The ISME Journal (2014) 8, 1428–1439
© 2014 International Society for Microbial Ecology All rights reserved 1751-7362/14
www.nature.com/ismej

ORIGINAL ARTICLE

Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters

Haiwei Luo[1], Brandon K Swan[2], Ramunas Stepanauskas[2], Austin L Hughes[3] and Mary Ann Moran[1]

[1]*Department of Marine Sciences, University of Georgia, Athens, GA, USA;* [2]*Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA and* [3]*Department of Biological Sciences, University of South Carolina, Columbia, SC, USA*

The vast majority of surface ocean bacteria are uncultivated. Compared with their cultured relatives, they frequently exhibit a streamlined genome, reduced G + C content and distinct gene repertoire. These genomic traits are relevant to environmental adaptation, and have generally been thought to become fixed in marine bacterial populations through selection. Using single-cell genomics, we sequenced four uncultivated cells affiliated with the ecologically relevant Roseobacter clade and used a composition-heterogeneous Bayesian phylogenomic model to resolve these single-cell genomes into a new clade. This lineage has no representatives in culture, yet accounts for ~35% of Roseobacters in some surface ocean waters. Analyses of multiple genomic traits, including genome size, G + C content and percentage of noncoding DNA, suggest that these single cells are representative of oceanic Roseobacters but divergent from isolates. Population genetic analyses showed that substitution of physicochemically dissimilar amino acids and replacement of G + C-rich to G + C-poor codons are accelerated in the uncultivated clade, processes that are explained equally well by genetic drift as by the more frequently invoked explanation of natural selection. The relative importance of drift vs selection in this clade, and perhaps in other marine bacterial clades with streamlined G + C-poor genomes, remains unresolved until more evidence is accumulated.

The ISME Journal (2014) **8,** 1428–1439; doi:10.1038/ismej.2013.248; published online 23 January 2014
Keywords: Roseobacter; single-cell genomics; GC content; natural selection; genetic drift; genome streamlining

Introduction

The Roseobacter clade is a monophyletic subfamily-level lineage in the Alphaproteobacteria that represents up to 20% of bacterial cells in marine environments (Moran *et al.*, 2007). Members of this clade are known for their large and G + C-rich genomes equipped with versatile metabolic capabilities that allow rapid response to transient, nutrient-enriched patches (Moran *et al.*, 2004; Buchan *et al.*, 2005; Newton *et al.*, 2010). Several recent metagenomic analyses, however, suggest that views of this lineage may be biased by the characteristics of the cultured strains. For example, natural populations of Roseobacters in surface ocean waters have a significantly lower proportion of genes for adhesion, cellular capsule formation and signal transduction (Luo *et al.*, 2012), as well as for motility and quorum sensing (Newton *et al.*, 2010), compared with the numerous genomes available from cultured strains. Moreover, one cultured Roseobacter member branching at the base of the phylogenomic tree, strain HTCC2255, has a small and G + C-poor genome with evidence for gene loss biased toward functions related to cell–cell interaction, transcriptional regulation and motility (Luo *et al.*, 2013). Collectively, these studies suggest that genome content and life history strategies of Roseobacter clade members are more diverse than originally thought, with potential implications for their role in heterotrophic processing of fixed carbon in the ocean (Luo *et al.*, 2013).

The genome streamlining hypothesis has been invoked to explain the reduced genomes characteristic of many marine bacterial taxa (Dufresne *et al.*, 2005; Giovannoni *et al.*, 2005; Dupont *et al.*, 2012), based on the idea that selection acts to reduce genome size because of the metabolic costs of maintaining DNA with no adaptive value. Streamlined marine bacteria are also often small in cell size (potentially conferring a surface-to-volume ratio advantage; Dufresne *et al.,* 2005), have genomes

Correspondence: MA Moran, Department of Marine Sciences, University of Georgia, Marine Sciences Building, Athens, GA 30602, USA.
E-mail: mmoran@uga.edu

Received 2 July 2013; revised 11 December 2013; accepted 14 December 2013; published online 23 January 2014

poor in G + C pairs (which require more nitrogen than A + T pairs; Grzymski and Dussaq, 2012; Swan *et al.,* 2013), have a low proportion of noncoding DNA (Giovannoni *et al*., 2005; Swan *et al*., 2013) and maintain fewer genes involved in transcription and signal transduction (Swan *et al*., 2013). The primary mechanism theorized to underlie genome streamlining in marine bacteria has been selection for low metabolic costs in the oligotrophic ocean environment (Dufresne *et al*., 2005; Giovannoni *et al*., 2005).

Purifying selection, which acts to eliminate deleterious mutations, is considered the most prevalent form of natural selection at the molecular level (Nei, 1987; Hughes, 1999, 2008; Nei *et al*., 2010). Evidence for purifying selection on protein-coding genes is typically provided by the observation that the number of synonymous (silent) substitutions per synonymous site (d_S) exceeds the number of nonsynonymous (amino acid changing) substitutions per nonsynonymous site (d_N) in a vast majority of genes, as nonsynonymous mutations are more likely to be deleterious than synonymous mutations (Hughes, 2008). Purifying selection may become relatively inefficient, however, in taxa where the effective population size is reduced. This can occur frequently in natural populations (for example, owing to population bottlenecks or founder effects) and it results in chance fixation of slightly deleterious mutations by genetic drift (Ohta, 1992), a mechanism referred to as relaxation of purifying selection (Hughes and Friedman, 2009; Hughes, 2009, 2013).

The purpose of this study is to use molecular evolutionary genetic approaches to understand the evolution of oceanic Roseobacters that show evidence of genome streamlining. Our focus is on the mechanisms giving rise to differences in their genome traits compared with the better-studied members with large, G + C-rich genomes equipped for exploiting nutrient-rich environments (Moran *et al*., 2004; Buchan *et al*., 2005; Newton *et al*., 2010). We turned to single-cell amplified genomes (SAGs) obtained by flow cytometric sorting of single bacterial cells from seawater for this analysis (Swan *et al*., 2013), as SAGs are not biased toward traits required for growth on rich, solid media and, unlike metagenomic assemblies, represent a single genome appropriate for population genetic analysis at the nucleotide and amino acid level. Our results indicate that genetic drift cannot be ruled out as a dominant evolutionary force giving rise to the ecologically important Roseobacter lineages containing smaller and G + C-poor genomes.

Materials and methods

Sampling site and procedure

Water samples for single-cell analyses were collected using Niskin bottles from the South Atlantic (10 m; 27 November 2007; 12°29′41.40″ S, 4°59′55.20″ W), HOT station ALOHA in the North Pacific (25 m; 9 September 2009; Cruise KN192-5 station 11; 22°45′00″ N, 158°00′00″ W) and the Gulf of Maine (1 m; 16 September 2009; 43°50′39.87″ N, 69°38′27.49″ W). Replicate aliquots of water (1 ml volume) were cryopreserved with 6% glycine betaine (Sigma) and stored at − 80 °C (Cleland *et al*., 2004). Uncultivated cells SCGC AAA015-O19 (South Atlantic), SCGC AAA298-K06 and SCGC AAA300-J04 (North Pacific) and SCGC AAA076-C03 (Gulf of Maine) were sequenced using methods in Swan *et al.* (2013).

