Genome Streamlining in a Cosmopolitan Oceanic Bacterium

Stephen J. Giovannoni,[1]* H. James Tripp,[1] Scott Givan,[2] Mircea Podar,[3] Kevin L. Vergin,[1] Damon Baptista,[3] Lisa Bibbs,[3] Jonathan Eads,[3] Toby H. Richardson,[3] Michiel Noordewier,[3] Michael S. Rappé,[4] Jay M. Short,[3] James C. Carrington,[2] Eric J. Mathur[3]

The SAR11 clade consists of very small, heterotrophic marine α-proteobacteria that are found throughout the oceans, where they account for about 25% of all microbial cells. *Pelagibacter ubique*, the first cultured member of this clade, has the smallest genome and encodes the smallest number of predicted open reading frames known for a free-living microorganism. In contrast to parasitic bacteria and archaea with small genomes, *P. ubique* has complete biosynthetic pathways for all 20 amino acids and all but a few cofactors. *P. ubique* has no pseudogenes, introns, transposons, extrachromosomal elements, or inteins; few paralogs; and the shortest intergenic spacers yet observed for any cell.

Pelagibacter ubique, strain HTCC1062, belongs to one of the most successful clades of organisms on the planet (*1*), but it has the smallest genome (1,308,759 base pairs) of any cell known to replicate independently in nature (Fig. 1). In situ hybridization studies show that these organisms occur as unattached cells suspended in the water column (*1*). They grow by assimilating organic compounds from the ocean's dissolved organic carbon (DOC) reservoir, and can generate metabolic energy either by a light-driven proteorhodopsin proton pump (*2*) or by respiration (*3*). The marine planktonic environment is poor in nutrients, and the availability of N, P, and organic carbon typically limits the productivity of microbial communities. *P. ubique* is arguably the smallest free-living cell that has been studied in a laboratory, and even its small genome occupies a substantial fraction (~30%) of the cell volume. The small size of the SAR11 clade cells fits a model proposed by Button (*4*) for natural selection acting to optimize surface-to-volume ratios in oligotrophic cells, such that the capacity of the cytoplasm to process substrates will be matched to steady-state membrane transport rates.

Surprisingly, this genome appears to encode nearly all of the basic functions of α-proteobacterial cells (Table 1). The small genome size is attributable to the nearly complete absence of nonfunctional or redundant DNA and the paring down of all but the most fundamental metabolic and regulatory functions. For example, *P. ubique* falls at the extreme end of the range for intergenic DNA regions, with a median spacer size of only three bases (Fig. 2). Intergenic DNA regions vary considerably among bacteria and archaea, even including parasites that have small genomes (*5*). No pseudogenes, phage genes, or recent gene duplications were found in *P. ubique*.

To further explore this trend, we investigated paralogous gene families by means of BLAST clustering with variable threshold limits. The genome had the smallest number of paralogous genes observed in any free-living cell (Fig. 1) (fig. S1). A steep slope in

[1]Department of Microbiology, [2]Center for Gene Research and Biotechnology, Oregon State University, Corvallis, OR 97331, USA. [3]Diversa Corporation, 4955 Directors Place, San Diego, CA 92121, USA. [4]Hawaii Institute of Marine Biology, School of Ocean and Earth Science and Technology, University of Hawaii, Post Office Box 1346, Kaneohe, HI 96744, USA.

*To whom correspondence should be addressed. E-mail: steve.giovannoni@oregonstate.edu

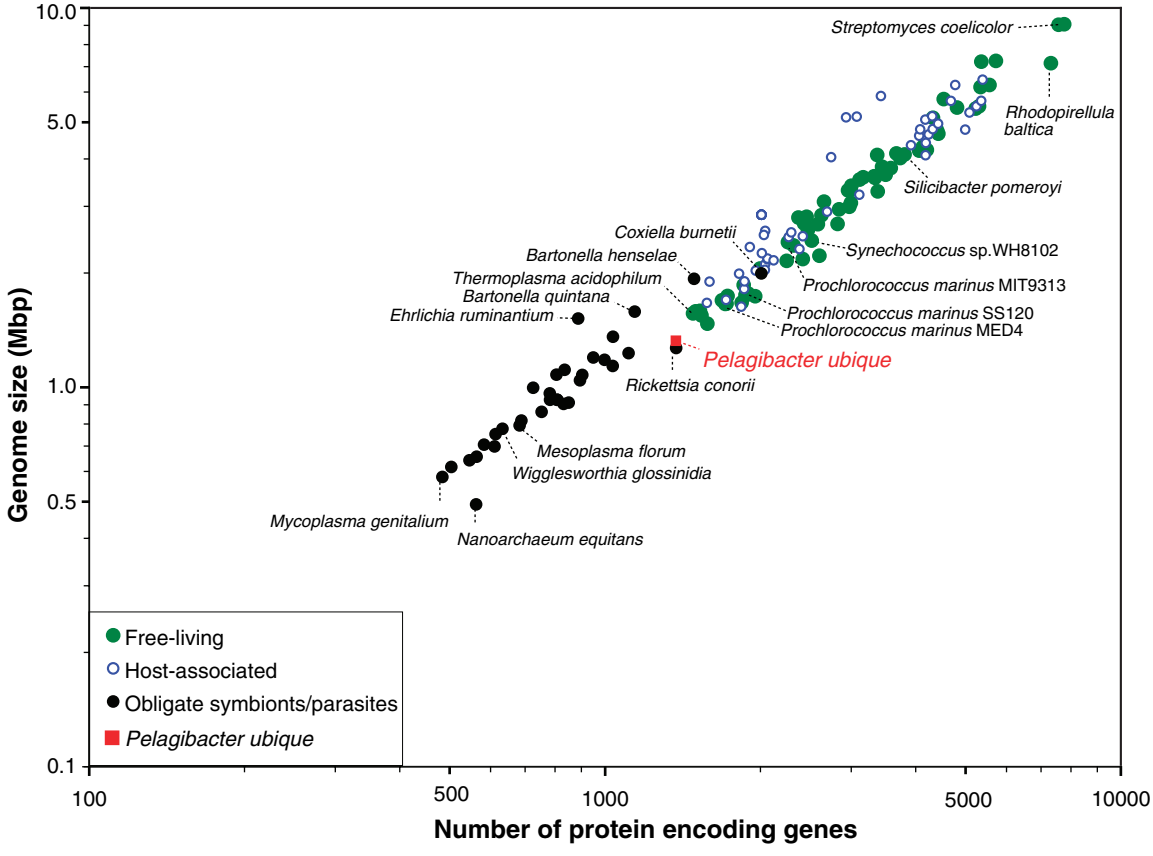


Fig. 1. Number of predicted protein-encoding genes versus genome size for 244 complete published genomes from bacteria and archaea. *P. ubique* has the smallest number of genes (1354 open reading frames) for any free-living organism.

the decline of potential paralogs with increasing gene pairwise similarity threshold, relative to other organisms, suggested that the few paralogs present in *P. ubique* are descended from relatively old duplication events, and that steady evolutionary pressure has constrained the expansion of gene families in this organism (fig. S2). Furthermore, there was no evidence of DNA originating from recent horizontal gene transfer events. The presence of DNA uptake and competence genes (*PilC*, *PilD*, *PilE*, *PilF*, *PilG*, *PilQ*, *comL*, and *cinA*) in the genome suggests that *P. ubique* has the ability to acquire foreign DNA. These data are consistent with the hypothesis that cells in some ecosystems are subject to powerful selection to minimize the material costs of cellular replication; this concept is known as streamlining (*5*).

Several hypotheses have been used to explain genome reduction in prokaryotes, particularly in parasites, which have the smallest cellular genomes known. The relaxation of positive selection for genes used in the biosynthesis of compounds that can be imported from the host, together with a bias favoring deletions over insertions in most or all bacteria, appear to account for genome reduction in many parasites and organelles (*5*). The streamlining hypothesis assumes that selection acts to reduce genome size because of the metabolic burden of replicating DNA with no adaptive value. Under this hypothesis, it is presumed that repetitive DNA arises when mechanisms that add DNA to genomes—for example, recombination and the propagation of self-replicating DNA (e.g., introns, inteins, and transposons)—overwhelm the simple economics of metabolic costs. However, evolutionary theory predicts that the probability that selection will act to eliminate DNA merely because of the metabolic cost of its synthesis will be greatest in very large populations of cells that do not experience drastic periodic declines (*6*).

The streamlining hypothesis has been used to explain genome reduction in *Prochlorococcus*, a photoautotroph that reaches population sizes in the oceans that are similar to those of *Pelagibacter* (*7*–*9*). *Prochlorococcus* genomes range from 1.66 to 2.41 million base pairs (Mbp). Many organisms with reduced genomes, including some pathogens, also have very low G:C to A:T ratios (*10*) (fig. S3), which can be attributed to biases in mutational frequencies, but alternatively might convey a selective advantage by lowering the nitrogen requirement for DNA synthesis, thereby reducing the cellular requirement for fixed forms of nitrogen (*7*). N and P are both proportionately important constituents of DNA that are frequently limiting in seawater. The *P. ubique* genome is 29.7% G+C. Of four complete *Prochlorococcus* genome sequences, the two that lack the DNA repair enzyme 6-0-methylguanine-DNA methyltransferase also have very low G:C to A:T ratios. In the absence of this enzyme, the extent of accepted G:C to A:T mutations increases; however, the *P. ubique* genome encodes this enzyme, which suggests that other factors are the cause of its low G:C to A:T ratio.

Annotation revealed a spare metabolic network encoding a variant of the Entner-Duodoroff pathway, a tricarboxylic acid (TCA) cycle, a glyoxylate bypass, and a typical electron transport chain (Table 1). Anapleurotic pathways for cellular constituents, other than five vitamins, appeared to be complete, but genes that would confer alternate metabolic lifestyles, motility, or other complexities of structure and function were nearly absent. Conspicuous exceptions were genes for carotenoid synthesis, retinal synthesis, and proteorhodopsin. *P. ubique* constitutively expresses a light-dependent retinylidine proton pump and is the first cultured bacterium to exhibit the gene that encodes it (*2*). The genome also contained genes for type II secretion (including adhesion) and type IV pilin biogenesis. Examination of gene distributions among metabolic categories (fig. S4) supported the conclusion that genome reduction in *P. ubique* has spared genes for core proteobacterial functions while reducing the proportion of the genome devoted to noncoding DNA. Relative to other α-proteobacterial genomes, the proportions of *P. ubique* genes encoding transport functions, biosynthesis of amino acids, and energy metabolism were high (table S3).

The sheer size of *Pelagibacter* populations indicates that they consume a large proportion of the labile DOC in the oceans. The global DOC pool is estimated to be 6.85×10^{17} g C (*11*), roughly equaling the mass of inorganic C in the atmosphere (*12*). Examination of the *P. ubique* genome revealed that about half of all transporters, and nearly all nutrient-uptake transporters, are members of the ATP-binding cassette (ABC) family (table S1). ABC transporters typically have high substrate affinities and therefore provide an advantage at the cost of ATP hydrolysis. Inferred transport functions included the uptake of a variety of nitrogenous compounds: ammonia, urea, basic amino acids, spermidine, and putrescine. Broad-specificity transporters for sugars, branched amino acids, dicarboxylic and tricarboxylic acids, and a number of common osmolytes (including glycine betaine, proline, mannitol, and 3-dimethylsulfoniopropionate) were found in the genome. Autoradiography with native populations of SAR11 has demonstrated high uptake activity for amino acids and 3-dimethylsulfoniopropionate (*13*). Hence, efficiency is achieved in a low-nutrient system by reliance on transporters with broad substrate ranges (*14*) and a number of specialized substrate targets, in particular, nitrogenous compounds and osmolytes.

Table 1. Metabolic pathways in *Pelagibacter*.

