Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype

J. Cameron Thrash[1,2,*], Ben Temperton[1], Brandon K. Swan[3], Zachary C. Landry[1], Tanja Woyke[4], Edward F. DeLong[5], Ramunas Stepanauskas[3] and Stephan J. Giovannoni[1]

1. Department of Microbiology, Oregon State University, Corvallis, OR 97331
2. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, 70803
3. Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544
4. DOE Joint Genome Institute, Walnut Creek, CA 94598
5. Departments of Civil & Environmental Engineering and Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

*To whom correspondence should be addressed: thrashc@lsu.edu

Abstract

Bacterioplankton of the SAR11 clade are the most abundant microorganisms in marine systems, usually representing 25% or more of the total microbial cells in seawater worldwide. SAR11 is divided into subclades with distinct spatiotemporal distributions (ecotypes), some of which appear to be specific to deep water. Here we examine the genomic basis for deep ocean distribution of one SAR11 bathytype (depth-specific ecotype), subclade Ic. Four single-cell Ic genomes, with estimated completeness of 58-91%, were isolated from 770 m at station ALOHA and compared with eight SAR11 surface genomes and metagenomic datasets. Subclade Ic genomes dominated metagenomic fragment recruitment below the euphotic zone. They had similar COG distributions, high local synteny, and shared a large number (69%) of orthologous clusters with SAR11 surface genomes, yet were distinct at the 16S rRNA gene and amino acid level, and formed a separate, monophyletic group in phylogenetic trees. Subclade Ic genomes were enriched in genes associated with membrane/cell-wall/envelope biosynthesis and showed evidence of unique phage defenses. The majority of subclade Ic-specfic genes were hypothetical, and some were highly abundant in deep ocean metagenomic data, potentially masking mechanisms for niche differentiation. However, the evidence suggests these organisms have a similar metabolism to their surface counterparts, and that subclade Ic adaptations to the deep ocean do not involve large variations in gene content, but rather more subtle differences previously observed deep ocean genomic data, like preferential amino acid substitutions, larger coding regions among SAR11 clade orthologs, larger intergenic regions, and larger estimated average genome size.

Keywords: bathytype/ecotype/metagenomics/SAR11/single-cell genomics/deep ocean

Introduction

Characterized by darkness, average temperatures of ~2-4˚C, increased hydrostatic pressure, and general oligotrophy, the relatively extreme environment of the deep ocean is also, ironically, the largest biome on Earth. The mesopelagic (200-1000 m) and bathypelagic (1000-4000 m) zones contain > 70% of marine microbial biomass (Arıstegui *et al.*, 2009) and these organisms play vital roles in global cycling of carbon, nitrogen, and other biogeochemical processes (Nagata *et al.*, 2010, Robinson *et al.*, 2010). In addition to microorganisms necessarily being adapted to cold and increased pressure there, the deep sea also contains more recalcitrant forms of carbon than at the surface (Arıstegui *et al.*, 2009, Nagata *et al.*, 2010, Robinson *et al.*, 2010). Cultivated isolates have revealed some microbial adaptations associated with life at depth, including increased intergenic spacer regions, rRNA gene indels, and higher abundances of membrane polyunsaturated fatty acids and surface-adhesion/motility genes (Lauro and Bartlett, 2008, Nagata *et al.*, 2010, Simonato *et al.*, 2006, Wang *et al.*, 2008).

However, many of the most abundant bacterial groups from the deep ocean remain uncultivated, for example the SAR202, SAR324, and SAR406 clades, which make up significant fractions of microbial communities at depth (DeLong *et al.*, 2006, Giovannoni *et al.*, 1996, Gordon and Giovannoni, 1996, Morris *et al.*, 2006, Morris *et al.*, 2012, Schattenhofer *et al.*, 2009, Treusch *et al.*, 2009, Varela *et al.*, 2008, Wright *et al.*, 1997). Thus, it remains uncertain how widespread the known adaptations of cultivated isolates are among deep ocean microorganisms. Metagenomic analyses have provided evidence for common genomic features in the deep ocean, such as increased proliferation of transposable elements and phage, amino acid content changes, and increased average genome size (DeLong *et al.*, 2006, Konstantinidis *et al.*, 2009). Single-cell genomic analyses provide another powerful means to understand the metabolism and evolution of organisms eluding cultivation-based techniques (Blainey, 2013, Lasken, 2013, Rinke *et al.*, 2013, Stepanauskas, 2012). This approach provided the first insight into the metabolism of several of these deep ocean clades, including SAR324, Arctic96BD-19,

and Agg47, and made the important discovery that at least some of these organisms are capable of chemoautotrophy (Swan *et al.*, 2011). The findings from single-cell genomics are consistent with widespread autotrophy genes in other dominant deep ocean microorganisms, such as the *Thaumarchaea* (Karner *et al.*, 2001, Pester *et al.*, 2011), and direct measurements of high levels of carbon fixation in the meso- and bathypelagic zones (Reinthaler *et al.*, 2010).

Another abundant group of microorganisms that populates the deep ocean is SAR11. Bacterioplankton of the SAR11 clade are the most numerous in marine systems, typically comprising ~25% of all prokaryotic cells (Morris *et al.*, 2002, Schattenhofer *et al.*, 2009). While the majority of research has focused on the SAR11 clade in the euphotic and upper mesopelagic zones, multiple studies have demonstrated evidence of substantial SAR11 populations deeper in the mesopelagic, as well as in the bathy-, and even hadopelagic (> 6000 m) realms (Eloe *et al.*, 2011a, Eloe *et al.*, 2011b, King *et al.*, 2013, Konstantinidis *et al.*, 2009, Martin-Cuadrado *et al.*, 2007, Quaiser *et al.*, 2010, Schattenhofer *et al.*, 2009, Swan *et al.*, 2011).

SAR11, or the “Pelagibacterales,” is a diverse group, spanning at least 18% 16S rRNA gene divergence, and is comprised of subclades with unique spatiotemporal distributions (ecotypes) that follow seasonal patterns (Carlson *et al.*, 2009, Field *et al.*, 1997, Giovannoni and Vergin, 2012, Grote *et al.*, 2012, Vergin *et al.*, 2013). All genome-sequenced representatives are characterized by small (1.3-1.4 Mbp), streamlined genomes with low GC content, few gene duplications, and an obligately aerobic, heterotrophic metabolism generally focused on oxidation of low molecular weight carbon compounds such as carboxylic and amino acids, osmolytes, and methylated compounds (Carini *et al.*, 2012, Grote *et al.*, 2012, Schwalbach *et al.*, 2010, Yilmaz *et al.*, 2011). Representatives spanning the known subclade diversity have an unusually high level of core genome conservation and gene synteny, however some subclade-specific genomic features have been identified (Grote *et al.*, 2012). The subclade V representative, HIMB59, encodes a complete glycolysis pathway and a variety of predicted sugar transporters. As

subclade V organisms bloom at the surface concurrently with the more numerically dominant subclade Ia ecotype (Vergin *et al.*, 2013), genetic machinery for the oxidation of sugars may provide a means of niche differentiation.

A recent study has pointed towards a deep SAR11 bathytype (depth-specific ecotype (Lauro and Bartlett, 2008)), phylogenetically distinct from the currently cultivated strains. This "subclade Ic" was represented by a single 16S clone library sequence that preferentially recruited pyrosequencing reads from depths of 200 m and below at the Bermuda Atlantic Time-series Study site (BATS) (Vergin *et al.*, 2013), and formed a monophyletic group with 16S sequences from single-cell genomes collected at 770 m at Station ALOHA. Here we present a comparative analysis of subclade Ic utilizing four single-amplified genomes (SAGs), metagenomes from euphotic, meso-, bathy-, and hadopelagic samples and eight pure-culture SAR11 genomes from three surface subclades. We tested the hypothesis that the subclade Ic genomes would have features that distinguish this bathytype from surface organisms to yield a better understanding of SAR11 adaptations to the ocean interior and of the genomic basis for SAR11 subclade differentiation by depth.

Materials and Methods

Comparative genomics

Single-cell separation, multiple displacement amplification (MDA), quality control, and SAG selection for sequencing based on MDA kinetics was all carried out as described previously (Swan *et al.*, 2011). More detailed descriptions are available in Supplemental Methods. Sequencing and assembly of the SAGs was carried out by the DOE Joint Genome Institute as part of a Community Sequencing Program grant 2011- 387. Genome annotations can be accessed using the Integrated Microbial Genomes (IMG) database (http://img.jgi.doe.gov).

SAG gene orthology with other SAR11 genomes was completed using the Hal pipeline (Robbertse *et al.*, 2011) and a series of custom filters, described in detail in Supplemental Methods. Post assembly quality control was assisted by examination of gene conservation across SAR11 strains. SAG genome completion was evaluated based on 599 single-copy genes present in all eight pure-culture SAR11 genomes. Overall SAG genome completion percentage was based on the percentage of these orthologs found in the SAGs (Table S1). Average amino acid identity (AAI) and local synteny between genomes were calculated with the scripts/methods of (Yelton *et al.*, 2011). Pairwise 16S rRNA gene identity was calculated with megablast using default settings. COG distribution among SAR11 genomes is part of data supplied by IMG (Table S1). Patterns of amino acid substitution between surface and deep-water strains of SAR11 were analyzed as described in (Konstantinidis *et al.*, 2009). Fold-change abundance of amino acids across similar and non-similar substitutions were calculated from all vs. all BLASTP output within homologous clusters. Intergenic spacer regions are provided as part of the IMG annotation process. Sizes and statistics for each set of intergenic regions were calculated using the fasta_length_counter.pl script. Distribution of intergenic regions was examined in R (http://www.R-project.org). Transposable elements were assessed using TBLASTN and the sequences collected by Brian Haas of the Broad Institute for the program TransposonPSI (http://transposonpsi.sourceforge.net). CRISPRs are detected as part of the automated IMG annotation process. A search for *cas* genes was conducted using 46 HMMs developed by Haft et al. (Haft *et al.*, 2005) and hmmsearch (Eddy) using default settings.

