Phylogenomics of Rhodobacteraceae Reveals Evolutionary Adaptation to Marine and Non-Marine Habitats

Meinhard Simon, Carmen Scheuner, Jan P. Meier-Kolthoff, Thorsten Brinkhoff, Irene Wagner-Döbler, Marcus Ulbrich, Hans-Peter Klenk, Dietmar Schomburg, Jörn Petersen, Markus Göker

Supplementary file 1

Preamble

This file is provides background information on the materials and methods at a level of detail unsuitable for the main manuscript.

Genome-scale phylogenetic analysis

Genome sampling

In order to address the questions raised in the main manuscript, we carried out phylogenomic analyses of 106 sequenced *Rhodobacteraceae* genomes including 73 roseobacters, 20 of the *Paracoccus*/*Rhodobacter* group and 13 genomes of the *Labrenzia*/*Stappia* group, which phylogenetically does not belong to *Rhodobacteraceae* and served as an outgroup. This large set of genomes allows for distinct gene selections ranging from the analysis of the core genes to the "full" supermatrix, as well as for distinct inference algorithms. Protein sequences from the 106 genomes available in November 2013 were retrieved from the IMG website (http://img.jgi.doe.gov/cgi-bin/w/main.cgi), the J. Craig Venter Institute (Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project; https://moore.jcvi.org/moore/) and from NCBI (supplementary file 1). The 13 strains taxonomically assigned to *Rhodobacteraceae* but rather placed within *Rhizobiales* in 16S rRNA gene analyses (Pujalte et al. 2014; Munoz et al. 2011) were used as outgroup. An extended dataset including 132 genomes (as of June 2014) was also analysed phylogenetically, using the recommended setting of the efficient GBDP approach, applied to the entire proteomes (Meier-Kolthoff et al. 2014; Auch et al. 2006). Digital DNA:DNA hybridization values calculated via nucleotide GBDP were used to check the species affiliation of all genomes (Meier-Kolthoff et al. 2013a; Auch et al. 2010).

Rationale. See introduction of main manuscript.

Generation of data matrices for phylogenetic analysis

The proteome sequences were phylogenetically investigated using the DSMZ phylogenomics pipeline as previously described (Andersson et al. 2011; Breider et al. 2014; Frank et al. 2014; Stackebrandt et al. 2014; Verbarg et al. 2014). Clusters of orthologs were determined with a re-implementation of the OrthoMCL algorithm (Li et al. 2003) using NCBI BLAST version 2.2.25 (Altschul et al. 1997) and in conjunction with MCL (van Dongen 2000) version 11-294 (http://micans.org/mcl/) under default settings. OrthoMCL clusters containing inparalogs were

reduced by selecting the most “central” sequence from each genome, aligned using MUSCLE version 3.8.31 (Edgar 2004), and the alignments filtered with the program scan_orphanerrs from the RASCAL package version 1.3.4 (Thompson et al. 2003) as well as GBLOCKS version 0.91b (Castresana 2000). Three main distinct supermatrices (concatenated alignments) were generated: (i) using the “core genes” only, i.e. those alignments that contain sequences from all genomes, (ii) a “full” matrix using all alignments comprising at least four sequences and (iii) the “full” matrix filtered with MARE (http://mare.zfmk.de) (Meusemann et al. 2010) under the constraint to not remove genomes. The core-gene matrices were also subjected to further reductions by restricting them to their 50, 100, 150 and 200 most conserved genes (up to 250 for the matrix generated without outgroup), with non-conservation measured as the average entropy of the alignment columns. To detect possible long-branch attraction artefacts (Bergsten 2005) analyses were also conducted after long-branch extraction (Siddall and Whiting 1999) by removing the 13 (distant) outgroup strains. As the gene compositions of the supermatrices depend on the set of strains, they were also generated anew after outgroup removal. Trees without outgroup were subjected to root estimation with Least-Squares Dating as implemented in LSD version 0.2 (To et al. 2015) as an alternative to outgroup rooting. For comparative purposes, 16S rRNA gene sequences were extracted from the genomes using RNAmmer version 1.2 (Lagesen and Hallin 2007) and their pairwise similarities determined as previously described (Meier-Kolthoff et al. 2013b).

