetic signal contained in r-proteins has been preserved during the diversification of *Proteobacteria*.

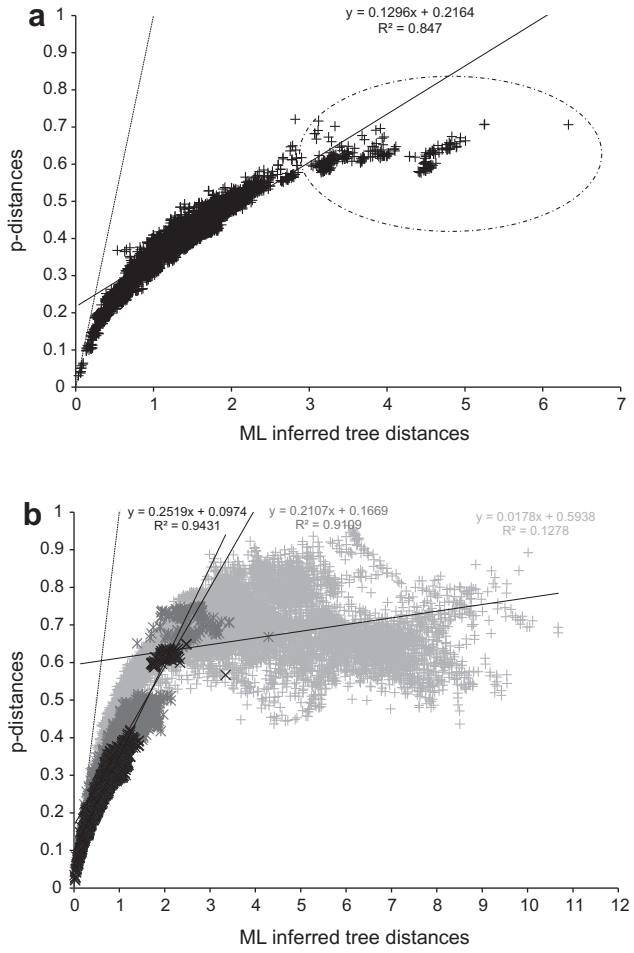

Fig. 1. Mutational saturation level of (a) FAA-137 and (b) FNT-137 at the first (black), second (dark grey) and third (light grey) codon positions. For each pair of sequences, the p-distance corresponding to the proportion of observed substitutions is plotted against to the ML distance deduced from the corresponding ML tree. The slope of the linear regression indicates the amount of mutational saturation: the lower the slope is, the greater the number of inferred multiple substitutions is. Conversely, the higher the slope is, the lower the number of inferred multiple substitutions is. The dash-line corresponds to the ideal case, for which at most one substitution occurred by position, meaning that no multiple substitution occurred during the evolution of sequences and thus that the level of saturation is close to zero.

The analysis of the FNT-137 supermatrix provided a very different picture (Fig. 1b). While a strong correlation was observed between the p- and the ML-distances at the first two codon positions (slope of the linear regression = 0.2107 and 0.2519, and R^2 = 0.9109 and R^2 = 0.9431, respectively), a very weak correlation and a great dispersal was observed at the third codon position (slope of the linear regression = 0.0178 and R^2 = 0.127). This reflects the fast evolutionary rate of the third codon position and its higher saturation with respect to the two other positions. Actually, while a maximum of 4.29 and 3.34 substitutions per site was estimated by ML at the first and second codon positions, respectively, up to 10.68 substitutions per site were inferred at the third codon position (Fig. 1b). In addition to higher evolutionary rates, the highest heterogeneity in term of G + C content was observed at the third codon position with respect to the two other positions due to its strong correlation with the genomic G + C content (Fig. 2). These results were expected because selective pressures are known to be more relaxed at the third codon position due to the redundancy of the genetic code. The combined effect of base composition heterogeneity and fast evolutionary rate may strongly bias tree reconstructions. This was recently illustrated in the case of *Plasmodium* (Davalos and Perkins, 2008) and turtles (Chiari et al., 2012). *Proteobacteria* are no exception, as shown by the ML phylogeny inferred with the FNT-137 supermatrix using the GTR + Γ_4 model, which was the best-fitted model proposed by TreeFinder (Supplementary Fig. S5). As all homogeneous and stationary models, the GTR model assumes that the sequences are at equilibrium and thus have the same base composition (see below). Fig. 2 shows that this assumption is strongly violated in our data. This may explain the artefactual clustering of unrelated sequences sharing similar base compositions, as illustrated by the grouping of low G + C espilonproteobacteria, low G + C alphaproteobacteria and *Bdellovibrio bacteriovorus* (*Deltaproteobacteria*) (Supplementary Fig. S5). Expectedly, the monophyly of these classes was recovered when the third codon position was removed from the analysis (Supplementary Fig. S6). Importantly, artefactual clustering can also occur even at small evolutionary scales as exemplified by the grouping of '*Ca.* Liberibacter asiaticus str. psy62' and *Bartonella grahamii* as4aup, two unrelated rhizobiales harbouring moderate G + C contents compared to other rhizobiales (Supplementary Fig. S5). These taxa are in fact related to *Sinorhizobium medicae* WSM419 and *Brucella suis* 1330, respectively (Fig. 3 and Supplementary Fig. S6). Expected relationships were also recovered by applying the non-homogeneous Galtier and Gouy (GG) model on the three and on the first two codon positions of the FNT-137 supermatrix (Supplementary Figs. S7 and S8), because this model, contrarily to most evolutionary Markovian models, allows the process of evolution to vary through time and which therefore models variations of base composition among lineages (Galtier and Gouy, 1998). The GG model is thus more able to discriminate homoplasies owing to compositional convergence from the true phylogenetic signal than homogeneous models such as GTR and thus to reduce the impact of mutational saturation on tree reconstructions. Altogether, these results strengthen the idea that the use of approaches and evolutionary models designed to overcome mutational saturation and compositional biases should be systematically considered for the inference of prokaryotic phylogenies, even at small evolutionary scales.

Our analyses showed that the phylogenetic signal contained in r-proteins has not been completely blurred by mutational saturation and compositional biases. Then, we asked whether this phylogenetic signal reflects the evolutionary history of *Proteobacteria* or if it has been obscured by HGTs. To address this question, we investigated the phylogeny of each r-protein in order to quantify the

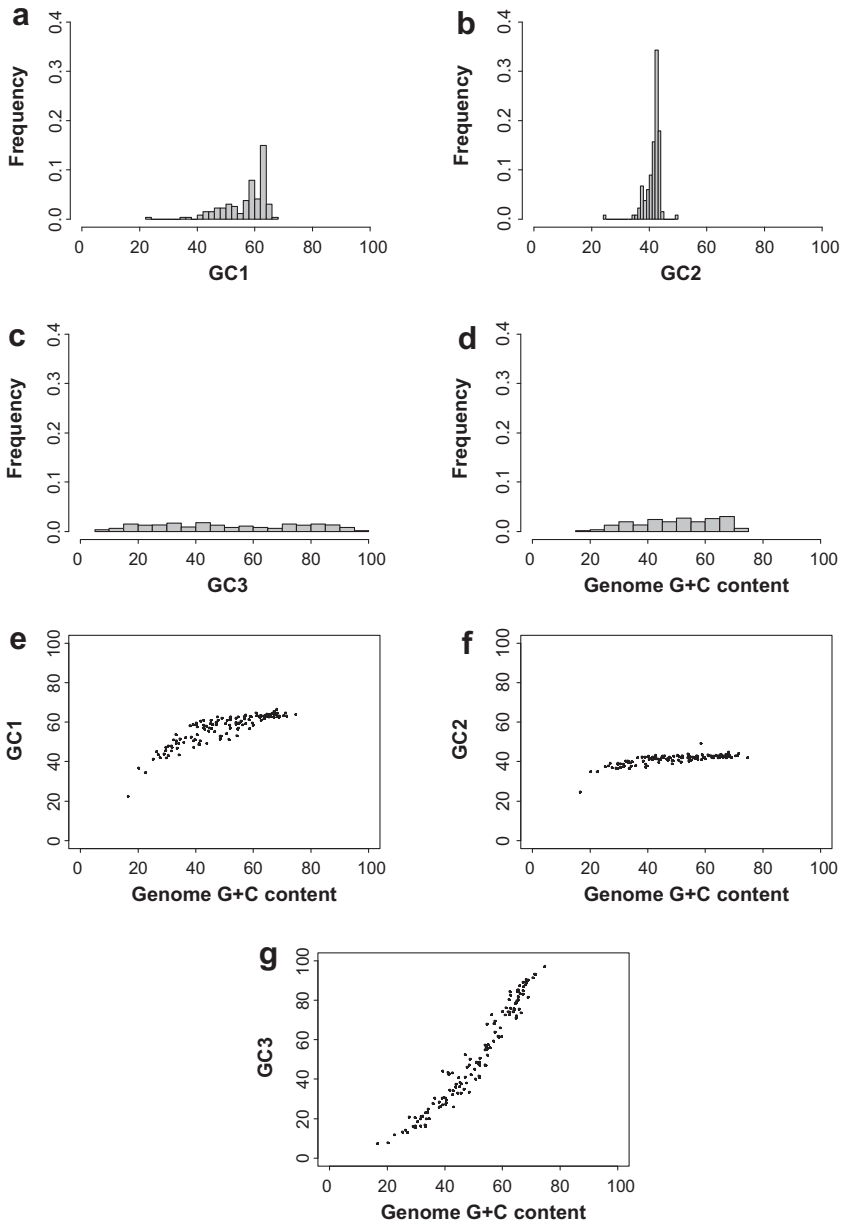

Fig. 2. Distribution of G + C frequencies at the three codon positions (a–c) and at the genome level (d). Correlation between the G + C content at each codon position and the whole genome G + C content (e–g).