All phylogenetic analyses, with the exception of proteorhodopsin, were completed by aligning sequences with MUSCLE (Edgar, 2004) and computing trees with RAxML (Stamatakis, 2006, Stamatakis *et al.*, 2008). Alignments for trees in Figures 1 and 5 were curated for poorly aligned sites using Gblocks (Castresana, 2000). ProtTest (Abascal *et al.*, 2005) was utilized to optimize amino acid substitution modeling for protein coding trees. The concatenated protein phylogeny of the SAR11 clade was completed using the Hal pipeline (Robbertse *et al.*, 2011),

144 including . The proteorhodopsin tree was computed using the iterative Bayesian
145 alignment/phylogeny program HandAlign (Westesson *et al.*, 2012). Detailed methodology for
146 every tree, along with the unaligned fasta files for each of the single gene trees and the super
147 alignment and model file for the concatenated protein tree provided in Supplemental
148 Information.
149
150 *Metagenomics*
151 DNA was extracted from microbial biomass collected from BATS in August 2002 across
152 a depth profile (0, 40, 80, 120, 160, 200, and 250 m) and sequenced using 454 pyrosequencing
153 (GS-FLX, Roche). Metagenomes from ALOHA are previously described in (Shi *et al.*, 2011).
154 Data was also analyzed from 454 metagenomic sequences collected from Eastern Tropical
155 South Pacific Oxygen Minimum Zone (Stewart *et al.*, 2012), the Puerto Rico Trench (Eloe *et al.*,
156 2011a), the Sea of Marmara (Quaiser *et al.*, 2010), and the Matapan-Vavilov Deep in the
157 Mediterranean Sea (Smedile *et al.*, 2013). All raw data was trimmed of low quality end
158 sequences using Lucy (Chou and Holmes, 2001) and de-replicated using CDHIT-454 (Fu *et al.*,
159 2012). Sanger-sequenced reads from 4000 m at ALOHA (Konstantinidis *et al.*, 2009) were also
160 analyzed but not compared with the 454 pyrosequenced reads. GOS (Brown *et al.*, 2012, Rusch
161 *et al.*, 2007a, Venter *et al.*, 2004) surface sequences were analyzed for temperature
162 dependence of subclade Ic abundance, but also not included in gene relative abundance
163 normalizations (Supplementary Information).
164 Comparative recruitment of metagenomic sequences was completed using a reciprocal
165 best BLAST (rbb) (e.g., Wilhelm *et al.,* 2007) of eight SAR11 isolate genomes (HTCC1062,
166 HTCC1002, HTCC9565, HTCC7211, HIMB5, HIMB114, IMCC9063, HIMB59) and the four
167 SAR11 SAGs. Each concatenated SAR11 genome sequence was searched against each
168 metagenome database with BLASTN on default settings. All hits to SAR11 genomes were then
169 searched against the entire IMG database (v400), containing the 12 SAR11 genome sequences

using BLASTN. The best hits to each genome after this reciprocal best blast were then normalized by gene length, the average number of sequences, and relative abundance of SAR11 per sample. Taxonomic relative abundance for SAR11 and non-SAR11 organisms was estimated with metagenomic best-blast hits to whole genome sequences in the IMG v400 database. The results presented in Figure 2 represent an aggregation of all normalized metagenomic recruitment for all genomes in a given subclade, divided by the total number of SAR11 hits in that sample.

Gene clusters that may putatively play a role in depth adaptation in subclade 1c were identified as follows: Metagenomic samples were classified as 'deep' (< 200 m) or 'surface' (≥ 200 m) and gene cluster abundance in surface and deep samples was determined by reciprocal best-BLAST. The R package DESeq (Anders & Huber, 2010) was used to identify genes that were statistically significantly enriched at depth and at the surface. Detailed workflows for the metagenomic analyses are available in Supplemental Information.

Results and Discussion

Subclade Ic relative abundance in metagenomic datasets

Previous results demonstrated an abundance of upper mesopelagic 16S rRNA gene sequences phylogenetically affiliated with a single clone branching between SAR11 subclades Ia/Ib and subclades IIa/IIb, termed subclade Ic (Vergin *et al.*, 2013) (Fig. 1). Phylogenetic evaluation of SAR11-type SAG 16S rRNA gene sequences demonstrated a congruent topology, with a monophyletic group of SAGs collected from mesopelagic samples corresponding to the subclade Ic position (Fig. S1). Four SAGs were selected to represent the breadth of the clade, determined by branch lengths (Fig. S1). The 16S rRNA gene sequences from the SAGs formed a monophyletic group with the subclade Ic clone from (Vergin *et al.*, 2013), basal to subclades Ia/b (Fig. 1). All four SAGs were isolated from a single station ALOHA sample taken at 770 m.

Recruitment of metagenomic 454 pyrosequences from Station ALOHA, the Eastern Subtropical Pacific oxygen minimum zone (ESTP OMZ), and BATS indicated a higher relative abundance of subclade Ic in the mesopelagic relative to the euphotic zone (Fig. 2, Figs. S2-4), and greater relative abundance in the 6000 m Puerto Rico Trench metagenomic dataset compared to other subclades (Fig. S5). The Sea of Marmara dataset showed similar distributions between subclade Ia (predominantly HTCC1062 type) and Ic (Fig. S6), and although the Matapan-Vavilov Deep dataset had very little recruitment to any SAR11 genome (Fig. S7), consistent with the previous analysis (Smedile *et al.*, 2013), those sequences that did recruit to SAR11 genomes were predominantly Ic-like. Longer Sanger shotgun-sequencing reads from 4000 m at Station ALOHA (Konstantinidis *et al.*, 2009) also demonstrated increased recruitment to the SAGs relative to other genomes in deeper water (Fig. S8). We tested whether the increased abundance at depth might be due to temperature dependence. Recruitment from the GOS dataset (Rusch *et al.*, 2007b, Venter *et al.*, 2004) (Brown *et al.*, 2012) consistently showed a dearth of subclade Ic abundance relative to Ia in surface waters around the globe, and did not support the conclusion that subclade Ic abundance at depth is driven by temperature (Supplementary Information).

Comparisons with surface SAR11 genomes

The SAGs had total assembly sizes between 0.81-1.40 Mbp spanning 81-151 scaffolds > 500 bp, GC content between 29-30%, and coded for 948-1621 genes (Table 1). Estimated genome completeness, using 599 SAR11-specific single-copy orthologs (Table S1), was between 58 and 91% with the corresponding estimated average genome size for the subclade Ic organisms at 1.42 ± 0.08 Mbp. Protein-coding orthologous clusters (OCs) for the SAGs and eight isolate SAR11 genomes were determined by all vs. all BLASTP and Markov clustering using the automated pipeline Hal (Robbertse *et al.*, 2011) and custom filters for length and synteny. Of the 3156 total OCs in the twelve SAR11 genomes, 1763 (56%) were present in at

least one SAG, and 69% of the OCs found in the SAGs were shared with between one and eight other SAR11 genomes. COG distribution among the SAGs was generally the same as in surface genomes, except for categories M and P (Figs. 3, S9, see below). The majority of Ic-specific genes were hypothetical (Table S1), although several notable Ic-specific genes were present (see below). As would be expected from a low percentage of unique genes in the SAGs, much of the metabolism of these organisms appeared to be similar to that of the surface strains, particularly the subclade Ia organisms. Collectively, the Ic subclade were predicted to be obligate aerobic organisms, with cytochrome *c* oxidase as the sole terminal oxidase, a complete tricarboxylic acid cycle, conserved lesions in several glycolytic pathways (Schwalbach *et al.*, 2010), a reliance on reduced sulfur compounds (Tripp *et al.*, 2008), and an abundance of pathways for the metabolism and oxidation of small organic molecules such as amino/carboxylic acids and one-carbon and methylated compounds (Grote *et al.*, 2012, Yilmaz *et al.*, 2011)Carini, 2012} (Table S1).

Also consistent with previous findings about the *Pelagibacterales* (Grote *et al.*, 2012), the Ic SAGs had an unusually high conservation of local synteny among SAR11 genes (Fig. 3). When compared among themselves, the Ic SAGs had less local synteny than most organisms at that level of 16S rRNA gene identity. However, we attributed this to the SAGs being incomplete and fragmented, because when the SAGs were compared to other SAR11 genomes, syntenic genes were a characteristically high percentage of the total shared genes. High amounts of local synteny may seem unlikely given predicted SAR11 recombination rates are among the highest measured for prokaryotes (Vergin *et al.*, 2007, Vos and Didelot, 2009), however, it was shown previously that much of the rearrangement within genomes occurs at operon boundaries, and thus local synteny is not disrupted (Wilhelm *et al.*, 2007). Further, the rates in (Vergin *et al.*, 2007) were restricted to closely related organisms within subclade Ia.

Although gene content and local gene order conservation between the isolate genomes and the SAGs was high, the SAGs were distinct at the amino acid level. A concatenated protein

247 phylogeny using 322 single-copy orthologs supported the 16S phylogeny, placing the subclade
248 Ic SAGs as a monophyletic sister group to the subclade Ia surface strains (Fig. 5A). The
249 divergence from other strains and the depth of branching within the subclade Ic supported
250 conceptualization of subclade Ic as a new genus of SAR11, separate from the subclade Ia, or
251 *Pelagibacter* genus (Grote *et al.*, 2012). Comparison of average amino acid identity (AAI) versus
252 16S rRNA gene identity was also in accordance with the metrics proposed by Konstantinidis
253 and Tiedje for delineation of genera (66-72% AAI) (Grote *et al.*, 2012, Konstantinidis and Tiedje,
254 2007) (Fig. 5B). Specific amino acid substitution patterns among orthologs shared between the
255 SAGs and the surface genomes showed relative increases in cysteine, isoleucine, lysine,
256 asparagine, arginine and tryptophan in the predicted subclade Ic protein sequences at the
257 expense of alanine, aspartatic acid, glutamic acid, methionine, glutamine, threonine and valine
258 (Figs. 6, S10).
259 Many of the previously reported features associated with deep-ocean adaptation in
260 microorganisms were not observed in the SAGs, such as rRNA gene insertions, increased
261 transposable elements, or genes for chemoautotrophy (see Supplemental Information for
262 detailed discussion). Nevertheless, there were still some distinguishing characteristics between
263 subclade Ic and surface strains at the whole genome level that were similar to or matching
264 those previously observed in deep ocean metagenomic datasets (DeLong *et al.*, 2006,
265 Konstantinidis *et al.*, 2009) and comparative genomics studies. The subclade Ic genomes had a
266 small, but statistically significantly increase in intergenic space (Fig. S11) and a slightly (but
267 statistically insignificant) higher estimated average genome size than that of current surface
268 genomes (1.42 ± 0.08 vs. 1.33 ± 0.07, Table S1). Also, consistent with (Konstantinidis *et al.*,
269 2009) and a general trend towards larger genomes in deeper samples, there were more gaps in
270 the surface strain ortholog alignments (Fig. S10), indicating nucleotide insertions and thus larger
271 coding regions in the subclade Ic open reading frames. Unlike the surface strains, three of the
272 four SAGs showed a statistically significant enrichment in category M, cell

