

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

2

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

5

6 1. Department of Microbiology, Oregon State University, Corvallis, OR 97331

7 2. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, 70803

8 3. Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544

9 4. DOE Joint Genome Institute, Walnut Creek, CA 94598

10 5. Departments of Civil & Environmental Engineering and Biological Engineering,  
11 Massachusetts Institute of Technology, Cambridge, MA 02139

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

13

14

15

16 **Abstract**

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

37

38

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

40

## 41 **Introduction**

42           Characterized by darkness, average temperatures of ~2-4°C, increased hydrostatic  
43 pressure, and general oligotrophy, the relatively extreme environment of the deep ocean is also,  
44 ironically, the largest biome on Earth. The mesopelagic (200-1000 m) and bathypelagic (1000-  
45 4000 m) zones contain > 70% of marine microbial biomass (Aristegui *et al.*, 2009) and these  
46 organisms play vital roles in global cycling of carbon, nitrogen, and other biogeochemical  
47 processes (Nagata *et al.*, 2010, Robinson *et al.*, 2010). In addition to microorganisms  
48 necessarily being adapted to cold and increased pressure there, the deep sea also contains  
49 more recalcitrant forms of carbon than at the surface (Aristegui *et al.*, 2009, Nagata *et al.*, 2010,  
50 Robinson *et al.*, 2010). Cultivated isolates have revealed some microbial adaptations associated  
51 with life at depth, including increased intergenic spacer regions, rRNA gene indels, and higher  
52 abundances of membrane polyunsaturated fatty acids and surface-adhesion/motility genes  
53 (Lauro and Bartlett, 2008, Nagata *et al.*, 2010, Simonato *et al.*, 2006, Wang *et al.*, 2008).

54           However, many of the most abundant bacterial groups from the deep ocean remain  
55 uncultivated, for example the SAR202, SAR324, and SAR406 clades, which make up significant  
56 fractions of microbial communities at depth (DeLong *et al.*, 2006, Giovannoni *et al.*, 1996,  
57 Gordon and Giovannoni, 1996, Morris *et al.*, 2006, Morris *et al.*, 2012, Schattenhofer *et al.*,  
58 2009, Treusch *et al.*, 2009, Varela *et al.*, 2008, Wright *et al.*, 1997). Thus, it remains uncertain  
59 how widespread the known adaptations of cultivated isolates are among deep ocean  
60 microorganisms. Metagenomic analyses have provided evidence for common genomic features  
61 in the deep ocean, such as increased proliferation of transposable elements and phage, amino  
62 acid content changes, and increased average genome size (DeLong *et al.*, 2006, Konstantinidis  
63 *et al.*, 2009). Single-cell genomic analyses provide another powerful means to understand the  
64 metabolism and evolution of organisms eluding cultivation-based techniques (Blainey, 2013,  
65 Lasken, 2013, Rinke *et al.*, 2013, Stepanauskas, 2012). This approach provided the first insight  
66 into the metabolism of several of these deep ocean clades, including SAR324, Arctic96BD-19,

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

72 Another abundant group of microorganisms that populates the deep ocean is SAR11.  
73 Bacterioplankton of the SAR11 clade are the most numerous in marine systems, typically  
74 comprising ~25% of all prokaryotic cells (Morris *et al.*, 2002, Schattenuhofer *et al.*, 2009). While  
75 the majority of research has focused on the SAR11 clade in the euphotic and upper  
76 mesopelagic zones, multiple studies have demonstrated evidence of substantial SAR11  
77 populations deeper in the mesopelagic, as well as in the bathy-, and even hadopelagic (> 6000  
78 m) realms (Eloe *et al.*, 2011a, Eloe *et al.*, 2011b, King *et al.*, 2013, Konstantinidis *et al.*, 2009,  
79 Martin-Cuadrado *et al.*, 2007, Quaiser *et al.*, 2010, Schattenuhofer *et al.*, 2009, Swan *et al.*,  
80 2011).

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

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

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

108

## 109 **Materials and Methods**

### 110 *Comparative genomics*

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

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

138 All phylogenetic analyses, with the exception of proteorhodopsin, were completed by  
139 aligning sequences with MUSCLE (Edgar, 2004) and computing trees with RAxML (Stamatakis,  
140 2006, Stamatakis *et al.*, 2008). Alignments for trees in Figures 1 and 5 were curated for poorly  
141 aligned sites using Gblocks (Castresana, 2000). ProtTest (Abascal *et al.*, 2005) was utilized to  
142 optimize amino acid substitution modeling for protein coding trees. The concatenated protein  
143 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

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

177 Gene clusters that may putatively play a role in depth adaptation in subclade 1c were  
178 identified as follows: Metagenomic samples were classified as 'deep' (< 200 m) or 'surface' ( $\geq$   
179 200 m) and gene cluster abundance in surface and deep samples was determined by reciprocal  
180 best-BLAST. The R package DESeq (Anders & Huber, 2010) was used to identify genes that  
181 were statistically significantly enriched at depth and at the surface. Detailed workflows for the  
182 metagenomic analyses are available in Supplemental Information.