Estimating genome size

The size of single-cell and isolate genomes was estimated using a polynomial regression model by regressing the number of nucleotides sequenced against the ratio of the number of conserved single-copy genes universally present in five fully sequenced Roseobacter genomes to the number of predicted protein-coding genes in each genome. Although the MDA method used to amplify the SAG genomes has been shown to be biased, it is nonetheless appropriate provided that the bias is random, which has been shown to be the case (Woyke *et al*., 2009). The average genome size of Roseobacters in the Global Ocean Survey (GOS) metagenomes (Rusch *et al*., 2007) was obtained by dividing the number of Roseobacter nucleotides by the number of Roseobacter genomes occurring in GOS. The former value was determined based on sequences with best hits to a Roseobacter gene using the RAPsearch2 software (Zhao *et al*., 2012) with an *E*-value of 10^{-3} to search all GOS reads against the NCBI nonredundant (nr) database. This software is more efficient compared with BLAST, while achieving a similar accuracy (Zhao *et al*., 2012). In this way, we identified 85 842 reads or 90 555 430 bp as belonging to the Roseobacter clade. The number of genomes was approximated by the total number of Roseobacter GOS nucleotides coding for one of the five single-copy genes (*atpD*, *dnaK*, *gyrB*, *recA* and *rpoB*) divided by the average length of that gene in Roseobacter isolate genomes. GOS reads associated with Roseobacters coding for these five genes were verified using RPS-BLAST searches (Altschul *et al*., 1997) against the Clusters of Orthologous Groups database (Tatusov *et al*., 1997), producing an estimate of 31.6 Roseobacter genome equivalents in the GOS database. This value was divided into the number of Roseobacter bases to estimate genome size. To obtain a distribution of genome size in GOS, we applied a resampling technique (with replacement) to the Roseobacter reads identified by RAPsearch2. Replicates were resampled 1000 times, and the average genome size was estimated for each as described above.

Estimating the fraction of noncoding DNA

The fraction of noncoding DNA in single-cell genomes and isolate genomes was estimated by dividing the total length of predicted noncoding DNA sequences by the number of nucleotides sequenced. To obtain the fraction of noncoding DNA in GOS roseobacters, we used the 5608 Roseobacter reads (including the paired-end reads; 5 941 949 bp) sampled by the d_{N} pipeline software and validated by phylogenetic reconstruction (Luo *et al.*, 2012). When these high-confidence Roseobacter reads were searched against the NCBI RefSeq database (Pruitt *et al.*, 2009) using the BLASTX software (Altschul *et al.*, 1997), we determined that 5 491 733 bp hit within a coding region using a cutoff of E-value <0.1 and bit score >40. The remaining base pairs represented noncoding DNA in the GOS Roseobacters. To obtain a distribution of noncoding DNA in GOS, we applied a resampling technique (with replacement) to the 5608 Roseobacter reads. A total of 1000 resamplings were carried out, and the fraction of noncoding DNA was calculated for each.

Estimating the relative abundance of the uncultivated Roseobacters in GOS

The fraction of Roseobacter reads in the GOS data set that were affiliated with the SAG-O19 clade (consisting of O19, K06 and J04) was estimated in two ways. First, we constructed a custom reference database by adding the predicted amino acid sequences of the four SAGs to the NCBI microbial RefSeq database (Pruitt *et al.*, 2009). GOS reads were queried against this database using the RAPsearch2 software (Zhao *et al.*, 2012) with a cutoff E-value of 10^{-3}, and sequences with a best hit to a Roseobacter gene were retained. To account for the lower genome coverage of the SAGs, we focused on 1396 orthologous gene families that covered at least one member of the SAG-O19 clade genomes and one of the remaining Roseobacter genomes, with orthologous families identified by Cluster of Orthologous Groups (Tatusov *et al.*, 1997). The proportion of best hits to SAG-O19 clade members in this subset was used as an indication of clade representation in the GOS data set. Because not all marine Roseobacter lineages are currently represented by public genome sequences, we also compared the 16S rRNA gene sequences between the SAG-O19 clade members and the GOS Roseobacters. The metagenomic reads that harbored a 16S rRNA gene with a best hit to a Roseobacter 16S rRNA gene were retrieved from the GOS metagenome. Those with a sequence identity $>99\%$ to the 16S rRNA gene of a SAG-O19 clade member were considered members of the clade.

Ortholog identification, alignment construction, model adequacy test

The four SAGs and 40 isolate sequences were used in genome tree construction with four outgroup genomes in other Alphaproteobacteria lineages. Orthologous genes were identified by a reciprocal BLAST search (Altschul *et al.*, 1997) for all possible genome pairs at the amino acid level using an E-value of 10^{-6}. Then, shared orthologous genes among the 48 genomes were identified by the MSOAR software (Chen *et al.*, 2005); MSOAR assigns orthologs between a pair of genomes by considering sequence similarity among homologs and minimizing the number of gene rearrangements (Chen *et al.*, 2005). Next, one genome was picked at random as the reference genome, and pairwise orthologs were linked to the reference genome. In this way, orthologs shared by all the genomes were identified (Luo *et al.*, 2011a, b, 2012, 2013). To identify false positive orthologous sets, each member in an orthologous set was searched against the Clusters of Orthologous Groups database (Tatusov *et al.*, 1997) using the RPS-BLAST software (Altschul *et al.*, 1997). Orthologous sets including members hitting to different Clusters of Orthologous Groups were not used further. Following this procedure, we identified 52 single-copy orthologous protein families that are shared by the 48 genomes.

As member genes in a family can be divergent, no single alignment method worked well for all of the data sets. Therefore, we used four different algorithms to align the member genes at the amino acid level. These four methods are Mafft (Katoh *et al.*, 2005), T-Coffee (Notredame *et al.*, 2000), Muscle (Edgar, 2004) and Prank (Löytynoja and Goldman, 2008). Each alignment was trimmed using the trimAl software (Capella-Gutiérrez *et al.*, 2009) with the parameters '-automated1 -resoverlap 0.55 -seqoverlap 60'. The quality of four trimmed alignments for each family was evaluated using a consistency score calculated by trimAl (Capella-Gutiérrez *et al.*, 2009), with a higher consistency score showing that the alignment produced by one method was more consistent than alignments produced by other methods. Overall, Mafft was selected for 37 data sets, and one of the other methods was chosen for each of the remaining data sets. The best alignments were concatenated for genome-based phylogenetic construction, resulting in a concatenated data set with 18 825 sites.

Model adequacy with respect to amino acid composition was assessed for each of the 52 protein families using χ^2-homogeneity statistic on posterior distributed samples, which were generated by posterior predictive simulation using the P4 Bayesian phylogenetic software package (Foster, 2004).