wall/membrane/envelope biogenesis (Fig. 5, Fig. S9). An increase in COG M genes was previously noted in the deep ocean *Photobacterium profundum* SS9 relative to mesophilic *Vibrionaceae* strains (Campanaro *et al.*, 2008) and in a deep water ecotype of *Alteromonas macleodii* (Ivars-Martínez *et al.*, 2008). COG M genes enriched in the SAGs include glycosyltransferases, methyltransferases, sugar epimerases, a sialic acid synthase, the cellular morphology gene *ccmA* (Hay *et al.*, 1999), and polysaccharide export proteins (Supplementary Information). The SAGs also showed a significant reduction of COG P genes for inorganic ion transport and metabolism that may reflect increased reliance on organic N and P sources. In support of this hypothesis, none of the SAGs had homologs of the phosphate metabolism genes *phoU*, *pstS*, *pstA*, or *pstC*, and while they had predicted ammonia permeases that clustered with ammonium transporters (clusters 150010.f.ok and 1500936.f.ok), none had genes annotated as an ammonium transporter. Furthermore, the SAGs had a unique pathway for purine degradation to ammonia (Fig. S12), including a 2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline (OHCU) decarboxylase that was specific to, and conserved in, all four SAGs, possibly indicating a clade-specific nitrogen salvage pathway.

There were also indications of unique phage interactions and defense mechanisms in subclade Ic compared to the surface strains, consistent with previous studies showing enrichment of phage genes at depth (Konstantinidis *et al.*, 2009, Martin-Cuadrado *et al.*, 2007). The SAGs had unique phage integrases and phage protein D genes (Table S1), and AAA240-E13 contained a predicted clustered regularly interspaced short palindromic repeat (CRISPR) region (Makarova *et al.*, 2011) on scaffold 14 (Fig. 7). A search for corresponding CRISPR-associated (*cas*) genes using HMMs developed by (Haft *et al.*, 2005, Makarova *et al.*, 2011) found some evidence for a *cas4*-like gene currently annotated as a hypothetical protein, conserved in three SAGs and HTCC9565 (Table S1, cluster 15001317). In AAA240-E13, this *cas4*-like protein was on scaffold 18 and thus not located directly nearby the CRISPR. Widespread Pelagiphage that infect at least a subset of the known surface strains have been

299 recently discovered (Zhao *et al.*, 2013), but this is the only CRISPR locus identified so far in
300 SAR11 genomes. Detailed analysis showed that this region had recruitment of metagenomic
301 sequences mostly from the mesopelagic Station ALOHA samples, indicating that the CRISPR is
302 relatively specific, geographically, with the majority of recruited sequences coming from
303 mesopelagic samples at ALOHA (Fig. 7). The observed increase in subclade Ic COG M genes
304 may also have a role in phage defense (Rodriguez-Valera *et al.*, 2009).
305
306 *Gene-specific relative abundance in metagenomic datasets*
307 We used metagenomic data to evaluate the relative importance of SAG genes *in situ*,
308 postulating that genes with little or no recruitment could be discounted as being present in fewer
309 organisms, whereas those with high levels of recruitment could be inferred as being the most
310 conserved, and therefore most important, to Ic-type organisms. Broadly, patterns of differential
311 gene abundance between the SAR11 subclades could be identified across datasets. In most of
312 the deep water samples, SAGs formed statistically significant grouping based on hierarchical
313 clustering of recruitment profiles, indicating that these genomes are highly similar based on
314 relative abundance of reciprocal best blast hits in deep-water environments (Fig. S13). The
315 normalized relative abundances of every gene for each SAG is reported in Table S1 for all
316 datasets. Thirty-nine clusters showed significantly higher relative abundance of metagenomic
317 sequence recruitment in deep water datasets (those at 200 m and below) compared to surface
318 datasets (Fig. 8, Supplementary Information). Only two of these clusters did not contain SAG
319 genes, whereas of the 42 clusters that were significantly more abundant in surface samples,
320 only two contained SAG genes- the rest were exclusive surface genomes. Half of these deep
321 abundance clusters were exclusive to the SAGs, the other half had some shared distribution
322 between the SAGs and surface genomes (Table S1).
323 Of the nineteen of these clusters that were specific to subclade Ic, nine were annotated
324 as hypothetical proteins. A subclade Ic-specific cluster of putative Fe-S oxidoreductases

contained multiple copies from each SAG, and all of the SAGs also had multiple copies of uncharacterized genes that clustered with single copies of predicted membrane occupation and recognition nexus (MORN) repeat genes from the subclade Ia genomes. The gene expansions for both these clusters suggested the proteins were important in the Ic subclade and in support of this hypothesis both were among the clusters significantly more abundant in deep metagenomic datasets (Table S1). A predicted adenosine deaminase, unique to the SAGs, was highly abundant in deep samples. This gene works upstream of xanthine dehydrogenase (also significantly more abundant) in purine degradation, and although not statistically significant, other elements of the putative subclade Ic-specific purine degradation pathway, including the OHCU decarboxylase, had high recruitment in deep samples compared to surface samples. Putative pillin assembly (*pilF*) genes, shared with other SAR11s, were also significantly more abundant in deep water samples, as were several methyltransferases, a Na+/proline symporter, and a high-affinity Fe^{2+}/Pb^{2+} permease.

Sulfite oxidase genes, conserved in three SAGs and shared only with HTCC9565, showed more recruitment in deep water samples, and were located directly adjacent to a cytochrome in the same configuration as the *sorAB* genes with proven sulfite oxidase activity in *Starkeya novella* ATCC 8083^{T} (Kappler *et al.*, 2000, Kappler *et al.*, 2012). The predicted AAA240-E13 sulfite oxidase had 33% identity with the *S. novella* SorA protein (blastp). Nearby were genes encoding for predicted Fe-S proteins, molybdopterin biosynthesis enzymes, and molybdenum cofactor synthesis (Mo and heme are required cofactors (Aguey-Zinsou *et al.*, 2003, Kappler *et al.*, 2000)), which also appeared qualitatively more abundant in deep water samples. This may therefore indicate a mechanism for sulfur chemolithotrophy in subclade Ic and HTCC9565. Utilization of partially-reduced sulfur compounds could also potentially explain the high abundance of SAR11 organisms and SAR11-type adenosine phosphosulfate reductase (*aprAB*) genes found in the ESTP OMZ, particularly at 200 m where dissolved oxygen is lowest and sulfur cycling has been identified (Fig. 2) (Canfield *et al.*, 2010, Stewart *et al.*, 2012). The

aprAB genes were found in all subclade Ia and two of the subclade Ic genomes (Table S1), and had high abundances in most of the deep water samples and higher abundance in deep vs. shallow samples in datasets from the same water column. Given the lack of additional genes in the assimilatory sulfate reduction pathway in most SAR11 organisms, (there was a predicted *sat* gene in HTCC9565 (Grote *et al.*, 2012)) *aprAB* have been proposed to play a role in taurine metabolism (Williams *et al.*, 2012), and may serve as a key sulfur cycling process for SAR11 in deep water as well. Our results indicate that the observed abundance of *aprAB* in the ESTP OMZ may be due to subclade Ic, rather than subclade Ia organisms.

Metagenomic relative abundance measurements allowed us to evaluate the potential importance of other notable genes found in the SAGs. Two of the SAGs, AAA288-G21 and AAA288-N07, contained predicted copies of proteorhodopsin- unexpected given the predominance of subclade Ic below the photic zone. The phylogeny of the proteorhodopsin genes generally matched the topology of the species tree (Fig. S14) and these loci showed modest recruitment in many of the samples for both strains (Table S1), indicating that the subclade Ic may cycle to the euphotic zone with enough frequency, as a population, for the physiological benefits of retaining proteorhodopsin to be realized. Many of the unique or unexpected SAG genes with annotations were located in hypervariable regions (genomic islands), where there was little or no recruitment of metagenomic sequences (Coleman, 2006, Grote *et al.*, 2012, Tully *et al.*, 2011, Wilhelm *et al.*, 2007) (Table S1). Two of the SAGs, AAA240-E13 and AAA288-E13 had copies of two predicted flagellar proteins, including a motor switch protein, a basal-body P-ring protein, located together, and AAA240-E13 additionally had a putative flagellar biosynthesis/type III secretory pathway protein. However, the first two genes showed no recruitment in any of the metagenomic datasets, and the third had recruitment in only one, indicating that they were unlikely to be a common trait among subclade Ic strains (Table S1). AAA240-E13 had the first mismatch repair (*mutS*) family homolog found in a SAR11 genome (Viklund *et al.*, 2012), but it too was located in a hypervariable region.

Summary

The results of our metagenomic analyses from a variety of locations strongly support the conclusion that the subclade Ic organisms are autochthonous to the deep ocean. However, this raises the question, what are the depths to which they are best adapted? Are subclade Ic SAR11 truly piezophilic (growth rates increasing with pressure from 1-500 atm (Madigan *et al.*, 2000)), or are they primarily adapted to the shallower mesopelagic zone (piezotolerant)? While the ALOHA 4000 m and PRT metagenomic analyses demonstrated subclade Ic organisms can be found in abysso- and hadopelagic realms, the lack of additional data from extreme deep water sites leaves the abundance of *Pelagibacterales* subclade Ic in such locations in question. Further, many previously identified features of both piezophilic isolates and deep ocean single-cell genomes (Lauro and Bartlett, Nagata *et al.*, 2010, Simonato *et al.*, 2006, Swan *et al.*, 2011) are absent in the SAR11 SAGs. While the incomplete state of the SAGs leaves open the possibility that these features may be contained in the unsequenced portion of the genomes, their absence in the nearly complete of AAA240-E13 SAG implies that even if present in some SAR11 Ic organisms, they are not universally conserved by the subclade. Alternatively, previously described features of deep ocean isolates may not be a commonality to all piezophiles, and some piezophilic adaptations may not be directly observable at the level of nucleic acid or protein sequence variation. For example, many, but not all, piezophiles contain polyunsaturated acids, and cold or high pressure adaption can also be achieved by changing the ratio of unsaturated to saturated monounsaturated fatty acids in membrane lipids (DeLong and Yayanos, 1985). Such properties are not readily predictable from genomes. Finally, since these SAGs were isolated from 770 m, a depth that does not usually represent a piezophilic environment, the possibility exists that the Ic subclade may have further bathytype divisions, including true piezophiles that occupy the deeper realms.