Rationale. See below.

Phylogenetic analysis

ML (Felsenstein 1981) and MP (Fitch 1977; Goloboff 2003) phylogenetic trees were inferred from the data matrices as previously described (Breider et al. 2014; Andersson et al. 2011; Frank et al. 2014; Stackebrandt et al. 2014; Verbarg et al. 2014). The Pthreads-parallelized RAxML package version 8.2.4 (Stamatakis et al. 2007) was used for ML, applying CAT for modelling rate heterogeneity, fast bootstrapping in conjunction with the autoMRE bootstopping criterion (Pattengale et al. 2010) and subsequent search for the best tree (Stamatakis et al. 2008). The best ML amino-acid substitution model (a single one for the entire matrix) was determined beforehand by comparing the resulting log likelihoods on an MP starting tree. Additionally, ML phylogenies for each supermatrix were calculated with ExaML version 3.0.7 (Stamatakis and Aberer 2013) using MP starting trees calculated with RAxML. Here, the best amino-acid substitution model was fitted with and used by ExaML for each gene separately. One hundred replicates of ordinary bootstrapping (modified to guarantee that the number of characters remain the same for each gene, not only for the entire data set) were conducted by re-identifying the best model in each replicate, as well as of partition bootstrapping (Siddall 2010) by randomly resampling entire genes with replacement instead of individual alignment columns and, accordingly, keeping the originally estimated models for each gene. Tree searches under MP were conducted with TNT version 1.1 (Goloboff et al. 2008) using ten random addition sequences, saving up to ten trees per replication, swapping trees with TBR and keeping only the best trees found, and also subjected to 100 rounds of ordinary and partition bootstrapping, respectively. Bootstrap support was calculated for all matrices except for the “full” matrix under ML for reasons of running time.

Rationale. The often higher resolution when using more genes (up to entire genomes) might argue for the “total evidence” approach (Kluge 1989; Lienau and Desalle 2009; Breider et al. 2014).

However, can single genes, yielding conflicting phylogenies, be reliably combined to supermatrices (Klenk and Göker 2010)? Even partially overlapping sets of genes might yield distinct topologies (Breider et al. 2014). Even the use of ordinary bootstrapping has been contested in this context (Siddall 2010). Moreover, applying a single substitution model to supermatrices comprising many genes might be oversimplified, as phylogenetic artefacts such as long-branch attraction (Felsenstein 1978; Bergsten 2005) not only affect Maximum Parsimony (MP) but also ML under too simplistic models. For these reasons, we assessed distinct gene selections, strain selections, inference algorithms, and bootstrapping approaches.

Comparisons between trees

The affiliation of the strains to the *Roseobacter* group (supplementary file 1) was inferred from the literature (Pujalte et al. 2014; Brinkhoff et al. 2008; Ivanova et al. 2010; Tang et al. 2012; Giebel et al. 2013; Riedel et al. 2013). To measure the conflict, if any, between the genomic data and the monophyly of this group, accordingly constrained ML and MP searches were conducted and site-wise ML and MP scores calculated from unconstrained and constrained trees (Felsenstein 2004). Wilcoxon and T-tests as implemented in R (R core team 2014) were applied to compare the scores, optionally summing them up per gene beforehand, thus treating one gene as a single character much like the partition bootstrap. For the ML trees, the approximately unbiased test as implemented in CONSEL (Shimodaira and Hasegawa 2001) was compared. Potential conflict with the 16S rRNA gene (sequences extracted from the genomes aligned using MUSCLE) was measured in the same manner, using constraints derived from the supermatrix trees. Major sublineages of *Rhodobacteraceae* were newly inferred from the phylogenomic trees in a non-arbitrary manner by detecting the maximally inclusive subtrees that were maximally supported under the different bootstrapping methods; these clades were numbered according to their size.