Genome-based phylogenetic construction using a composition-heterogeneous Bayesian model

To account for a systematic error associated with compositional bias in the amino acid sequences, the concatenated protein sequence was recoded into the following six Dayhoff groups that correspond to PAM and most other matrices (Hrdy *et al.*, 2004): (cysteine), (alanine, serine, threonine, proline, glycine), (asparagine, aspartic acid, glutamic acid, glutamine), (histidine, arginine, lysine), (methionine, isoleucine, leucine, valine), (phenylalanine, tyrosine, tryptophan). This recoding scheme has been used to improve topological estimation in the presence of compositional heterogeneity in phylogenomic studies (Cox *et al*., 2008; Foster *et al*., 2009; Nesnidal *et al*., 2010). We applied Bayesian nonstationary models including the node-discrete composition heterogeneity (NDCH) model and the node-discrete rate matrix heterogeneity (NDRH) model. The NDCH model allows different composition on different branches, and the NDRH model allows different rate matrices on different branches (Foster, 2004). These models are computationally expensive but specifically designed to account for compositional heterogeneity. They are implemented in the P4 Bayesian phylogenetic software package (Foster, 2004).

The Dayhoff-recoded concatenated data set was analyzed using multiple configurations of NDCH and NDRH, general time-reversible substitution matrix plus four Gamma-distributed rate categories and using the polytomy prior (Lewis *et al*., 2005). Ten replicate runs were performed for each configuration of NDCH and NDRH. In each replicate run, one cold and three heated MCMC chains were run for a total of 1 500 000 generations, with trees sampled every 1000 generations. The first 500 000 generations were discarded as 'burn-in'. The model adequacy with respect to composition was assessed using χ^2-homogeneity statistic on posterior distributed samples, which were generated by posterior predictive simulation in P4. This test rejected the stationary model (1 composition vector plus 1 general time-reversible rate matrix across the tree). Model selection was performed using the Bayes factor, which is the ratio of the marginal likelihood of the two models being compared. Mathematically, it is expressed as twice the difference in the logs of the marginal likelihoods. A value >10 is considered significant to favor the better model (Kass and Raftery, 1995). The marginal likelihoods were estimated using equation 16 in Newton and Raftery (Newton and Raftery, 1994) based on the harmonic mean of the likelihoods from posterior distributions. This test selected a nonstationary model of NDCH(8) + NDRH(2), that is, a model with eight composition vectors and two general time-reversible rate matrices. The average standard deviation of split support was <0.01, suggesting that convergence was reached. A majority-rule consensus tree was constructed from the post-burn-in trees. This conservative method resolved three SAGs (O19, K06 and J04) into a novel clade (designated SAG-O19 clade) in which no cultured representatives were found.

Estimating the rates of conservative and radical amino acid replacement

Orthologous gene families among the 48 genomes were identified using the OrthoMCL software (Li *et al*., 2003). In each family, duplicates in the same genomes were discarded such that each family presumably contains single-copy orthologous genes. Families that contained <30 genomes of Roseobacter isolates and less than two of the three SAGs in the SAG-O19 clade were also discarded. This data filtering procedure resulted in 585 families for subsequent analyses. Genes were aligned using Mafft (Katoh *et al*., 2005) at the amino acid level, and DNA sequences were imposed on the alignments.

The 20 amino acids were classified using two widely accepted schemes based on their physicochemical properties: (I) charge (Hughes *et al*., 1990), and (II) polarity and volume (Miyata *et al*., 1979). Conservative and radical amino acid replacements were defined as amino acid substitutions within and between groups, respectively. Rates of conservative (d_C) and radical (d_R) replacements, that is, the number of conservative nonsynonymous nucleotide substitutions per conservative nonsynonymous site and the number of radical nonsynonymous nucleotide substitutions per radical nonsynonymous site, between two orthologous genes were determined using the HON-new software (Zhang, 2000). To account for the transitional bias, the transition/transversion ratio, determined separately for each gene family using the MEGA-CC software (Kumar *et al*., 2012), was incorporated. In d_C and d_R calculations, the two genomes in the basal HTCC2255 clade were not included, because this lineage has distinct genome characteristics and does not share an evolutionary path with all other Roseobacters (Luo *et al*., 2013).

Results and Discussion

Roseobacter single-cell genome characterization

Partial genomes of four SAGs were recovered from the Pacific and Atlantic Oceans (Table 1), represented by 0.6–2.0 Mbp of assembled DNA per genome. These SAGs were determined to represent cells with reduced genome sizes (estimated to be 2.6–3.5 Mbp) using a polynomial regression model based on the presence of single-copy genes conserved in cultured Roseobacters (Figure 1). One of the SAGs, C03, shares an identical 16S rRNA gene and $>97\%$ average nucleotide identity with the strain HTCC2255, which branches at the base of the Roseobacter phylogeny and has genome content consistent with a free-living strategy (Luo *et al*., 2013). Genome content of the other three SAGs (O19, K06, J04) is also significantly different from the cultured, presumably patch-adapted Roseobacters (χ^2-test; $P<0.001$). In particular, genes involved in transcriptional regulation and replication/recombination/repair are significantly less common in these three SAGs (Xipe analysis, Rodriguez-Brito *et al.,* 2006; 99% confidence interval, 1000 replicates), functional categories found to be underrepresented

Table 1 Characteristics of single-cell amplified genomes

SAG ID (SCGC-)	*Clade*	*Source*	*Assembled nucleotides (Mb)*	*No. of Contigs*	*Identified coding sequences*	*Predicted genome size (Mb)*	*Estimated genome coverage (%)*
AAA300-J04	SAG-O19	North Pacific	0.62	77	651	2.65	23
AAA298-K06	SAG-O19	North Pacific	1.70	231	1766	3.50	48
AAA015-O19	SAG-O19	South Atlantic	1.70	159	1783	3.10	55
AAA076-C03	HTCC2255	Gulf of Maine	2.00	107	1925	2.64	76

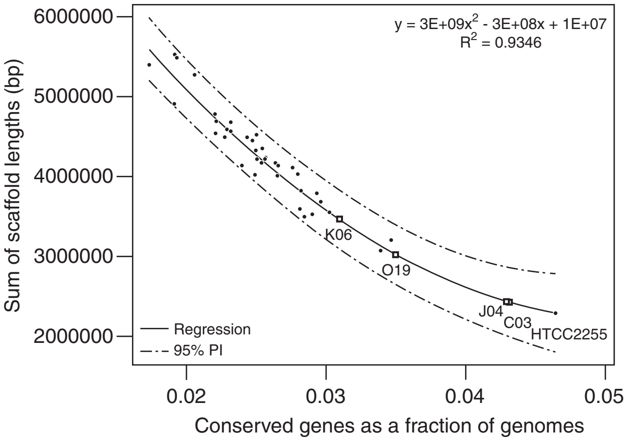

Figure 1 Regression model for Roseobacter SAG genome size estimation based on genome statistics for 40 cultured Roseobacter strains. The *x* axis shows the ratio of the number of conserved single-copy genes universally present in fully sequenced Roseobacter genomes to the number of predicted protein-coding genes in a genome; the *y* axis is the number of nucleotides sequenced. The data were fit to a polynomial regression model ($R^2 = 0.93$), and the model was used to estimate the genome sizes of the four SAGs (C03, O19, K06 and J04), which were found to be 2.64, 3.10, 3.50 and 2.65 Mb, respectively. The prediction interval (PI) is also shown.

in streamlined marine bacterial genomes (Lauro *et al.*, 2009; Yooseph *et al.*, 2010; Luo *et al.*, 2013; Swan *et al.*, 2013). They also have an atypically low G + C content, averaging 39% (± 1%) compared with a mean of 60% (± 4%) for cultured Roseobacters (Figure 2b), and a smaller fraction of noncoding DNA (0.075 ± 0.004) compared with the cultured Roseobacters (0.109 ± 0.022) (Figure 2c), two characteristics frequently linked with bacterial genome reduction (Giovannoni *et al.*, 2005; McCutcheon and Moran, 2012).