The evidence herein suggests these are a piezotolerant subclade, with metabolism similar to that of surface subclades focused on aerobic oxidation of organic acids, amino acids, and C1 and methylated compounds- universal products of metabolism that are expected to be found in all biomes- and may contain mechanisms for nitrogen salvage and sulfur chemolithotrophy unusual in most surface SAR11 genomes. They also appear to have been evolving as an environmentally isolated subclade for long enough to show distinct signatures at the genome level. Thus, we can affirm our hypothesis- the subclade Ic SAGs did contain genomic features that distinguished them from the surface SAR11 genomes, although these features were generally more subtle than large-scale gene content variations. They had larger intergenic regions and larger coding regions in SAR11 clade orthologs, had a slightly larger estimated average genome size, were distinct phylogenetically and at the amino acid content level, were enriched and depleted in COG M and P genes compared to other SAR11 genomes, respectively, and contained clade-specific hypothetical genes with increased relative-abundances in deep water samples. Further examination of such hypothetical genes and cultivation successes with deep ocean SAR11 strains will help provide a mechanistic explanation for how the features described by this study contribute to the predominance of subclade Ic organisms in deeper water.

Acknowledgements

This work was supported by the Gordon and Betty Moore Foundation (S.J.G. and E.F.D.), the U.S. Department of Energy Joint Genome Institute (JGI) Community Supported Program grant 2011-387 (R.S., B.K.S, E.F.D, S.J.G), National Science Foundation (NSF) Science and Technology Center Award EF0424599 (E.F.D.), NSF awards EF-826924 (R.S.), OCE-821374 (R.S.) and OCE-1232982 (R.S. and B.K.S.), and is based on work supported by the NSF under Award no. DBI-1003269 (J.C.T.). Sequencing was conducted by JGI and supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The authors thank Christopher M. Sullivan and the Oregon State University Center for Genome Research and Biocomputing, as well as the Louisiana State University Center for Computation and Technology for vital computational resources. We also thank Kelly C. Wrighton and Laura A. Hug for critical discussions about single-cell genomics, metagenomics and metabolic reconstruction.

The authors declare no conflict of interest in publication of this work.

Supplementary information is available at The ISME Journal's website.

References

Abascal F, Zardoya R, Posada D (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21:** 2104-2105.

Aguey-Zinsou K-F, Bernhardt PV, Kappler U, McEwan AG (2003). Direct Electrochemistry of a Bacterial Sulfite Dehydrogenase. *J Am Chem Soc* **125:** 530-535.

Anders S, Huber W (2010). Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Arıstegui J, Gasol JM, Duarte CM, Herndl GJ (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnol Oceanogr* **54:** 1501-1529.

Blainey PC (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* **37:** 407-427.

Brown MV, Lauro FM, DeMaere MZ, Les M, Wilkins D, Thomas T *et al.* (2012). Global biogeography of SAR11 marine bacteria. *Mol Sys Biol* **8:** 1-13.

Campanaro S, Treu L, Valle G (2008). Protein evolution in deep sea bacteria: an analysis of amino acids substitution rates. *BMC Evol Biol* **8:** 313.

Canfield DE, Stewart FJ, Thamdrup B, De Brabandere L, Dalsgaard T, DeLong EF *et al.* (2010). A Cryptic Sulfur Cycle in Oxygen-Minimum-Zone Waters off the Chilean Coast. *Science* **330:** 1375-1378.

Carini P, Steindler L, Beszteri S, Giovannoni SJ (2012). Nutrient requirements for growth of the extreme oligotroph 'Candidatus Pelagibacter ubique' HTCC1062 on a defined medium. *ISME J*.

Carlson CA, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K (2009). Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* **3:** 283-295.

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

Chou HH, Holmes MH (2001). DNA sequence quality trimming and vector removal. *Bioinformatics* **17:** 1093-1104.

Coleman ML (2006). Genomic Islands and the Ecology and Evolution of Prochlorococcus. *Science* **311:** 1768-1770.

DeLong EF, Yayanos AA (1985). Adaptation of the membrane lipids of a deep-sea bacterium to changes in hydrostatic pressure. *Science* **228:** 1101-1103.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496.

Eddy SR (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* **7:** e1002195.

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

Eloe EA, Fadrosh DW, Novotny M, Zeigler Allen L, Kim M, Lombardo M-J *et al.* (2011a). Going Deeper: Metagenome of a Hadopelagic Microbial Community. *PLOS ONE* **6:** e20388.

Eloe EA, Shulse CN, Fadrosh DW, Williamson SJ, Allen EE, Bartlett DH (2011b). Compositional differences in particle‐associated and free‐living microbial assemblages from an extreme deep‐ocean environment. *Environ Microbiol Rep* **3:** 449-458.

Field K, Gordon D, Wright T, Rappe M, Urbach E, Vergin K *et al.* (1997). Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl Environ Microbiol* **63:** 63-70.

Fu L, Niu B, Zhu Z, Wu S, Li W (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28:** 3150-3152.

Giovannoni SJ, Rappe MS, Vergin KL, Adair NL (1996). 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria. *P Natl Acad Sci USA* **93:** 7979-7984.

Giovannoni SJ, Vergin KL (2012). Seasonality in ocean microbial communities. *Science* **335:** 671-676.

Gordon DA, Giovannoni SJ (1996). Detection of stratified microbial populations related to Chlorobium and Fibrobacter species in the Atlantic and Pacific oceans. *Appl Environ Microbiol* **62:** 1171-1177.

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.

Haft DH, Selengut J, Mongodin EF, Nelson KE (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLOS Comput Biol* **1:** e60.

Hay NA, Tipper DJ, Gygi D, Hughes C (1999). A novel membrane protein influencing cell shape and multicellular swarming of Proteus mirabilis. *J Bacteriol* **181:** 2008-2016.

Ivars-Martínez E, Martin-Cuadrado A-B, D'Auria G, Mira A, Ferriera S, Johnson J *et al.* (2008). Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISME J* **2:** 1194-1212.

Kappler U, Bennett B, Rethmeier J, Schwarz G, Deutzmann R, McEwan AG *et al.* (2000). Sulfite: Cytochrome c Oxidoreductase from Thiobacillus novellus. *J Biol Chem* **275:** 13202-13212.

Kappler U, Davenport K, Beatson S, Lucas S, Lapidus A, Copeland A *et al.* (2012). Complete genome sequence of the facultatively chemolithoautotrophic and methylotrophic alpha Proteobacterium Starkeya novella type strain (ATCC 8093(T)). *Stand Genomic Sci* **7:** 44-58.

Karner MB, DeLong EF, Karl DM (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409:** 507-510.

King GM, Smith CB, Tolar B, Hollibaugh JT (2013). Analysis of composition and structure of coastal to mesopelagic bacterioplankton communities in the northern gulf of Mexico. *Front Microbiol* **3:** 438.

Konstantinidis KT, Tiedje JM (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* **10:** 504-509.

Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009). Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North Pacific Subtropical Gyre. *Appl Environ Microbiol* **75:** 5345-5355.

Lasken RS (2013). Single-cell sequencing in its prime. *Nat Biotechnol* **31:** 211-212.

Lauro FM, Bartlett DH (2008). Prokaryotic lifestyles in deep sea habitats. *Extremophiles* **12:** 15-25.

Madigan MT, Martinko JM, Parker J (2000). *Brock Biology of Microorganisms*, 9th edn. Prentice-Hall.

Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P *et al.* (2011). Evolution and classification of the CRISPR/Cas systems. *Nat Rev Micro* **9:** 467-477.

Martin-Cuadrado A-B, López-García P, Alba J-C, Moreira D, Monticelli L, Strittmatter A *et al.* (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLOS ONE* **2:** e914.

Morris R, Rappé M, Connon S, Vergin K, Siebold WA, Carlson CA *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420:** 806-810.

Morris RM, Longnecker K, Giovannoni SJ (2006). Pirellula and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. *Environ Microbiol* **8:** 1361-1370.

Morris RM, Frazar CD, Carlson CA (2012). Basin-scale patterns in the abundance of SAR11 subclades, marine Actinobacteria (OM1), members of the Roseobacter clade and OCS116 in the South Atlantic. *Environ Microbiol* **14:** 1133-1144.

Nagata T, Tamburini C, Arístegui J, Baltar F, Bochdansky AB, Fonda-Umani S *et al.* (2010). Emerging concepts on microbial processes in the bathypelagic ocean – ecology, biogeochemistry, and genomics. *Deep-Sea Res II* **57:** 1519-1536.

Pester M, Schleper C, Wagner M (2011). The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr Opin Microbiol* **14:** 300-306.

Quaiser A, Zivanovic Y, Moreira D, López-García P (2010). Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *ISME J* **5:** 285-304.

Reinthaler T, van Aken HM, Herndl GJ (2010). Major contribution of autotrophy to microbial carbon cycling in the deep North Atlantic's interior. *Deep-Sea Res II* **57:** 1572-1580.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*.

Robbertse B, Yoder RJ, Boyd A, Reeves J, Spatafora JW (2011). Hal: an Automated Pipeline for Phylogenetic Analyses of Genomic Data. *PLOS Currents Tree of Life* **3:** RRN1213.

Robinson C, Steinberg DK, Anderson TR, Arístegui J, Carlson CA, Frost JR *et al.* (2010). Mesopelagic zone ecology and biogeochemistry–a synthesis. *Deep-Sea Res II* **57:** 1504-1518.