Pathway	Prediction
Glycolysis	Uncertain
TCA cycle	Present
Glyoxylate shunt	Present
Respiration	Present
Pentose phosphate cycle	Present
Fatty acid biosynthesis	Present
Cell wall biosynthesis	Present
Biosynthesis of all 20 amino acids	Present
Heme biosynthesis	Present
Ubiquinone	Present
Nicotinate and nicotinamide	Present
Folate	Present
Riboflavin	Present
Pantothenate	Absent
B_6	Absent
Thiamine	Absent
Biotin	Absent
B_{12}	Absent
Retinal	Present

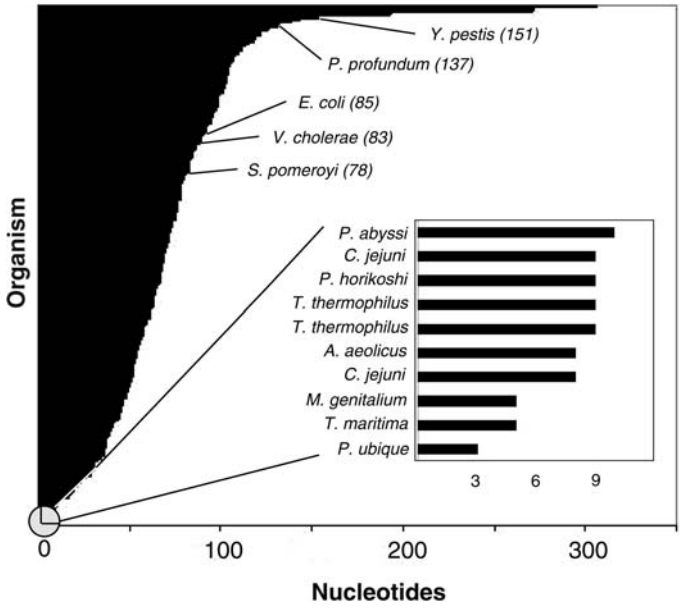


Fig. 2. Median size of intergenic spacers for bacterial and archaeal genomes. Inset shows expanded view of range for organisms with the smallest intergenic spacers.

The genome encoded two sigma factors, the heat shock factor σ^{32} and a σ^{70} (*rpo*D), but no homolog of *rpo*N, the gene for the nitrogen starvation factor σ^{54} (table S2). Only four two-component regulatory systems were identified, three of which match the only two-component regulatory systems in *Rickettsia* (*15*). The presence of homologs to *PhoR*/*PhoB*/*PhoC*, *NtrY*/*NtrX*, and *envZ*/*OmpR* suggested regulated responses to phosphate limitation, N limitation, and osmotic stress. The only additional two-component system, *RegB*/*RegA*, has been implicated in the regulation of cellular oxidation/reduction processes in phototrophic α-proteobacteria (*16*). A gene encoding a ferric iron uptake regulator was also present.

In its simplicity the *P. ubique* genome is unique among other heterotrophic marine bacteria, such as *Vibrio* sp. (*17*), *Pseudoalteromonas* (*18*), *Shewanella* (*19*), and *Silicibacter* (*20*), which have considerably larger genomes (4.0 to 5.3 Mbp) and global regulatory systems that enable them to implement a variety of metabolic strategies in response to environmental variation. We hypothesize that *P. ubique* makes use of the ambient DOC field (*21*), whereas heterotrophic bacterioplankton with larger genomes are poised to rapidly exploit pulses of nutrients (*22*) at the expense of replication efficiency during the intervening periods (*23*). This hypothesis is consistent with the observation that *P. ubique* has a single ribosomal RNA (rRNA) operon and a low growth rate (0.40 to 0.58 cell divisions per day) that does not vary in response to nutrient addition. In contrast, heterotrophic marine bacteria with large genomes have some of the highest recorded growth rates and are very responsive to nutrient concentration.

Like some other α-proteobacteria and especially archaea, HTCC1062 has an alternate thymidylate synthase for thymine synthesis,

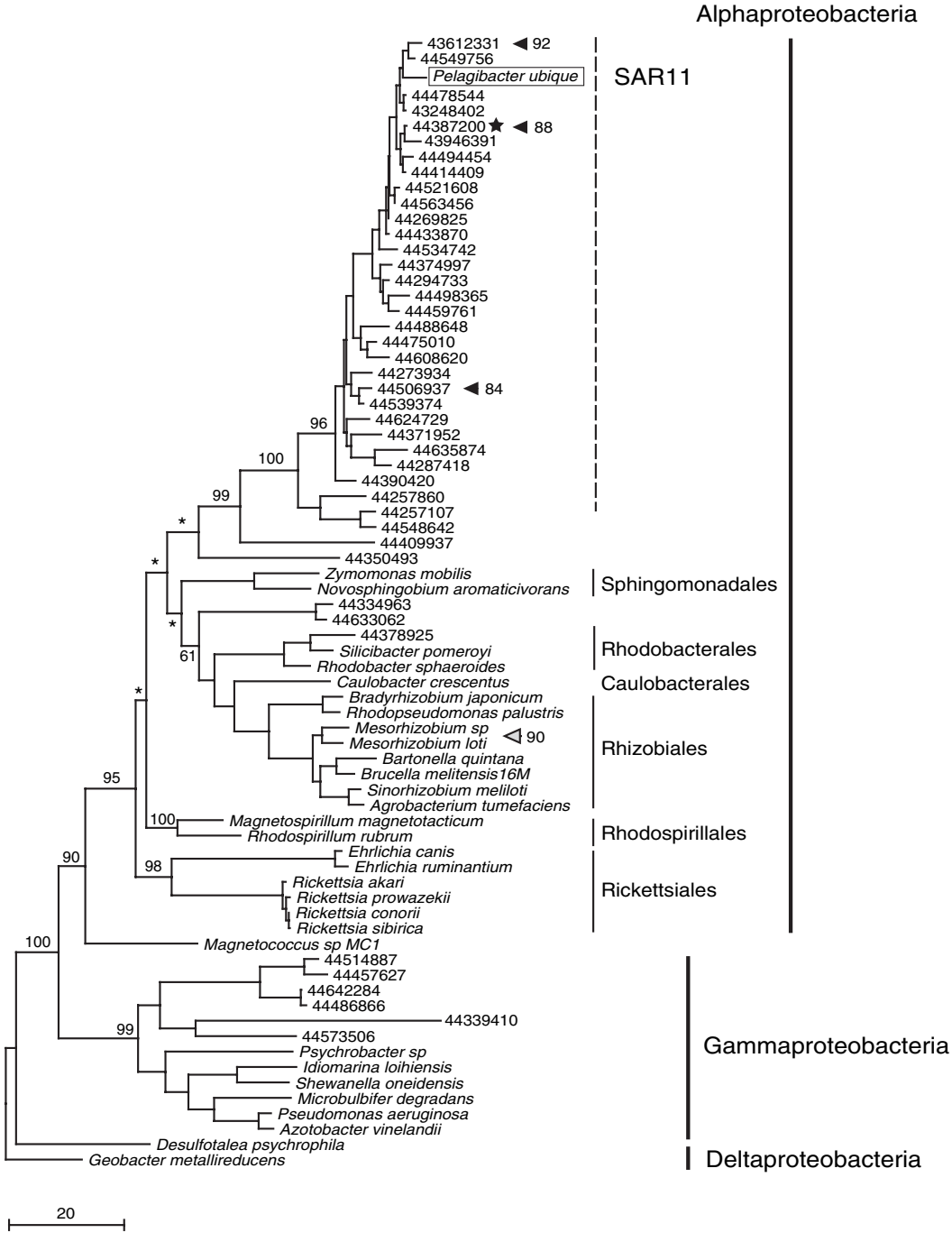


Fig. 3. Maximum likelihood phylogenetic tree for the gene encoding RNA polymerase subunit B. Sequences represented by accession numbers are environmental sequences from the Sargasso Sea (*19*). The sequence indicated by a star is part of the 5.7-kb contig IBEA_CTG_2159647 that is part of a conserved gene cluster also present in *Pelagibacter ubique*. Numbers indicated by solid arrowheads represent amino acid percentage identity to the *Pelagibacter* gene. For comparison, the identity between two species of *Mesorhizobium* is also indicated (open arrowhead). Bootstrap support (100 maximum-likelihood replicates) is indicated for the major clades (* if less than 50).

thyX (*24*). As in other strains that lack the most common thymidylate synthase (thyA) but have thyX, HTCC1062 also lacks the dihydrofolate reductase folA (*25*). Evidence suggests that the gene encoding thyX can substitute for folA (*24*). A full glycolytic pathway was not reconstructed because of the confounding diversity of glycolytic pathways (*26*). Five enzymes in the canonical glycolytic pathway were not seen, including two key enzymes involved in allosteric control: phosphofructokinase and pyruvate kinase. An enzyme thought to substitute for pyruvate kinase (*27*), known as PPDK (pyruvate-phosphate dikinase), was found. Some but not all of the enzymes for the nonphosphorylated Entner-Duodoroff pathway, considered more ancient than canonical glycolysis (*26*, *28*), were detected, as well as a complete pathway for gluconeogenesis, also considered more ancient than canonical glycolysis (*29*). Sugar transporters with best BLAST hits to maltose/trehalose transport were found, so presumably a complete glycolytic pathway does function in this cell.

Whole-genome shotgun (WGS) sequence data from the Sargasso Sea segregated at high similarity values, relative to other α-proteobacteria and proteobacteria, in a BLASTN analysis of the *P. ubique* genome (fig. S4). Sequence diversity prevented Venter *et al.* (*19*) from reconstructing SAR11 genomes from the Sargasso Sea WGS data set, although SAR11 rRNA genes accounted for 380 of 1412 16*S* rRNA genes and gene fragments they recovered (26.9%), and the library was estimated to encode the equivalent of about 775 SAR11 genomes. Three Sargasso Sea contiguous sequences (contigs) that were long (5.6 to 22.5 kb) and highly similar to the *P. ubique* genome were analyzed in detail. Genes on these contigs were syntenous with genes from the *P. ubique* genome, with amino acid sequence identities ranging from 68 to 96% (fig. S5). Phylogenetic analysis of four conserved genes from these contigs (those encoding RNA polymerase subunit B, Fig. 3; elongation factor G, fig. S6; DNA gyrase subunit B, fig. S7; and ribosomal protein S12, fig. S8) showed them to be associated with large, diverse environmental clades that branched within the α-proteobacteria. We hypothesize that evolutionary divergence within the SAR11 clade and the accumulation of neutral variation are the most likely explanations for the natural heterogeneity in SAR11 genome sequences.

Metabolic reconstruction failed to resolve why *P. ubique* will not grow on artificial media. When cultured in seawater, it attains cell densities similar to populations in nature, typically 10^5 to 10^6 ml^{-1} depending on the water sample (*3*). No evidence of quorum-sensing systems was found in the genome, and experimental additions of nutrients supported the results from metabolic reconstruction, which suggests that an unusual growth factor may play a role in the ecology of this organism.

P. ubique has taken a tack in evolution that is distinctly different from that of all other heterotrophic marine bacteria for which genome sequences are available. Evolution has divested it of all but the most fundamental cellular systems such that it replicates under limiting nutrient resources as efficiently as possible, with the outcome that it has become the dominant clade in the ocean.