amount of HGT that has affected their evolutionary history. To do so, we used Prunier, a recently developed statistical approach of gene tree reconciliation (Abby et al., 2010). The r-protein S22 was not taken into account due to its very restricted taxonomic distribution (Supplementary Table S1). The analysis of the 53 remaining r-proteins revealed 68 HGT events, representing 1.28 HGTs per protein family in average (Table 1 and Supplementary Table S2). More precisely, 26 r-proteins were devoid of HGT, whereas one, two, three and four HGTs were inferred for 15, six, two and one r-proteins, respectively. In the case of L28, L33, and L36 more than four HGTs (i.e., five, seven, and 19, respectively) were detected. However, for these three proteins Prunier failed to identify a suitable scenario of HGT (Table 1). Ignoring these three potentially artefactual HGT scenarios, 37 HGT events were inferred, representing 0.7 HGT events per gene family in average. This number was roughly twice as low in a similar analysis performed with a sampling of 52 species (not shown). This indicated that HGTs have rarely affected the evolutionary history of r-proteins in *Proteobacteria* and that most of the topological inconsistences observed in phylogenies of single r-proteins result from a lack of phylogenetic signal and not from HGT. This also confirmed previous studies showing that proteins involved in large complexes are rarely successfully transferred (Cohen et al., 2011; Jain et al., 1999; Leigh et al., 2011) and that r-proteins can be used to investigate the evolutionary history of *Proteobacteria*.

3.3. The root of Proteobacteria

The ML tree inferred with the FAA-137 supermatrix (LG + Γ_4 + I model, Fig. 3) was overall consistent with the trees inferred with the FNT-137 supermatrix (GG + Γ_5 model, Supplementary Figs. S7 and S8). As expected, it strongly supported the monophyly of each class (Fig. 3): *Epsilonproteobacteria* (BV = 100%), *Deltaproteobacteria* (BV = 93%), *Alphaproteobacteria* (BV = 93%), as well as the

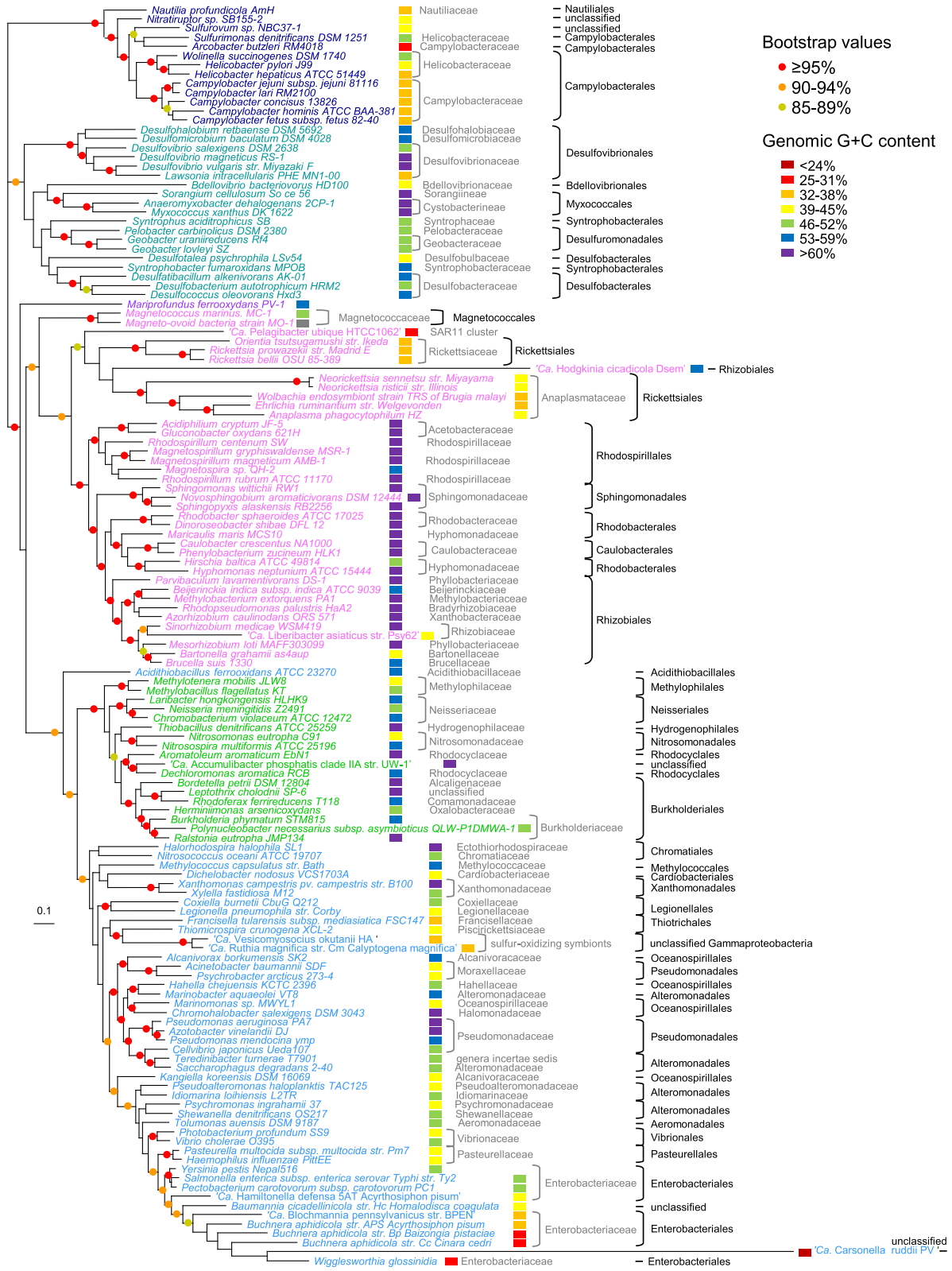

Fig. 3. Unrooted ML phylogeny of a subset of 137 *Proteobacteria* based on the FAA-137supermatrix (5,124 amino acids positions). The tree was inferred with the LG + Γ_4 + I model, as suggested by ProtTest v2.4. The scale bar represents the average number of substitutions per site. Coloured circles correspond to bootstrap values ranges (100 replicates of the original dataset). For clarity, supports lower than 85% are not shown. *Epsilonproteobacteria* are in dark blue, *Deltaproteobacteria* in green–blue, *Alphaproteobacteria* in pink, *Zetaproteobacteria* in purple, *Betaproteobacteria* in light green and *Gammaproteobacteria* in light blue. Orders and families are depicted in black and in light grey, respectively, according to current taxonomy. Coloured rectangles correspond to genomic G + C contents.

monophyly of *Beta* + *Gammaproteobacteria* (BV = 91%). It also strongly supported the separation of *Epsilon*- and *Deltaproteobacteria* from other proteobacteria (BV = 98%). Regarding *Zetaproteobacteria*, in the FAA-137 tree *Mariprofundus ferrooxydans* PV-1 represents a distinct lineage emerging as the sister-group of the clade formed by *Alpha*, *Beta* and *Gammaproteobacteria*. However, the support

Table 1
Number of HGTs inferred by Prunier in each r-protein family. We only show results inferred for the eleven root positions (out of 271) that minimise the number of HGTs. They correspond to root number 71–77, and 91–94 according to Supplementary Table S2. These eleven roots are located within *Epsilonproteobacteria* or in the branch separating *Epsilonproteobacteria* from other proteobacterial classes. The * designs the three families suspected to yield artefactual scenarios.

r-Protein	Number of HGT	r-protein	Number of HGT	r-Protein	Number of HGT
L10	1	**L28***	**7**	S13	0
L1	2	L29	0	S14	0
L11	2	L30	0	S15	0
L13	1	L3	1	S16	2
L14	0	L31	0	S17	0
L15	1	L32	0	S18	1
L16	1	**L33***	**5**	S19	0
L17	0	L34	0	S20	1
L18	0	L35	3	S2	2
L19	1	**L36***	**19**	S21	4
L20	0	L4	1	S3	2
L2	1	L5	0	S4	3
L21	1	L6	0	S5	0
L22	0	L7	1	S6	0
L23	1	L9	2	S7	0
L24	0	S10	0	S8	1
L25	1	S11	0	S9	0
L27	0	S12	0		

Total number of HGTs: 68 (corresponding to 1.28 HGT per r-protein family).
Total number of HGTs excluding the three doubtful scenarios: 37 (corresponding to 0.7 HGT per r-protein family).

for this position is weak (BV = 66%, Fig. 3). Therefore, according to this phylogeny, we cannot exclude that *Zetaproteobacteria* may be the sister-lineage of *Beta* + *Gammaproteobacteria* or of *Alphaproteobacteria*. Further analyses using additional markers or the inclusion of new representatives of this class when they become available are needed to precisely determine the position of *Zetaproteobacteria* with respect to these proteobacteria classes.