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F *et al.* (2009). Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* **7:** 828-836.

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

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

Schattenhofer M, Fuchs BM, Amann R, Zubkov MV, Tarran GA, Pernthaler J (2009). Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environ Microbiol* **11:** 2078-2093.

Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ (2010). The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. *Environ Microbiol* **12:** 490-500.

Shi Y, Tyson GW, Eppley JM, DeLong EF (2011). Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J* **5:** 999-1013.

Simonato F, Campanaro S, Lauro FM, Vezzi A, D'Angelo M, Vitulo N *et al.* (2006). Piezophilic adaptation: a genomic point of view. *J Biotechnol* **126:** 11-25.

Smedile F, Messina E, La Cono V, Tsoy O, Monticelli LS, Borghini M *et al.* (2013). Metagenomic analysis of hadopelagic microbial assemblages thriving at the deepest part of Mediterranean Sea, Matapan-Vavilov Deep. *Environ Microbiol* **15:** 167-182.

Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690.

Stamatakis A, Hoover P, Rougemont J (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst Biol* **57:** 758-771.

Stepanauskas R (2012). Single cell genomics: an individual look at microbes. *Curr Opin Microbiol***:** 1-8.

Stewart FJ, Ulloa O, DeLong EF (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol* **14:** 23-40.

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D *et al.* (2011). Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. *Science* **333:** 1296-1300.

Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ *et al.* (2011). Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci Rep* **1:** 1-9.

Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA *et al.* (2009). Seasonality and vertical structure of microbial communities in an ocean gyre. *ISME J* **3:** 1148-1163.

Tripp HJ, Kitner JB, Schwalbach MS, Dacey JWH, Wilhelm LJ, Giovannoni SJ (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452:** 741-744.

Tully BJ, Nelson WC, Heidelberg JF (2011). Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* **14:** 254-267.

Varela MM, Van Aken HM, Herndl GJ (2008). Abundance and activity of Chloroflexi‐type SAR202 bacterioplankton in the meso‐and bathypelagic waters of the (sub) tropical Atlantic. *Environ Microbiol* **10:** 1903-1911.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappé MS, Giovannoni SJ (2007). High intraspecific recombination rate in a native population of Candidatus Pelagibacter ubique (SAR11). *Environ Microbiol* **9:** 2430-2440.

Vergin KL, Beszteri B, Monier A, Thrash JC, Temperton B, Treusch AH *et al.* (2013). High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J***:** 1-11.

Viklund J, Ettema TJG, Andersson SGE (2012). Independent Genome Reduction and Phylogenetic Reclassification of the Oceanic SAR11 Clade. *Mol Biol Evol* **29:** 599-615.

Vos M, Didelot X (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3:** 199-208.

Wang F, Wang J, Jian H, Zhang B, Li S, Wang F *et al.* (2008). Environmental Adaptation: Genomic Analysis of the Piezotolerant and Psychrotolerant Deep-Sea Iron Reducing Bacterium Shewanella piezotolerans WP3. *PLOS ONE* **3:** e1937.

686 Westesson O, Barquist L, Holmes I (2012). HandAlign: Bayesian multiple sequence alignment,
687 phylogeny and ancestral reconstruction. *Bioinformatics* **28:** 1170-1171.
688
689 Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ (2007). Natural variation in SAR11
690 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* **2:** 27.
691
692 Williams TJ, Long E, Evans F, DeMaere MZ, Lauro FM, Raftery MJ *et al.* (2012). A
693 metaproteomic assessment of winter and summer bacterioplankton from Antarctic Peninsula
694 coastal surface waters. *ISME J***:** 1-18.
695
696 Wright TD, Vergin KL, Boyd PW, Giovannoni SJ (1997). A novel delta-subdivision
697 proteobacterial lineage from the lower ocean surface layer. *Appl Environ Microbiol* **63:** 1441-
698 1448.
699
700 Yelton AP, Thomas BC, Simmons SL, Wilmes P, Zemla A, Thelen MP *et al.* (2011). A Semi-
701 Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and
702 Poorly Annotated Bacterial and Archaeal Genes. *PLOS Comput Biol* **7:** e1002230.
703
704 Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G *et al.* (2011).
705 The genomic standards consortium: bringing standards to life for microbial ecology. *ISME J* **5:**
706 1565-1567.
707
708 Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC *et al.* (2013).
709 Abundant SAR11 viruses in the ocean. *Nature* **494:** 357-360.
710
711

712 **Figure Legends**

713

714 Figure 1. Maximum-likelihood tree of 16S rRNA genes for the SAR11 clade in the context of
715 other *Alphaproteobacteria*. Genome sequenced strains are in bold, with subclade Ic sequences
716 in red and other SAR11 sequences in blue. Bootstrap values (n=1000) are indicated at the
717 nodes; scale bar represents 0.06 changes per position.

718

719 Figure 2. Relative abundance of SAR11 subclades based on reciprocal best blast recruitment of
720 metagenomic sequences.

721

722 Figure 3. Local synteny in SAR11 genomes. The percentage of genes in conserved order
723 relative to the total number of shared genes (Gene order conservation) vs. average normalized
724 bit score of the shared amino acid content. Red dots are all pairwise comparisons of SAR11
725 genomes, the total in a given area indicated by n. Data is overlaid on that from (Yelton *et al.*,
726 2011) (open grey circles).

727

728 Figure 4. A) Maximum likelihood tree of the SAR11 clade using 322 concatenated proteins.
729 Subclade Ic highlighted in blue. All nodes had 100% bootstrap support unless otherwise
730 indicated. Scale bar indicates changes per position. Root was inferred from (Grote *et al.*, 2012,
731 Thrash *et al.*, 2011). B) Average amino acid identity vs. 16S rRNA gene identity. Colors
732 correspond to values in each cell according to the key. Dashed line indicates genus-level
733 boundaries according to (Konstantinidis and Tiedje, 2007). Note, AAA240-E13 has only a partial
734 16S rRNA gene sequence, all others are full-length (See SI).

735

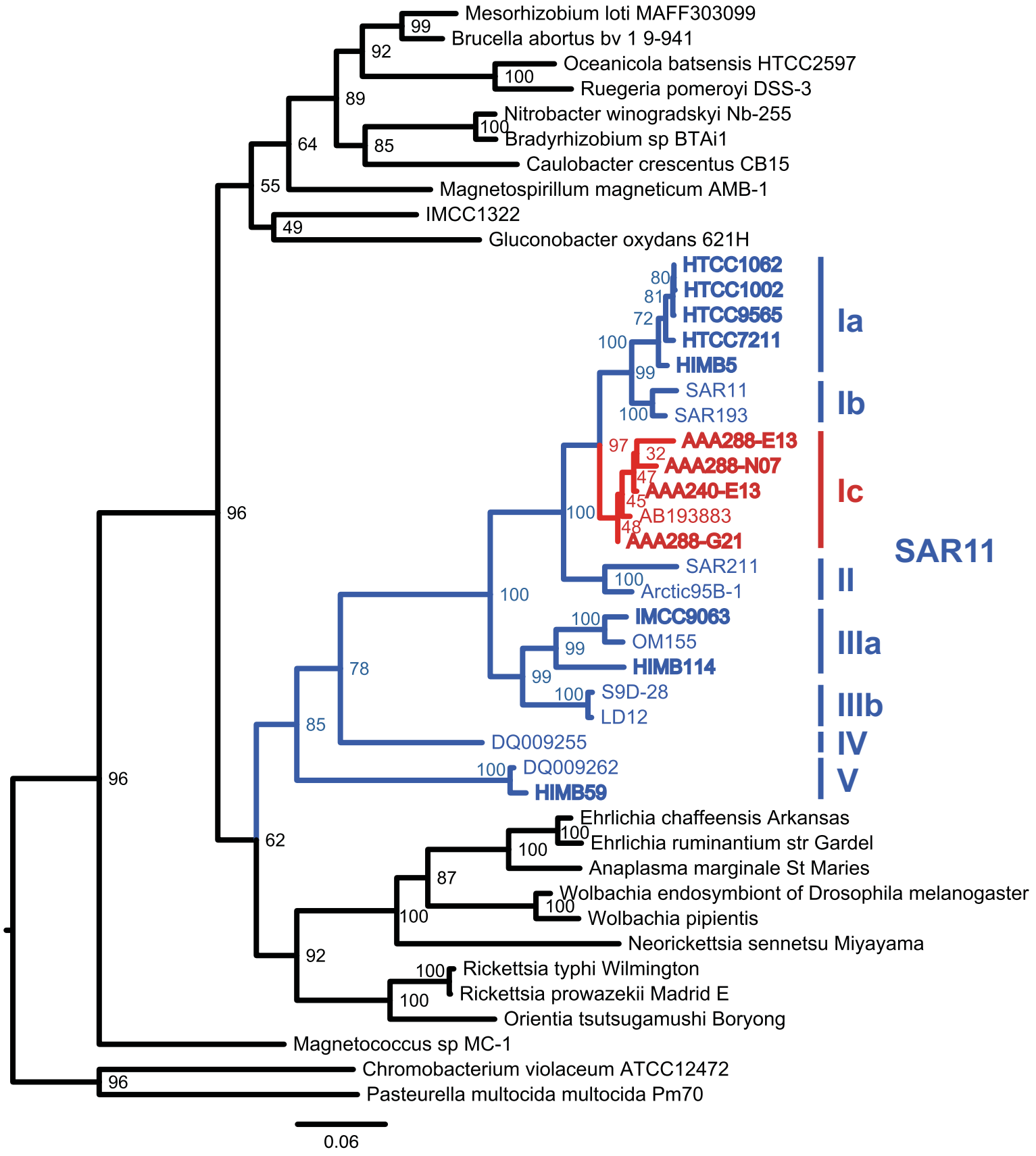
736 Figure 5. COG distribution as a percentage of total genes assigned to COGs. Y-axis:
737 percentage of genes, x-axis: COG categories. Colors correspond to the genomes according to