We compared characteristics of oceanic Roseobacter sequences obtained from metagenomic data with the SAG and isolate genomes. Over 5600 Roseobacter reads retrieved from the GOS metagenome with high confidence using the d_N pipeline software (Luo *et al.*, 2012) had a bimodal distribution of G + C content, with a major peak centered at 42% and a secondary peak at 54% (Figure 2a). The oceanic genomes had an estimated size of 2.87 (± 0.15) Mb (Figure 2d), and percent noncoding DNA content of 0.076 (± 0.002) (Figure 2c). All three of these characteristics are consistent with those of the SAGs while divergent from the isolates, and suggest a cultivation bias against Roseobacter clade members with streamlined genomes. Thus, the SAG genomes provide an opportunity to explore the evolution of a fraction of this lineage that has previously been largely inaccessible to molecular genetic analysis. We note, however, that the size fractionation scheme used in the GOS study (0.1–0.8 μm size range; Rusch *et al.* 2007) was biased against large and particle-associated Roseobacter cells, and these may better resemble the cultured strains. Roseobacters in the surface ocean are estimated to be about equally divided between free-living and particle-associated cells (Luo *et al.*, 2012).

Phylogenomic tree construction

A robust phylogenomic tree was constructed to address the evolutionary history of the Roseobacter clade. The initial RAxML-based trees were incongruent depending on whether or not SAGs O19, K06 and J04 were included (Supplementary Figure S1), despite using an identical set of amino acid characters in both cases. Inconsistencies in the trees included an unresolved branching order of the five major Roseobacter clades (Newton *et al.*, 2010), as shown by low bootstrap support, and an extraordinarily long branch for SAG C03 in the HTCC2255 clade (Supplementary Figure S1A), conflicting with its identical 16S rRNA gene sequence and > 97% average nucleotide identity to the HTCC2255 genome.

A wide range in G + C content such as that found in the Roseobacter genomes (37–70%) is known to cause compositional bias in the amino acid sequences (Gu *et al.*, 1998; Foster and Hickey, 1999; Singer and Hickey, 2000; Swan *et al.*, 2013). Such compositional bias introduces systematic errors in phylogenetic reconstruction using standard approaches that assume homogeneity of amino acid (or nucleotide) composition among taxa (Galtier and Gouy, 1995; Jermiin *et al.*, 2004; Herbeck *et al.*, 2005; Foster *et al.*, 2009; Nesnidal *et al.*, 2010). Possible disparity in amino acid composition across the lineages was explored using a posterior predictive simulation, and subsequent χ^2-homogeneity tests showed that half of the 52 orthologous protein groups shared by all Roseobacter genomes significantly violated the homogeneous assumption with respect to amino acid composition across the tree (Supplementary Table S1). To address the tree-wide compositional variation, the P4 Bayesian

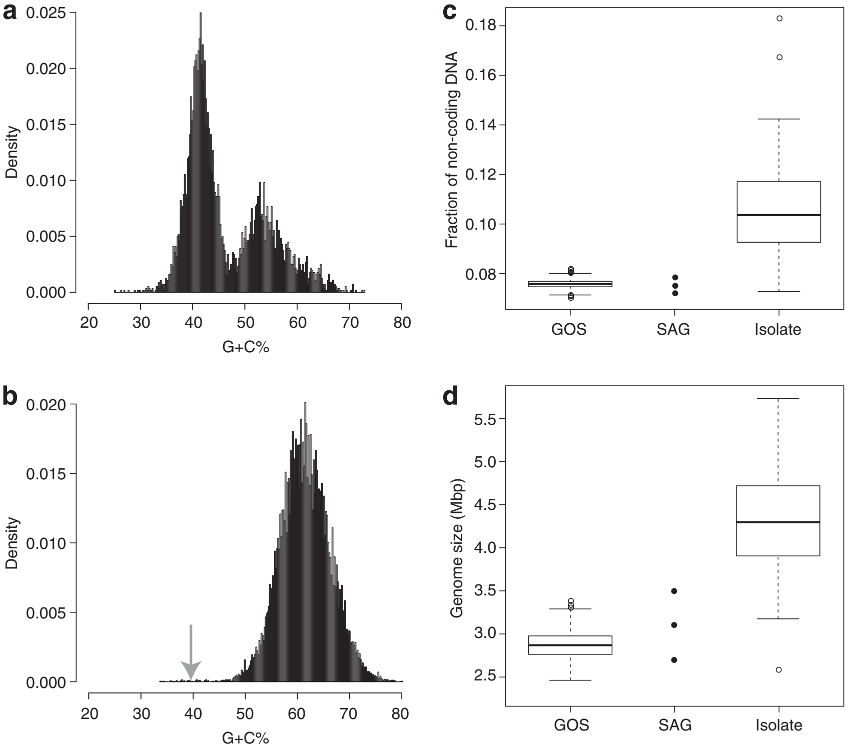

Figure 2 Genomic characteristics of the three SAG-O19 clade members compared with metagenomic roseobacter sequences and isolates. Distribution of G + C content sampled from Global Ocean Survey (GOS) metagenomic data sets (**a**) and cultured Roseobacter genomes (**b**). For panel a, the GOS roseobacter reads ($n = 5608$) were identified using the d_N pipeline software, which assigns metagenomic reads to a microbial clade with high confidence (Luo *et al.*, 2012). For panel b, the cultured roseobacter genomes were randomly sheared to generate an *in silico* metagenome, and then analyzed as described for panel a. The gray arrow in panel b indicates the average G + C content of the SAGs. The fraction of noncoding DNA (**c**) and estimated average genome size (**d**) in Roseobacter GOS reads, SAGs and isolates. The box-and-whiskers plots indicate the median (horizontal line), the boundary of the first (Q1) and third quartile (Q3) (box) and the range from Q1–1.5(Q3–Q1) to Q3 + 1.5(Q3–Q1) (whiskers); circles indicate outliers. The details for noncoding DNA fraction and genome size estimation can be found in the Materials and methods.

phylogenetic software package (Foster, 2004) was selected for phylogenomic tree construction as it implements the NDCH model, which allows for varying composition over the tree, and the NDRH model, which allows different rate matrices on different branches. The P4 phylogenomic tree resolved the earlier issues and was supported by high posterior probabilities on all internal branches (Figure 3). In this tree, SAGs O19, K06 and J04 formed a new clade without any cultured representatives (Figure 3; hereafter, the SAG-O19 clade). In addition, members of the SAG-O19 clade show extraordinarily long branches in the P4 tree (Figure 1), suggesting an enhanced evolutionary rate of this uncultivated clade.