Based on protein signatures, it was proposed that the root of *Proteobacteria* separates *Thiobacteria* (i.e., *Delta* and *Epsilonproteobacteria*) from the three other classes (Gupta, 2000) or that sulphur oxidation performed by *Thiobacteria* was ancestral (Cavalier-Smith, 2002). Based on molecular phylogenies, either *Deltaproteobacteria* (Ciccarelli et al., 2006) or *Epsilonproteobacteria* (Gupta, 2000; Yutin et al., 2012) appeared as the first emerging class, albeit frequently with non-significant supports. However, these works aimed at reconstructing global phylogenies of *Bacteria* without addressing specifically the question of the root of *Proteobacteria*. To our knowledge, the precise location of the root of *Proteobacteria* has not been carefully investigated and remained to be elucidated. The usual approach to root a phylogenetic tree is based on the use of outgroups. However, this increases the risk of LBA because the branch separating the ingroup from the outgroup is usually longer than the internal branches of the ingroup (Philippe and Laurent, 1998). Non-homogeneous and non-stationary models of evolution, as the GG model (see Materials and Methods and below), represent an alternative way to root phylogenies without the use of outgroups (Yang and Roberts, 1995). In fact, homogeneous and stationary models, such as GTR or LG assume that the overall sequence composition does not change through time and that the process is at equilibrium from the root to the leaves. These models are reversible, in the sense that there is no direction of evolution along the inferred trees (Felsenstein, 2004), such that the root can be placed wherever on the tree without influencing the likelihood (Yang, 2006). In contrast, non-homogeneous and non-stationary models do not assume the reversibility hypothesis and assign specific base frequencies to the root so that its position influences the likelihood of the tree (Boussau and Gouy, 2006; Yang and Roberts, 1995). Because of the features of non-homogeneous and non-stationary models mentioned above, it is possible to use them in order to determine the ML position of the root of a tree for which the topology is known. Here, we used this approach to address the question of the root of *Proteobacteria* with the GG model, which models the variation of base composition in time by assigning to each branch of the tree its own G + C equilibrium frequency, as well as on the root. To do so, we considered the three ML topologies inferred with the FAA-137 supermatrix and the LG + Γ_4 + I model (Fig. 3), and with the first two codon positions of the FNT-137 supermatrix using the GTR + Γ_4 and the GG + Γ_5 models (Supplementary Figs. S6 and S8). For each topology, nine root positions were tested, corresponding to all possible placements of the root on the internal branches connecting the proteobacterial classes (Table 2). The three topologies provided very similar results, regarding the rank of the nine roots and the conclusions of the AU test. More precisely, the best likelihood was associated to a rooting on the branch separating *Epsilonproteobacteria* from all other proteobacterial classes. It is worth noting that while four alternative roots were rejected by AU tests (their AU values were below 0.05, indicated with * in Table 2), a rooting on the branch leading either to *Deltaproteobacteria*, to *Delta* + *Epsilonproteobacteria*, to *Acidithiobacillus*, or to *Alpha* + *Delta* + *Epsilonproteobacteria*, even if less likely, was not significantly statistically rejected (AU values above 0.05, Table 2).

Another approach to root trees without using outgroup is based on the phylogenetic signal carried by HGT. Indeed, it has been shown recently that the pattern of HGT can be exploited to discriminate among putative root positions in species trees, because when the species tree is rooted in a group of species whose ancestor has received a gene by HGT, the number of transfers needed to reconcile the gene trees with the species tree increases dramatically (Abby et al., 2010, 2012). We compared the number of HGTs inferred in r-protein families for all of the 271 possible locations of the root in the FAA-137 tree (Supplementary Table S2). Interestingly, the positions of the root that minimised the number of HGT placed *Epsilonproteobacteria* (or an espilonproteobacterial lineage) as the first diverging lineage within *Proteobacteria* (Supplementary Table S2). Other rootings implied much more HGT events (Supplementary Fig. S9). Similar results were obtained with a more restricted taxonomic sampling (52 species) of proteobacteria (not shown).

Altogether, the use of non-homogeneous and non-reversible models, and patterns of HGT favours a rooting of *Proteobacteria* at the base of *Epsilonproteobacteria*.

3.4. The phylogenetic position of fast evolving proteobacterial lineages

Proteobacteria contain a number of lineages whose phylogenetic position is difficult to determine (Moran et al., 2008). This is, for instance, the case of obligate endosymbionts of insects, such as the gammaproteobacterium '*Ca.* Carsonella ruddii PV' and the alphaproteobacterium '*Ca.* Hodgkinia cicadicola'. '*Ca.* Hodgkinia cicadicola' is an obligate endosymbiont of the cicada *Diceroprocta semicincta* which harbours one of the smallest genomes known to date (144 Kb), and is known to be very fast evolving (McCutcheon et al., 2009). Interestingly, while most bacterial symbionts with highly reduced genomes harbour high A + T rich genomic sequences, the genome of '*Ca.* H. cicadicola' is G + C rich, suggesting that strong selective pressures counteract the natural mutational bias toward A + T (Van Leuven and McCutcheon, 2012). McCutcheon and collaborators hypothesised a possible relationship between '*Ca.* H. cicadicola' and *Rickettsiales*, but phylogenetic analyses of SSU rRNA and protein markers favoured a link with *Rhizobiales* (McCutcheon et al., 2009). The ML trees of FAA-137 and FNT-137 strongly support the former hypothesis because '*Ca.* H. cicadicola' grouped with Rickettsiales and SAR11 (BV > 85%,

Table 2
Results of the AU test for the position of the proteobacterial root. The root is located on the branch connecting the two groups separated by the vertical bar. The nine tested positions are ranked according to their likelihood computed by nhPhyML. The trees used as input are the ML trees based on the FAA-137 with the LG + Γ_4 + I model (1) and on the first two codon positions of the FNT-137 supermatrix inferred either with the GTR + Γ_4 (2) or the GG + Γ_5 (3) model. E: *Epsilonproteobacteria*, D: *Deltaproteobacteria*, T: *Acidithiobacillus*, B: *Betaproteobacteria*, A: *Alphaproteobacteria*, G: *Gammaproteobacteria*.

Position of the root	Rank	ΔLn *L* (1)	AU *p-value*	Rank	ΔLn *L* (2)	AU *p-value*	Rank	ΔLn *L* (3)	AU *p-value*
E|T,B,G,A,D	1	0	0.937	1	0	0.930	1	0	0.961
D|E,T,B,G,A	2	32.6	0.215	2	32.0	0.182	2	27.6	0.246
T|E,D,B,G,A	3	36.3	0.188	3	36.2	0.188	3	31.3	0.169
E,D|T,B,G,A	4	37.2	0.299	4	37.1	0.119	4	32.2	0.087
T,B,G|E,D,A	5	43.2	0.152	5	39.2	0.204	5	37.5	0.118
B,G|T,E,D,A	6	49.8	**0.036***	7	50.2	0.059	6	44.8	**0.018***
A|B,G,T,E,D	7	56.2	0.092	6	46.0	0.051	7	51.2	**0.043***
B|A,G,T,E,D	8	80.1	0.084	8	51.4	**0.044***	8	75.1	0.091
G|T,Ac,B,E,D	9	83.2	0.086	9	51.9	**0.037***	9	78.2	**0.041***

* *p-Value* < 0.05.

Fig. 3, and BV > 95% Supplementary Figs. S6–S8). However, the long branches harboured by these species suggested the possibility of a LBA. To investigate this hypothesis, we reanalysed the phylogeny of *Alphaproteobacteria* with the CAT model implemented in PhyloBayes (Lartillot and Philippe, 2004), which is less prone to tree reconstruction artefacts such as the LBA (Lartillot et al., 2007). In contrast with ML trees, the Bayesian tree inferred with the CAT + Γ_4 model strongly rejected the grouping of '*Ca.* H. cicadicola' with *Rickettsiales* (Posterior Probability (PP) = 0.97, Fig. 4a), the former being displaced to the apical part of the alphaproteobacterial tree (PP = 0.98, Fig. 4a). However, according to this tree, the precise position of '*Ca.* H. cicadicola' relatively to *Rhodobacterales*, *Rhizobiales* and *Caulobacterales* could not be determined (Fig. 4a). A similar result was obtained when the multiple alignment was recorded according to the four Dayhoff's amino acid categories (not shown). This confirmed that the grouping of '*Ca.* H. cicadicola' with the *Rickettsiales* and SAR11 in the FAA-137 and F137-NT ML trees likely resulted from a LBA.