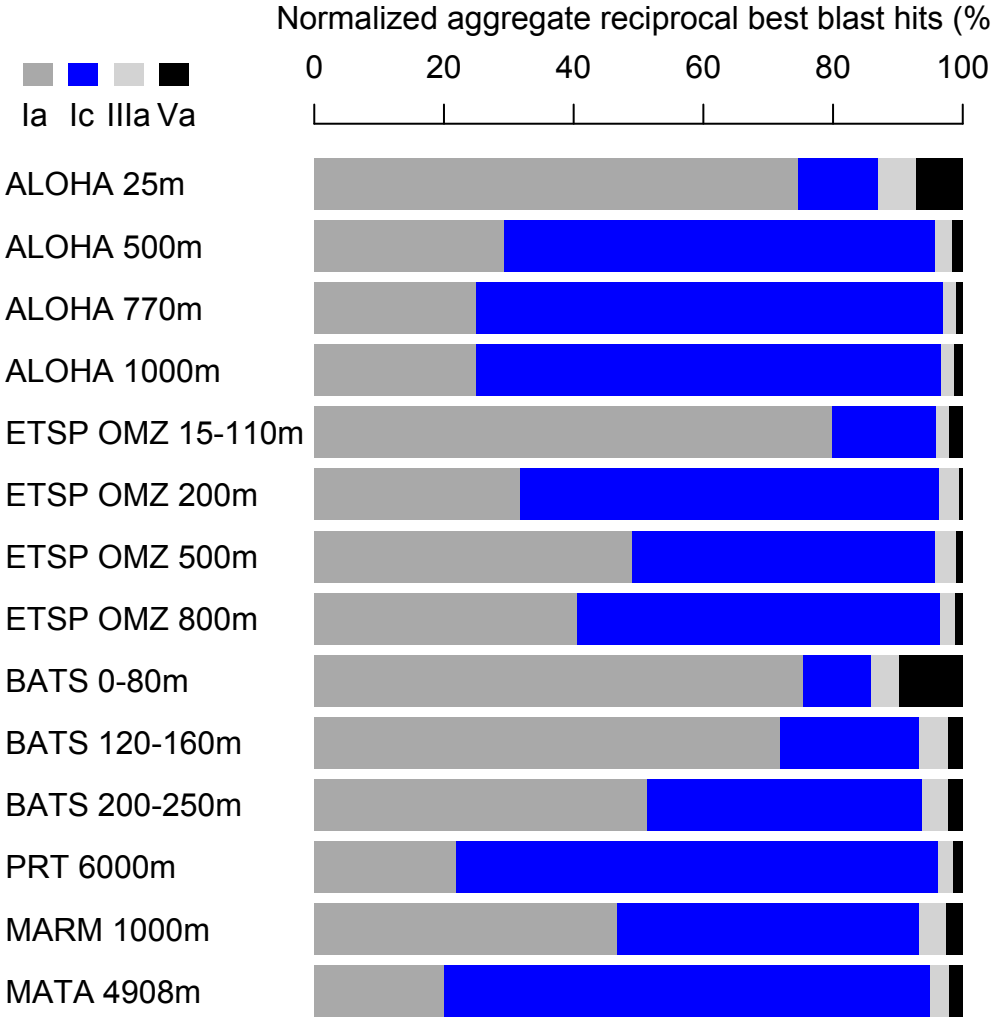
the key. Asterisks indicate categories with differential distribution in the SAGs relative to the isolate genomes. E- Amino acid metabolism and transport; G- Carbohydrate metabolism and transport; D- Cell division and chromosome partitioning; N- Cell motility and secretion; M- Cell wall/membrane/envelope biogenesis; B- Chromatin structure and dynamics; H- Coenzyme metabolism; Z- Cytoskeleton; V- ; C- Energy production and conversion; S- Unknown function; R- General function prediction only; P- Inorganic ion transport and metabolism; U- Intracellular trafficking and secretion; I- Lipid metabolism; F- Nucleotide transport and metabolism; O- Posttranslational modification, protein turnover, chaperones; L- DNA replication, recombination, and repair; Q- Secondary metabolite biosynthesis, transport and catabolism; T- Signal transduction mechanisms; K- Transcription; J- Translation.

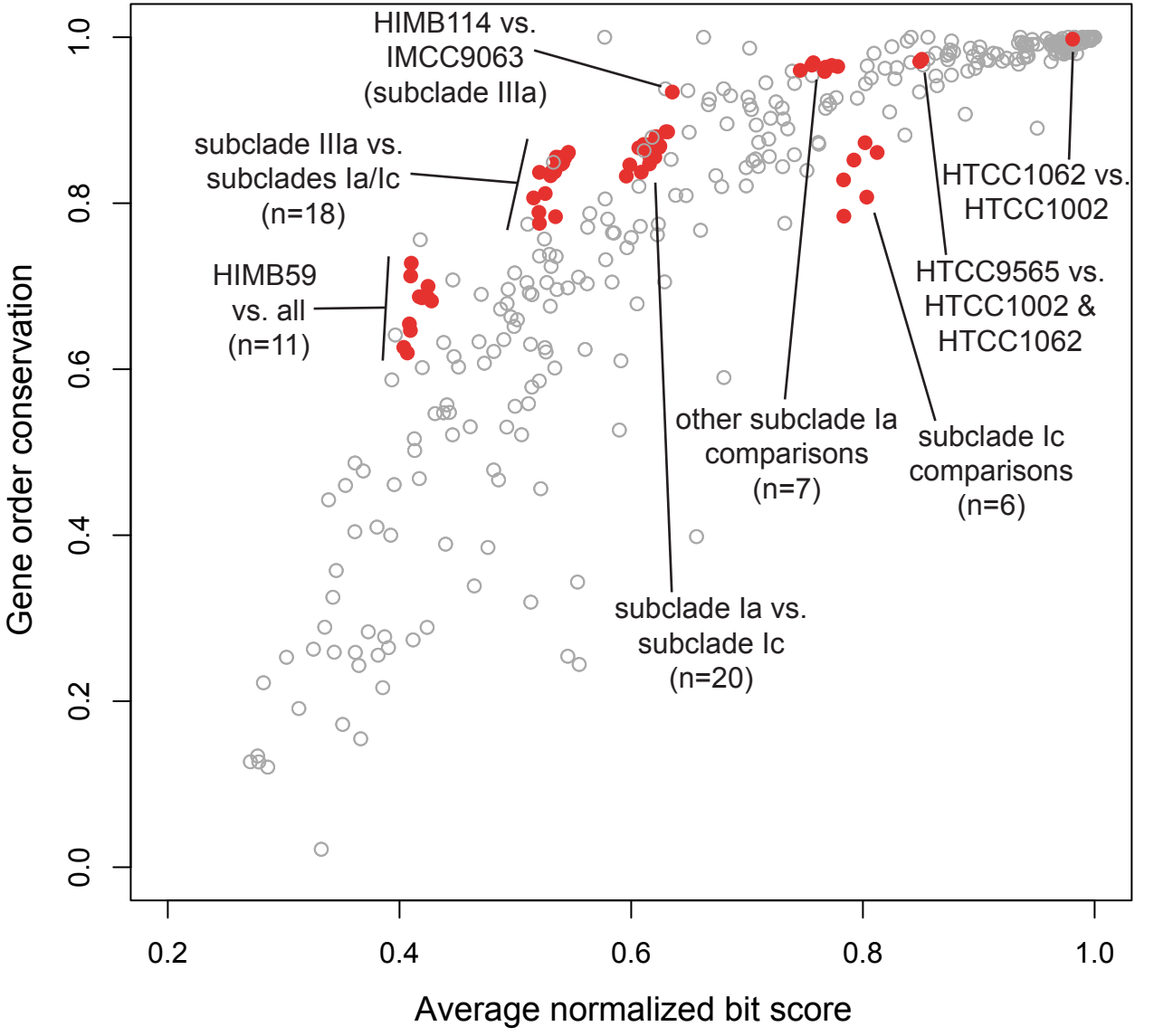
Figure 6. Fold-change in amino acid substitutions between the SAGs and the surface genomes. Pair-wise substitutions were quantified based on BLAST alignments of homologs between surface genomes and SAGs. X- unknown codons.

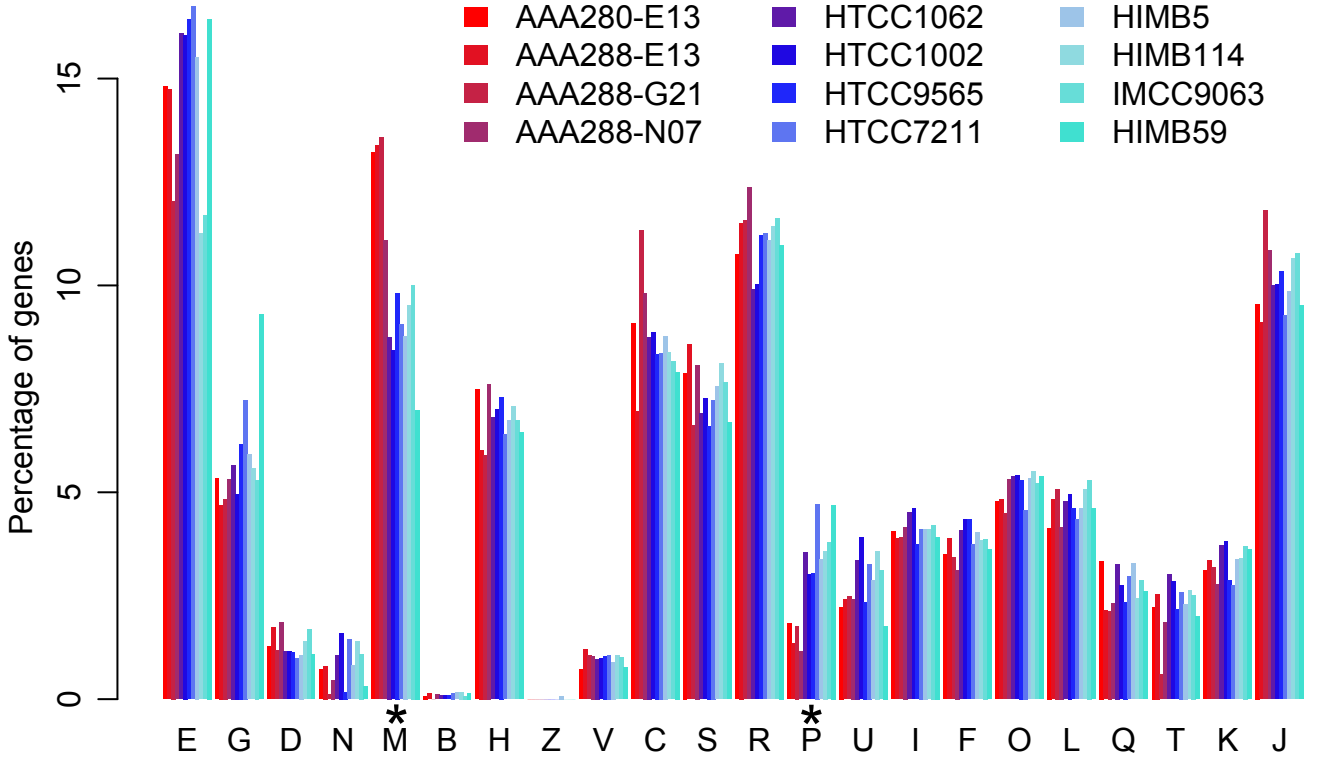
Figure 7. Recruitment of metagenomic sequences to the predicted CRISPR region. Upper box represents a magnification of the genomic region on scaffold 14 indicated in the title. Each line is a metagenomic sequence with reciprocal best hits (rbhs) to this region, organized by % identity (y-axis) and sample (color). Those samples not appearing in the analysis either had only rbhs < 50bp or no rbhs.