Enhanced evolutionary rate of dissimilar amino acid substitutions in the uncultivated clade

Using approaches based on genome binning and 16S rRNA gene analysis (see Materials and methods), the SAG-O19 clade is estimated to account for ∼35% of the free-living Roseobacter cells (<0.8 μm) in GOS metagenomic samples of surface ocean waters (Supplementary Figure S2). Comparing nucleotide substitution patterns may shed light on the nature of the evolutionary forces giving rise to this abundant lineage. A common approach to diagnosing the efficiency of purifying selection compares the genomic average of the ratio of d_N to d_S, among different lineages (Clark *et al.*, 1999; Wernegreen and Moran, 1999; Hu and Blanchard, 2009; Kuo *et al.*, 2009). Because nonsynonymous mutations are more likely to be deleterious than synonymous mutations, an elevated genome-wide value indicates that more deleterious mutations have accumulated owing to relaxation of purifying selection (Ohta, 1992). A major limitation of this approach, however, is that it only applies to closely related lineages; for more divergent sequences, synonymous sites are saturated with substitutions, making the estimate of d_S unreliable (Ochman *et al.*, 1999; Luo and Hughes, 2012). In the case of the Roseobacter SAG-O19 clade, d_N/d_S is not applicable owing to saturation at the synonymous sites.

Using an alternative approach, we examined the type of nonsynonymous substitutions in the SAG-O19 clade. Conservative substitutions involving amino acids with similar physicochemical properties are more common than radical substitutions involving amino acids with substantial differences in physicochemical properties (Zuckerkandl and Pauling, 1965;

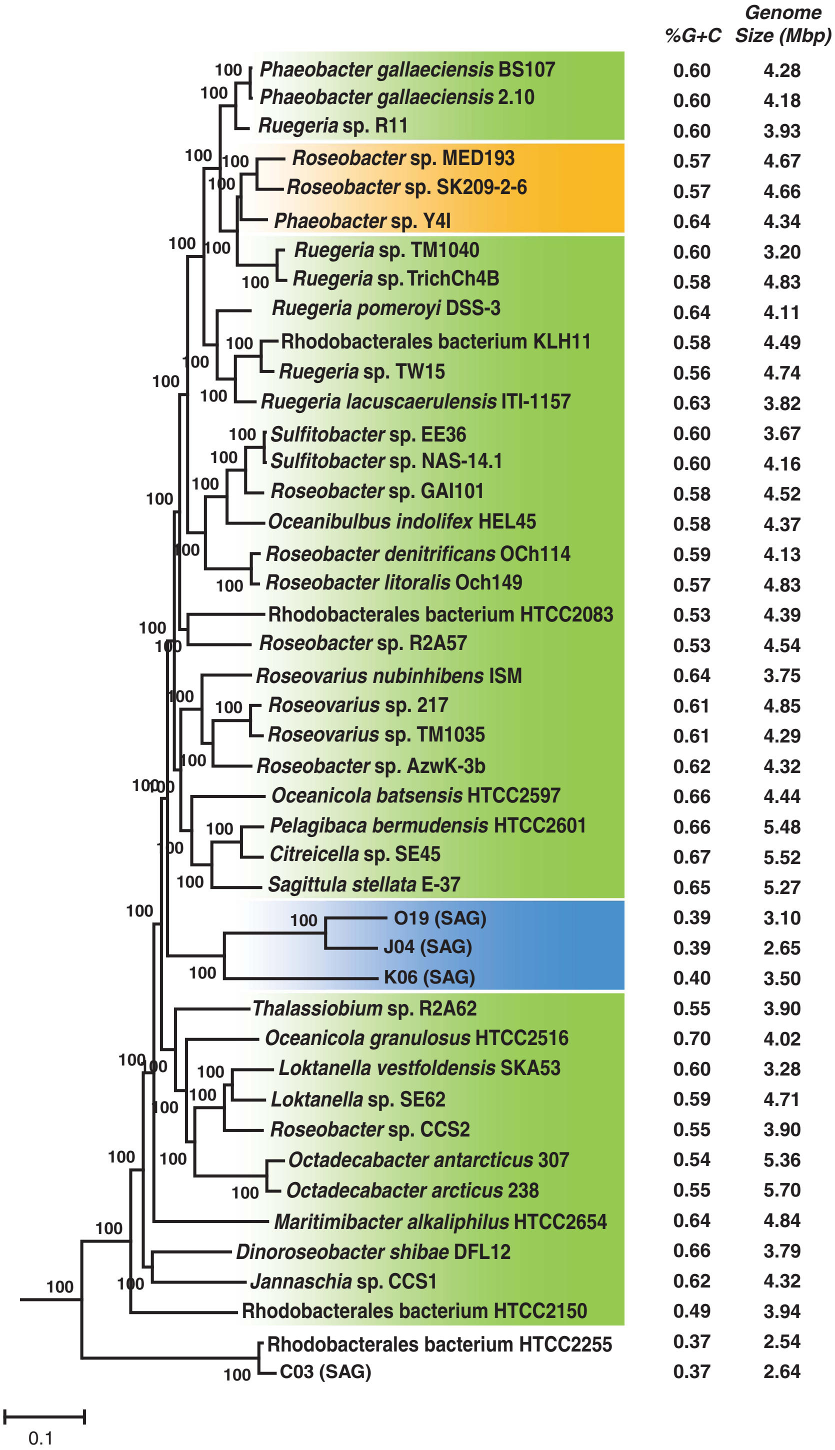
%G+C
Genome Size (Mbp)
Phaeobacter gallaeciensis BS107 0.60 4.28
Phaeobacter gallaeciensis 2.10 0.60 4.18
Ruegeria sp. R11 0.60 3.93
Roseobacter sp. MED193 0.57 4.67
Roseobacter sp. SK209-2-6 0.57 4.66
Phaeobacter sp. Y4I 0.64 4.34
Ruegeria sp. TM1040 0.60 3.20
Ruegeria sp. TrichCh4B 0.58 4.83
Ruegeria pomeroyi DSS-3 0.64 4.11
Rhodobacterales bacterium KLH11 0.58 4.49
Ruegeria sp. TW15 0.56 4.74
Ruegeria lacuscaerulensis ITI-1157 0.63 3.82
Sulfitobacter sp. EE36 0.60 3.67
Sulfitobacter sp. NAS-14.1 0.60 4.16
Roseobacter sp. GAI101 0.58 4.52
Oceanibulbus indolifex HEL45 0.58 4.37
Roseobacter denitrificans OCh114 0.59 4.13
Roseobacter litoralis Och149 0.57 4.83
Rhodobacterales bacterium HTCC2083 0.53 4.39
Roseobacter sp. R2A57 0.53 4.54
Roseovarius nubinhibens ISM 0.64 3.75
Roseovarius sp. 217 0.61 4.85
Roseovarius sp. TM1035 0.61 4.29
Roseobacter sp. AzwK-3b 0.62 4.32
Oceanicola batsensis HTCC2597 0.66 4.44
Pelagibaca bermudensis HTCC2601 0.66 5.48
Citreicella sp. SE45 0.67 5.52
Sagittula stellata E-37 0.65 5.27
O19 (SAG) 0.39 3.10
J04 (SAG) 0.39 2.65
K06 (SAG) 0.40 3.50
Thalassiobium sp. R2A62 0.55 3.90
Oceanicola granulosus HTCC2516 0.70 4.02
Loktanella vestfoldensis SKA53 0.60 3.28
Loktanella sp. SE62 0.59 4.71
Roseobacter sp. CCS2 0.55 3.90
Octadecabacter antarcticus 307 0.54 5.36
Octadecabacter arcticus 238 0.55 5.70
Maritimibacter alkaliphilus HTCC2654 0.64 4.84
Dinoroseobacter shibae DFL12 0.66 3.79
Jannaschia sp. CCS1 0.62 4.32
Rhodobacterales bacterium HTCC2150 0.49 3.94
Rhodobacterales bacterium HTCC2255 0.37 2.54
C03 (SAG) 0.37 2.64
100
0.1