We also investigated the phylogenetic position of '*Ca.* C. ruddii', a psyllid endosymbiont, which was described as an intermediate evolutionary state between organism and organelle (Tamames et al., 2007). It harbours a very reduced (160 Kb) and G + C poor (16.6%) genome (Nakabachi et al., 2006). Due to extreme evolutionary rates and compositional biases (both at the nucleic and amino acid levels, see above), the phylogenetic position of this bacterium remains uncertain (Williams et al., 2010). In the ML tree of FAA-137, this gammaproteobacterium robustly emerged within *Enterobacteriaceae* (BV = 91%, Fig. 3) and more precisely within a large group of obligate endosymbionts of insects including *Wigglesworthia* (a symbiont of tsetse flies), *Buchnera* and '*Ca.* Hamiltonella' (two aphid symbionts), '*Ca.* Baumannia' (a symbiont of sharpshooters), and *Blochmannia* (a symbiont of Ants) (BV = 91%, Fig. 3). This suggests that these obligate endosymbionts of insects originated from a single endosymbiosis event, a hypothesis which contradicts a recent phylogenetic analysis suggesting that at least four lineages of obligate endosymbionts of insects emerged independently from free living species during the diversification of *Enterobacteriaceae* (Husnik et al., 2011). According to that study: (i) *Sodalis*, *Baumannia*, *Blochmannia* and *Wigglesworthia* could be related to *Pectobacterium* and *Dickeya*; (ii) *Buchnera* to a large group encompassing *Erwinia* and *Pantoea*, its closest relatives, but also *Escherichia*, *Salmonella* and other lineages; (iii) *Hamiltonella* and *Regiella* to *Yersinia* and *Serratia*, and (iv) *Riesia* and *Arsenophonus* to *Xenorhabdus*, *Proteus,* and *Photorhabdus* (Husnik et al., 2011). Beside '*Ca.* C. ruddii', our taxonomic sampling, which was not designed to address specifically the question of the origin of obligate endosymbionts of insects, encompassed representatives of the first three groups. Because of their very long branches, the grouping of these obligate endosymbionts in the FAA-137 and FNT-137 tree was suspect and prompted us to investigate the possibility of an LBA. Similar to the case of '*Ca.* H. cicadicola', we reanalysed the phylogeny of *Gammaproteobacteria* with the CAT + Γ_4. In agreement with ML phylogenies, the Bayesian tree strongly support the monophyly of the enterobacteriales obligate endosymbionts of insects with a significant statistical support (PP = 0.97), to the notable exception of '*Ca.* Carsonella ruddii PV', which was robustly displaced outside of *Enterobacteriales* (PP = 0.95) and yet grouped with *Thiomicrospira crunogena* XCL-2 and the two sulphur-oxidising symbionts, albeit with a non-significant support (PP = 0.51) (Fig. 4b). This indicate that the grouping of '*Ca.* Carsonella ruddii PV' with the enterobacteriales obligate endosymbionts of insects in the FAA-137 and FNT-137 ML trees resulted from an LBA. The recoding of the multiple alignment according to the four Dayhoff's amino acid families provided similar results but did not allow clarifying the phylogenetic position of '*Ca.* Carsonella ruddii PV' (not shown). To further investigate the relationships among enterobacteriales obligate endosymbionts of insects, we removed '*Ca.* Carsonella ruddii PV' from the alignment. The Bayesian trees inferred with the CAT + Γ_4 and CAT + REC4 + Γ_4 were well resolved and overall consistent (Fig. 5a and b, respectively). Interestingly, while '*Ca.* Hamiltonella defensa 5AT' emerged with other enterobacteriales obligate endosymbionts of insects in the former (PP = 1, Fig. 5a), it grouped robustly with *Yersinia* when the amino acids were recoded (PP = 1, Fig. 5b), in agreement with the study of Husnik and colleagues (Husnik et al., 2011). This indicated that the grouping of the '*Ca.* Hamiltonella defensa 5AT' with other enterobacteriales obligate endosymbionts of insects in the F137-AA, F137-NT ML trees and in the gammaproteobacterial Bayesian phylogeny inferred without amino acid recoding was artefactual. Regarding the other endosymbionts, we could not separate *Baumannia*, *Blochmannia*, and *Wigglesworthia* from *Buchnera* (Fig. 5b). This could mean that their separation was artefactual in the study of Husnik, or that our taxonomic sampling is not sufficient to address this specific question. Indeed, contrarily to the analysis of Husnik et al. our analysis did not aim at dissecting in-depth the relationships among obligate endosymbiontic and free living enterobacteriales, which explains our limited taxonomic sampling for this order (10 species). However, even with such a very restricted taxonomic sampling we showed that the phylogenetic signal carried by r-proteins is sufficient to overcome (at least partially) LBA resulting from the very fast evolutionary rates of the alphaproteobacterial and gammaproteobacterial obligate endosymbionts of insects.

3.5. The evolutionary history within proteobacterial classes

The ML tree of *Proteobacteria* inferred with FAA-137 (Fig. 3) and the Bayesian trees of each proteobacterial classes (inferred without and with amino acid recoding, Figs. 5–7) were overall consistent

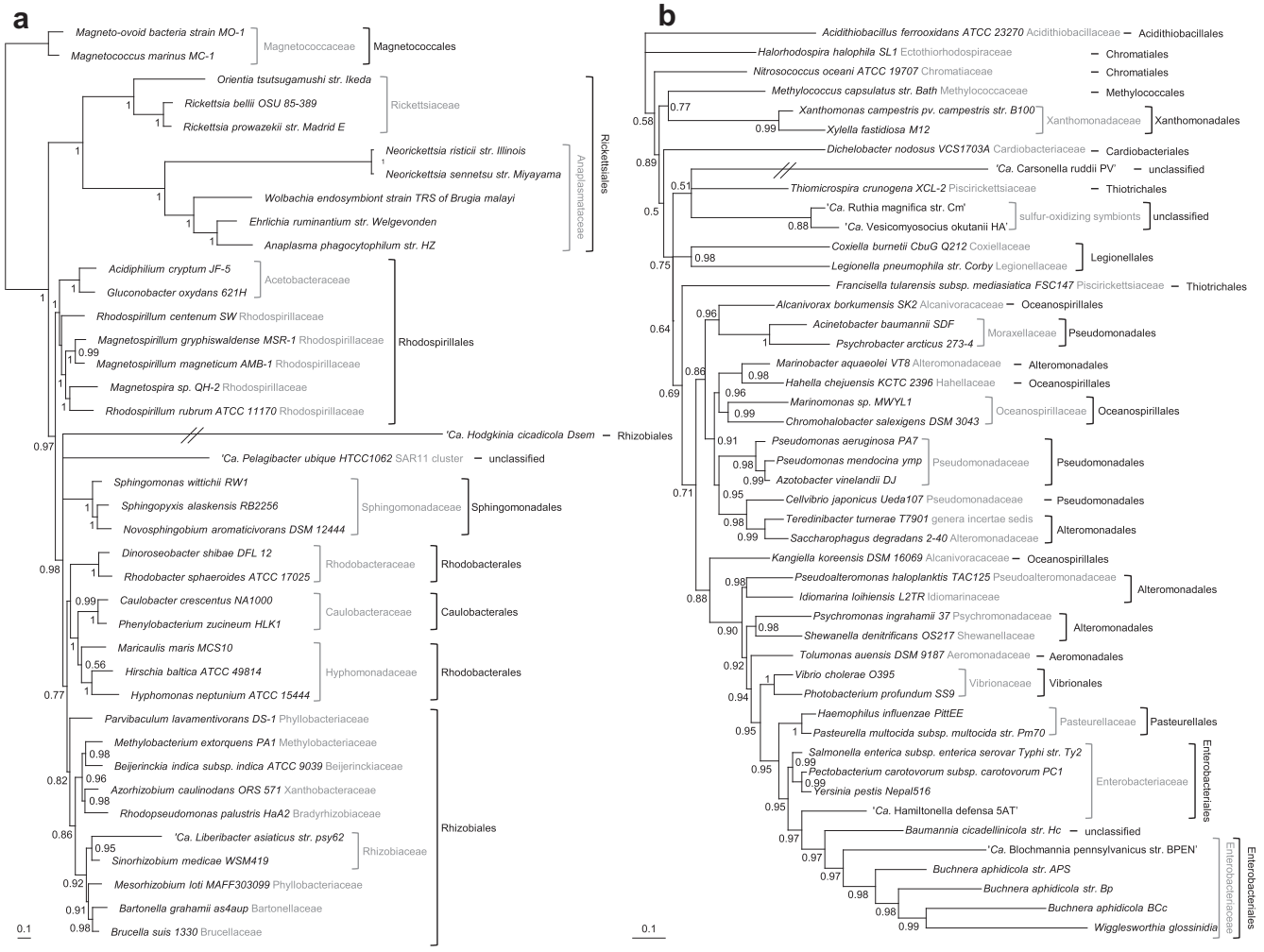

Fig. 4. Bayesian phylogenies of *Alphaproteobacteria* (a) and *Gammaproteobacteria* (b) inferred using the FAA-137 supermatrix (5,124 amino acids positions) with the CAT + Γ_4 model. The scale bars represent the average number of substitutions per site. The statistical supports correspond to posterior probabilities estimated with PhyloBayes. Orders and families are depicted in black and in light grey, respectively, according to current taxonomy. The trees were rooted according to Fig. 3.

except for the position of the fast evolving obligate symbionts (see above). The monophyly of nearly all orders was recovered except within *Gammaproteobacteria* (see below). More precisely, the relationships within *Epsilonproteobacteria* were strongly supported, albeit they were not all in agreement with the current taxonomy. Indeed, *Arcobacter butzleri* (*Campylobacteraceae*) and *Sulfurimonas denitrificans* (*Helicobacteraceae*) grouped robustly with the unclassified epsilonproteobacterium *Sulfurovorum sp.* (BV = 87%, PP = 1.0 and 0.96, Fig. 3 and 6a and b, respectively), and not with other *Campylobacteraceae* or *Helicobacteraceae*. This is in agreement with the report of genomic similarities shared between *Arcobacter butzleri* and *Sulfurimonas denitrificans* (Miller et al., 2007), and suggests that *Sulfurovorum* sp., *A. butzleri*, and *S. denitrificans* represent a new family within *Epsilonproteobacteria*, distinct from *Campylobacteraceae*, *Helicobacteraceae*, and *Nautiliaceae*. The main differences among the three trees concerned the position of this group, which formed the sister-lineage of other *Campylobacteraceae* and *Helicobacteraceae* in the Bayesian tree inferred without amino acid recoding (PP = 1.0, Fig. 6a) or was more related to *Campylobacteraceae* when amino acids were recoded (PP = 0.98, Fig. 6b), whereas it position was unresolved in the FAA-137 ML tree (BV < 85%, Fig. 3).