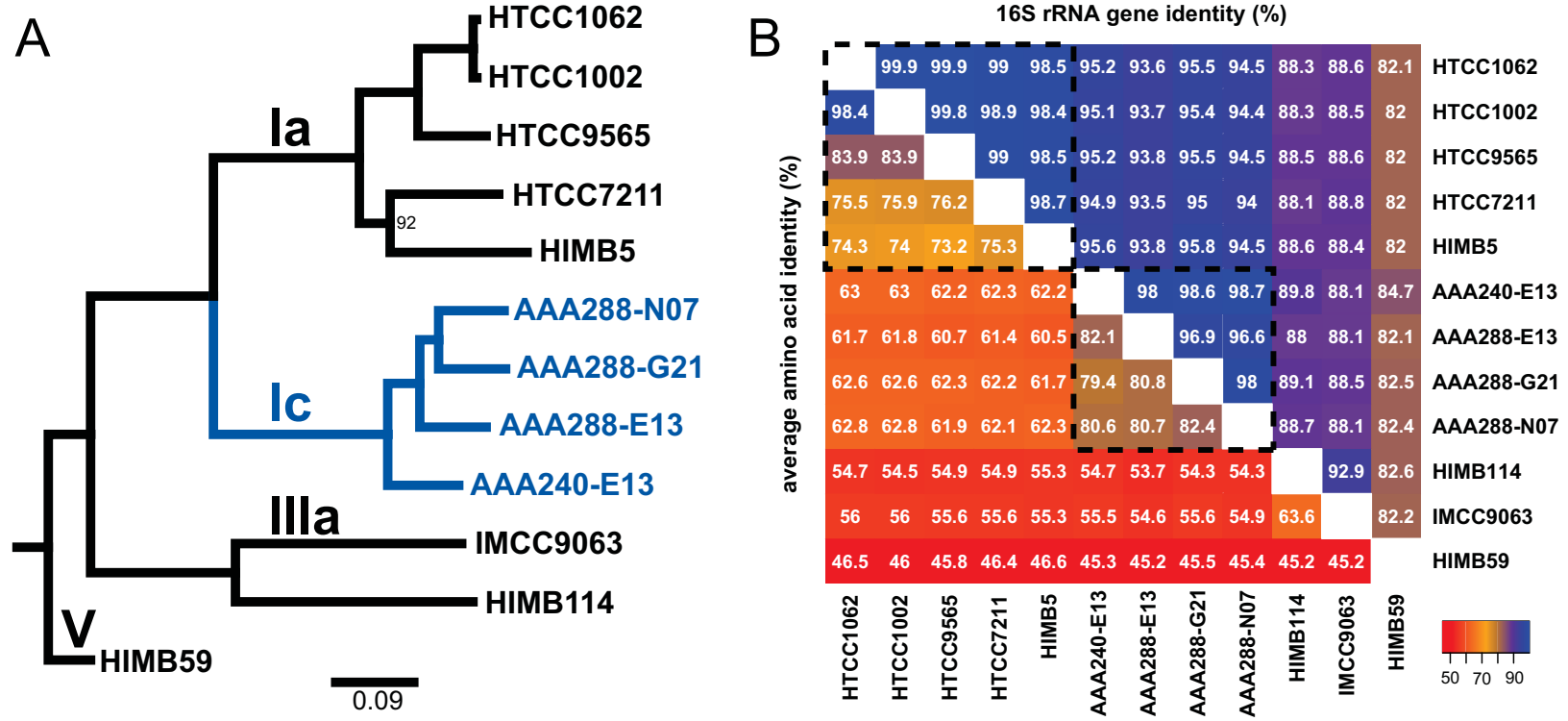
Figure 8. Plot of normalized mean vs. log-fold change for surface vs. deep gene clusters.

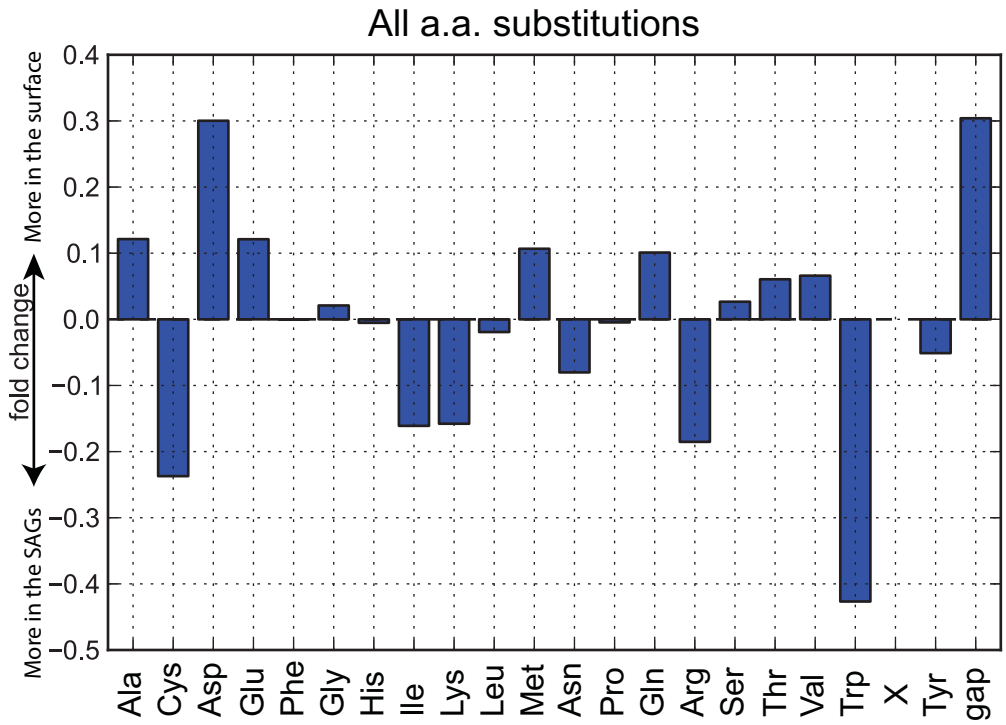

Mesorhizobium loti MAFF303099
Brucella abortus bv 1 9-941
Oceanicola batsensis HTCC2597
Ruegeria pomeroyi DSS-3
Nitrobacter winogradskyi Nb-255
Bradyrhizobium sp BTAi1
Caulobacter crescentus CB15
Magnetospirillum magneticum AMB-1
IMCC1322
Gluconobacter oxydans 621H
HTCC1062
HTCC1002
HTCC9565
HTCC7211
HIMB5
SAR11
SAR193
AAA288-E13
AAA288-N07
AAA240-E13
AB193883
AAA288-G21
SAR211
Arctic95B-1
IMCC9063
OM155
HIMB114
S9D-28
LD12
DQ009255
DQ009262
HIMB59
Ia
Ib
Ic
II
IIIa
IIIb
IV
V
SAR11
Ehrlichia chaffeensis Arkansas
Ehrlichia ruminantium str Gardel
Anaplasma marginale St Maries
Wolbachia endosymbiont of Drosophila melanogaster
Wolbachia pipientis
Neorickettsia sennetsu Miyayama
Rickettsia typhi Wilmington
Rickettsia prowazekii Madrid E
Orientia tsutsugamushi Boryong
Magnetococcus sp MC-1
Chromobacterium violaceum ATCC12472
Pasteurella multocida multocida Pm70
0.06