References and Notes

1. R. M. Morris *et al.*, *Nature* **420**, 806 (2002).
2. S. J. Giovannoni *et al.*, *Nature*, in press.
3. M. S. Rappé, S. A. Connon, K. L. Vergin, S. J. Giovannoni, *Nature* **418**, 630 (2002).
4. D. K. Button, *Appl. Environ. Microbiol.* **57**, 2033 (1991).
5. A. Mira, H. Ochman, N. A. Moran, *Trends Genet.* **17**, 589 (2001).
6. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
7. A. Dufresne, L. Garczarek, F. Partensky, *Genome Biol.* **6**, R14 (2005).
8. B. Strehl, J. Holtzendorff, F. Partensky, W. R. Hess, *FEMS Microbiol. Lett.* **181**, 261 (1999).
9. G. Rocap *et al.*, *Nature* **424**, 1042 (2003).
10. D. W. Ussery, P. F. Hallin, *Microbiology* **150**, 749 (2004).
11. D. A. Hansell, C. A. Carlson, *Global Biogeochem. Cycles* **12**, 443 (1998).
12. D. A. Hansell, C. A. Carlson, *Deep Sea Res.* **48**, 1649 (2001).
13. R. R. Malmstrom, R. P. Kiene, M. T. Cottrell, D. L. Kirchman, *Appl. Environ. Microbiol.* **70**, 4129 (2004).
14. D. K. Button, B. Robertson, E. Gustafson, X. Zhao, *Appl. Environ. Microbiol.* **70**, 5511 (2004).
15. S. G. Andersson *et al.*, *Nature* **396**, 133 (1998).
16. S. Elsen, L. R. Swem, D. L. Swem, C. E. Bauer, *Microbiol. Mol. Biol. Rev.* **68**, 263 (2004).
17. E. G. Ruby *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 3004 (2005).
18. B. D. Lanoil, L. M. Ciufettii, S. J. Giovannoni, *Genome Res.* **6**, 1160 (1996).
19. J. C. Venter *et al.*, *Science* **304**, 66 (2004); published online 4 March 2004 (10.1126/science.1093857).
20. M. A. Moran *et al.*, *Nature* **432**, 910 (2004).
21. C. A. Carlson, H. W. Ducklow, A. F. Michaels, *Nature* **371**, 405 (1994).
22. F. Azam, *Science* **280**, 694 (1998).
23. J. A. Klappenbach, J. M. Dunbar, T. M. Schmidt, *Appl. Environ. Microbiol.* **66**, 1328 (2000).
24. H. Myllykallio *et al.*, *Science* **297**, 105 (2002); published online 23 May 2002 (10.1126/science.1072113).
25. H. Myllykallio, D. Leduc, J. Filee, U. Liebl, *Trends Microbiol.* **11**, 220 (2003).
26. T. Dandekar, S. Schuster, B. Snel, M. Huynen, P. Bork, *Biochem. J.* **343**, 115 (1999).
27. R. E. Reeves, R. A. Menzies, D. S. Hsu, *J. Biol. Chem.* **243**, 5486 (1968).
28. E. Melendez-Hevia, T. G. Waddell, R. Heinrich, F. Montero, *Eur. J. Biochem.* **244**, 527 (1997).
29. R. S. Ronimus, H. W. Morgan, *Archaea* **1**, 199 (2003).

30. Supported by NSF grant EF0307223, Diversa Corporation, the Gordon and Betty Moore Foundation, and the Oregon State University Center for Gene Research and Biotechnology. We thank S. Wells, M. Hudson, D. Barofsky, M. Staples, J. Garcia, B. Buchner, P. Sammon, K. Li, and J. Ritter for technical assistance and J. Heidelberg for advice about genome assembly. We also acknowledge the crew of the R/V Elakha for assistance with sample and seawater collections, the staff of the Central Services Laboratory at Oregon State University for supplementary sequence analyses, and the staff of the Mass Spectrometry Laboratory at Oregon State University for proteomic analyses. The sequence reported in this study has been deposited in GenBank under accession number CP000084.

Supporting Online Material

www.sciencemag.org/cgi/content/full/309/5738/1242/DC1
Materials and Methods
Tables S1 to S3
Figs. S1 to S9
References

26 April 2005; accepted 11 July 2005
10.1126/science.1114057

Contact-Dependent Inhibition of Growth in *Escherichia coli*

Stephanie K. Aoki, Rupinderjit Pamma, Aaron D. Hernday, Jessica E. Bickham, Bruce A. Braaten, David A. Low*

Bacteria have developed mechanisms to communicate and compete with each other for limited environmental resources. We found that certain *Escherichia coli*, including uropathogenic strains, contained a bacterial growth-inhibition system that uses direct cell-to-cell contact. Inhibition was conditional, dependent upon the growth state of the inhibitory cell and the pili expression state of the target cell. Both a large cell-surface protein designated Contact-*d*ependent *i*nhibitor A (CdiA) and two-partner secretion family member CdiB were required for growth inhibition. The CdiAB system may function to regulate the growth of specific cells within a differentiated bacterial population.

Bacteria communicate with each other in multiple ways, including the secretion of signaling molecules that enable a cell population to determine when it has reached a certain density or that a potential partner is present for conjugation (*1*, *2*). Cellular communication can also occur through contact between cells, as has been shown for *Myxococcus xanthus*, which undergoes a complex developmental pathway (*3*, *4*). Here we describe a different type of intercellular interaction in which bacterial growth is regulated by direct cell-to-cell contact.

Wild-type *Escherichia coli* isolate EC93 inhibited the growth of laboratory *E. coli* K-12

Molecular, Cellular, and Developmental Biology, University of California–Santa Barbara (UCSB), Santa Barbara, CA 93106, USA.

*To whom correspondence should be addressed. E-mail: low@lifesci.ucsb.edu

Supporting Online Material

Genome Streamlining in a Cosmopolitan Oceanic Bacterium

Stephen J. Giovannoni, H. James Tripp, Scott Givan, Mircea Podar, Kevin L. Vergin, Damon Baptista, Lisa Bibbs, Jonathan Eads, Toby H. Richardson, Michiel Noordewier, Michael S. Rappé, Jay Short, James C. Carrington and Eric J. Mathur

Supporting Notes

Amount of N & P saved by reduction of G+C of genomic DNA from 50% to 30%:
One atom of nitrogen is saved by converting a G-C base pair to an A-T base pair.
The nitrogen savings at 20% A-T content vs. 50% A-T content is 30% fewer nitrogen atoms per *P. ubique* genome. In a genome of $1.3 * 10^6$ bp, this amounts to 390,000 nitrogen atoms saved per cell. Assuming a cell density of $5 \times 10^8/\ell$, this is a nitrogen savings of 216 picomoles per liter. The result, 216 picomoles of N per liter, may seem low to non-oceanographers, but it may be significant in an environment where N compounds are often at nanomolar concentrations or undetectable (*1*).

Supporting Methods

Cultivation: Cells were cultivated as described by Rappé et al., on medium LNHM, with the addition of 1 µM retinal. For cells grown in the light, cool-white light of 24 µmole photons/m^2/sec was supplied in a 14 h light/10 h dark cycle.

Extraction of DNA. 80 L of cultured cells were collected by filtration through 0.2 µm Supor filters and stored at –80 C in sucrose lysis buffer (SLB, 20 mM EDTA, 400 mM NaCl, 0.75 M sucrose, and 50 mM Tris-HCl, pH 9.0) until extraction. Extraction and purification were as described (*2*). Briefly, proteinase K and SDS were added to final concentrations of 100 µg/ml and 1%, respectively, and filters were incubated at 37 and 55 C for 30 min each. Cell lysates were extracted with buffered phenol and chloroform and ethanol precipitated. Nucleic acids were resuspended in TE and further purified by ultracentrifugation through a cesium trifluoroacetate gradient. DNA fractions were precipitated with isopropanol, re-precipitated with ethanol, and combined. Purity was assessed by LH-PCR as described (*3*). Briefly, purified DNA was used as template in a PCR reaction using FAM-labeled 27F-B and 519R primers. PCR products were cleaned, separated on an ABI capillary 3730 Genetic Analyzer, and analyzed using Genescan software. DNA samples with only one peak corresponding to the HTCC1062 16S rRNA gene were used for library construction.

Library Construction and Sequencing. For genomic sequencing, 4 µg of genomic DNA was cut with 6-base recognition site restriction enzymes and cloned into a phage lambda vector (*4*). The library was amplified once then in vivo excised to form a phagemid library. Approximately 13,000 end reads were generated from plasmid prepped DNA using an ABI 3700 automated sequencer and the ABI Prism BigDye Sequencing kits. The data was delivered to Oregon State University for assembly.

Assembly, Gap Closure and Quality Control. 13,344 individual sequence reads were used to assemble the HTCC1062 genome. PHRED, PHRAP and CONSED were used to assign quality scores, trim vector sequences, assemble contigs, and manually resolve sequence ambiguities (*5-7*). Following 3 rounds of AUTOFINISH (*8*), approximately 18 gaps were identified, five of which were spanned by known plasmids and, thus, were closed by sequencing the remainder of that plasmid. The remaining 13 gaps were closed by PCR using methods described in (*9*). Briefly, primers were designed approximately 100 bases from the ends of each contig. Primers were pooled and HTCC1062 DNA was amplified by PCR. Positive products were sequenced. No gaps were larger than 4 kb, so all gaps were closed by conventional PCR. The closed genomic sequence was analyzed by Consed (*5*) for low quality sequence, single covered areas, and ambiguous sequence which were investigated by subsequent PCR analysis of genomic HTCC1062 DNA using primers designed to span the questionable areas.

Annotation. GenDB was used as the annotation database (6). Potential genes were identified using Glimmer, version 2.13 (*10*). Original gene assignments were refined based upon putative ribosome-binding sites (*11*) and similarity to other alphaproteobacterial proteins. Functional assignments were manually assigned based upon the results from the following analyses: BLASTP vs the SwissProt database (*12*), BLASTP vs proteins from other alphaproteobacteria, BLASTP vs the NCBI nr protein database, HMM search against the Pfam database (*13*), InterPro (*14*) searches, and TMHMM (*15*). To detect instances of horizontal gene transfer the program PyPhy was used to investigate phylogenetic relationships to nearest neighbors for all genes (*16*).

Liquid chromatography/tandem mass spectrometry (LC/MC/MC) of the HTCC1062 proteome resulted in the identification of 426 different proteins, resulting in the confirmation of 11 proteins which otherwise would have been placed in the conserved hypothetical category, and nine proteins which otherwise would have been listed as hypothetical proteins (*17*).

The replication origin was predicted by analyzing GC bias using the program Genskew (*18*) (Fig S9).

Comparative genome analysis. Genomic data (genome size, GC content, predicted genes) for 142 bacteria and archaea was obtained from the NCBI database. The species were classified into: 1) free living; 2) host associated associated, including commensal organisms and opportunistic pathogens; and 3) obligate parasites and symbionts that absolutely require a host. The genomes are listed by category below.

1. Free living: *Acinetobacter sp. ADP1, Aeropyrum pernix, Aquifex aeolicus, Archaeoglobus fulgidus, Azoarcus sp. EbN1, Bacillus licheniformis ATCC14580, Bacillus cereus ATCC14579, Bacillus halodurans, Bacillus subtilis, Bdellovibrio bacteriovorus, Caulobacter crescentus, Chlorobium tepidum, Chromobacterium violaceum, Clostridium acetobutylicum, Clostridium perfringens, Clostridium tetani E88, Corynebacterium glutamicum, Dehalococcoides ethenogenes, Deinococcus radiodurans, Desulfotalea psychrophila, Desulfovibrio vulgaris, Geobacter sulfurreducens, Gloeobacter violaceus, Gluconobacter oxydans, Idiomarina loihiensis, Lactococcus lactis, Legionella pneumophila, Listeria monocytogenes, Methanobacterium*

thermoautotrophicum, Methanococcu maripaludis, Methanococcus jannaschii, Methanopyrus kandleri, Methanosarcina acetivorans, Methanosarcina mazei, Methylococcus capsulatus, Nitrosomonas europaea, Nostoc sp, Oceanobacillus iheyensis, Picrophillus torridus, Rhodopirellula baltica, Prochlorococcus marinus CCMP1375, Prochlorococcus marinus MED4, Prochlorococcus marinus MIT9313, Pseudomonas aeruginosa, Pseudomonas putida KT2440, Pyrobaculum aerophilum, Pyrococcus abyssi, Pyrococcus horikoshii, Rhodopseudomonas palustris CGA009, Shewanella oneidensis, Silicibacter pomeroyi, Streptomyces avermitilis, Streptomyces coelicolor, Sulfolobus solfataricus, Sulfolobus tokodaii, Symbiobacterium thermophilum, Synechococcus sp. WH8102, Synechocystis sp. PCC6803, Thermoanaerobacter tengcongensis, Thermoplasma acidophilum, Thermoplasma volcanium, Thermosynechococcus elongatus, Thermotoga maritima, Zymomonas mobilis.