Similarly to *Epsilonproteobacteria*, a robust phylogeny of *Deltaproteobacteria* emerged from the ML and Bayesian analyses (Fig. 3 and 6c and d). More precisely, the monophyly of *Myxococcales* was supported (BV = 96%, PP = 1.0 and PP = 0.62) as well as their grouping with *Bdellovibrionales* (BV = 83%, PP = 1.0 in both Bayesian trees). Furthermore, the monophyly of *Geobacteraceae*, *Desulfobacteraceae*, *Desulfuromonadales*, and *Desulfovibrionales* was strongly recovered (BV $\geqslant$ 97% and PP = 1.0 in both Bayesian trees). The monophyly of *Desulfovibrionaceae* was significantly supported in the ML tree (BV = 97%, Fig. 3) and in the Bayesian phylogeny inferred without amino acid recoding (PP = 1.0, Fig. 6c), whereas *Desulfovibrio salexigens* grouped with *Desulfomicrobium* and *Desulfohalobium*when amino acid were recoded, albeit with a non-significant support (PP = 0.90, Fig. 6d), which could reflect an insufficient phylogenetic signal. A lack of phylogenetic signal could also explain the weak support for *Desulfobacteriales*, whereas the non-monophyly of *Syntrophobacterales* represented here by *Syntrophus aciditrophicus SB* (Syntrophaceae) and *Syntrophobacter fumaroxidans MPOB* (*Syntrophobacteraceae*) was not recovered in the ML tree and in the recoded Bayesian tree, albeit with a non-significant support (Fig. 3 and 6d), and strongly rejected in the Bayesian tree inferred without recoding (PP = 1.0, Fig. 6c).

In the case of *Alphaproteobacteria*, the phylogenetic analysis of r-proteins confirmed the close relationship between *Magnetococcus marinus* MC-I and the strain Magneto-ovoid MO-I strain (BV = 100%, Fig. 3, PP = 1.0 and 0.99, Fig. 7) (Lefevre et al., 2009), as well as the early branching of *Magnetococcales* with respect to other alphaproteobacterial orders (BV = 93%, Fig. 3) (Spring et al.,

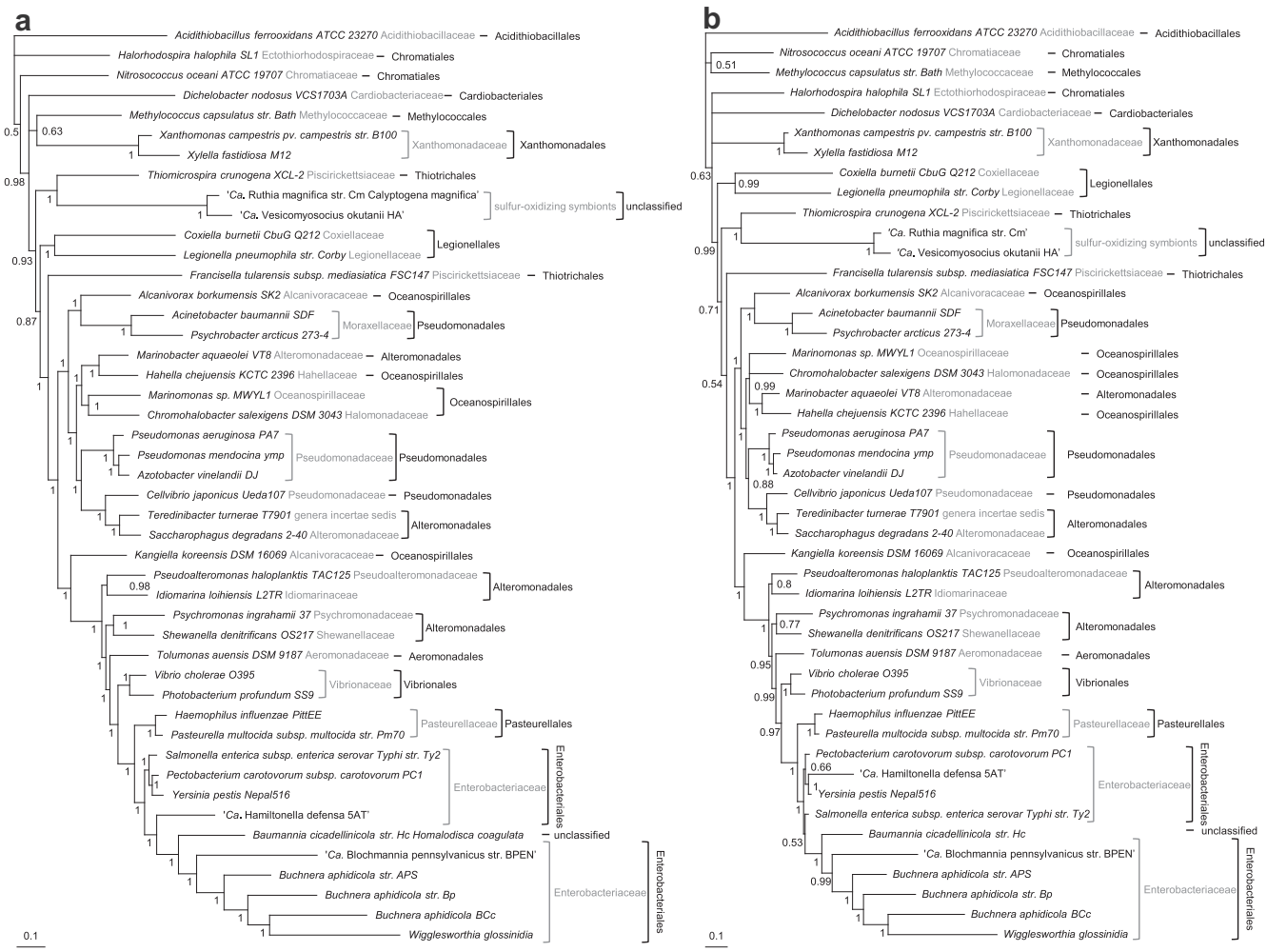

Fig. 5. Bayesian phylogenies of *Gammaproteobacteria* inferred using the FAA-137 supermatrix (5,124 amino acids positions) with the CAT + Γ_4 model (a) and the CAT + REC4 + Γ_4 model (b). The scale bars represent the average number of substitutions per site. The statistical supports correspond to posterior probabilities estimated with PhyloBayes. Orders and families are depicted in black and in light grey, respectively, according to current taxonomy. The trees were rooted according to Fig. 3.

1998), strengthening the recent proposal that they represent a proteobacterial lineage of high taxonomic rank (Bazylinski et al., 2013). More generally, the monophyly of most alphaproteobacterial orders was recovered (Figs. 3 and 7). One exception concerned *Rhodobacterales* due to the robust grouping of *Caulobacterales* with or within *Hyphomonadaceae* (BV = 100%, PP = 1.0 and 0.99, Figs. 3 and 7a and b, respectively). Such a clustering has been observed previously in protein and SSU rRNA trees (Badger et al., 2005; Lee et al., 2005; Thrash et al., 2011; Williams et al., 2007) and is supported by genomic and biological features. This suggests that the boundaries of *Caulobacterales* and *Rhodobacterales* should be revised. An interesting point concerned the phylogenetic position of the SAR11 lineage with respect to Rickettsiales. SAR11 represented here by '*Ca.* Pelagibacter ubique' (Giovannoni et al., 2005) is a major component of ocean surface waters (Morris et al., 2002; Rappe et al., 2002; Steindler et al., 2011). The phylogenetic position of SAR11 remains controversial: some analyses suggested that this group is related to *Rickettsiales* (Rappe et al., 2002; Thrash et al., 2011) and thus to mitochondria, whereas others supported a relationship with free-living marine and soil alphaproteobacteria and explained the phylogenetic proximity observed between SAR11 and *Rickettsiales* as the result of compositional biases (Brindefalk et al., 2011; Rodriguez-Ezpeleta and Embley, 2012; Viklund et al., 2011). In the FAA-137 ML tree, *Rickettsiales* and SAR11 grouped together (BV = 88%) and represented the second diverging order within *Alphaproteobacteria* (Fig. 3). Interestingly, the grouping of *Rickettsiales* and SAR11 was strongly rejected in the Bayesian trees inferred with the CAT + Γ_4 or with the CAT + REC4 + Γ_4 model, due to the displacement of SAR11 to the apical part of the trees (PP = 1.0 and PP = 0.96, Fig. 7a and b, respectively). This indicated that the grouping of SAR11 with *Rickettsiales* in ML trees results from an LBA due to the fast evolutionary rates of these two lineages. However, based on these analyses, we could not indicate with precision the position of SAR11 with respect to *Sphingomonadales*, *Rhodobacterales*, *Caulobacterales* and *Rhizobiales*. To tackle this issue, a broader taxonomic sampling of this lineage and of *Alphaproteobacteria* in general would be required. However, even with a restricted taxonomic sampling, the phylogenetic signal carried by r-proteins allowed strengthening the hypothesis that SAR11 and *Rickettsiales* have different origins.