183

## 184 **Results and Discussion**

### 185 *Subclade 1c relative abundance in metagenomic datasets*

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

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

211

#### 212 *Comparisons with surface SAR11 genomes*

213 The SAGs had total assembly sizes between 0.81-1.40 Mbp spanning 81-151 scaffolds  
214 > 500 bp, GC content between 29-30%, and coded for 948-1621 genes (Table 1). Estimated  
215 genome completeness, using 599 SAR11-specific single-copy orthologs (Table S1), was  
216 between 58 and 91% with the corresponding estimated average genome size for the subclade  
217 Ic organisms at  $1.42 \pm 0.08$  Mbp. Protein-coding orthologous clusters (OCs) for the SAGs and  
218 eight isolate SAR11 genomes were determined by all vs. all BLASTP and Markov clustering  
219 using the automated pipeline Hal (Robbertse *et al.*, 2011) and custom filters for length and  
220 synteny. Of the 3156 total OCs in the twelve SAR11 genomes, 1763 (56%) were present in at

221 least one SAG, and 69% of the OCs found in the SAGs were shared with between one and  
222 eight other SAR11 genomes. COG distribution among the SAGs was generally the same as in  
223 surface genomes, except for categories M and P (Figs. 3, S9, see below). The majority of Ic-  
224 specific genes were hypothetical (Table S1), although several notable Ic-specific genes were  
225 present (see below). As would be expected from a low percentage of unique genes in the  
226 SAGs, much of the metabolism of these organisms appeared to be similar to that of the surface  
227 strains, particularly the subclade Ia organisms. Collectively, the Ic subclade were predicted to be  
228 obligate aerobic organisms, with cytochrome c oxidase as the sole terminal oxidase, a complete  
229 tricarboxylic acid cycle, conserved lesions in several glycolytic pathways (Schwalbach *et al.*,  
230 2010), a reliance on reduced sulfur compounds (Tripp *et al.*, 2008), and an abundance of  
231 pathways for the metabolism and oxidation of small organic molecules such as amino/carboxylic  
232 acids and one-carbon and methylated compounds (Grote *et al.*, 2012, Yilmaz *et al.*, 2011)Carini,  
233 2012} (Table S1).

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

245         Although gene content and local gene order conservation between the isolate genomes  
246 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 1c SAGs as a monophyletic sister group to the subclade 1a surface strains (Fig. 5A). The  
249 divergence from other strains and the depth of branching within the subclade 1c supported  
250 conceptualization of subclade 1c as a new genus of SAR11, separate from the subclade 1a, 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 1c protein sequences at the  
257 expense of alanine, aspartic 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 1c 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 1c 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 ( $1.42 \pm 0.08$  vs.  $1.33 \pm 0.07$ , 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 1c open reading frames. Unlike the surface strains, three of the  
272 four SAGs showed a statistically significant enrichment in category M, cell

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

288         There were also indications of unique phage interactions and defense mechanisms in  
289 subclade Ic compared to the surface strains, consistent with previous studies showing  
290 enrichment of phage genes at depth (Konstantinidis *et al.*, 2009, Martin-Cuadrado *et al.*, 2007).  
291 The SAGs had unique phage integrases and phage protein D genes (Table S1), and AAA240-  
292 E13 contained a predicted clustered regularly interspaced short palindromic repeat (CRISPR)  
293 region (Makarova *et al.*, 2011) on scaffold 14 (Fig. 7). A search for corresponding CRISPR-  
294 associated (*cas*) genes using HMMs developed by (Haft *et al.*, 2005, Makarova *et al.*, 2011)  
295 found some evidence for a *cas4*-like gene currently annotated as a hypothetical protein,  
296 conserved in three SAGs and HTCC9565 (Table S1, cluster 15001317). In AAA240-E13, this  
297 *cas4*-like protein was on scaffold 18 and thus not located directly nearby the CRISPR.  
298 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

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

338         Sulfite oxidase genes, conserved in three SAGs and shared only with HTCC9565,  
339 showed more recruitment in deep water samples, and were located directly adjacent to a  
340 cytochrome in the same configuration as the *sorAB* genes with proven sulfite oxidase activity in  
341 *Starkeya novella* ATCC 8083<sup>T</sup> (Kappler *et al.*, 2000, Kappler *et al.*, 2012). The predicted  
342 AAA240-E13 sulfite oxidase had 33% identity with the *S. novella* SorA protein (blastp). Nearby  
343 were genes encoding for predicted Fe-S proteins, molybdopterin biosynthesis enzymes, and  
344 molybdenum cofactor synthesis (Mo and heme are required cofactors (Aguey-Zinsou *et al.*,  
345 2003, Kappler *et al.*, 2000)), which also appeared qualitatively more abundant in deep water  
346 samples. This may therefore indicate a mechanism for sulfur chemolithotrophy in subclade Ic  
347 and HTCC9565. Utilization of partially-reduced sulfur compounds could also potentially explain  
348 the high abundance of SAR11 organisms and SAR11-type adenosine phosphosulfate reductase  
349 (*aprAB*) genes found in the ESTP OMZ, particularly at 200 m where dissolved oxygen is lowest  
350 and sulfur cycling has been identified (Fig. 2) (Canfield *et al.*, 2010, Stewart *et al.*, 2012). The

