Strong Purifying Selection Is Associated with Genome Streamlining in Epipelagic *Marinimicrobia*

Carolina Alejandra Martinez-Gutierrez and Frank O. Aylward *

Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia

*Corresponding author: E-mail: faylward@vt.edu.

Accepted: September 8, 2019

Abstract

Marine microorganisms inhabiting nutrient-depleted waters play critical roles in global biogeochemical cycles due to their abundance and broad distribution. Many of these microbes share similar genomic features including small genome size, low % G + C content, short intergenic regions, and low nitrogen content in encoded amino acid residue side chains (N-ARSC), but the evolutionary drivers of these characteristics are unclear. Here, we compared the strength of purifying selection across the *Marinimicrobia*, a candidate phylum which encompasses a broad range of phylogenetic groups with disparate genomic features, by estimating the ratio of nonsynonymous and synonymous substitutions (d*N*/d*S*) in conserved marker genes. Our analysis reveals that epipelagic *Marinimicrobia* that exhibit features consistent with genome streamlining have significantly lower d*N*/d*S* values when compared with their mesopelagic counterparts. We also found a significant positive correlation between median d*N*/d*S* values and % G + C content, N-ARSC, and intergenic region length. We did not identify a significant correlation between d*N*/d*S* ratios and estimated genome size, suggesting the strength of selection is not a primary factor shaping genome size in this group. Our findings are generally consistent with genome streamlining theory, which postulates that many genomic features of abundant epipelagic bacteria are the result of adaptation to oligotrophic nutrient conditions. Our results are also in agreement with previous findings that genome streamlining is common in epipelagic waters, suggesting that microbes inhabiting this region of the ocean have been shaped by strong selection together with prevalent nutritional constraints characteristic of this environment.

Key words: genome streamlining, purifying selection, evolutionary genomics, d*N*/d*S* ratio, *Marinimicrobia*, uncultured picoplankton.

Bacteria and Archaea play key roles in marine biogeochemical cycles and are a dominant force that drives global nutrient transformations (Azam et al. 1983; Falkowski et al. 2008). Our understanding of microbial diversity in the ocean has been transformed in the last few decades due to the discovery of several marine microbial lineages that are among the most numerically abundant life forms on Earth (Giovannoni and Stingl 2005). Work on some of these abundant lineages succeeded in culturing representatives that could then be studied extensively in the laboratory, such as *Prochlorococcus marinus* (Chisholm et al. 1992) and heterotrophic bacterioplankton belonging to the *Pelagibacteriales* (Rappé et al. 2002), and *Roseobacter* groups (Luo and Moran 2014), but many other dominant microbial lineages have not been brought into pure culture and require cultivation-independent methods for analysis (DeLong and Karl 2005).

Previous research of *Prochlorococcus marinus* and *Pelagibacter ubique* genomes provided some of the earliest insights into the ecology and evolution of these dominant planktonic microbial lineages (Rocap et al. 2003; Giovannoni et al. 2005). It was quickly noted that both groups had small genomes that contained short intergenic regions and encoded among the fewest genes of any free-living organism (Giovannoni et al. 2005). These characteristics were explained through the proposed theory of genome streamlining, which states that genome simplification is an adaptation to consistently oligotrophic conditions, and that the loss of unnecessary genes and their corresponding transcriptional, translational, and regulatory burdens is advantageous (Giovannoni et al. 2014). Genome streamlining theory is supported by the observation that many streamlined genomes also have lower % GC content and subsequently contain fewer codons encoding nitrogen-rich amino acids (Grzymski and Dussaq 2012; Mende et al. 2017), which is expected to be advantageous in nutrient-depleted conditions found in the open ocean (Giovannoni et al. 2014). Genome streamlining

therefore corresponds to multiple characteristics, and recent cultivation-independent studies have confirmed that many of them are present in the genomes of a variety of marine lineages in addition to *Prochlorococcus* and *Pelagibacter* (Dupont et al. 2012; Ghai et al. 2013; Swan et al. 2013; Luo et al. 2014; Getz et al. 2018), suggesting that common evolutionary drivers shape diverse bacterioplankton groups in the ocean.

Although the term "genome streamlining" implies adaptation under oligotrophic nutrient conditions, it remains a possibility that these genomic signatures are nonadaptive or potentially the result of genetic drift (Batut et al. 2014). For example, it has long been known that many endosymbiotic bacteria contain small genomes with short intergenic regions and low % GC content, but in these cases a small effective population size (*Ne*) and correspondingly high genetic drift are likely responsible for these features (Charlesworth 2009; Kuo et al. 2009). Although it remains unlikely that marine free-living bacteria have small effective population sizes comparable to those of endosymbiotic bacteria, it has been argued that population bottlenecks in the distant evolutionary past of some marine lineages may be responsible for aspects of their present genomic architecture (Luo et al. 2017). Moreover, recent work has also shown that weakly deleterious mutations and low recombination rates can substantially lower the efficacy of purifying selection in bacterial genomes (Price and Arkin 2015), implying that the large abundances of marine bacteria may not translate directly into high selection.