Concerning *Betaproteobacteria*, the ML phylogeny inferred with FAA-137 and the two Bayesian trees restricted to *Betaproteobacteria* revealed very few discrepancies (Figs. 3 and 6e–f). In particular, the monophyly of all orders and families was strongly recovered. We confirm the emergence of '*Ca.* Accumulibacter phosphatis' within *Rhodocyclaceae*/*Rhodocyclales*, in agreement with previous studies (Hesselmann et al., 1999), and suggest that it represents a *bonafide* representative of this taxon. The main difference between the three trees concerned the position of *Thiobacillus denitrificans* (*Hydrogenophilaceae*/*Hydrogenophilales*), which grouped with *Nitrosomonadales* in both FAA-137 ML and non-recoded Bayesian trees (BV = 69% and PP = 0.98, Figs. 3 and 6e), whereas

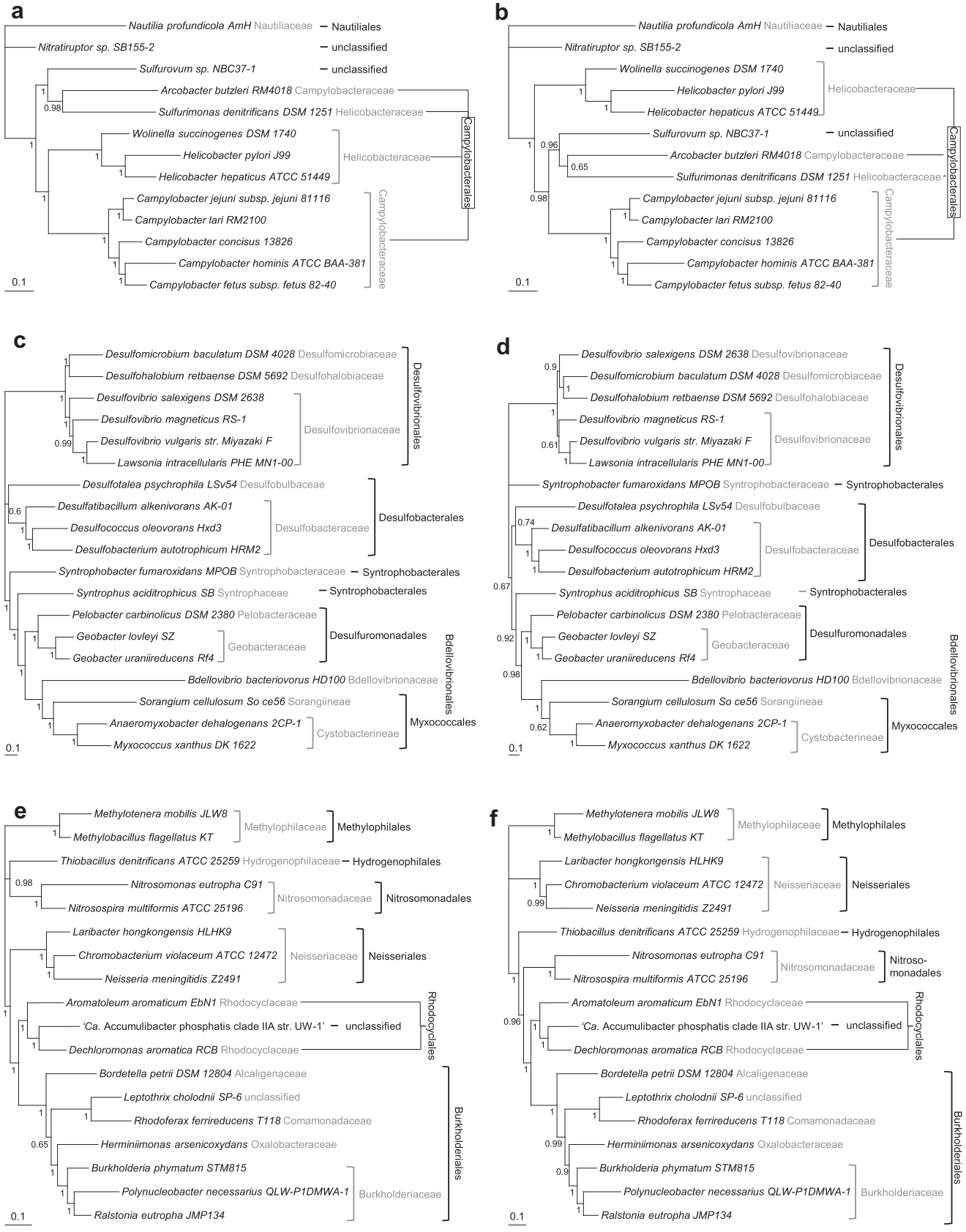

Fig. 6. Bayesian phylogenies of *Epsilonproteobacteria* (a and b) and *Deltaproteobacteria* (c and d) and *Betaproteobacteria* (e and f) inferred using the FAA-137 supermatrix (5,124 amino acids positions) with the CAT + Γ_4 model (a, c and e) and with the CAT + REC4 + Γ_4 model (b, d and f). The scale bars represent the average number of substitutions per site. The statistical supports correspond to posterior probabilities estimated with PhyloBayes. Orders and families are depicted in black and in light grey, respectively, according to current taxonomy. The trees were rooted according to Fig. 3.

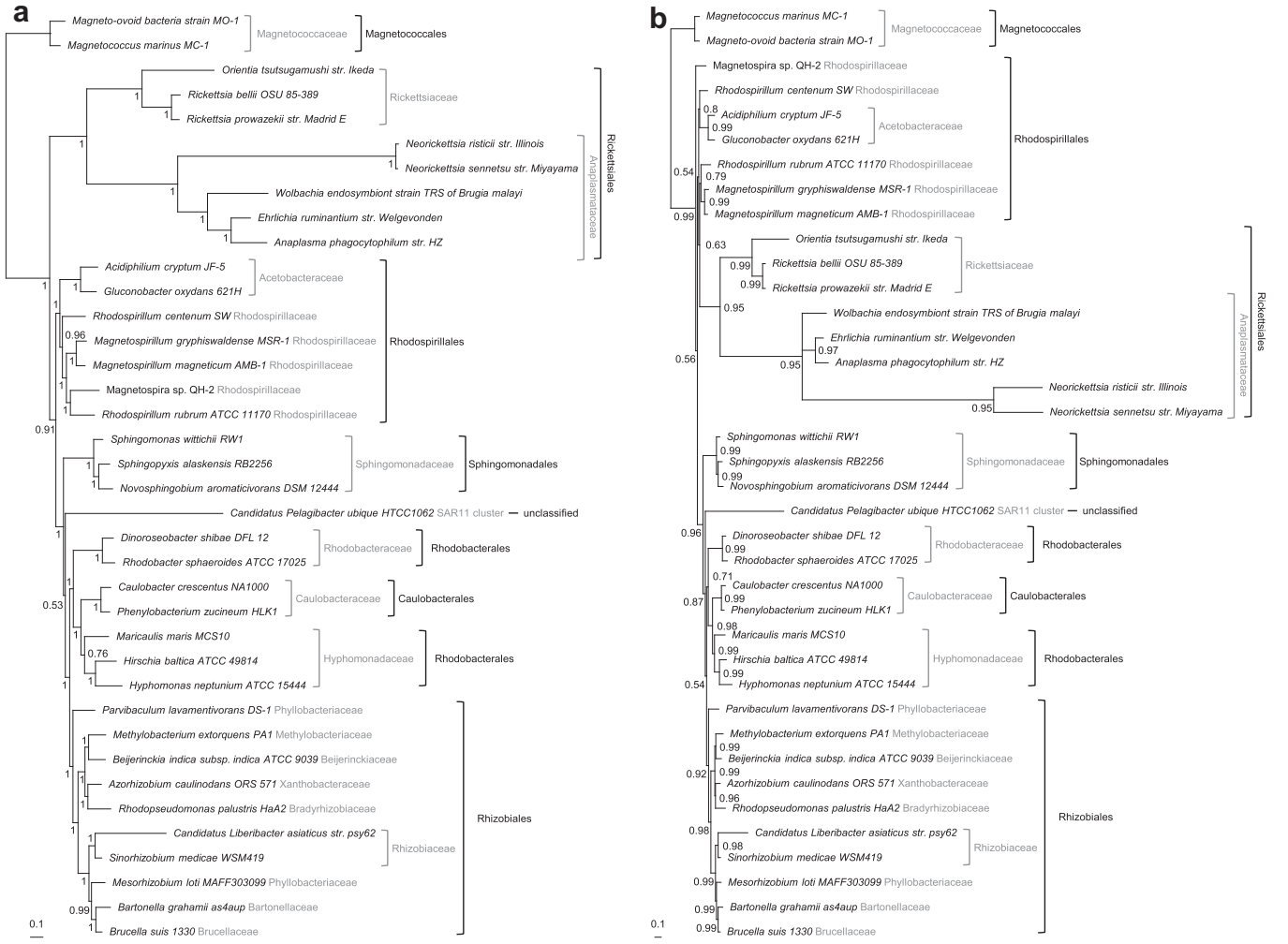

Fig. 7. Bayesian phylogenies of *Alphaproteobacteria* inferred using the FAA-137 supermatrix (5,124 amino acids positions) with the CAT + Γ_4 model (a) and the CAT + REC4 + Γ_4 model (b). The scale bars represent the average number of substitutions per site. The statistical supports correspond to posterior probabilities estimated with PhyloBayes. Orders and families are depicted in black and in light grey, respectively, according to current taxonomy. The trees were rooted according to Fig. 3.

it represented an isolated lineage when amino acids were recoded (Fig. 6f). Finally, while *Nitrosomonadales* formed the sister-group of *Burkholderiales* and *Rhodocyclales* in the ML and recoded Bayesian tree (BV = 89% and PP = 1.0, Figs. 3 and 6f), *Neisseriales* occupied this position in the Bayesian tree inferred without amino acid recoding (Fig. 6e).