2. Host-associated and opportunistic pathogens: *Agrobacterium tumefaciens C58, Bacillus thuringiensis, Bacillus anthracis Ames, Bacteroides thetaiotaomicron, Bifidobacterium longum, Bordetella bronchiseptica, Bordetella parapertussis, Burkholderia pseudomallei, Campylobacte jejuni, Corynebacterium diphtheriae, Enterococcus faecalis, Escherichi coli K12, Escherichia coli O157H7, Francisella tularensis, Fusobacterium nucleatum, Haemophilus ducreyi , Helicobacter hepaticus, Helicobacter pylori, Lactobacillus johnsonii, Leifsonia xyli, Mannheimia succiniciproducens, Mycobacterium bovis, Mycobacterium tuberculosis CDC155, Neisseria meningitidis, Pasteurella multocida, Photorhabdus luminescens, Porphyromonas gingivalis, Propionibacterium acnes, Pseudomonas syringae syringae, Ralstonia solanacearum, Salmonella typhi Ty2, Salmonella typhimurium LT2, Shigella flexneri 2a 2457T, Staphylococcus aureus Mu50, Staphylococcus epidermidis ATCC 12228, Streptococcus agalactiae 2603, Streptococcus mutans, Streptococcus pneumoniae R6, Streptococcus pyogenes MGAS315, Vibrio cholerae, Vibrio parahaemolyticus, Vibrio vulnificus CMCP6, Wolinella succinogenes, Xanthomonas axonopodis citri, Xanthomonas campestris, Xylella fastidiosa Temecula1, Yersinia pestis KIM.*

3. Obligate parasites or symbionts*: Anaplasma marginale str. St. Maries, Bartonella henselae, Bartonella quintana, Blochmannia floridanus, Borrelia burgdorferi, Borrelia garinii, Buchnera aphidicola APS, Buchnera aphidicola Bp, Buchnera aphidicola Sg, Chlamydia muridarum, Chlamydia trachomatis, Chlamydophila caviae, Chlamydophila pneumoniae, Coxiella burnetii, Ehrlichia ruminantium, Mesoplasma florum, Mycoplasma mobile, Mycoplasma gallisepticum, Mycoplasma genitalium, Mycoplasma penetrans, Mycoplasma pneumoniae, Mycoplasma pulmonis, Nanoarchaeum equitans, Onion yellows phytoplasma, Rickettsia conorii, Rickettsia prowazekii, Treponema pallidum, Tropheryma whipplei TW08/27, Tropheryma whipplei Twist, Ureaplasma urealyticum, Wigglesworthia glossinidia,*
Wolbachia endosymbiont TRS of Brugia malayi.

To identify the families of related protein encoding genes in all the individual genomes, we used BLASTCLUST, which performs a BLAST pairwise comparison followed by single-linkage clustering of the statistically significant matches. Various computational definitions and approaches have been used to identify paralogous genes in an organism, a high accuracy requiring in depth phylogenetic analysis. The BLAST parameters used in pairwise protein comparisons were the BLOSUM62 matrix, gap opening 11, gap extension 1 and e-value threshold 1e-6. The minimal length of the sequence coverage in the pairwise comparison was set

to 50% or 90% and the minimal sequence similarity threshold was varied stepwise from 30% to 90%. By increasing the stringency of the analysis we wanted to eliminate any potential outliers due to domain fusions or other recombination events and to see the effects on the overall numbers of gene families and gene members across all genomes. Also, the degree of sequence similarity is related to the timing of the duplication event that resulted in the paralogous genes and the rate of sequence divergence ("molecular clock") for those particular genes.

Using the most permissive threshold settings (30% sequence similarity and 50% length coverage), the three largest clusters of paralogs for *Pelagibacter* were the ATP-binding subunit of ABC transporters (22 genes), short chain dehydrogenases (13 genes) and aldehyde dehydrogenases (7 genes). Under the most stringent setting that identified paralogs in that genome (70% sequence similarity at 50-90% length coverage), only two clusters with two paralogs each were found: the fimbrial protein pilin (C134_0936 and C134_1216, 75% sequence identity) and the cold shock DNA binding domain (C134_0477 and C134_1274, 73% sequence identity).

Boussau (*20*) published a figure showing the number of genes per functional category for the following alphaproteobacteria: *Mesorhizobium loti, Bradyrhizobium japonicum, Sinorhizobium melloti, Caulobacter crescentus, Agrobacterium tumefaciens, Brucella suis, Brucella melitensis, Bartonella henselae, Bartonella quintana, Rickettsia conorii, Rickettsia prowazekii*. The number of genes per functional category for alphaproteobacteria was determined from the regression charts published by Boussau. These were entered into a spreadsheet program with the corresponding *Pelagibacter ubique* data and the linear regression was recalculated. The points at which the recalculated regression lines crossed the *Pelagibacter ubique* genome size are given as the predicted number of genes in supplemental table S3, which are compared to the actual number and reported as percent above trend for the selected categories. When the number of genes in functional categories for *P. ubique* is compared to the Boussau linear regressions of alpha proteobacterial genes by category against genome size, a striking overabundance of *P. ubique* genes is seen for energy metabolism, amino acid biosynthesis, and transport. The number of *P. ubique* genes in these categories is roughly 170% of predicted when *P. ubique* is added to the regression analysis. This is consistent with the hypothesis that *P. ubique* is more likely free-living than parasitic because self-sufficiency requires more of the overabundant genes.

Comparison with environmental DNA sequences (Sargasso Sea data). The Sargasso Sea environmental DNA database was examined using BLASTN with *Pelagibacter* genome sequence serving as a query. Similar searches were performed against a collection of all published alphaproteobacterial genomes, and a collection of all other proteobacterial genomes. The results were filtered to exclude hits that had sequence length coverage of less than 500 nt and pairwise identity values below 70%. The results are plotted in Fig. S4A. Three large contigs were selected for more detailed analysis. A comparison of gene order and percent identity to the homologous *Pelagibacter* genes are shown in Fig S4B. For phylogenetic tree construction, we selected several genes from the contigs that are relatively conserved and have been used as taxonomic markers, including RNA polymerase B subunit, translation elongation factor G, ribosomal protein S12 and DNA gyrase subunit B. We searched for additional close relatives of those *Pelagibacter* genes in the Sargasso Sea dataset using BLASTP and we also obtained their homologs from the available alphaproteobacterial genomes and representatives of other

proteobacteria. Sequences were aligned using ClustalW (*19*) or HMMalign (http://hmmer.wustl.edu/). Alignments were curated by hand in Bioedit. Some environmental sequences that were very short were excluded and the portions of the protein sequences that could not be confidently aligned were masked out. Maximum likelihood phylogenetic trees were calculated with proml (PHYLIP 3.6)(Felsenstein, J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle)* using the JTT matrix, equal rates and 5 rounds of random sequence addition followed by global rearrangements.

Supporting Tables

Table S1. *P. ubique* transport proteins

Type	Family	Component	Gene	Function
a-channels				
Porins	OmpA		C134_0598	outer membrane channel
	TonB	TolQ/TolR/TolB	C134_0594-597	active transport from OM
Electrochemica	TRK	TrkH, TrkA	C134_0211, 0949, 0950	K+ uptake
	NCS2		C134_0183	nucleobase:cation symport
	Amt		C134_0049-0050	ammonium transport
	AmtB		C134_0818	ammonium transport[1]
	Amt		C134_1310	ammonium transport
	TTT		C134_1201-1203	tricarboxylic acid/Na symport
	SulP		C134_1190	sulfate import
	NhaA		C134_1005	Na+/H+ antiporter
	MgtE		C134_1089	magnesium
	SSS		C134_0048	unknown molecule/Na+ sympo
	SSS		C134_0811	urea/Na+ symport[1]
	SSS		C134_0316	glyoxylate or acetate/Na+ sym
	MatE		C134_1073, C134_1228	Ion antiport/efflux, induced by DNA damage
	DMT		C134_0187	unknown
	DMT		C134_0260	unknown
	DMT		C134_0696	unknown
	DMT		C134_1022	unknown
	DMT		C134_1367	unknown
	MFS		C134_0213	muropeptide transport
	MFS		C134_0274	unknown
	MFS		C134_0695	unknown
	MFS		C134_0830	unknown
	LysE		C134_1325, C134_0800	unknown
ATP-dependen	ABC		C134_1175-1179	phosphate
	ABC		C134_0353	arsenate
	ABC		C134_1236-1238	Fe3+
	ABC		C134_0267-C134_0271	molybdenum/tungsten
	ABC-TRAP		C134_0864-0868	dicarboxylate
	ABC		C134_0797-799	glycine betaine, proline
	ABC-TRAP		C134_1297-1299; 1301-	glycine betaine, proline
	ABC		C134_1290-1293	mannitol/chloroaromatic
	ABC-TRAP	PBP/MSP/ATPa	C134_0264-C134_0271	sugar
	ABC	PBP/MSP/ATPa	C134_0769-0772	sugar
	ABC	PBP/MSP/ATPa	C134_0805-0807	taurine
	ABC	PBP/MSP/ATPa	C134_1334-1337	spermidine/putrescine

	ABC	PBP/MSP/ATPa	C134_1346-1362	branched chain amino acids
	ABC	PBP/MSP/ATPa	C134_0655--0659	branched chain amino acids
	ABC	PBP/MSP/ATPa	C134_0953-0957	general L amino acid transport
	ABC	2MSP/PBP	C134_1208-1210	His/Glu/Gln/Arg transport
	ABC	ATPase	C134_0495	sulfate/thiosulfate
	ABC-MsbA	ATPase/MSP	C134_0147	transport to OM
	ABC-LPT	PBP/MSP/ATPa	C134_0848-0850	lysophospholipase L1 biosynth
	ABC-MsbA	MSP/ ATPase	Ç134_0147	lipid export to OM
	ABC	ATPase	C134_0201	unknown
	ABC	MSP	C134_0501	unknown
	ABC	MSP	C134_0749	unknown
	ABC	MSP/ATPase	C134_0812-0813	unknown[1]
	ABC	MSP/ATPase	C134_0903, 0905	unknown
	ABC	MSP	C134_1069	unknown
Other			C134_0623	small multidrug resistance prot
			C134_0786	small multidrug resistance prot
	AzlC		C134_1235	homolog, branched chain amin acid import

1. Possible operon including urea and ammonium transport, a protease, an unidentified ABC transporter, and pyruvate amination to alanine.