In this study, we focused our analyses on the candidate phylum *Marinimicrobia*, a predominantly marine group that comprises diverse globally abundant lineages involved in distinct biogeochemical processes (Hawley et al. 2017; Getz et al. 2018). Formerly referred to as clade SAR406 or Marine Group A, the *Marinimicrobia* span a broad range of distinct marine lineages that are poorly understood, in part due to difficulties in cultivating representatives of this phylum. Advances in metagenomics and single-cell sequencing have yielded a large number of draft genomes from this group, however, and several recent studies have shed light on the important role of different *Marinimicrobia* lineages to carbon and nitrogen cycling in the ocean (Wright et al. 2014; Aylward et al. 2015; Zhang et al. 2016; Bertagnolli et al. 2017; Thrash et al. 2017; Plominsky et al. 2018). In a recent study, we compiled a set of draft *Marinimicrobia* genomes that have been sequenced using cultivation-independent methods (Getz et al. 2018), and we leverage this set here to analyze the evolutionary genomics of this group. The 211 genomes we used here belong to *Marinimicrobia* that inhabit both epipelagic and mesopelagic waters across the global ocean and comprise a broad range of phylogenetic diversity (38% average amino acid identity among marker genes in the CheckM marker set; see Materials and Methods).

The *Marinimicrobia* are an ideal group to test genome streamlining theory because streamlined genomic traits have evolved multiple times independently in this phylum (Getz et al. 2018). Moreover, streamlined genomic characteristics are linked to the environment in which *Marinimicrobia* are found; epipelagic *Marinimicrobia* tend to have genomes with low % GC content, short intergenic spacers, and relatively low nitrogen and high carbon-encoded amino acids, while mesopelagic *Marinimicrobia* generally lack these features (fig. 1). Higher levels of purifying selection in epipelagic *Marinimicrobia* would therefore be consistent with genome streamlining theory, because it would indicate that these genomic features are not due to genetic drift. Conversely, lower levels of purifying selection in epipelagic *Marinimicrobia* would suggest that their genomes are shaped by a process analogous to that experienced by endosymbiotic bacteria.

In order to test our hypothesis, we estimated the ratio of nonsynonymous and synonymous substitutions (d*N*/d*S*) of conserved marker genes in *Marinimicrobia*. In general, d*N*/d*S* values <1 are indicative of purifying selection, and the relative strength of selection can be compared across groups using this metric, with lower values implying higher levels of purifying selection (Kryazhimskiy and Plotkin 2008; Kuo et al. 2009). To ensure that our results could be accurately compared across divergent clades, we used two sets of marker genes that are broadly shared among Bacteria, which we refer to here as the EMBL (Sunagawa et al. 2013) and CheckM (Parks et al. 2015) gene sets. We observed a general trend in which epipelagic genomes exhibited lower median d*N*/d*S* values (fig. 2 and supplementary data set S1, Supplementary Material online). Although less pronounced, we also observed higher median d*S* values in epipelagic *Marinimicrobia* (supplementary fig. S1, Supplementary Material online). The d*N*/d*S* values we obtained are far lower than one, which is consistent with the expectation that conserved phylogenetic marker genes experience purifying selection in order to maintain protein function. Our observation of lower median d*N*/d*S* values in epipelagic *Marinimicrobia* was strongly supported by statistical analyses of both the CheckM and the EMBL marker gene sets (Mann–Whitney *U* test, $P < 0.005$ in both cases; fig. 2). Our findings suggest that *Marinimicrobia* found in epipelagic waters experience higher levels of purifying selection than those inhabiting mesopelagic waters.

It has been shown that d*N*/d*S* values are dependent on the time scale in which comparisons are performed (Rocha et al. 2006; Balbi et al. 2009). To test if our results were consistent across different time scales, we created two sets of d*N*/d*S* values based on their corresponding d*S* values, which is a reflection of sequence divergence; one set corresponded to d*S* values greater than the mean (more divergent comparisons) while the other set corresponded to those lower than the mean (less divergent comparisons). We compared epipelagic versus mesopelagic d*N*/d*S* values for both marker sets, and found that epipelagic *Marinimicrobia* had lower median d*N*/d*S* values in all cases ($P < 0.005$), indicating that the time