Rationale. Analyses of the 16S rRNA gene, from which the notation of a "*Roseobacte*r clade" were derived, suffer from a lack of resolution (Buchan et al. 2005; Newton et al. 2010; Breider et al. 2014). For instance, in a recent review (Pujalte et al. 2014), five 16S rRNA gene-derived sublineages within *Rhodobacteracea*e, among them the *Roseobacte*r group, were presented, but without statistical support for the according branches and without an explicit rationale for this number of sublineages. In more thorough phylogenetic analyses that applied bootstrapping, the 16S rRNA gene topologies were only poorly resolved within *Rhodobacteraceae* and did not yield statistical support for a "*Roseobacte*r clade" (Breider et al. 2014).

Analysis of character evolution

Phylogeny-aware correlation analysis

The BayesDiscrete module of BayesTraits (Pagel et al. 2004) version 2.0 was used for detecting phylogenetic correlations between pairs of discrete (binary) traits (Pagel 1994). For each pair of genes obtained as described below, the likelihood ratio between the model for independent and correlated evolution was tested against a chi-squared distribution with one degree of freedom (Pagel 1994) and $\alpha = 0.01$. The analyses were performed using the BayesDiscrete ML method under default parameters (with the exception of 'mltries' set to 100) and the rooted ML phylogenies as reference trees. For selected characters, the ratio of the sum of the rates of change indicating co-occurrence of marine habitat and genomic feature (q21, q24, q31 and q34) to the overall sum of the

rates of change was calculated to verify the tendency of change. To correct for the influence of only partially sequenced genomes, distinct samplings were analysed: (i) all genomes; (ii) without *Paracoccus denitrificans* SD1 (plasmids not sequenced) and *Ruegeria mobilis* F1926 (>1000 contigs) and (iii) also without *Oceanicola* sp. S124 and *Rhodovulum* sp. PH10 (both >200 contigs). Only results stable with respect to topology and genome sampling were considered further. Strains for which gene information was lacking were specifically removed from each pairwise comparison. The evolution of selected genes was visualized using the Ancestral State Reconstruction package (v2.75) of Mesquite v2.75 (Maddison and Maddison 2011) under ordered MP.

Rationale. See the main manuscript for the presentation of the BayesTraits tests conducted with these genomic features.

Interpretation of habitats

Assignment to habitats (supplementary file 1) was based on isolation sources as available from the literature. Because of otherwise incomplete information, only marine and non-marine habitats were distinguished. Habitats with a salt concentration comparable to the marine environment, such as saline soil or water from a hypersaline lake, were considered equivalent. Saline, hypersaline and marine habitats are often commonly reported for closely related strains, as, e.g., in *Rhodovulum* (Hiraishi et al. 1994) and *Thiomicrospira* (Brinkhoff and Muyzer 1997).

Rationale. See Supplementary file 2.

Enzyme, pathway and COG annotation

The EnzymeDetector (Quester and Schomburg 2011) was used for an initial overview of the available enzyme annotations of the 106 genomes from NCBI (Sayers et al. 2010), KEGG (Kanehisa and Goto 2000) and UniProt/SwissProt (The UniProt Consortium 2013). These data were improved by manually compiled strain-specific enzyme information from BRENDA and AMENDA (Schomburg et al. 2013), sequence pattern searches with BrEPS (Bannert et al. 2010) and a sequence-based similarity analysis by BLAST (Altschul et al. 1990) against annotated enzymes provided by UniProt (UniProt Consortium 2013). Further, orthology data from the KEGG (Kanehisa and Goto 2000), PATRIC (Gillespie et al. 2011) and PFAM (Finn et al. 2014) databases were requested for the available annotations. To validate the completeness of each proteome, its proportion of enzymes was calculated (see supplementary file 1). A binary (presence/absence) matrix of the complete *Rhodobacteraceae* enzyme pool was created for evaluation, as previously done for the carbohydrate and amino-acid catabolism of the *Roseobacter* group (Drüppel et al. 2014; Wiegmann et al. 2014). Furthermore, these enzymes were mapped on MetaCyc pathways (Caspi et al. 2012) to create a binary pathway matrix (Chang et al. 2015). In order to get a general overview it was initially assumed that a pathway was present if at least 75% of the enzymes were present. For pathways discussed in the paper this was manually refined considering the fact that certain enzymes are essential for the whole pathway to be functional. The COG content of the genomes was taken from the gene annotations identified with Prodigal (Hyatt et al. 2010) as part of the Integrated Microbial Genomes Expert Review (IMG/ER) annotation pipeline (Mavromatis et al. 2009) were used. IMG translates the predicted CDSs and uses them to search the NCBI nonredundant database, UniProt, TIGR-Fam, Pfam, PRIAM, KEGG, COG and InterPro databases.