Table S2. *P. ubique* regulatory proteins

Gene	Protein	Class	Function
C134_0606	σ^{32}	sigma factor	heat shock transcription factor
C134_0037	σ^{70}	sigma factor	vegetative transcription factor
C134_0089/C134_008	envZ/OmpR*	sensor/regulator	osmolarity
C134_0198/C134_019	Unidentified*	sensor	unknown
C134_0946/C134_094	ntrY/ntrX	sensor/regulator	N regulation
C134_0447	RegB	sensor	redox
C134_0203	RegA	regulator	redox
C134_0363	Unidentified	regulator	unknown
C134_1180/C134_117	PhoR/PhoB	sensor/regulator	Activates high-affinity phosphate uptak
C134_0382	Fur	Negative regulator	Iron uptake
C134_0516	Unidentified	?	regulation of amino-acid metabolism
C134_0741	sufD	?	regulation of nitrogen and sulphur utili
C134_0738	Unidentified	?	regulation of nitrogen and sulphur utili
C134_0297	PhoE	?	regulation of phosphate utilization
C134_1135	Unidentified	?	regulation of C-compound and carbohydrate utilization
C134_0824	recX	?	recombination and DNA repair
C134_0423	Unidentified	?	transcriptional control
C134_0138	MarR family	?	transcriptional control
C134_0064	Unidentified	?	transcriptional control
C134_0087	petP	?	transcriptional control
C134_0047	Unidentified	?	transcriptional control
C134_0273	Unidentified	?	transcriptional control
C134_0860	Unidentified	?	transcriptional control
C134_0974	Unidentified	?	transcriptional control
C134_0964	Unidentified	?	transcriptional control
C134_1242	Unidentified	?	transcriptional control
C134_1034	MerR family	?	transcriptional control
C134_0958	Unidentified	?	transcriptional control
C134_0768	NAGC-like	?	transcriptional control
C134_1243	Unidentified	?	transcriptional control
C134_1175	PhoU	?	transcriptional control
C134_1248	DNA-binding	?	transcriptional control
C134_0881	clpX	?	protein targeting, sorting and translocat

*sensor and regulatory element adjacent on same strand.
** sensor and regulatory element on opposite strands.

Table S3. Number of Genes in Selected Functional Categories

	Pelagibacter	Final Regression	% Above Trend
Energy Metabolism	207	124	166.94
Transport and Binding Proteins	162	98	165.31
Amino Acid Biosynthesis	92	53	173.58

Supporting Figure Legends

Fig. S1. Number of paralogous gene families vs. predicted proteome sizes for bacteria. Gene clustering, a measurement of paralogous gene families, was determined using the program BLASTCLUST, with the threshold set at 30% sequence similarity over 50% of the sequence length and e value =$1e^{-6}$. *P. ubique* has both the smallest number of predicted proteins and the smallest number of gene families found in a free living bacterium.

Fig. S2. The number of paralogous gene families in microbial genomes plotted as a function of the BLAST sequence similarity threshold. Paralogous gene families were defined with the program BLASTCLUST as evalue=1e-6, for 50 % the gene length, with varying similarity thresholds. The steeper slope of *P. ubique* suggests a decline in fixation of more recent duplication events in comparison to the other marine bacteria. *Nanoarchaeum* and *Rickettsia conorii* are parasites with highly reduced genomes. This data is consistent with several models. Since gene duplication and divergence is a major avenue by which new functions evolve, reduced pressure for evolutionary change could explain this evidence. It can also be explained as a response to the pressure of streamlining evolution.

Fig. S3. Proteome size vs. GC content for published microbial genomes.

Fig. S4. Linear regressions of the percentage of genes by functional category vs. genome size for selected alphaproteobacterial genomes. The data and format are the same as presented by Boussau (*20*), with the addition of the *P. ubique* data for comparison. A, the six largest functional categories; B, the remaining eight functional categories.

Fig. S5. A. BLASTN of the *P. ubique* genome against Venter Sargasso Sea database, other alphaproteobacteria, and other bacteria, filtered to show hits longer than 500 bp and with 70% identity and higher. B. Gene content of three long contigs identified by blast to have high similarity to *P. ubique* (indicated on the plot by stars). The black arrows indicate genes that have the same position in *P. ubique*, forming conserved gene clusters. The red arrows indicate genes that are not present in *P. ubique* or are located in other regions of the chromosome. The blue numbers indicate percentage identity between the *P. ubique* and the Sargasso Sea genes.

Fig. S6. Maximum likelihood tree of translation elongation factor G from *P. ubique*, related sequences from the Venter Sargasso Sea database, and selected proteobacteria. The red numbers indicate percent identity to the *P. ubique* gene over the available sequence span. For comparison, the identity between two *Mesorhizobium* species is indicated also. The Sargasso sequence marked with a star is part of contig IBEA_CTG_2159647.

Fig. S7. Maximum likelihood tree of DNA gyrase, subunit B from *P. ubique*, related sequences from the Venter Sargasso Sea database, and selected alphaproteobacteria. The red numbers indicate percent identity to the *P. ubique* gene over the available sequence span. The Sargasso sequence marked with a star is part of contig IBEA_CTG_2157419.

Fig. S8. Maximum likelihood tree of ribosomal protein S12 from *P. ubique*, related sequences from the Venter Sargasso Sea database, and selected alphaproteobacteria. The red numbers

indicate percent identity to the *P. ubique* gene over the available sequence span. The Sargasso sequence marked with a star is part of contig IBEA_CTG_2159647.

Fig. S9. Replication origin prediction for *P. ubique* (*18*).

Supporting Figures

Fig. S1

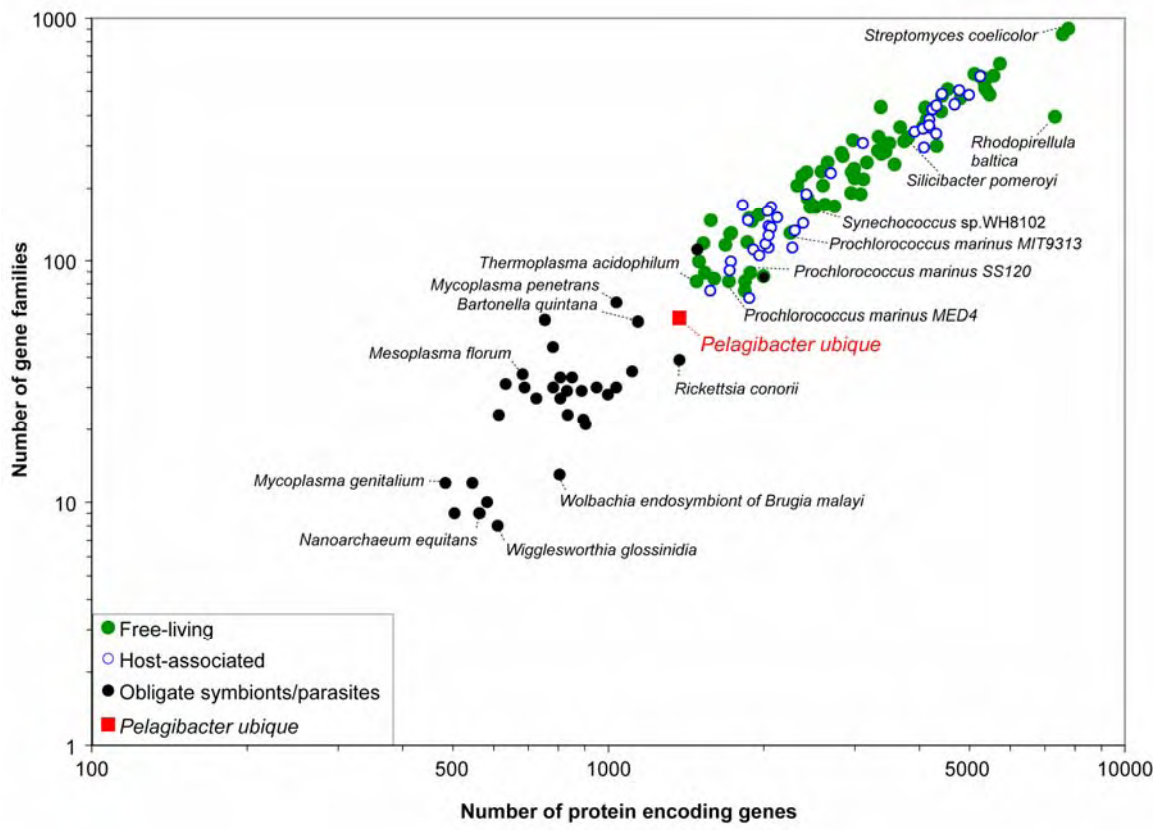


Fig. S2

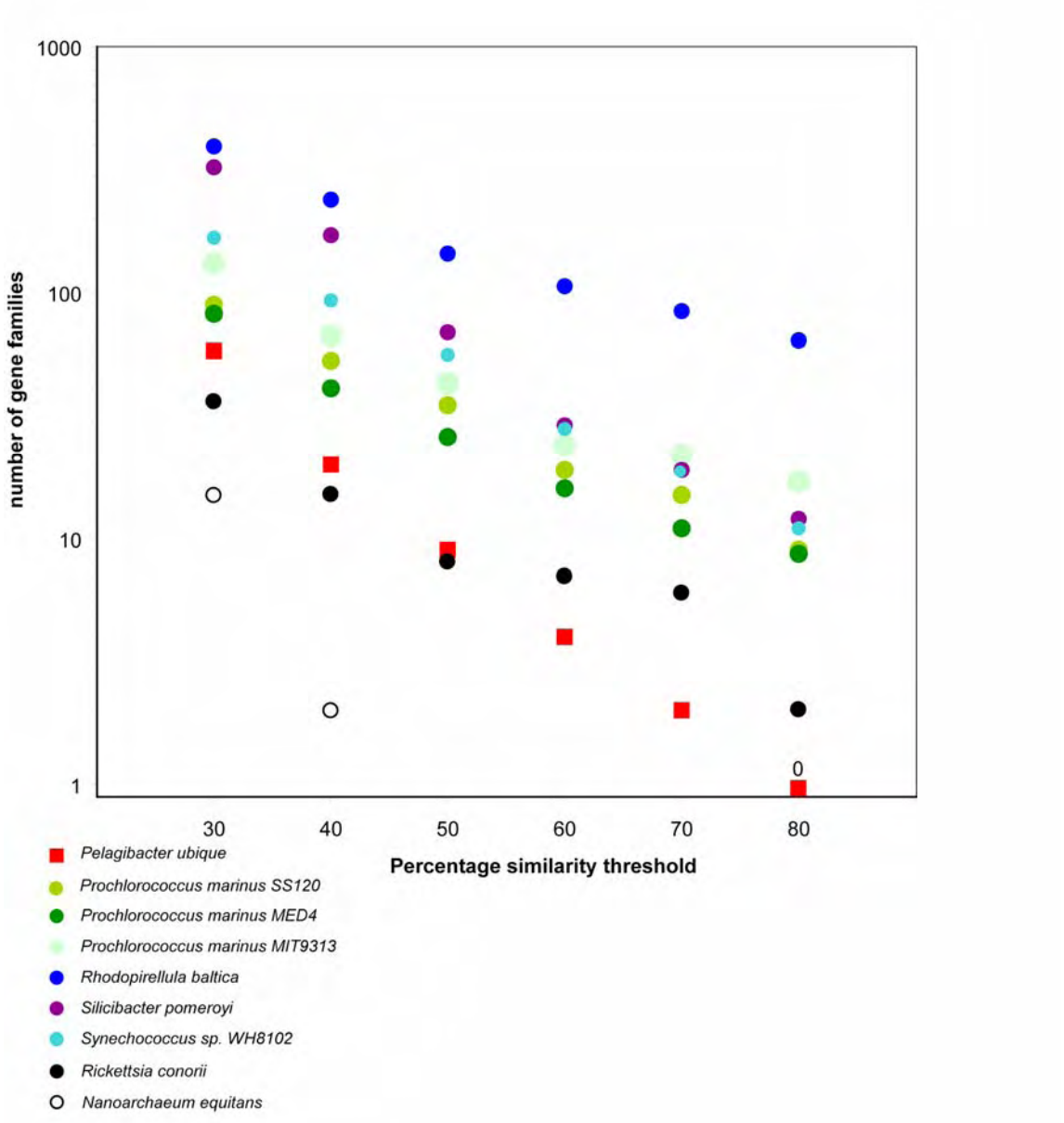

1000
100
10
1
number of gene families
0
30
40
50
60
70
80
Percentage similarity threshold
Pelagibacter ubique
Prochlorococcus marinus SS120
Prochlorococcus marinus MED4
Prochlorococcus marinus MIT9313
Rhodopirellula baltica
Silicibacter pomeroyi
Synechococcus sp. WH8102
Rickettsia conorii
Nanoarchaeum equitans