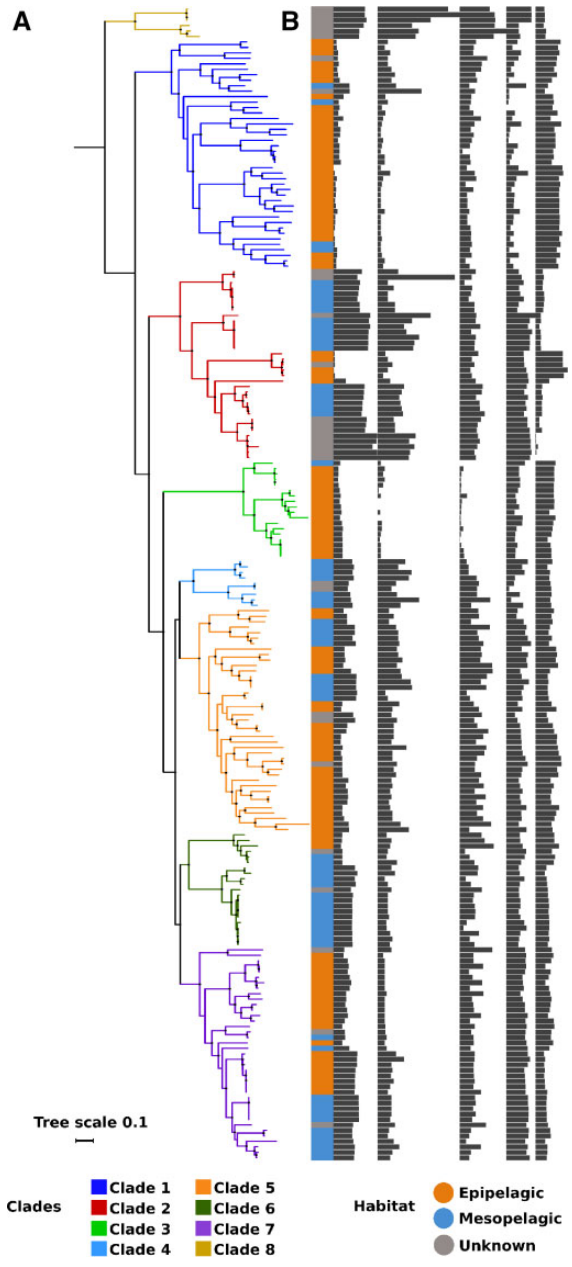

Fig. 1.—Representation of phylogeny, habitat classification, and genomic features of Marinimicrobia. (*a*) Maximum likelihood phylogenetic tree of the 211 genomes constructed using amino acid sequences of 120 highly conserved marker genes. (*b*) Habitat classification based on Getz et al. (2018) and genomic features of *Marinimicrobia* genomes. Abbreviations: H, Habitat; GC, % GC content (range, 27–55%); IGR, median intergenic region length (range, 7–78 nucleotides); EGS, estimated genome size (range 1–4.4 Mb); N-ARSC (range, 0.3–0.34); C-ARSC (range, 3–3.2). Black points on branches represent support values >0.95.

dependence of d*N*/d*S* values is not responsible for our findings.

We also explored the correlation between the strength of selection and several genomic features associated with streamlining. For this purpose, we generated several habitat-specific clusters of closely related *Marinimicrobia* and plotted their median d*N*/d*S* ratio against average genomic characteristics within that cluster (see Materials and Methods). In general, we found that features consistent with genome streamlining were correlated with low d*N*/d*S* values (fig. 3 and supplementary fig. S2, Supplementary Material online), with the strongest correlations observed for % GC content ($\rho = 0.71$, fig. 3*a*), nitrogen content in amino acid residue side chains (N-ARSC; $\rho = 0.54$, fig. 3*b*), and median intergenic region length ($\rho = 0.68$, fig. 3*e*). The low N-ARSC values in epipelagic *Marinimicrobia* are consistent with previous findings, and are likely a product of nutrient limitation in surface waters (Getz et al. 2018). Similarly, we identified a negative correlation between d*N*/d*S* values and the carbon content of amino acid residue side chains (C-ARSC; $\rho = -0.59$, fig. 3*d*), which is also consistent with higher carbon availability in epipelagic waters. The weakest correlation we observed was between d*N*/d*S* values and estimated genome size ($\rho = 0.24$, fig. 3*e*). To confirm these trends, we also performed a multivariate analysis of the d*N*/d*S* values and genomic features of the *Marinimicrobia* genome clusters, and the results of this analysis confirmed the tendency of streamlined epipelagic genomes to have lower d*N*/d*S* values (fig. 4 and supplementary fig. S3, Supplementary Material online).