Rationale. See the main manuscript for the presentation of the BayesTraits tests conducted with these genomic features.

Annotation of plasmids, chromids, and flagellar gene clusters

Replication systems of extrachromosomal elements were identified via BLASTP and TBLASTN searches with the RepA-I (YP_006564734.1), RepB-I (YP_006564673.1) and DnaA-like I (YP_006575239.1) replicases from *Phaeobacter inhibens* DSM 17395 and the RepABC-1 equivalent from *Dinoroseobacter shibae* DSM 16493^T (YP_001542300.1). The individual extrachromosomal element compatibility groups (Petersen 2011) were determined with phylogenetic analyses of the replicases as previously described (Petersen et al. 2009; Petersen et al. 2011). The flagellar gene clusters (FGCs) were identified via BLASTP search at the IMG platform of JGI with the FliF query sequence of *Dinoroseobacter shibae* DSM 16493^T. The presence of the complete superoperon was checked individually based on an inspection of the adjacent genes. The identification of the three flagellar types (*fla1*, *fla2*, *fla3*) based on the structural composition of the individual FGCs and the presence of diagnostic genes (Frank et al. 2015a).

Rationale. Genome plasticity, i.e. the ability to take up DNA and integrate it into the genome, seems to be one explanation for the adaptability and diversity within the *Roseobacter* group (Luo and Moran 2014). Applying a likelihood-based ancestral genome content reconstruction method, Luo et al. (2013) predicted a genome reduction from a large common ancestral roseobacter genome with two subsequent episodes of genome innovation and expansion via lateral gene transfer. Extant roseobacters exhibit genome sizes ranging from 2.22 to 5.5 Mbp (Voget et al. 2015). A distinct signature is the great flexibility of the genome architecture of strains affiliated to this group. Strictly pelagic strains contain a single chromosome with genome streamlining features and between zero and two extrachromosomal replicons (ECRs; Voget et al. 2015). Surface-associated strains harbour at least one, but mostly between three and up to twelve coexisting ECRs in *Marinovum algicola* with 30% of the genomic information encoded on ECRs (Pradella et al. 2010; Voget et al. 2015). Extrachromosomal elements comprise stably maintained chromids with a codon usage comparable to that of the chromosome and genuine plasmids with a deviant codon usage that are subjected to conjugative transfer mediated by characteristic type-IV secretion systems (Harrison et al. 2010; Petersen et al. 2013). Four different replication systems (RepA, RepB, RepABC, DnaA-like) with about 20 phylogenetically distinguishable compatibility groups have been identified in the *Roseobacter* group so far (Petersen et al. 2013), but a comprehensive analysis of the genome architecture is not yet available for the entire *Rhodobacteraceae* family. Further, it is unknown whether the genome architecture is distinct for marine and non-marine strains of this family.