Figure 3 Bayesian phylogenomic tree of the Roseobacter clade using a composition-heterogeneous model in the P4 software package based on a concatenation of 52 single-copy orthologous protein sequences. The scale bar indicates the number of substitutions per site. The value near each internal branch is the posterior probability for that branch. The tree is rooted using four genomes from sister clades in the Alphaproteobacteria (*Hyphomonadaceae*, *Caulobacterales*, *Rhizobiales*); the outgroups are not shown. The SAG clade is highlighted in blue, an example control clade is highlighted in orange and the remaining members used to calculate d_R/d_C ratios are indicated in green. The basal lineage consisting of HTCC2255 and SAG C03 were not included in the analyses. G + C content and genome size are indicated in the two right columns, the latter estimated using a regression model for incomplete genomes (Figure 1). Plasmids, if any, are not included in genome size calculation.

Dayhoff *et al.*, 1972). This bias is attributed to a higher intensity of purifying selection on changes that affect protein function. The radical (d_R) and conservative (d_C) substitution rates in Roseobacter genomes were computed by classifying amino acids into three groups according to charge (positive, negative and neutral; Hughes *et al.*, 1990), and computing the number of radical (those between charge groups) nonsynonymous substitutions per radical nonsynonymous site (d_R) and the number of conservative (within a charge group) nonsynonymous substitutions per conservative nonsynonymous site (d_C) (Hughes *et al.*, 1990; Zhang, 2000). The genomic average of the d_R/d_C ratio has been used previously to compare the efficiency of purifying selection among different lineages, with an inflated mean value indicative of relaxed purifying selection (Zhang, 2000; Eyre-Walker *et al.*, 2002; Hughes and Friedman, 2009; Wernegreen, 2011). The advantages of this approach are that it is applicable to data sets with sequences that are too divergent for synonymous site analysis and provides a population genetics environment to address ancient events.

The pattern of nonsynonymous substitution in the three SAG-O19 clade genomes was compared with control isolate clades similarly consisting of three genomes. Six different control clades were designated, all of which yielded qualitatively the same results (Figure 4); detailed results are discussed only for the MED193/SK209-2-6/Y4I control group here. For over 500 orthologs, the mean d_C and d_R was computed for the following two sets of comparisons: (1) between the SAG-O19 clade and the remaining genomes (excluding the control clade and the basal HTCC2255 lineage; see Materials and methods and Figure 3) and (2) between the control clade and the remaining genomes (excluding the SAG-O19 clade and the basal HTCC2255 lineage). As anticipated, the mean d_C was significantly greater than mean d_R (0.297 ± 0.004 vs 0.287 ± 0.004 for the comparisons involving the SAG-O19 clade; 0.209 ± 0.003 vs 0.192 ± 0.003 for comparisons involving the control clade; paired *t*-test $P<0.001$ for both), confirming the expected bias toward conservative amino acid changes. However, the mean ratio d_R/d_C was significantly greater in the SAG-O19 clade (0.970 ± 0.007) compared with the control clades (0.916 ± 0.007 for the MED193/SK209-2-6/Y4I comparison; paired *t*-test; $P<0.001$; Figure 4), suggesting that the bias toward conservative amino acid changes is weaker for the SAG-O19 clade. We conducted a similar analysis that classified amino acids by volume and polarity (Table 2; Miyata *et al.* 1979) rather than charge, and this also showed a greater rate of physicochemically dissimilar amino acid changes in the SAG-O19 clade relative to the cultured Roseobacter clades (data not shown). Therefore, analyses of d_R/d_C are consistent with the hypothesis that relaxation of purifying selection is a dominant evolutionary force giving rise to the clade.

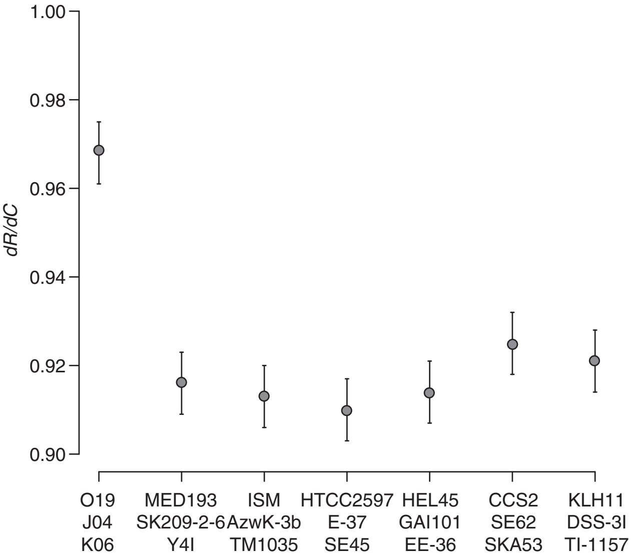

Figure 4 The ratio of radical (d_R) to conservative (d_C) substitution rates based on physicochemical properties (shown here, charge) of amino acids for the SAG-O19 clade and six control clades. Bars indicate one standard deviation of the mean. For the control clades, only strain names are shown; full names can be found in Figure 3.

Enhanced evolutionary rate of G + C content change in the uncultivated clade

As members of the SAG-O19 clade have a substantially lower G + C content compared with most other Roseobacter lineages, we examined whether this difference is caused by a change in selective pressure. The 20 amino acids were classified into three groups based on the G + C content at the first two sites of a codon (as natural selection largely operates on nucleotide substitutions resulting in the replacement of amino acids): high G + C, low G + C and neutral (Table 3). Using the same analysis design as above, we compared the pattern of codon replacement for the SAG-O19 clade and for each of the six different control clades with the remaining Roseobacter genomes. In the comparisons involving the SAG-O19 clade, mean d_{within} (within group; 0.323 ± 0.004) was significantly greater than mean $d_{between}$ (between group; 0.267 ± 0.004; paired *t*-test; $P<0.001$). Similarly, in the comparisons involving the MED193/SK209-2-6/Y4I control clade, mean d_{within} (0.242 ± 0.003) was significantly greater than

Table 2 Classification of amino acids by two independent schemes based on physicochemical properties

Classification by charge (Hughes *et al.*, 1990)	
Positive	R, H, K
Negative	D, E
Neutral	A, N, C, Q, G, I, L, M, F, P, S, T, W, Y, V
Classification by volume and polarity (Miyata *et al.*, 1979)	
Special	C
Neutral and small	A, G, P, S, T
Polar and relative small	N, Q, Q, E
Polar and relative large	R, H, K
Nonpolar and relatively small	I, L, M, V
Nonpolar and relatively large	F, W, Y

Table 3 Grouping of amino acids based on the G + C content of the first two positions in a codon

Classification by G + C content of the first two positions in a codon	
High G + C	P, A, G, R
Low G + C	F, I, M, Y, N, K, L
Neutral	V, S, T, H, Q, D, E, C, W

mean d_{between} (0.174 ± 0.003; paired *t*-test; $P < 0.001$). Thus, in both sets of comparisons, amino acid changes had a preference for codons with similar G + C content. However, the mean ratio $d_{\text{between}}/d_{\text{within}}$ was significantly greater in the comparisons involving the SAG-O19 clade (0.811 ± 0.005) than those involving all six control clades (0.708 ± 0.004 for the MED193/SK209-2-6/Y4I comparison; paired *t*-test; $P < 0.001$), suggesting a greater rate of substitution toward G + C-poor codons in the SAG-O19 clade.