Several previous studies compared a broad array of bacterial lineages and found that d*N*/d*S* ratios are generally higher in smaller genomes, suggesting that genetic drift is a prominent evolutionary force in genome reduction (Kuo et al. 2009; Novichkov, et al. 2009b; Sela et al. 2016). We did not identify a strong relationship between estimated genome size and d*N*/d*S* ratios in the *Marinimicrobia*, suggesting that this trend may not hold for this group. Many abundant marine lineages remain poorly represented in sequenced genome repositories due to difficulties in cultivation, and the evolutionary factors shaping their genomes are therefore relatively unexplored. There is evidence of widespread genome streamlining in abundant marine lineages in the ocean (Swan et al. 2013), and as more genomes become available it will be possible to rigorously evaluate the strength of purifying selection on these groups and its possible impact on genome size. The theory of genome streamlining predicts that streamlining may occur across a range of genome sizes due to different genetic repertoires that are necessary to thrive in different environments or ecological niches (Giovannoni et al. 2014), and it is therefore unclear if we would expect streamlined epipelagic *Marinimicrobia* to have substantially smaller genomes overall. An additional complication of the present study is that the genomes analyzed are incomplete owing to their sequencing via metagenomic or single-cell sequencing efforts, and we extrapolated genome sizes from completeness estimates. The lack of a significant correlation between d*N*/d*S* values and genome size must therefore be interpreted

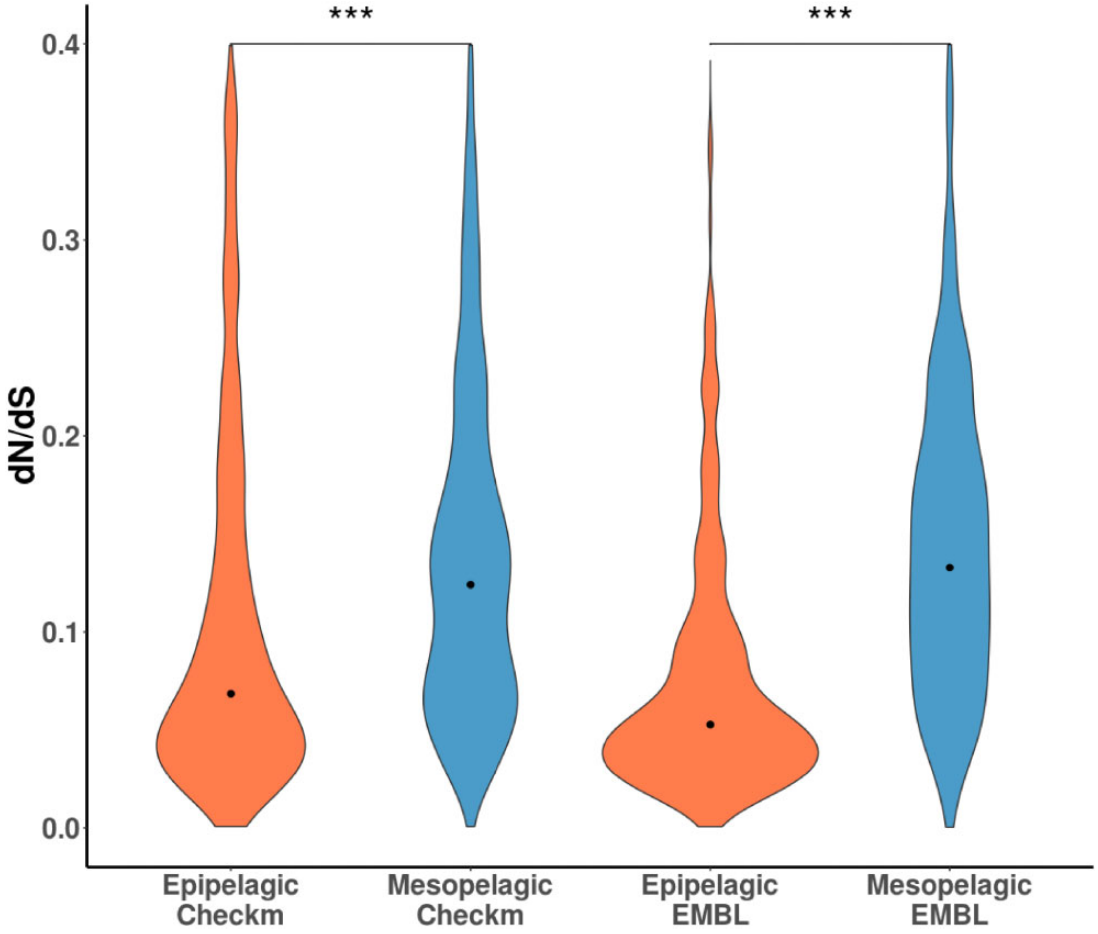

Fig. 2.—Violin plot representing median d*N*/d*S* values of epipelagic and mesopelagic *Marinimicrobia*. Statistical significance of differences between d*N*/d*S* values of the compared groups according to a nonpaired, one-sided Mann–Whitney–Wilcoxon test is denoted by: (***) for $P < 0.005$.

with caution, and further studies will be needed to examine this in more detail.