Tests for phylogenetic inertia

Numeric features of the genome sequences ranging from the overall genome size to the number of single COGs were tested for phylogenetic inertia (Diniz-Filho et al. 1998). For Pearson and Kendall correlation analyses ($\alpha = 0.01$) with continuous genomic features, ML phylogenies were transformed into eigenvectors using the AxPcoords program (Stamatakis et al. 2007) and the significant eigenvectors determined by comparing with a broken-stick distribution (Legendre and Legendre 1998). For testing phylogenetic inertia of discrete genomic characters such as the number of ECRs of a certain type, their score was determined under ordered MP in conjunction with the

distinct trees and compared to the score of the same character after permuting the tips of the tree; 1000 replications were applied to obtain a tail probability ($\alpha = 0.01$).

Rationale. See the main manuscript for interpretations of phylogenetic inertia, or lack thereof.

Genome-derived oligotrophy index

Among the genomic features characteristic for oligotrophic and copiotrophic lifestyles (Lauro et al. 2009), those based on COG classes seemed to be least affected by annotation quality and evolutionary fluctuations of gene content. As they also differentiated best between a set of six oligotrophic (*Oceanicola batsensis* HTCC2597^{T}, *Planktomarina temperata* RCA23^{T}, "*Rhodobacterales* bacterium" HTCC2255, "*Rhodobacteraceae* bacterium" HTCC2150, "*Rhodobacteraceae* bacterium" HTCC2083, *Sulfitobacter* sp. NAS-14.1) and three copiotrophic test strains (*Leisingera caerulea* 13^{T}, *Phaeobacter inhibens* DSM 17395, *Ruegeria* sp. TM1040), the index log(#I + #Q + 1) - log(#K + #N + #T + #V + 1) was chosen as a genomic proxy for oligotrophy, i.e. the log-transformed relation of the number of COGs in the classes I and Q, typical for oligotrophs, to the number of COGs in the classes K, N, T and V, typical for copiotrophs (Lauro et al. 2009).

Rationale. Pelagic roseobacters such as *Planktomarina temperata* are well adapted to an oligotrophic life style (Voget et al. 2015). This is not only evident from their streamlined genomes but also from a comparison of distinct clusters of orthologous groups (COG) with other oligotrophic and copiotrophic bacteria such as *Sphingopyxis alaskensis* RB 2256 (*Sphingomonadaceae*, *Alphaproteobacteria*) and *Photobacterium angustum* S14 (*Vibrionaceae*, *Gammaproteobacteria*). In oligotrophic bacteria COG categories for motility (N), transcription (K), defence mechanisms (V) and signal transduction (T) constitute lower fractions, whereas COG categories for lipid transport and metabolism (I) and secondary metabolites, biosynthesis, transport and catabolism (Q) constitute higher fractions as compared to copiotrophic bacteria (Lauro et al. 2009). However, it has not been examined systematically whether other roseobacters and even other *Rhodobacteraceae* with greatly different genome sizes and different physiologies comply with this distinction of oligotrophic and copiotrophic life styles.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

2. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.

3. Anderson I, Scheuner C, Göker M, Mavromatis K, Hooper SD, Porat I, Klenk H-P, Ivanova N, Kyrpides NC. 2011. Novel insights into the diversity of catabolic metabolism from ten haloarchaeal genomes. *PLoS One* 6:e20237.

4. Auch AF, Henz SR, Holland BR, Göker M. 2006. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. *BMC Bioinformatics* 7:350.

5. Auch AF, Klenk H-P, Göker M. 2010. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci*. 2:142–148.

6. Bannert C, Welfle a, Aus dem Spring C, Schomburg D. 2010. BrEPS: a flexible and automatic protocol to compute enzyme-specific sequence profiles for functional annotation. *BMC Bioinformatics* 11:589.

7. Bergsten J, A. 2005. A review of long-branch attraction. *Cladistics*. 21:163–193.

8. Breider S, Scheuner C, Schumann P, Fiebig A, Petersen J, Pradella S, Klenk HP, Brinkhoff T, Göker M. 2014. Genome-scale data suggest reclassifications in the Leisingera-Phaeobacter cluster including proposals for *Sedimentitalea* gen. nov. and *Pseudophaeobacter* gen. nov. *Front Microbiol*. 5:416.