In contrast to *Betaproteobacteria*, the phylogeny of *Gammaproteobacteria* showed strong discrepancies with the current taxonomy. First of all, in the FAA-137 tree, *Acidithiobacillales* (represented here by *Acidithiobacillus ferrooxidans* str. ATCC 23270) robustly branched at the base of the group composed of Beta- and other *Gammaproteobacteria* (BV = 91%). This position is in agreement with the recent analysis of 356 protein families from 108 gammaproteobacterial proteomes (Williams et al., 2010). Our analysis strengthens this observation and indicates that *Acidithiobacillales* are neither *Beta-* nor *Gammaproteobacteria*, but form a distinct lineage. This means that the taxonomic affiliation of *Acidithiobacillus* (and thus of *Acidithiobacillales*) to *Gammaproteobacteria* based on the phylogenetic analysis of a few SSU rRNA sequences using distance methods (Kelly and Wood, 2000) must be reconsidered, either through the creation of a new class or revision of the boundaries between *Beta-* and *Gammaproteobacteria*. Strong discrepancies with the current taxonomy were also observed for three major orders: the Alteromonadales, the *Pseudomonadales* and the *Oceanospirillales*. These lineages branched in the central part of the gammaproteobacterial tree, i.e., after the divergence of *Chromatiales*, *Methylococcales*, *Cardiobacterales*, *Xanthomonadales*, *Legionellales*, and *Thiotrichales*, but before the diversification of *Vibrionales*, *Pasteurellales*, *Aeromonadales* and *Enterobacteriales*. A careful examination of the FAA-137 ML and Bayesian trees revealed that the *Oceanospirillales* families (i.e., *Alcanivoracaceae*, *Hahellaceae*, *Oceanospirillaceae*, and *Halomonadaceae*) form four unrelated lineages, with the *Alcanivoracaceae* family being split into two (Figs. 3 and 5). A similar situation was observed for *Alteromonadales* that formed four distinct lineages and for *Pseudomonadales* which were split in three unrelated families. These observations are significantly supported by high BV and PP and are in agreement with the recent study of Williams et al. (2010). This situation requires urgent in-depth investigations aiming at revisiting the taxonomy of these families and orders. In contrast, the most apical part of the gammaproteobacterial tree was well resolved and in agreement with the current taxonomy. The monophyly of *Enterobacteriales* (including the unclassified '*Ca.* Baumannia cicadellinicola'), of *Pasteurellales* and of *Vibrionales* was recovered, as well as the sister relationship between *Enterobacteriales* and *Pasteurellales* (all BV > 90%, Fig. 3, and all PP = 1.0, Fig. 5). In contrast, the basal part of the phylogeny of *Gammaproteobacteria* was moderately resolved. While the monophyly of *Xanthomonadales*, and *Legionellales* was recovered and well supported in all trees, the monophyly of *Thiotrichales*, and *Chromatiales* was weakly supported in the FAA-137 ML tree

(Fig. 3), and not recovered in Bayesian trees (Fig. 5). However, in these trees, the unclassified sulphur oxidising symbionts robustly clustered within *Thiomicrospira* (Thiotrichales) (Figs. 3 and 5), suggesting that they belong to the same lineage. Finally, the relationships among the basal branching orders were not resolved, leaving open the question of the early steps of the diversification of *Gammaproteobacteria*.

4. Conclusions

Using *Proteobacteria* as a case study, we showed that ribosomal proteins are highly conserved, easily identifiable and have been rarely lost, duplicated and/or horizontally transferred. Their combination allows assembling relatively large supermatrices, which contain a moderate level of mutation saturation at the protein level but also at the nucleic acid level, provided that the third codon position was removed. Importantly, the use of accurate evolutionary models allows overcoming most of the tree reconstruction artefacts linked to fast evolving lineages and/or compositional biases, which are highly problematic in systematic and taxonomy studies. Finally, by comparing our results with previously published studies, we showed that the phylogenetic signal contained in r-proteins is a good proxy of the phylogenetic signal contained in larger sets of conserved proteobacterial genes, while allowing applying ML and Bayesian approaches in acceptable computational time. The phylogenies based on r-proteins allowed us to robustly infer most of the relationships among orders and families of *Proteobacteria*, to assign a number of unclassified proteobacterial lineages to existing taxa and to point out a number of discrepancies with the current proteobacterial taxonomy that deserve further consideration. Given the ever increasing availability of complete genome sequences, and although additional studies and further developments are required, we anticipate that r-proteins will likely represent the next generation standard for prokaryotic systematics.

Acknowledgments

C.B.-A. was funded by an ATIP from the CNRS and is member of the Institut Universitaire de France. M.G was the recipient of a PhD grant from the French Ministère de l'Education Nationale. H.G.R. was supported by a post-doctoral fellowship from CNRS. R.P. was supported by a grant from the Agence Nationale de la Recherche (ANR-07-BLAN-02). This project was supported by the ANR-07-BLAN-02 (BioSuf) and ANR-10-BINF-01-01 (Ancestrome) grants. We thank the PRABI (Pôle Rhône-Alpes de Bioinformatique) for providing computing facilities. We thank Long-Fey Wu for sharing magneto-ovoid strain MO-I unpublished data. We would like to acknowledge Manolo Gouy, Laura Eme, Céline Petitjean, Ji Boyang, Rym Agrebi, and especially Simonetta Gribaldo for stimulating discussions.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.02.013.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.
Abby, S.S., Tannier, E., Gouy, M., Daubin, V., 2010. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinform. 11, 324.
Abby, S.S., Tannier, E., Gouy, M., Daubin, V., 2012. Lateral gene transfer as a support for the tree of life. Proc. Natl. Acad. Sci. USA 109, 4962–4967.
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 database search programs. Nucl. Acids Res. 25, 3389–3402.
Badger, J.H., Eisen, J.A., Ward, N.L., 2005. Genomic analysis of Hyphomonas neptunium contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders 'Rhodobacterales' and Caulobacterales. Int. J. Syst. Evol. Microbiol. 55, 1021–1026.
Bazylinski, D.A., Williams, T.J., Lefevre, C.T., Berg, R.J., Zhang, C.L., Bowser, S.S., Dean, A.J., Beveridge, T.J., 2013. *Magnetococcus marinus* gen. nov., sp. nov., a marine, magnetotactic bacterium that represents a novel lineage (*Magnetococcaceae* fam. nov., *Magnetococcales* ord. nov.) at the base of the Alphaproteobacteria. Int. J. Syst. Evol. Microbiol. 63, 801–808.
Bergsten, J., 2005. A review of long-branch attraction. Cladistics 21, 163–193.
Boussau, B., Gouy, M., 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55, 756–768.
Brindefalk, B., Ettema, T.J., Viklund, J., Thollesson, M., Andersson, S.G., 2011. A phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial relatives unrelated to the SAR11 clade. PloS One 6, e24457.
Brochier, C., Bapteste, E., Moreira, D., Philippe, H., 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Gen.: TIG 18, 1–5.
Brochier, C., Philippe, H., Moreira, D., 2000. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Gen.: TIG 16, 529–533.
Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J., 2001. Universal trees based on large combined protein sequence data sets. Nat. Gen. 28, 281–285.
Cavalier-Smith, T., 2002. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int. J. Syst. Evol. Microbiol. 52, 7–76.
Chen, K., Roberts, E., Luthey-Schulten, Z., 2009. Horizontal gene transfer of zinc and non-zinc forms of bacterial ribosomal protein S4. BMC Evol. Biol. 9, 179.
Chiari, Y., Cahais, V., Galtier, N., Delsuc, F., 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10, 65.
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.
Cohen, O., Gophna, U., Pupko, T., 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol. Biol. Evol. 28, 1481–1489.
Davalos, L.M., Perkins, S.L., 2008. Saturation and base composition bias explain phylogenomic conflict in Plasmodium. Genomics 91, 433–442.
Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.
DeLuca, T.F., Cui, J., Jung, J.Y., St Gabriel, K.C., Wall, D.P., 2012. Roundup 2.0: enabling comparative genomics for over 1800 genomes. Bioinformatics 28, 715–716.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32, 1792–1797.
Embley, T.M., van der Giezen, M., Horner, D.S., Dyal, P.L., Bell, S., Foster, P.G., 2003. Hydrogenosomes, mitochondria and early eukaryotic evolution. IUBMB Life 55, 387–395.
Emerson, D., Rentz, J.A., Lilburn, T.G., Davis, R.E., Aldrich, H., Chan, C., Moyer, C.L., 2007. A novel lineage of proteobacteria involved in formation of marine Fe-oxidizing microbial mat communities. PloS One 2, e667.
Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.
Felsenstein, J., 2004. Inferring phylogenies. Sunderland, Massachusetts.
Galtier, N., Gouy, M., 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15, 871–879.
Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., Rappe, M.S., Short, J.M., Carrington, J.C., Mathur, E.J., 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309, 1242–1245.
Gribaldo, S., Brochier, C., 2009. Phylogeny of prokaryotes: does it exist and why should we care? Res. Microbiol. 160, 513–521.
Gribaldo, S., Philippe, H., 2002. Ancient phylogenetic relationships. Theor. Popul. Biol. 61, 391–408.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Gupta, R.S., 2000. The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol. Rev. 24, 367–402.
Hesselmann, R.P., Werlen, C., Hahn, D., van der Meer, J.R., Zehnder, A.J., 1999. Enrichment, phylogenetic analysis and detection of a bacterium that performs enhanced biological phosphate removal in activated sludge. Syst. Appl. Microbiol. 22, 454–465.
Husnik, F., Chrudimsky, T., Hypsa, V., 2011. Multiple origins of endosymbiosis within the Enterobacteriaceae (gamma-Proteobacteria): convergence of complex phylogenetic approaches. BMC Biol. 9, 87.
Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. USA 96, 3801–3806.
Ji, B., Zhang, S.D., Arnoux, P., Rouy, Z., Alberto, F., Philippe, N., Murat, D., Zhang, W.J., Rioux, J.B., Ginet, N., Sabaty, M., Mangenot, S., Pradel, N., Tian, J., Yang, J., Zhang, L., Zhang, W., Pan, H., Henrissat, B., Coutinho, P.M., Li, Y., Xiao, T., Medigue, C., Barbe, V., Pignol, D., Talla, E., Wu, L.F., 2013. Comparative genomic analysis