It is important to note that the results we present here do not imply that strong selection directly leads to low % GC content, low N-ARSC, or other streamlined features, since the genomic changes that result from strong selection depend on prevailing environmental factors. For example, strong selection on genomes in mesopelagic waters would not be predicted to lead to a decrease in the N-ARSC of encoded proteins, since nitrogen is more abundant in deeper waters and this evolutionary transition would not be advantageous. The disparate genomic features of epipelagic and mesopelagic *Marinimicrobia* are therefore likely the result of differential nutrient availability and environmental factors along the water column; surface waters are relatively depleted in nitrogen and phosphorus, while mesopelagic waters contain more of these nutrients but less photosynthetically derived carbon (Karl 2002; Moore et al. 2013). Other factors in addition to selection under different environmental conditions may also play a role in genome streamlining; for example, it has been hypothesized that an increase in mutation rate may lead to some of the genomic features of both endosymbiotic bacteria and abundant marine bacteria (Marais et al. 2008). In our phylogeny of the *Marinimicrobia*, genomes that contain streamlined genomic features in Clades 2 and 3 (red and light green in fig. 1, respectively) are associated with long branches that may be indicative of increased mutation rates. This link between long branches and genome streamlining must be made with caution, however, because long branches do not definitively demonstrate increased mutation rates, and even so it is unclear if this would lead to other genomic features of streamlining. Nevertheless, given the complexity of these genome evolutionary processes, it is likely that multiple factors are responsible for the trends we observe here.

An important caveat of the d*N*/d*S* ratio is that it only provides insight into the strength of recent selective pressure and therefore cannot be used to infer the selective strength experienced by lineages in the past. Other streamlined lineages such as the *Pelagibacterales* and *Prochlorococcus* are thought to have underwent genome reduction in the distant past, and it is therefore difficult to assess the strength of selection on these ancestral genomes during these transitions. Some