9. Brinkhoff T, Muyzer G. 1997. Increased species diversity and extended habitat range of sulfur-oxidizing *Thiomicrospira* spp. *Appl Environ Microbiol*. 63:3789–96.

10. Brinkhoff T, Giebel HA, Simon M. 2008. Diversity, ecology, and genomics of the *Roseobacter* clade: A short overview. Arch Microbiol. 189:531–539.

11. Buchan A, González JM, Moran MA. 2005. Overview of the marine *Roseobacter* lineage. *Appl Environ Microbiol*. 71:5665–5677.

12. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA et al.. 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 40:D742–D753.

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

14. Chang A, Schomburg I, Placzek S, Jeske L, Ulbrich M, Xiao M, Sensen CW, Schomburg D. 2015. BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic Acids Res. 43:D439–D446.

15. Diniz-Filho JAF, De Sant'Ana CER, Bini LM. 1998. An eigenvector method for estimating phylogenetic inertia. *Evolution*. 52:1247–1262.

16. Drüppel K, Hensler M, Trautwein K, Koßmehl S, Wöhlbrand L, Schmidt-Hohagen K, Ulbrich M, Bergen N, Meier-Kolthoff JP, Göker M, et al. 2014. Pathways and substrate-specific regulation of amino acid degradation in *Phaeobacter inhibens* DSM 17395 (archetype of the marine Roseobacter clade). *Environ Microbiol*. 16:218–238.

17. Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

18. Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Biol*. 27:401–410.

19. Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17:368–376.

20. Felsenstein J. 2004. Inferring Phylogenies. Sunderland, Massachusetts: Sinauer Associates.

21. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: The protein families database. *Nucleic Acids Res*. 42:D222–D230.

22. Fitch WM. 1977. Toward defining the course of evolution: minimum change on a specified tree topology. Syst Zool. 20:406–16.

23. Frank O, Pradella S, Rohde M, Scheuner C, Klenk H-P, Göker M, Petersen J. 2014. Complete genome sequence of the *Phaeobacter gallaeciensis* type strain CIP 105210T (= DSM 26640T = BS107T). *Stand Genomic Sci*. 9: 914–932.

24. Frank O, Göker M, Pradella S, Petersen J. 2015a. Ocean's twelve: Flagellar and biofilm chromids in the multipartite genome of *Marinovum algicola* DG898 exemplify functional compartmentalization in Proteobacteria. Environ Microbiol. 17: 4019–4034.

25. Giebel HA, Kalhoefer D, Gahl-Janssen R, Choo YJ, Lee K, Cho JC, Tindall B, Rhiel E, Beardsley C, Aydogmus OO, et al. 2013. *Planktomarina temperata* gen. nov., sp. nov., belonging to the globally distributed RCA cluster of the marine *Roseobacter* clade, isolated from the German Wadden Sea. *Int J Syst Evol Microbiol*. 63:4207–4217.

26. Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, Nordberg EK, Scott M, Schulman JR, Snyder EE, Sullivan DE, Wang C, Warren A, Williams KP, Xue T, Yoo HS, Zhang C, Zhang Y, Will R, Kenyon RW, Sobral BW. 2011. Patric: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun*. 79:4286–4298.

27. Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylogenetic analysis. *Cladistics* 25:774–786.

28. Goloboff PA. 2003. Parsimony, likelihood, and simplicity. *Cladistics* 19:91–103.

29. Harrison PW, Lower RPJ, Kim NKD, Young JPW. 2010. Introducing the bacterial “chromid”: Not a chromosome, not a plasmid. *Trends Microbiol*. 18:141–148.

30. Hiraishi a., Ueda Y. 1994. Intrageneric structure of the genus *Rhodobacter*: Transfer of *Rhodobacter sulfidophilus* and related marine species to the genus *Rhodovulum* gen. nov. *Int J Syst Bacteriol*. 44:15–23.

31. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 11:Article119.