Fig. S3

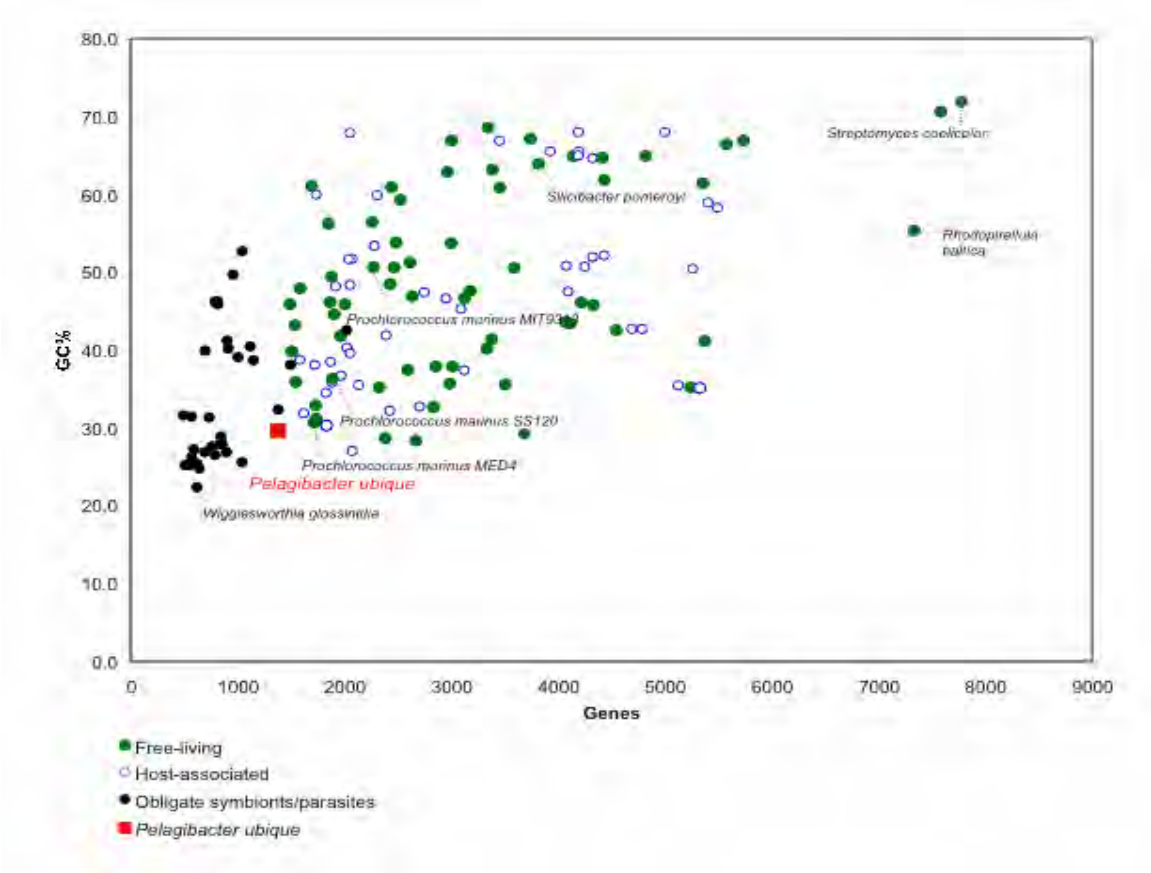

80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
GC%
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Genes
Streptomyces coelicolor
Silicibacter pomeroyi
Rhodopirellula baltica
Prochlorococcus marinus MIT9313
Prochlorococcus marinus SS120
Prochlorococcus marinus MED4
Pelagibacter ubique
Wigglesworthia glossinidia
Free-living
Host-associated
Obligate symbionts/parasites
Pelagibacter ubique

Fig. S4.

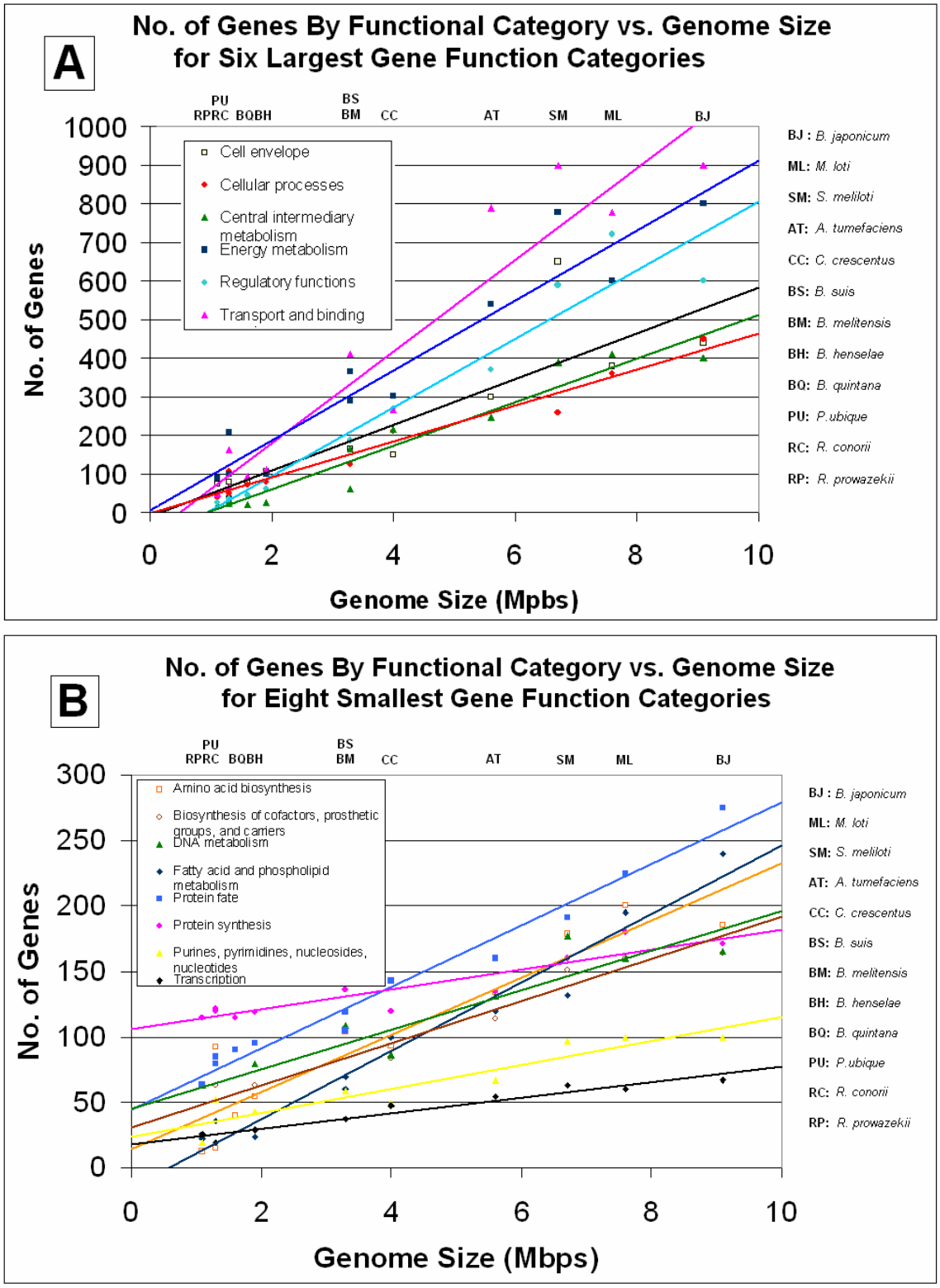

A
No. of Genes By Functional Category vs. Genome Size
for Six Largest Gene Function Categories
PU
RPRC BQBH
BS
BM
CC
AT
SM
ML
BJ
No. of Genes
Genome Size (Mpbs)
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Regulatory functions
Transport and binding
BJ : B. japonicum
ML: M. loti
SM: S. meliloti
AT: A. tumefaciens
CC: C. crescentus
BS: B. suis
BM: B. melitensis
BH: B. henselae
BQ: B. quintana
PU: P.ubique
RC: R. conorii
RP: R. prowazekii
B
No. of Genes By Functional Category vs. Genome Size
for Eight Smallest Gene Function Categories
No. of Genes
Genome Size (Mbps)
Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, and carriers
DNA metabolism
Fatty acid and phospholipid metabolism
Protein fate
Protein synthesis
Purines, pyrimidines, nucleosides, nucleotides
Transcription

Fig. S5

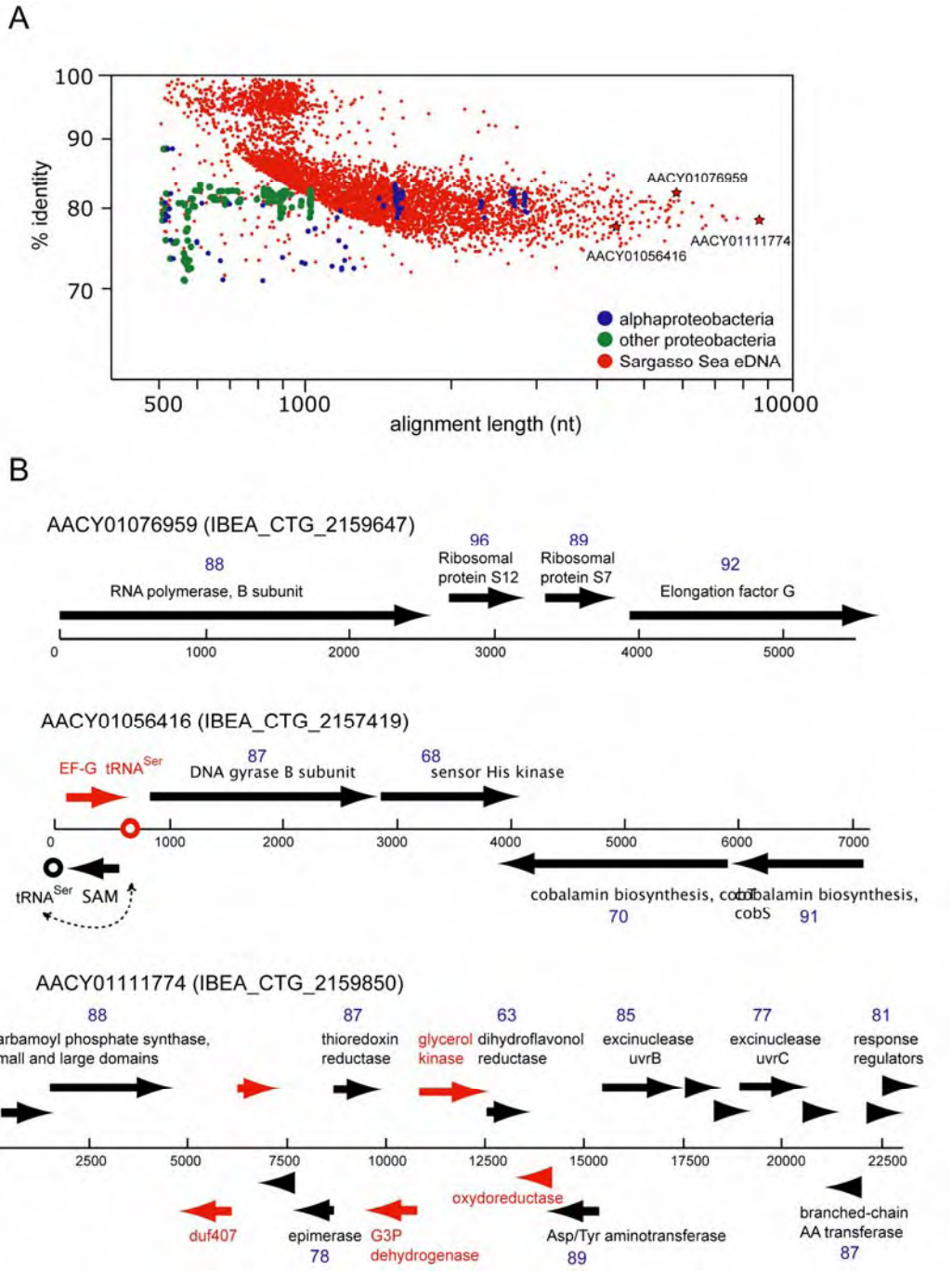

A
100
90
80
70
% identity
AACY01076959
AACY01111774
AACY01056416
alphaproteobacteria
other proteobacteria
Sargasso Sea eDNA
500
1000
10000
alignment length (nt)
B
AACY01076959 (IBEA_CTG_2159647)
88
RNA polymerase, B subunit
96
Ribosomal protein S12
89
Ribosomal protein S7
92
Elongation factor G
0
1000
2000
3000
4000
5000
AACY01056416 (IBEA_CTG_2157419)
EF-G
tRNASer
87
DNA gyrase B subunit
68
sensor His kinase
0
1000
2000
3000
4000
5000
6000
7000
tRNASer
SAM
cobalamin biosynthesis, cobS
70
cobalamin biosynthesis,
cobS
91
AACY01111774 (IBEA_CTG_2159850)
88
carbamoyl phosphate synthase, small and large domains
87
thioredoxin reductase
63
glycerol kinase
dihydroflavonol reductase
85
excinuclease uvrB
77
excinuclease uvrC
81
response regulators
0
2500
5000
7500
10000
12500
15000
17500
20000
22500
oxydoreductase
duf407
epimerase
78
G3P dehydrogenase
Asp/Tyr aminotransferase
89
branched-chain AA transferase
87