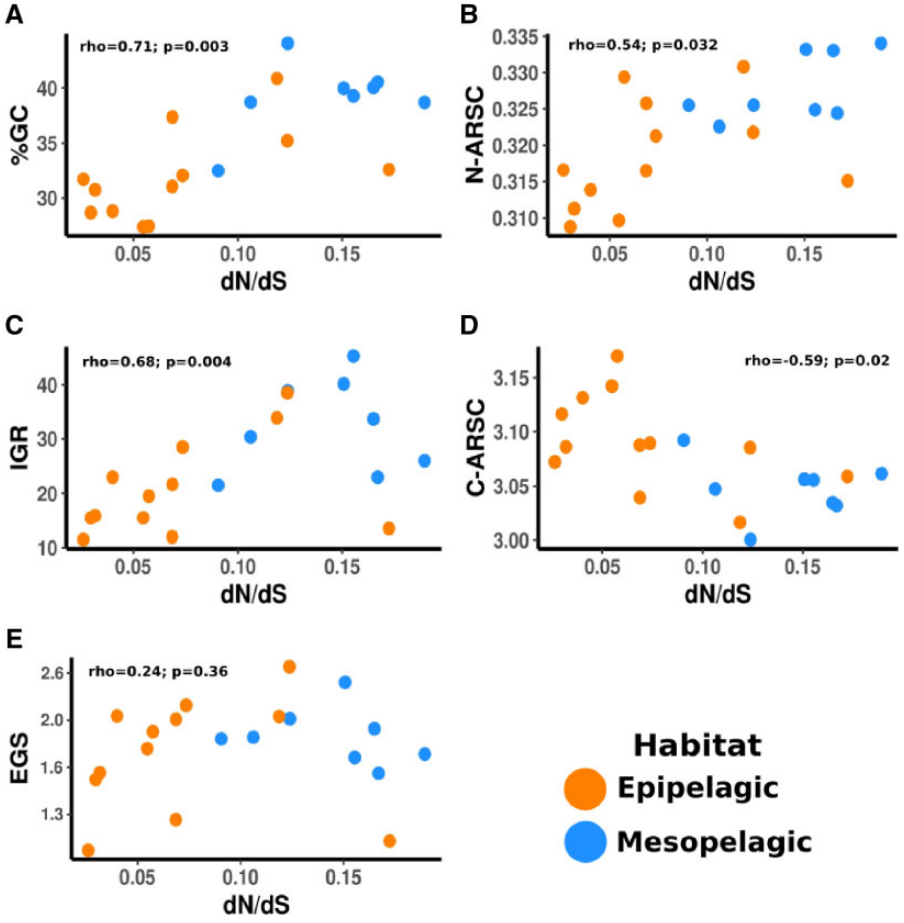

Fig. 3.—Scatter plots showing the relationship between median d*N*/d*S* values and streamlined genomic features of the *Marinimicrobia* genome clusters. Median d*N*/d*S* values were calculated using the CheckM marker gene set. (*a*) GC content versus d*N*/d*S*; (*b*) N-ARSC versus d*N*/d*S*; (*c*) Median intergenic regions length (bp) versus d*N*/d*S*; (*d*) C-ARSC versus d*N*/d*S*; (*e*) estimated genome size (log bp) versus d*N*/d*S*. Spearman correlations were performed for each variables pair and details can be found on the main text. Details for the genome clusters can be found in supplementary data set S2, Supplementary Material online.

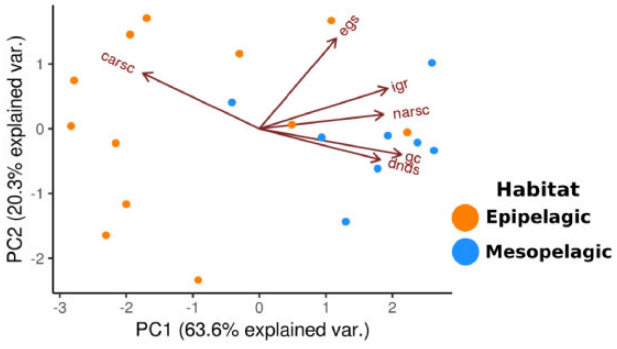

Fig. 4.—PCA analysis displaying the Euclidean distance among *Marinimicrobia* genomes. Abbreviations: gc, %GC content; narsc: N-ARSC; igr, intergenic regions length; d*N*d*S*, d*N*/d*S* ratio; egs, estimated genome size; CARSC, C-ARSC.

studies have suggested that genetic drift due to possible population bottlenecks drove these genomic changes (Luo et al. 2017), while other studies have argued that strong purifying selection was the primary driver (Sun and Blanchard 2014). In contrast to these other streamlined groups, the *Marinimicrobia* appear to have experienced multiple independent genome transition events relatively recently in their evolutionary history (Getz et al. 2018), and comparison of the selective pressures across disparate clades with similar genomic features therefore provides insight into more recent selective regimes that led to current genomic architectures. Overall our results suggest that although parasitic and endosymbiotic bacteria share some genomic features with streamlined bacteria, these features are the product of distinct evolutionary paths.

Materials and Methods

Marinimicrobia Genomes Used

We analyzed a set of 211 *Marinimicrobia* genomes derived from a previous study (Getz et al. 2018). This data set included genomes from GenBank (Sayers et al. 2019), the Integrated

Microbial Genomes database (IMG; Markowitz and Kyrpides 2007), and from two different studies in which Metagenome-Assembled Genomes (MAGs) were generated (Delmont et al. 2018; Tully et al. 2018). The data set employed by Getz et al. was complemented with the genomes SCGC_AD-604-D17, SCGC_AD-606-A07, SCGC_AD-615_E22 from another recent study (Plominsky et al. 2018). Methods for quality filtering, estimation of genome completeness and contamination, and the calculation of genomic features have been described previously (Getz et al. 2018).

Phylogenetic Reconstruction

To reconstruct the *Marinimicrobia* phylogeny, we predicted proteins from genomes using Prodigal v2.6.2 (Hyatt et al. 2010) and identified phylogenetic marker genes using HMMER3 (Eddy 2011). We constructed a phylogeny from an amino acid alignment created from the concatenation of 120 marker genes that have been previously used for phylogenetic reconstructions of Bacteria (Parks et al. 2015). The trusted cutoffs were used in all HMMER3 searches with the "cut_tc" option in hmmsearch. We used the standard_fasttree workflow included in the ETE Toolkit which includes ClustalOmega for alignment (Sievers and Higgins 2018), trimAl for alignment trimming (Capella-Gutierrez et al. 2009), and FastTree for phylogenetic estimation (Price et al. 2010). The different branches obtained were classified into clades based on previously published results (Getz et al. 2018). We visualized the resulting tree in the interactive Tree of Life (iTOL; Letunic and Bork 2016; https://itol.embl.de/tree/4537914239725156208868 3).

d*N*/d*S* Ratio Calculation and Filtering

To estimate the strength of purifying selection, we used the ratio of nonsynonymous and synonymous substitutions (d*N*/d*S*). When considering values <1, lower values are a sign of higher purifying selection while higher values are a sign of higher genetic drift (low purifying selection). To calculate genome-wide d*N*/d*S* ratios, we used two sets of conserved marker genes that would be expected to be found in most genomes. The first one consists of 120 phylogenetic marker genes that are highly conserved in Bacteria, which we also used for phylogenetic reconstruction (Parks et al. 2015). The second set consists of 40 phylogenetic marker genes used in phylogenetic reconstructions, which we refer to as the EMBL set due to its development in the European Molecular Biology Laboratory (Sunagawa et al. 2013).