32. Ivanova EP, Webb H, Christen R, Zhukova NV, Kurilenko VV, Kalinovskaya NI, Crawford RJ. 2010. *Celeribacter neptunius* gen. nov., sp. nov., a new member of the class Alphaproteobacteria. *Int J Syst Evol Microbiol*. 60:1620–1625.

33. Kanehisa M, Goto S. 2000. Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28:27–30.

34. Klenk H-P, Göker M. 2010. En route to a genome-based classification of Archaea and Bacteria? *Syst Appl Microbiol*. 33:175–182.

35. Kluge AG. A 1989. Concern for evidence and a phylogenetic hypothesis of relationships among epicrates (Boidae, Serpentes). Syst Zool. 38:7–25.

36. Lagesen K, Hallin P. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 35: 3100–3108.

37. Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, DeMaere MZ, Ting L, Ertan H, Johnson J, et al. 2009. The genomic basis of trophic strategy in marine bacteria. Proc Natl Acad Sci USA. 106:15527–15533.

38. Legendre P, Legendre L. 1998. Numerical Ecology. 2nd ed. Elsevier, Amsterdam.

39. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.

40. Lienau KE, Desalle R. 2009. Evidence, content and corroboration and the tree of life. *Acta Biotheor*. 57:187–199.

41. Luo H, Moran MA. 2014. Evolutionary ecology of the marine *Roseobacter* clade. *Microbiol Mol Biol Rev*. 78:573–587.

42. Luo H, Csuros M, Hughes AL, Moran MA. 2013. Evolution of divergent life history strategies in marine alphaproteobacteria. MBio4:e00373–13.

43. Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. Mesquite website: http:// mesquiteproject.org.

44. Mavromatis K, Ivanova NN, Chen I-M a, Szeto E, Markowitz VM, Kyrpides NC. 2009. The DOE-JGI standard operating procedure for the annotations of microbial genomes. *Stand Genomic Sci*. 1:63–67.

45. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. 2013a. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14:Article60.

46. Meier-Kolthoff JP, Göker M, Spröer C, Klenk HP. 2013b. When should a DDH experiment be mandatory in microbial taxonomy? *Arch Microbiol*. 195:413–418.

47. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. 2014. Highly parallelized inference of large genome-based phylogenies. *Concurr Comp Pr E*. 26SI:1715–1729.

48. Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kück P, Ebersberger I, Walzl M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wägele JW, Misof B. 2010. A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol*. 27:2451–2464.

49. Munoz R, Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer KH, Glöckner FO, Rosselló-Móra R. 2011. Release LTPs104 of the all-species t. *Syst Appl Microbiol*. 34:169–170.

50. Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE, Howard EC, King E, Oakley CA, Reisch CR, et al. 2010. Genome characteristics of a generalist marine bacterial lineage. *ISME J*. 4:784–798.

51. Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*. 53:673–684.

52. Pagel M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proc R Soc B Biol Sci. 255:37–45.

53. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. 2010. How many bootstrap replicates are necessary? *J Comput Biol*. 17:337–354.

54. Petersen J, Brinkmann H, Berger M, Brinkhoff T, Päuker O, Pradella S. 2011. Origin and evolution of a novel DnaA-like plasmid replication type in rhodobacterales. *Mol Biol Evol*. 28:1229–1240.

55. Petersen J, Brinkmann H, Pradella S. 2009. Diversity and evolution of repABC type plasmids in Rhodobacterales. *Environ Microbiol*. 11:2627–2638.

56. Petersen J, Frank O, Göker M, Pradella S. 2013. Extrachromosomal, extraordinary and essential - The plasmids of the Roseobacter clade. *Appl Microbiol Biotechnol*. 97:2805–2815.

57. Petersen J. 2011. Phylogeny and compatibility: Plasmid classification in the genomics era. *Arch Microbiol*. 193:313–321.

58. Pradella S, Päuker O, Petersen J. 2010. Genome organisation of the marine *Roseobacter* clade member *Marinovum algicola*. *Arch Microbiol*. 192: 115–126.