Fig. S6

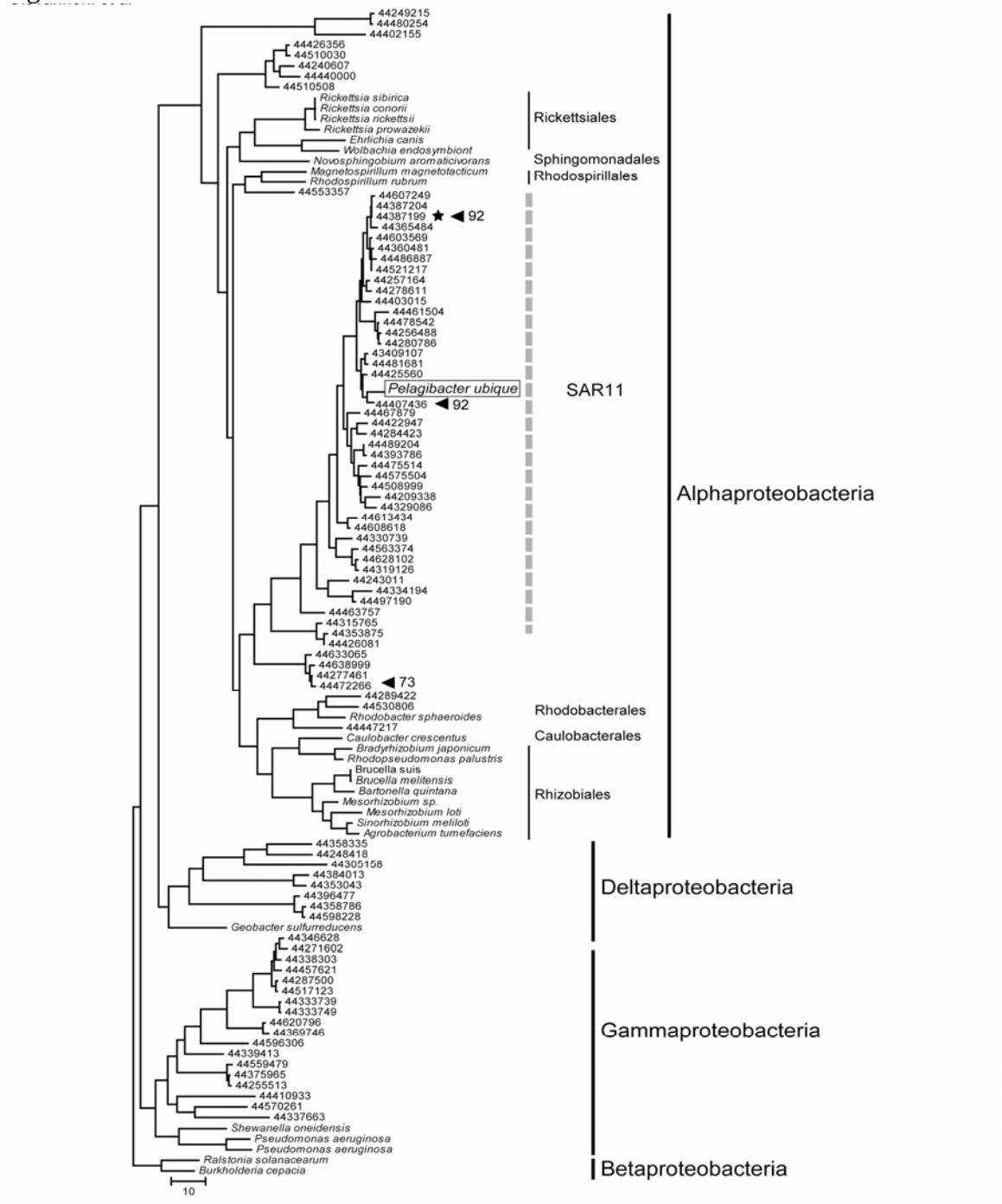

44249215
44480254
44402155
44426356
44510030
44240607
44440000
44510508
Rickettsia sibirica
Rickettsia conorii
Rickettsia rickettsii
Rickettsia prowazekii
Ehrlichia canis
Wolbachia endosymbiont
Novosphingobium aromaticivorans
Magnetospirillum magnetotacticum
Rhodospirillum rubrum
44553357
Rickettsiales
Sphingomonadales
Rhodospirillales
44607249
44387204
44387199
92
44365484
44603569
44360481
44486887
44521217
44257164
44278611
44403015
44461504
44478542
44256488
44280786
43409107
44481681
44425560
Pelagibacter ubique
SAR11
44407436
92
44467879
44422947
44284423
44489204
44393786
44475514
44575504
44508999
44209338
44329086
44613434
44608618
44330739
44563374
44628102
44319126
44243011
44334194
44497190
44463757
44315765
44353875
44426081
44633065
44638999
44277461
44472266
73
Alphaproteobacteria
44289422
44530806
Rhodobacter sphaeroides
44447217
Caulobacter crescentus
Bradyrhizobium japonicum
Rhodopseudomonas palustris
Brucella suis
Brucella melitensis
Bartonella quintana
Mesorhizobium sp.
Mesorhizobium loti
Sinorhizobium meliloti
Agrobacterium tumefaciens
Rhodobacterales
Caulobacterales
Rhizobiales
44358335
44248418
44305158
44384013
44353043
44396477
44358786
44598228
Geobacter sulfurreducens
Deltaproteobacteria
44346628
44271602
44338303
44457621
44287500
44517123
44333739
44333749
44620796
44369746
44596306
44339413
44559479
44375965
44255513
44410933
44570261
44337663
Shewanella oneidensis
Pseudomonas aeruginosa
Pseudomonas aeruginosa
Ralstonia solanacearum
Burkholderia cepacia
Gammaproteobacteria
Betaproteobacteria
10

Fig. S7

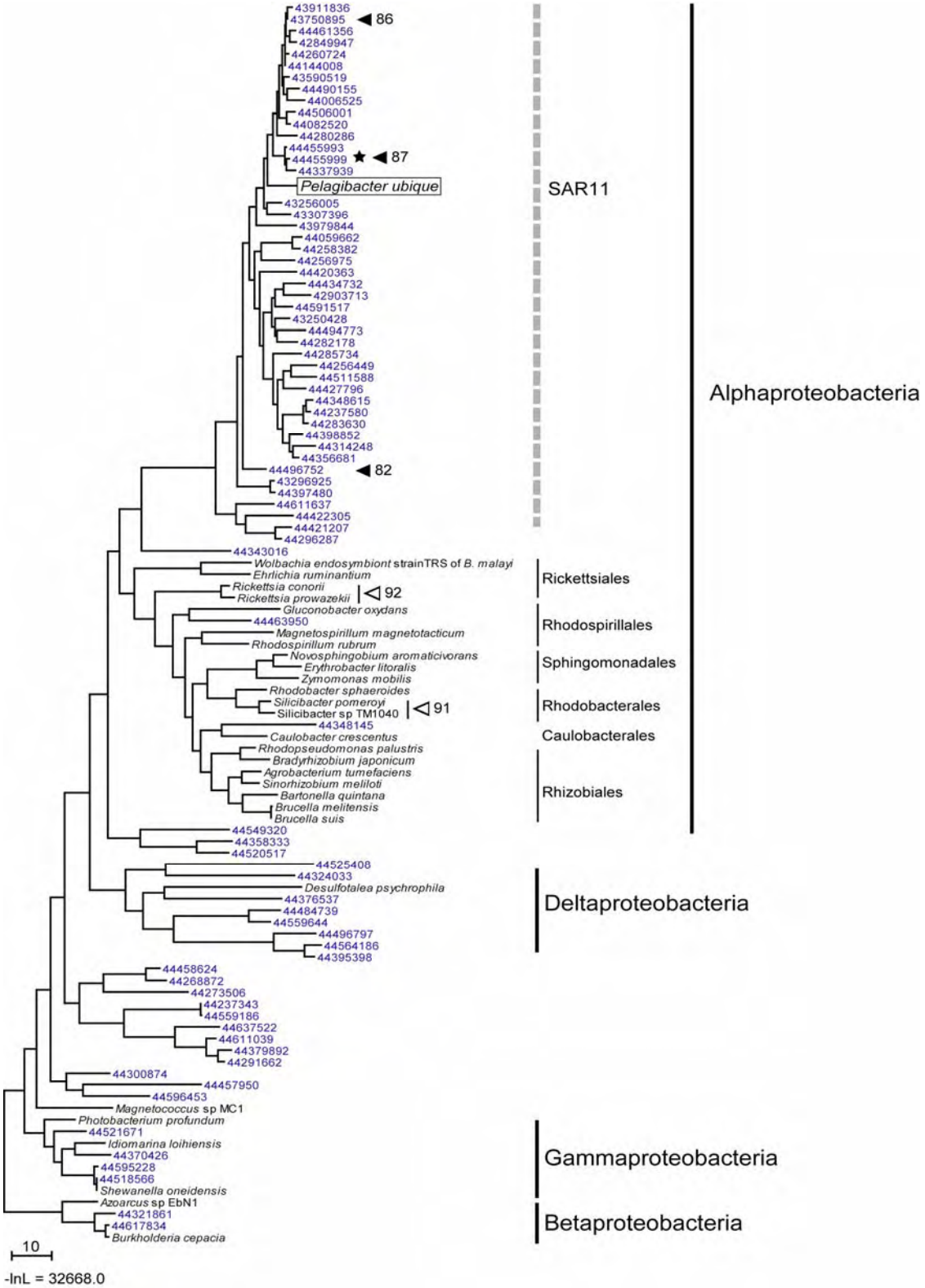

43911836
43750895
86
44461356
42849947
44260724
44144008
43590519
44490155
44006525
44506001
44082520
44280286
44455993
44455999
87
44337939
Pelagibacter ubique
SAR11
43256005
43307396
43979844
44059662
44258382
44256975
44420363
44434732
42903713
44591517
43250428
44494773
44282178
44285734
44256449
44511588
44427796
44348615
44237580
44283630
44398852
44314248
44356681
44496752
82
43296925
44397480
44611637
44422305
44421207
44296287
Alphaproteobacteria
44343016
Wolbachia endosymbiont strainTRS of B. malayi
Ehrlichia ruminantium
Rickettsia conorii
Rickettsia prowazekii
92
Rickettsiales
Gluconobacter oxydans
44463950
Magnetospirillum magnetotacticum
Rhodospirillum rubrum
Rhodospirillales
Novosphingobium aromaticivorans
Erythrobacter litoralis
Zymomonas mobilis
Sphingomonadales
Rhodobacter sphaeroides
Silicibacter pomeroyi
Silicibacter sp TM1040
91
Rhodobacterales
44348145
Caulobacter crescentus
Caulobacterales
Rhodopseudomonas palustris
Bradyrhizobium japonicum
Agrobacterium tumefaciens
Sinorhizobium meliloti
Bartonella quintana
Brucella melitensis
Brucella suis
Rhizobiales
44549320
44358333
44520517
44525408
44324033
Desulfotalea psychrophila
44376537
44484739
44559644
44496797
44564186
44395398
Deltaproteobacteria
44458624
44268872
44273506
44237343
44559186
44637522
44611039
44379892
44291662
44300874
44457950
44596453
Magnetococcus sp MC1
Photobacterium profundum
44521671
Idiomarina loihiensis
44370426
44595228
44518566
Shewanella oneidensis
Azoarcus sp EbN1
Gammaproteobacteria
44321861
44617834
Burkholderia cepacia
Betaproteobacteria
10
-lnL = 32668.0