For both marker gene sets, we predicted proteins from each genome using Prodigal and then annotated the marker genes of interest using the hmmsearch tool of HMMER3 with model-specific cutoffs. We aligned the amino acid sequences for each annotated gene coming from *Marinimicrobia* genomes separately using ClustalOmega, and the resulting alignments converted into codon alignments using PAL2NAL (Suyama et al. 2006). Maximum-likelihood approximation (codeML) within the PAML 4.9 h package (Yang 2007) was used through Biopython in order to perform d*N*/d*S* pairwise comparisons within the clades previously established (Getz et al. 2018). We removed d*N*/d*S* values with d*S* ≥ 1, which implies that synonymous substitutions are near saturation. Moreover, to avoid comparing sequences from genomes that may be part of the same population, we also excluded comparisons for which d*N* $= 0$ and d*S* ≤ 0.01. Additionally, we discarded all d*N*/d*S* values ≥ 10 on the grounds that these were largely artifactual. Lastly, because we wished to compare d*N*/d*S* values from *Marinimicrobia* that reside in different habitats, we only included d*N*/d*S* values where the pair of compared genomes were from the same habitat (epipelagic and mesopelagic). All values used can be found in supplementary data set S1, Supplementary Material online.

Genome Clustering

To compare d*N*/d*S* values with other genomic features, it was first necessary to generate clusters of closely related genomes. For this, *Marinimicrobia* genomes were compared using the MASH program (Ondov et al. 2016), which rapidly identifies similarities in the k-mer profiles of genomes and provides statistical measures of nucleotide similarity. Comparisons that yielded MASH e-values $<$1e-100 were retained and used to link closely related genomes, and final genome clusters were generated using a single-linkage clustering algorithm in R. Median d*N*/d*S* values for all clusters were calculated and then plotted against average genome features within that cluster (% GC content, estimated genome size, median intergenic region length, estimated genome size, N-ARSC, and C-ARSC; see figs. 3 and 4). Clusters that had fewer than 10 total d*N*/d*S* measurements were excluded from further analyses. Details for the genome clusters can be found in supplementary data set S2, Supplementary Material online. This approach of using clusters of related genomes to estimate group-specific d*N*/d*S* values is similar to previously used methods (Kuo et al. 2009; Novichkov et al. 2009a).

Statistical Analyses

In order to investigate the strength of selection acting on epipelagic and mesopelagic *Marinimicrobia*, genomes were classified into epipelagic and mesopelagic based on their biogeographic distribution (Getz et al. 2018). For statistical analysis, we loaded the filtered d*N*/d*S* values into R and performed comparisons using the Mann–Whitney *U* test (wilcox.test() function). Additionally, to investigate the relationship between median d*N*/d*S* and genomic features associated to each cluster, we applied the "cor.test" function using the Spearman method. Comparisons and correlation plots were visualized through the ggplot2 package (Wilkinson 2011). We also explored the distance between epipelagic and

mesopelagic *Marinimicrobia* genomes employing the genomic features and median d*N*/d*S* values through a PCA analysis with the "prcomp" function available on R. Euclidean distance was visualized using the "ggbiplot" function within the ggplot2 package (Wilkinson 2011).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. This work was supported by grants from the Institute for Critical Technology and Applied Science and the NSF (IIBR-1918271), a Sloan Research Fellowship in Ocean Sciences, and a Simons Early Career Award in Marine Microbial Ecology and Evolution to F.O.A.