Acceleration of replacement from G + C-rich to G + C-poor codons in the uncultivated SAG-O19 clade could have both selection and drift explanations. The low G + C content of marine bacterioplankton genomes has frequently been attributed to adaptation via selection for nitrogen conservation in the oligotrophic ocean, achieved by lowering cell quotas through the preferential use of nitrogen-poor A/T nucleobases in the genome and nitrogen cost minimization in the proteome (Bragg and Hyder, 2004; Martin-Cuadrado *et al.*, 2008; Grzymski and Dussaq, 2012). On the other hand, there is an inherent mutational bias from G/C to A/T in bacteria (Lind and Andersson, 2008; Balbi *et al.*, 2009; Hershberg and Petrov, 2010; Hildebrand *et al.*, 2010). Thus, the presence of lineages with high G + C content suggests that purifying selection has acted to maintain the G + C content (Balbi *et al.*, 2009; Hershberg and Petrov, 2010; Hildebrand *et al.*, 2010; Raghavan *et al.*, 2012), and, conversely, lineages with low G + C content could be caused by random fixation as a result of relaxation of purifying selection. In the case of Roseobacters, the mean ratio $d_{\text{between}}/d_{\text{within}}$ was significantly greater in the comparisons involving the G + C-poor SAG-O19 clade than those involving all six control clades, a result that could be explained by relaxation of purifying selection in the SAG-O19 clade.

If low G + C content is under selection in this clade, the attribution of a drift component to the high d_R/d_C is less strongly supported because selective pressure to decrease G + C content may interfere with selection for conservation of amino acid physicochemical properties. In other words, the fitness cost of radical changes could be balanced by the benefit of decreasing G + C content. However, the assumption that low G + C content in marine bacteria is favored by selection is controversial. For instance, members of the LD12 clade are found in freshwater where N is not limiting, yet have the same low G + C content as their marine SAR11 relatives (~30%; Giovannoni *et al.*, 2005), suggesting that factors other than selection for N thrift must have acted to shape these G + C-poor genomes (Ghai *et al.*, 2011).

Conclusion

Streamlined genomes are found in several other dominant marine bacterial lineages, such as SAR11 (Giovannoni *et al.*, 2005), SAR86 (Dupont *et al.*, 2012) and *Prochlorococcus* (Rocap *et al.*, 2003). Although there is no direct estimate of effective population sizes for these lineages, it has generally been assumed that they are quite large, thereby minimizing the power of genetic drift (Morris *et al.*, 2012) and implicating selective forces in shaping characteristics such as genome size (Dufresne *et al.*, 2005; Giovannoni *et al.*, 2005; Hu and Blanchard, 2009; Grote *et al.*, 2012; Morris *et al.*, 2012) and nucleotide composition (Giovannoni *et al.*, 2005; Martin-Cuadrado *et al.*, 2008). Indeed, small genomes may enhance the metabolic efficiency and lead to smaller cells with increased surface-to-volume ratios, both of which are advantageous to bacteria living in oligotrophic oceans (Dufresne *et al.*, 2005; Giovannoni *et al.*, 2005). Further, cells with G + C-poor genomes use less nitrogen in both genomes and proteomes compared with cells with G + C-rich genomes (Bragg and Hyder, 2004; Grzymski and Dussaq, 2012). Other explanations for genome streamlining in marine bacteria include changes in bacterial lifestyle, making previously necessary genes dispensable. This relaxation of purifying selection on nonessential genes (Luo *et al.*, 2011a, 2013) does not require a reduction in effective population size, and thus does not conflict with the selection hypothesis.

Here, we addressed the evolutionary history of a streamlined clade in the Roseobacter lineage through analysis of genomes from single, uncultivated cells, finding evidence for accelerated evolution toward nonconservative amino acid changes and low G + C content that suggests a possible role for genetic drift. The seemingly contradictory explanations of drift and selection can be reconciled if these evolutionary forces acted at different periods during the clade's evolutionary history. The conceptual differences between these evolutionary processes, including whether the fixation of genomic traits originates through random fixation of neutral or even slightly deleterious mutations or through directional evolution toward fixation of advantageous variants, is fundamental to our understanding of microbial adaptation in the ocean.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank A Rivers, B Satinsky and B Durham for comments on an earlier version of the manuscript, P Foster for advice on P4 phylogenetic analysis, S Sharma and Y Huang for help with bioinformatics, C English for assistance with graphics, three anonymous reviewers for constructive comments and the Georgia Advanced Computing Resource Center at the University of Georgia for computational resources. This research was funded by grants from the Gordon and Betty Moore Foundation and the National Science Foundation (MCB-0702125).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Balbi KJ, Rocha EPC, Feil EJ. (2009). The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella spp. Mol Biol Evol* **26**: 345–355.

Bragg JG, Hyder CL. (2004). Nitrogen versus carbon use in prokaryotic genomes and proteomes. *P Roy Soc B-Biol Sci* **271**: S374–S377.

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

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Chen X, Zheng J, Fu Z, Nan P, Zhong Y, Lonardi S *et al.* (2005). Assignment of orthologous genes via genome rearrangement. *IEEE T Comp Biol Bioinform* **2**: 302–315.

Clark MA, Moran NA, Baumann P. (1999). Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol Biol Evol* **16**: 1586–1598.

Cleland D, Krader P, McCree C, Tang J, Emerson D. (2004). Glycine betaine as a cryoprotectant for prokaryotes. *J Microbiol Methods* **58**: 31–38.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. (2008). The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci USA* **105**: 20356–20361.

Dayhoff M, Eck R, Park C. (1972). A model of evolutionary change in proteins. In: Dayhoff M. (ed). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation: Silver Spring, MD, USA, pp 89–100.

Dufresne A, Garczarek L, Partensky F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* **6**: R14.

Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Alexander Richter R, Valas R *et al.* (2012). Genomic insights to SAR86 an abundant and uncultivated marine bacterial lineage. *ISME J* **6**: 1186–1199.

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

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol* **19**: 2142–2149.

Foster PG. (2004). Modeling compositional heterogeneity. *Syst Biol* **53**: 485–495.

Foster PG, Cox CJ, Embley TM. (2009). The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos Tr Roy Soc B* **364**: 2197–2207.

Foster PG, Hickey DA. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* **48**: 284–290.

Galtier N, Gouy M. (1995). Inferring phylogenies from DNA sequences of unequal base compositions. *Proc Natl Acad Sci USA* **92**: 11317–11321.