Fig. S8

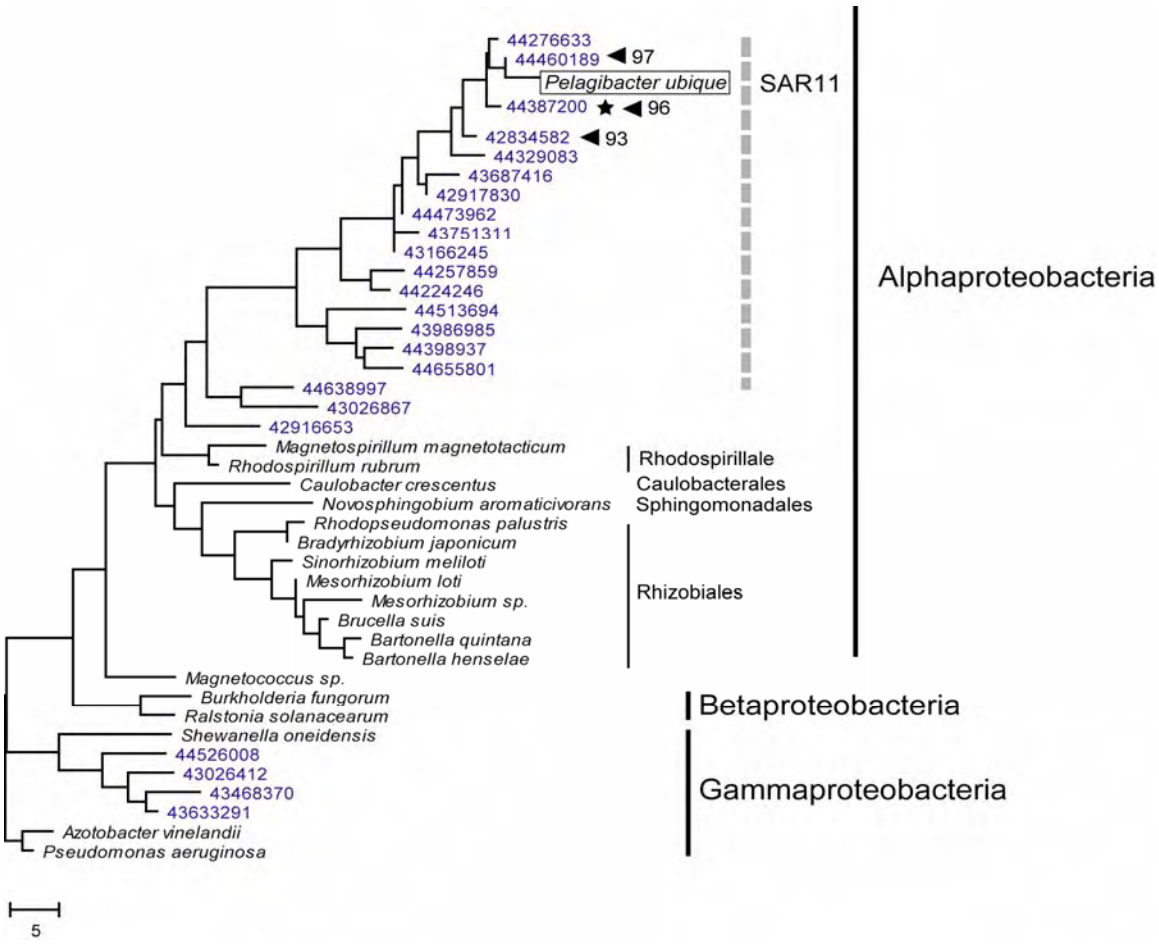
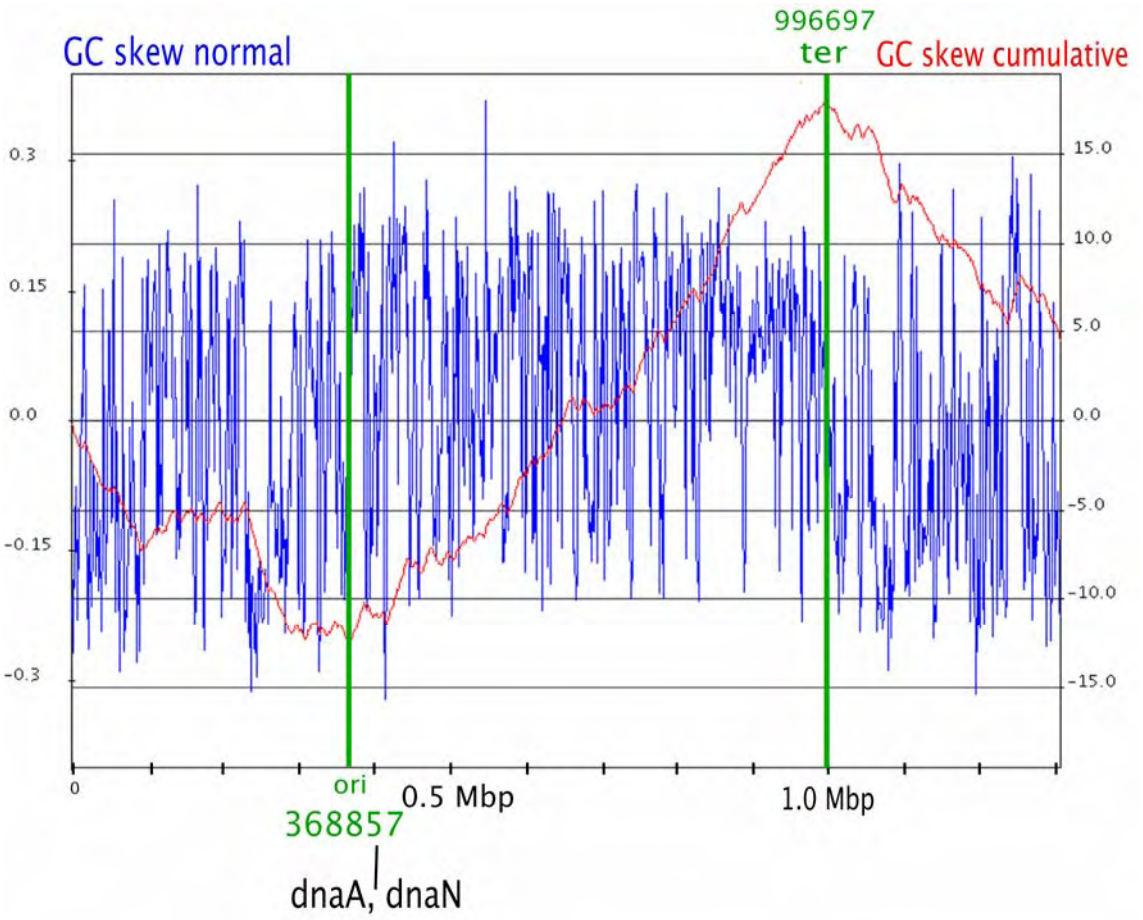

44276633
44460189
97
Pelagibacter ubique
44387200
96
42834582
93
44329083
43687416
42917830
44473962
43751311
43166245
44257859
44224246
44513694
43986985
44398937
44655801
44638997
43026867
42916653
Magnetospirillum magnetotacticum
Rhodospirillum rubrum
Caulobacter crescentus
Novosphingobium aromaticivorans
Rhodopseudomonas palustris
Bradyrhizobium japonicum
Sinorhizobium meliloti
Mesorhizobium loti
Mesorhizobium sp.
Brucella suis
Bartonella quintana
Bartonella henselae
Magnetococcus sp.
Burkholderia fungorum
Ralstonia solanacearum
Shewanella oneidensis
44526008
43026412
43468370
43633291
Azotobacter vinelandii
Pseudomonas aeruginosa
5
SAR11
Alphaproteobacteria
Rhodospirillale
Caulobacterales
Sphingomonadales
Rhizobiales
Betaproteobacteria
Gammaproteobacteria

Fig. S9


GC skew normal
996697
ter
GC skew cumulative
0.3
0.15
0.0
-0.15
-0.3
15.0
10.0
5.0
0.0
-5.0
-10.0
-15.0
0
ori
0.5 Mbp
1.0 Mbp
368857
dnaA, dnaN

Supporting References

1. D. K. Steinberg, *et al.*, *Deep-Sea Research II* **48**, 1405 (2001).
2. S. J. Giovannoni, E. F. DeLong, T. M. Schmidt, N. R. Pace, *Appl. Environ. Microbiol.* **56**, 2572 (1990).
3. M. T. Suzuki, S. J. Giovannoni, *Appl Environ Microbiol* **62**, 625 (1996).
4. J. M. Short, J. M. Fernandez, J. A. Sorge, W. D. Huse, *Nucleic Acids Res* **16**, 7583 (1988).
5. D. Gordon, C. Abajian, P. Green, *Genome Res* **8**, 195 (1998).
6. B. Ewing, P. Green, *Genome Res* **8**, 186 (1998).
7. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res* **8**, 175 (1998).
8. D. Gordon, C. Desmarais, P. Green, *Genome Res* **11**, 614 (2001).
9. H. Tettelin, D. Radune, S. Kasif, H. Khouri, S. L. Salzberg, *Genomics* **62**, 500 (1999).
10. A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res* **27**, 4636 (1999).
11. B. E. Suzek, M. D. Ermolaeva, M. Schreiber, S. L. Salzberg, *Bioinformatics* **17**, 1123 (2001).
12. B. Boeckmann, *et al.*, *Nucleic Acids Res* **31**, 365 (2003).
13. A. Bateman, *et al.*, *Nucleic Acids Res* **32**, D138 (2004).
14. N. J. Mulder, *et al.*, *Nucleic Acids Res* **33 Database Issue**, D201 (2005).
15. E. L. Sonnhammer, G. von Heijne, A. Krogh, *Proc Int Conf Intell Syst Mol Biol* **6**, 175 (1998).
16. T. Sicheritz-Ponten, S. G. Andersson, *Nucleic Acids Res* **29**, 545 (2001).
17. M. D. Stapels, J. C. Cho, S. J. Giovannoni, D. F. Barofsky, *J Biomol Tech* **15**, 191 (2004).
18. J. Song, A. Ware, S. L. Liu, *BMC Genomics* **4**, 17 (2003).
19. R. Chenna, *et al.*, *Nucleic Acids Res* **31**, 3497 (2003).
20. B. Boussau, E. O. Karlberg, A. C. Frank, B. A. Legault, S. G. Andersson, *Proc Natl Acad Sci U S A* **101**, 9722 (2004).