Literature Cited

Aylward FO, et al. 2015. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. Proc Natl Acad Sci U S A. 112(17):5443–5448.
Azam F, et al. 1983. The ecological role of water-column microbes in the sea. Mar Ecol Prog Ser. 10:257–263.
Balbi KJ, Rocha EPC, Feil EJ. 2009. The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella* spp. Mol Biol Evol. 26(2):345–355.
Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol. 12(12):841–850.
Bertagnolli AD, Padilla CC, Glass JB, Thamdrup B, Stewart FJ. 2017. Metabolic potential and in situ activity of marine *Marinimicrobia* bacteria in an anoxic water column. Environ Microbiol. 19(11):4392–4416.
Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.
Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10(3):195–205.
Chisholm SW, et al. 1992. *Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll A and B. Arch Microbiol. 157(3):297–300.
Delmont TO, et al. 2018. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol. 3(7):804–813.
DeLong EF, Karl DM. 2005. Genomic perspectives in microbial oceanography. Nature 437(7057):336–342.
Dupont CL, et al. 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. 6(6):1186–1199.
Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol. 7(10):e1002195.
Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320(5879):1034–1039.
Getz EW, Tithi SS, Zhang L, Aylward FO. 2018. Parallel evolution of genome streamlining and cellular bioenergetics across the marine radiation of a bacterial phylum. MBio 9(5):9.
Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. 2013. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Sci Rep. 3:2471.
Giovannoni SJ, et al. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309(5738):1242–1245.
Giovannoni SJ, Stingl U. 2005. Molecular diversity and ecology of microbial plankton. Nature 437(7057):343–348.
Giovannoni SJ, Thrash CJ, Temperton B. 2014. Implications of streamlining theory for microbial ecology. ISME J. 8(8):1553–1565.
Grzymski JJ, Dussaq AM. 2012. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. 6(1):71–80.
Hawley AK, et al. 2017. Diverse *Marinimicrobia* bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. Nat Commun. 8(1):1507.
Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11(1):119.
Karl DM. 2002. Nutrient dynamics in the deep blue sea. Trends Microbiol. 10(9):410–418.
Kryazhimskiy S, Plotkin JB. 2008. The population genetics of d*N*/d*S*. PLoS Genet. 4(12):e1000304.
Kuo C-H, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. 19(8):1450–1454.
Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44(W1):W242–W245.
Luo H, Huang Y, Stepanauskas R, Tang J. 2017. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nat Microbiol. 2:17091.
Luo H, Moran MA. 2014. Evolutionary ecology of the marine *Roseobacter* clade. Microbiol Mol Biol Rev. 78(4):573–587.
Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. 2014. Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters. ISME J. 8(7):1428–1439.
Marais GAB, Calteau A, Tenaillon O. 2008. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica 134(2):205–210.
Markowitz VM, Kyrpides NC. 2007. Comparative genome analysis in the integrated microbial genomes (IMG) system. Methods Mol Biol. 395:35–56.
Mende DR, et al. 2017. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol. 2(10):1367–1373.
Moore CM, et al. 2013. Processes and patterns of oceanic nutrient limitation. Nat Geosci. 6(9):701–710.
Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I. 2009a. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. Nucleic Acids Res. 37(Database):D448–D454.
Novichkov PS, Wolf YI, Dubchak I, Koonin EV. 2009b. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol. 191(1):65–73.
Ondov BD, et al. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17(1):132.
Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25(7):1043–1055.
Plominsky AM, et al. 2018. Metabolic potential and in situ transcriptomic profiles of previously uncharacterized key microbial groups involved in coupled carbon, nitrogen and sulfur cycling in anoxic marine zones. Environ Microbiol. 20(8):2727–2742.
Price MN, Arkin AP. 2015. Weakly deleterious mutations and low rates of recombination limit the impact of natural selection on bacterial genomes. MBio 6(6):e01302–e01315.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 5(3):e9490.
Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418(6898):630–633.
Rocap G, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424(6952):1042–1047.
Rocha EPC, et al. 2006. Comparisons of d*N*/d*S* are time dependent for closely related bacterial genomes. J Theor Biol. 239(2):226–235.
Sayers EW, et al. 2019. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 47(D1):D23–D28.
Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A. 113(41):11399–11407.
Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 27(1):135–145.
Sun Z, Blanchard JL. 2014. Strong genome-wide selection early in the evolution of *Prochlorococcus* resulted in a reduced genome through the loss of a large number of small effect genes. PLoS One 9(3):e88837.
Sunagawa S, et al. 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 10(12):1196–1199.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34(Web Server issue):W609–W612.
Swan BK, et al. 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc Natl Acad Sci U S A. 110(28):11463–11468.
Thrash JC, et al. 2017. Metabolic roles of uncultivated bacterioplankton lineages in the Northern Gulf of Mexico 'Dead Zone'. mBio 8:e01017–17.
Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 5:170203.
Wilkinson L. 2011. ggplot2: elegant graphics for data analysis by Wickham, H. Biometrics 67(2):678–679.
Wright JJ, et al. 2014. Genomic properties of marine group A bacteria indicate a role in the marine sulfur cycle. ISME J. 8(2):455–468.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Zhang W, et al. 2016. Genomic and transcriptomic evidence for carbohydrate consumption among microorganisms in a cold seep brine pool. Front Microbiol. 7:1825.

Associate editor: Emmanuelle Lerat