59. Pujalte MJ, Lucena T, Ruvira MA, Arahal DR, Macián MC. 2014. The Family Rhodobacteraceae. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (eds). The Prokaryotes – Alphaproteobacteria and Betaproteobacteria. Fourth Ed.: Springer, Berlin; pp. 545–577.

60. Quester S, Schomburg D. 2011. EnzymeDetector: an integrated enzyme function prediction tool and database. *BMC Bioinformatics* 12:Article376.

61. R Core Team. R: 2014. A Language and Environment for Statistical Computing. Vienna, Austria;

62. Riedel T, Fiebig A, Petersen J, Gronow S, Kyrpides NC, Göker M, Klenk H-P. 2013. Genome sequence of the *Litoreibacter arenae* type strain (DSM 19593(T)), a member of the *Roseobacter* clade isolated from sea sand. *Stand Genomic Sci*. 9:117–127.

63. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko AR, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J. 2010. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 38:D5–16.

64. Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A, et al. 2013. BRENDA in 2013: Integrated reactions, kinetic data, enzyme function data, improved disease classification: New options and contents in BRENDA. *Nucleic Acids Res*. 41:D764–D772.

65. Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *BMC Bioinformatics* 17:1246–1247.

66. Siddall ME, Whiting MF. 1999. Long-branch abstractions. *Cladistics* 15:9–24.

67. Siddall ME. 2010. Unringing a bell: Metazoan phylogenomics and the partition bootstrap. *Cladistics* 26:444–452.

68. Stackebrandt E, Scheuner C, Göker M, Schumann P. 2014. The Family *Intrasporangiaceae*. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (eds). The Prokaryotes – Actinobacteria. Fourth Ed. Springer Heidelberg, pp. 397–424.

69. Stamatakis A, Aberer AJ. 2013. Novel parallelization schemes for large-scale likelihood-based phylogenetic inference. Proceedings - IEEE 27th International Parallel and Distributed Processing Symposium, pp. 1195–1204.

70. Stamatakis A, Auch AF, Meier-Kolthoff J, Göker M. 2007. AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa. *BMC Bioinformatics* 8:Article405.

71. Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol*. 57:758–771.

72. Tang K, Liu K, Jiao N. 2012. Draft genome sequence of *Oceaniovalibus guishaninsula* JLT2003T. *J Bacteriol*. 194:6683.

73. The UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*. 41:D43–D47.

74. Thompson JD, Thierry J-CC, Poch O. 2003. RASCAL: rapid scanning and correction of multiple sequence alignments. *BMC Bioinformatics* 19:1155–1161.

75. To TH, Jung M, Lycett S, Gascuel O. 2015. Fast dating using least-squares criteria and algorithms. Systematic Biology, advance access (doi:10.1093/sysbio/syv068).

76. Verbarg S, Göker M, Scheuner C, Schumann P, Stackebrandt E. 2014. The families *Erysipelotrichaceae* emend., *Coprobacillaceae* fam. nov., and *Turicibacteraceae* fam. nov. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F, (eds). The Prokaryotes – Firmicutes and Tenericutes. Fourth Ed. Springer Heidelberg; pp. 79–105.

77. Voget S, Wemheuer B, Brinkhoff T, Vollmers J, Dietrich S, Giebel H-A, Beardsley C, Sardemann C, Bakenhus I, Billerbeck S, et al. 2015. Adaptation of an abundant *Roseobacter* RCA organism to pelagic systems revealed by genomic and transcriptomic analyses. *ISME J*. 9:371–384.

78. Wiegmann K, Hensler M, Wöhlbrand L, Ulbrich M, Schomburg D, Rabus R. 2014. Carbohydrate catabolism in *Phaeobacter inhibens* DSM 17395, a member of the marine *Roseobacter* clade. *Appl Environ Microbiol*. 80:4725–4737.

79. Van Dongen SM. 2000. Graph clustering by flow simulation. PhD Thesis. University of Utrecht, The Netherlands.