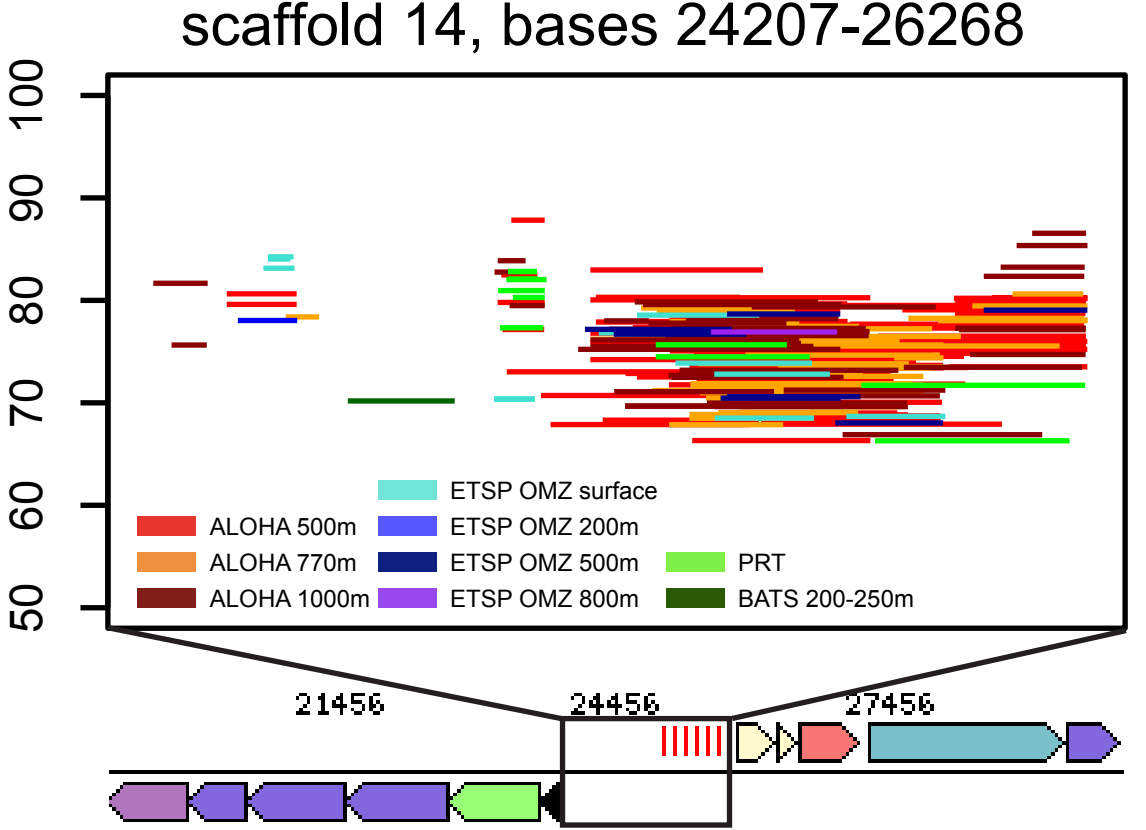

Normalized aggregate reciprocal best blast hits (%)
0
20
40
60
80
100
Ia
Ic
IIIa
Va
ALOHA 25m
ALOHA 500m
ALOHA 770m
ALOHA 1000m
ETSP OMZ 15-110m
ETSP OMZ 200m
ETSP OMZ 500m
ETSP OMZ 800m
BATS 0-80m
BATS 120-160m
BATS 200-250m
PRT 6000m
MARM 1000m
MATA 4908m

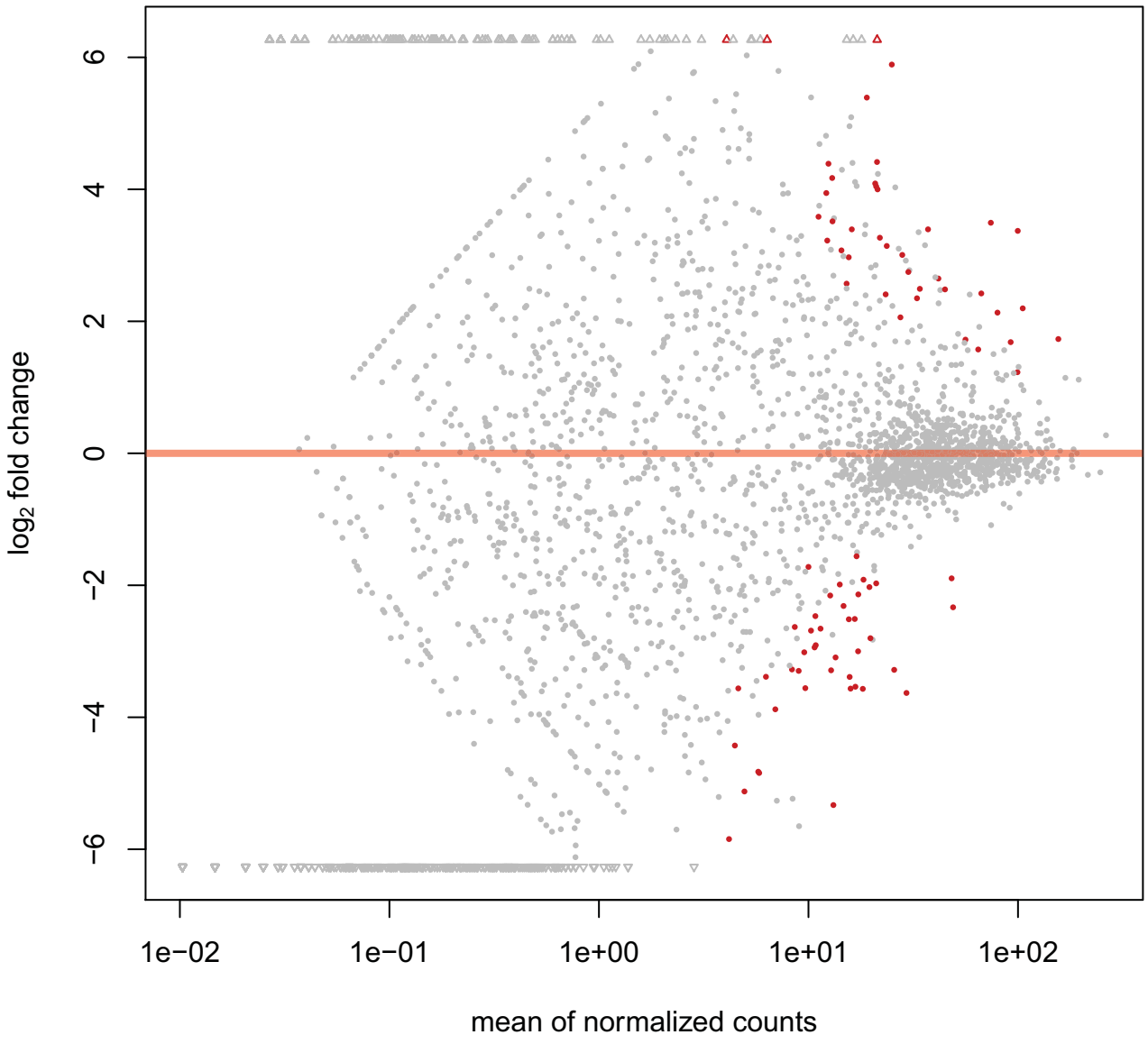

HIMB114 vs.
IMCC9063
(subclade IIIa)
subclade IIIa vs.
subclades Ia/Ic
(n=18)
HIMB59
vs. all
(n=11)
subclade Ia vs.
subclade Ic
(n=20)
other subclade Ia
comparisons
(n=7)
HTCC1062 vs.
HTCC1002
HTCC9565 vs.
HTCC1002 &
HTCC1062
subclade Ic
comparisons
(n=6)
Gene order conservation
0.0
0.2
0.4
0.6
0.8
1.0
Average normalized bit score
0.2
0.4
0.6
0.8
1.0


AAA280-E13
AAA288-E13
AAA288-G21
AAA288-N07
HTCC1062
HTCC1002
HTCC9565
HTCC7211
HIMB5
HIMB114
IMCC9063
HIMB59
Percentage of genes
0
5
10
15
E
G
D
N
M
B
H
Z
V
C
S
R
P
U
I
F
O
L
Q
T
K
J
*
*


A
HTCC1062
HTCC1002
Ia
HTCC9565
HTCC7211
92
HIMB5
AAA288-N07
AAA288-G21
Ic
AAA288-E13
AAA240-E13
IIIa
IMCC9063
HIMB114
V
HIMB59
0.09
B
16S rRNA gene identity (%)
average amino acid identity (%)
HTCC1062 99.9 99.9 99 98.5 95.2 93.6 95.5 94.5 88.3 88.6 82.1
HTCC1002 98.4 99.8 98.9 98.4 95.1 93.7 95.4 94.4 88.3 88.5 82
HTCC9565 83.9 83.9 99 98.5 95.2 93.8 95.5 94.5 88.5 88.6 82
HTCC7211 75.5 75.9 76.2 98.7 94.9 93.5 95 94 88.1 88.8 82
HIMB5 74.3 74 73.2 75.3 95.6 93.8 95.8 94.5 88.6 88.4 82
AAA240-E13 63 63 62.2 62.3 62.2 98 98.6 98.7 89.8 88.1 84.7
AAA288-E13 61.7 61.8 60.7 61.4 60.5 82.1 96.9 96.6 88 88.1 82.1
AAA288-G21 62.6 62.6 62.3 62.2 61.7 79.4 80.8 98 89.1 88.5 82.5
AAA288-N07 62.8 62.8 61.9 62.1 62.3 80.6 80.7 82.4 88.7 88.1 82.4
HIMB114 54.7 54.5 54.9 54.9 55.3 54.7 53.7 54.3 54.3 92.9 82.6
IMCC9063 56 56 55.6 55.6 55.3 55.5 54.6 55.6 54.9 63.6 82.2
HIMB59 46.5 46 45.8 46.4 46.6 45.3 45.2 45.5 45.4 45.2 45.2
HTCC1062
HTCC1002
HTCC9565
HTCC7211
HIMB5
AAA240-E13
AAA288-E13
AAA288-G21
AAA288-N07
HIMB114
IMCC9063
HIMB59
50 70 90


All a.a. substitutions
More in the surface
fold change
More in the SAGs
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
Ala
Cys
Asp
Glu
Phe
Gly
His
Ile
Lys
Leu
Met
Asn
Pro
Gln
Arg
Ser
Thr
Val
Trp
X
Tyr
gap


scaffold 14, bases 24207-26268
100
90
80
70
60
50
ALOHA 500m
ALOHA 770m
ALOHA 1000m
ETSP OMZ surface
ETSP OMZ 200m
ETSP OMZ 500m
ETSP OMZ 800m
PRT
BATS 200-250m
21456
24456
27456


log2 fold change
6
4
2
0
−2
−4
−6
1e−02
1e−01
1e+00
1e+01
1e+02
mean of normalized counts

Tables

Table 1. Subclade Ic SAG genome characteristics

Genome	AAA240-E13	AAA288-E13	AAA288-G21	AAA288-N07	other SAR11[#]
Number of scaffolds	151	106	139	81	-
Assembly size (Mbp)	1.40	0.81	0.91	0.95	-
Est. genome completeness (%)	91	58	67	70	-
Est. genome size (Mbp)	1.55	1.41	1.36	1.37	1.29-1.41*
GC content (%)	29	29	30	29	29-32
Number of genes	1621	948	1103	1110	1357-1576
Number of genes (prot. cod.)	1581	923	1074	1083	1321-1541

[#]Values from (Grote *et al.*, 2012) and IMG, *actual (not estimated) sizes