provides insights into the evolution and niche adaptation of marine *Magnetospira* sp. QH-2 strain. Environ. Microbiol.
Jobb, G., von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 18.
Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Briefings Bioinform. 9, 286–298.
Kelly, D.P., Wood, A.P., 2000. Reclassification of some species of *Thiobacillus* to the newly designated genera *Acidithiobacillus* gen. nov., *Halothiobacillus* gen. nov. and *Thermithiobacillus* gen. nov. Int. J. Syst. Evol. Microbiol. 50, 489–500.
Kelly, S., Wickstead, B., Gull, K., 2010. Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. In: Proceedings. Biological Sciences/The Royal Society.
Kersters, K., Devos, P., Gillis, M., Swings, J., Vandamme, P., Stackebrandt, E., 2006. Introduction to the proteobacteria. In: Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.H., Stackebrandt, E. (Eds.), In the Prokaryotes: A Handbook on the Biology of Bacteria. Springer, New York, pp. 3–37.
Kuzniar, A., van Ham, R.C., Pongor, S., Leunissen, J.A., 2008. The quest for orthologs: finding the corresponding gene across genomes. Trends Gen.: TIG 24, 539–551.
Lang, B.F., Gray, M.W., Burger, G., 1999. Mitochondrial genome evolution and the origin of eukaryotes. Annu. Rev. Genet. 33, 351–397.
Lang, J.M., Darling, A.E., Eisen, J.A., 2013. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PloS One 8, e62510.
Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.
Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7 (Suppl. 1), S4.
Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.
Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109.
Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320.
Lee, K.B., Liu, C.T., Anzai, Y., Kim, H., Aono, T., Oyaizu, H., 2005. The hierarchical system of the 'Alphaproteobacteria': description of *Hyphomonadaceae* fam. nov., *Xanthobacteraceae* fam. nov. and *Erythrobacteraceae* fam. nov. Int. J. Syst. Evol. Microbiol. 55, 1907–1919.
Lefevre, C.T., Bernadac, A., Yu-Zhang, K., Pradel, N., Wu, L.F., 2009. Isolation and characterization of a magnetotactic bacterial culture from the Mediterranean Sea. Environ. Microbiol. 11, 1646–1657.
Leigh, J.W., Schliep, K., Lopez, P., Bapteste, E., 2011. Let them fall where they may: congruence analysis in massive phylogenetically messy data sets. Mol. Biol. Evol. 28, 2773–2785.
Lerat, E., Daubin, V., Moran, N.A., 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol. 1, E19.
Lopez-Garcia, P., Moreira, D., 2008. Tracking microbial biodiversity through molecular and genomic ecology. Res. Microbiol. 159, 67–73.
Maki, Y., Yoshida, H., Wada, A., 2000. Two proteins, YfiA and YhbH, associated with resting ribosomes in stationary phase *Escherichia coli*. Genes Cells 5, 965–974.
Marthey, S., Aguileta, G., Rodolphe, F., Gendrault, A., Giraud, T., Fournier, E., Lopez-Villavicencio, M., Gautier, A., Lebrun, M.H., Chiapello, H., 2008. FUNYBASE: a FUNgal phYlogenomic dataBASE. BMC Bioinform. 9, 456.
Matte-Tailliez, O., Brochier, C., Forterre, P., Philippe, H., 2002. Archaeal phylogeny based on ribosomal proteins. Mol. Biol. Evol. 19, 631–639.
McCutcheon, J.P., McDonald, B.R., Moran, N.A., 2009. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Gen. 5, e1000565.
Miller, W.G., Parker, C.T., Rubenfield, M., Mendz, G.L., Wosten, M.M., Ussery, D.W., Stolz, J.F., Binnewies, T.T., Hallin, P.F., Wang, G., Malek, J.A., Rogosin, A., Stanker, L.H., Mandrell, R.E., 2007. The complete genome sequence and analysis of the epsilonproteobacterium *Arcobacter butzleri*. PloS One 2, e1358.
Moran, N.A., McCutcheon, J.P., Nakabachi, A., 2008. Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. 42, 165–190.
Morris, R.M., Rappe, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., Giovannoni, S.J., 2002. SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420, 806–810.
Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., Hattori, M., 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 314, 267.
Philippe, H., 1993. MUST, a computer package of management utilities for sequences and trees. Nucl. Acids Res. 21, 5264–5272.
Philippe, H., Douady, C.J., 2003. Horizontal gene transfer and phylogenetics. Curr. Opin. Microbiol. 6, 498–505.
Philippe, H., Forterre, P., 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49, 509–523.
Philippe, H., Laurent, J., 1998. How good are deep phylogenetic trees? Curr. Opin. Gen. Develop. 8, 616–623.
Philippe, H., Sörhannus, U., Baroin, A., Perasso, R., Gasse, F., Adoutte, A., 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J. Evol. Biol. 7, 247–265.
Puigbo, P., Wolf, Y.I., Koonin, E.V., 2010. The tree and net components of prokaryote evolution. Gen. Biol. Evol. 2, 745–756.
Rappe, M.S., Connon, S.A., Vergin, K.L., Giovannoni, S.J., 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418, 630–633.
Rodriguez-Ezpeleta, N., Embley, T.M., 2012. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. PloS One 7, e30520.
Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51, 492–508.
Shimodaira, H., Hasegawa, M., 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17, 1246–1247.
Spring, S., Lins, U., Amann, R., Schleifer, K.H., Ferreira, L.C., Esquivel, D.M., Farina, M., 1998. Phylogenetic affiliation and ultrastructure of uncultured magnetic bacteria with unusually large magnetosomes. Arch. Microbiol. 169, 136–147.
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.
Steindler, L., Schwalbach, M.S., Smith, D.P., Chan, F., Giovannoni, S.J., 2011. Energy starved Candidatus Pelagibacter ubique substitutes light-mediated ATP production for endogenous carbon respiration. PloS One 6, e19725.
Suarez, S., Ferroni, A., Lotz, A., Jolley, K.A., Guérin, P., Leto, J., Dauphin, B., Jamet, A., Maiden, M.C., Nassif, X., Armengaud, J., 2013. Ribosomal proteins as biomarkers for bacterial identification by mass spectrometry in the clinical microbiology laboratory. J. Microbiol. Meth. 94, 390–396.
Swithers, K.S., Gogarten, J.P., Fournier, G.P., 2009. Trees in the web of life. J. Biol. 8, 54.
Tamames, J., Gil, R., Latorre, A., Pereto, J., Silva, F.J., Moya, A., 2007. The frontier between cell and organelle: genome analysis of Candidatus Carsonella ruddii. BMC Evol. Biol. 7, 181.
Tamura, H., Hotta, Y., Sato, H., 2013. Novel accurate bacterial discrimination by MALDI-time-of-flight MS based on ribosomal proteins coding in S10-spc-alpha operon at strain level S10-GERMS. J. Am. Soc. Mass Spectrom. 24, 1185–1193.
Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M., 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA 102, 13950–13955.
Thrash, J.C., Boyd, A., Huggett, M.J., Grote, J., Carini, P., Yoder, R.J., Robbertse, B., Spatafora, J.W., Rappe, M.S., Giovannoni, S.J., 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Scientific Reports 1, 13.
Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M.E., Frapy, E., Garry, L., Ghigo, J.M., Gilles, A.M., Johnson, J., Le Bouguenec, C., Lescat, M., Mangenot, S., Martinez-Jehanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M.A., Pichon, C., Rouy, Z., Ruf, C.S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Medigue, C., Rocha, E.P., Denamur, E., 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Gen. 5, e1000344.
Van Leuven, J.T., McCutcheon, J.P., 2012. An AT mutational bias in the tiny GC-rich endosymbiont genome of Hodgkinia. Gen. Biol. Evol. 4, 24–27.
Viklund, J., Ettema, T.J., Andersson, S.G., 2011. Independent Genome Reduction and Phylogenetic Reclassification of the Oceanic SAR11 Clade. Mol. Biol. Evol.
Wang, Z., Wu, M., 2013. A phylum-level bacterial phylogenetic marker database. Mol. Biol. Evol. 30, 1258–1262.
Williams, K.P., Gillespie, J.J., Sobral, B.W., Nordberg, E.K., Snyder, E.E., Shallom, J.M., Dickerman, A.W., 2010. Phylogeny of gammaproteobacteria. J. Bacteriol. 192, 2305–2314.
Williams, K.P., Sobral, B.W., Dickerman, A.W., 2007. A robust species tree for the alphaproteobacteria. J. Bacteriol. 189, 4578–4586.
Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA 74, 5088–5090.
Wu, D., Jospin, G., Eisen, J.A., 2013. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PloS One 8, e77033.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford, England.
Yang, Z., Roberts, D., 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12, 451–458.
Yutin, N., Puigbò, P., Koonin, E.V., Wolf, Y.I., 2012. Phylogenomics of Prokaryotic Ribosomal Proteins. PloS One 7, e36972.