Ghai R, Rodriguez-Valera F, McMahon KD, Toyama D, Rinke R, Cristina Souza de Oliveira T *et al.* (2011). Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS One* **6**: e23785.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin K, Batista D *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242–1245.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ *et al.* (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio* **3**: e00252–00212.

Grzymski JJ, Dussaq AM. (2012). The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J* **6**: 71–80.

Gu X, Hewett-Emmett D, Li W-H. (1998). Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* **102/103**: 383–391.

Herbeck JT, Degnan PH, Wernegreen JJ. (2005). Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the Enterobacteriales. (γ-Proteobacteria). *Mol Biol Evol* **22**: 520–532.

Hershberg R, Petrov DA. (2010). Evidence that mutation Is universally biased towards AT in bacteria. *PLoS Genet* **6**: e1001115.

Hildebrand F, Meyer A, Eyre-Walker A. (2010). Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet* **6**: e1001107.

Hrdy I, Hirt RP, Dolezal P, Bardonova L, Foster PG, Tachezy J *et al.* (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **432**: 618–622.

Hu J, Blanchard JL. (2009). Environmental sequence data from the Sargasso Sea reveal that the characteristics of genome reduction in *Prochlorococcus* are not a harbinger for an escalation in genetic drift. *Mol Biol Evol* **26**: 5–13.

Hughes AL. (1999). *Adaptive evolution of genes and genomes*. Oxford University Press: NY, USA.

Hughes AL. (2008). Near neutrality: leading edge of the neutral theory of molecular evolution. *Ann NY Acad Sci* **1133**: 162–179.

Hughes AL. (2009). Relaxation of purifying selection on live attenuated vaccine strains of the family. *Paramyxoviridae. Vaccine* **27**: 1685–1690.

Hughes AL. (2013). Accumulation of slightly deleterious mutations in the mitochondrial genome: a hallmark of animal domestication. *Gene* **515**: 28–33.

Hughes AL, Friedman R. (2009). More radical amino acid replacements in primates than in rodents: support for the evolutionary role of effective population size. *Gene* **440**: 50–56.

Hughes AL, Ota T, Nei M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol Biol Evol* **7**: 515–524.

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol* **53**: 638–643.

Kass RE, Raftery AE. (1995). Bayes factors. *J Amer Stat Assoc* **90**: 773–795.

Katoh K, Kuma K-i, Toh H, Miyata T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**: 511–518.

Kumar S, Stecher G, Peterson D, Tamura K. (2012). MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **28**: 2685–2686.

Kuo C-H, Moran NA, Ochman H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19**: 1450–1454.

Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* **106**: 15527–15533.

Lewis P, Holder M, Holsinger K. (2005). Polytomies and Bayesian phylogenetic inference. *Syst Biol* **54**: 241–253.

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

Lind PA, Andersson DI. (2008). Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci USA* **105**: 17878–17883.

Löytynoja A, Goldman N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**: 1632–1635.

Luo H, Csürös M, Hughes AL, Moran MA. (2013). Evolution of divergent ecological strategies in marine Alphaproteobacteria. *mBio* **4**: mBio 00373–13.

Luo H, Hughes AL. (2012). dN/dS does not show positive selection drives separation of polar-tropical SAR11 populations. *Mol Syst Biol* **8**: 625.

Luo H, Friedman R, Tang J, Hughes AL. (2011a). Genome reduction by deletion of paralogs in the marine cyanobacterium. *Prochlorococcus. Mol Biol Evol* **28**: 2751–2760.

Luo H, Löytynoja A, Moran MA. (2012). Genome content of uncultivated marine *Roseobacters* in the surface ocean. *Environ Microbiol* **14**: 41–51.

Luo H, Tang J, Friedman R, Hughes AL. (2011b). Ongoing purifying selection on intergenic spacers in group A streptococcus. *Infect Genet Evol* **11**: 343–348.

Martin-Cuadrado A-B, Rodriguez-Valera F, Moreira D, Alba JC, Ivars-Martinez E, Henn MR *et al.* (2008). Hindsight in the relative abundance metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J* **2**: 865–886.

McCutcheon JP, Moran NA. (2012). Extreme genome reduction in symbiotic bacteria. *Nat Rev Micro* **10**: 13–26.

Miyata T, Miyazawa S, Yasunaga T. (1979). Two types of amino acid substitutions in protein evolution. *J Mol Evol* **12**: 219–236.

Moran MA, Belas R, Schell MA, Gonzalez JM, Sun F, Sun S *et al.* (2007). Ecological genomics of marine Roseobacters. *Appl Environ Microbiol* **73**: 4559–4569.

Moran MA, Buchan A, Gonzalez JM, Heidelberg JF, Whitman WB, Kiene RP *et al.* (2004). Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432**: 910–913.

Morris JJ, Lenski RE, Zinser ER. (2012). The black queen hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**: e00036–00012.

Nei M. (1987). *Molecular evolutionary genetics*. Columbia University Press: NY, USA.

Nei M, Suzuki Y, Nozawa M. (2010). The neutral theory of molecular evolution in the genomic era. *Ann Rev Genom Hum G* **11**: 265–289.

Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B. (2010). Compositional heterogeneity and phylogenomic inference of metazoan relationships. *Mol Biol Evol* **27**: 2095–2104.

Newton MA, Raftery AE. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *J R Stat Soc* **56**: 3–56.

Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE *et al.* (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME J* **4**: 784–798.

Notredame C, Higgins D, Heringa J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205–217.

Ochman H, Elwyn S, Moran NA. (1999). Calibrating bacterial evolution. *Proc Natl Acad Sci USA* **96**: 12638–12643.

Ohta T. (1992). The nearly neutral theory of molecular evolution. *Ann Rev Ecol Syst* **23**: 263–286.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. (2009). NCBI Reference Sequences: current status policy and new initiatives. *Nucleic Acids Res* **37**: D32–D36.

Raghavan R, Kelkar YD, Ochman H. (2012). A selective force favoring increased G+C content in bacterial genes. *Proc Natl Acad Sci USA* **109**: 14504–14507.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.

Rodriguez-Brito B, Rohwer F, Edwards RA. (2006). An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**: 162.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Singer GAC, Hickey DA. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* **17**: 1581–1588.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of marine bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**: 11463–11468.

Tatusov RL, Koonin EV, Lipman DJ. (1997). A genomic perspective on protein families. *Science* **278**: 631–637.

Wernegreen JJ. (2011). Reduced selective constraint in endosymbionts: elevation in radical amino acid replacements occurs genome-wide. *PLoS One* **6**: e28905.

Wernegreen JJ, Moran NA. (1999). Evidence for genetic drift in endosymbionts. (*Buchnera*): analyses of protein-coding genes. *Mol Biol Evol* **16**: 83–97.

Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H *et al.* (2009). Assembling the marine metagenome, one cell at a time. *PLoS One* **4**: e5299.

Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M *et al.* (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**: 60–66.

Zhang J. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J Mol Evol* **50**: 56–68.

Zhao Y, Tang H, Ye Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**: 125–126.

Zuckerkandl E, Pauling L. (1965). Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel H. (eds). *Evolving Genes and Proteins*. Academic Press: NY, USA, pp 97–116.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)