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

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

377

378 *Summary*

379           The results of our metagenomic analyses from a variety of locations strongly support the  
380 conclusion that the subclade Ic organisms are autochthonous to the deep ocean. However, this  
381 raises the question, what are the depths to which they are best adapted? Are subclade Ic  
382 SAR11 truly piezophilic (growth rates increasing with pressure from 1-500 atm (Madigan *et al.*,  
383 2000)), or are they primarily adapted to the shallower mesopelagic zone (piezotolerant)? While  
384 the ALOHA 4000 m and PRT metagenomic analyses demonstrated subclade Ic organisms can  
385 be found in abysso- and hadopelagic realms, the lack of additional data from extreme deep  
386 water sites leaves the abundance of *Pelagibacterales* subclade Ic in such locations in question.  
387 Further, many previously identified features of both piezophilic isolates and deep ocean single-  
388 cell genomes (Lauro and Bartlett, Nagata *et al.*, 2010, Simonato *et al.*, 2006, Swan *et al.*, 2011)  
389 are absent in the SAR11 SAGs. While the incomplete state of the SAGs leaves open the  
390 possibility that these features may be contained in the unsequenced portion of the genomes,  
391 their absence in the nearly complete of AAA240-E13 SAG implies that even if present in some  
392 SAR11 Ic organisms, they are not universally conserved by the subclade. Alternatively,  
393 previously described features of deep ocean isolates may not be a commonality to all  
394 piezophiles, and some piezophilic adaptations may not be directly observable at the level of  
395 nucleic acid or protein sequence variation. For example, many, but not all, piezophiles contain  
396 polyunsaturated acids, and cold or high pressure adaption can also be achieved by changing  
397 the ratio of unsaturated to saturated monounsaturated fatty acids in membrane lipids (DeLong  
398 and Yayanos, 1985). Such properties are not readily predictable from genomes. Finally, since  
399 these SAGs were isolated from 770 m, a depth that does not usually represent a piezophilic  
400 environment, the possibility exists that the Ic subclade may have further bathytype divisions,  
401 including true piezophiles that occupy the deeper realms.

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

419

## 420 **Acknowledgements**

421 This work was supported by the Gordon and Betty Moore Foundation (S.J.G. and E.F.D.), the  
422 U.S. Department of Energy Joint Genome Institute (JGI) Community Supported Program grant  
423 2011-387 (R.S., B.K.S, E.F.D, S.J.G), National Science Foundation (NSF) Science and  
424 Technology Center Award EF0424599 (E.F.D.), NSF awards EF-826924 (R.S.), OCE-821374  
425 (R.S.) and OCE-1232982 (R.S. and B.K.S.), and is based on work supported by the NSF under  
426 Award no. DBI-1003269 (J.C.T.). Sequencing was conducted by JGI and supported by the  
427 Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

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

433

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

435

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

437

438 **References**

- 439 Abascal F, Zardoya R, Posada D (2005). ProtTest: selection of best-fit models of protein  
440 evolution. *Bioinformatics* **21**: 2104-2105.  
441
- 442 Aguey-Zinsou K-F, Bernhardt PV, Kappler U, McEwan AG (2003). Direct Electrochemistry of a  
443 Bacterial Sulfite Dehydrogenase. *J Am Chem Soc* **125**: 530-535.  
444
- 445 Anders S, Huber W (2010). Differential expression analysis for sequence count data. *Genome*  
446 *Biol* **11**: R106.  
447
- 448 Aristegui J, Gasol JM, Duarte CM, Herndl GJ (2009). Microbial oceanography of the dark  
449 ocean's pelagic realm. *Limnol Oceanogr* **54**: 1501-1529.  
450
- 451 Blainey PC (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS*  
452 *Microbiol Rev* **37**: 407-427.  
453
- 454 Brown MV, Lauro FM, DeMaere MZ, Les M, Wilkins D, Thomas T *et al.* (2012). Global  
455 biogeography of SAR11 marine bacteria. *Mol Sys Biol* **8**: 1-13.  
456
- 457 Campanaro S, Treu L, Valle G (2008). Protein evolution in deep sea bacteria: an analysis of  
458 amino acids substitution rates. *BMC Evol Biol* **8**: 313.  
459
- 460 Canfield DE, Stewart FJ, Thamdrup B, De Brabandere L, Dalsgaard T, DeLong EF *et al.* (2010).  
461 A Cryptic Sulfur Cycle in Oxygen-Minimum-Zone Waters off the Chilean Coast. *Science* **330**:  
462 1375-1378.  
463
- 464 Carini P, Steindler L, Beszteri S, Giovannoni SJ (2012). Nutrient requirements for growth of the  
465 extreme oligotroph 'Candidatus Pelagibacter ubique' HTCC1062 on a defined medium. *ISME J*.  
466
- 467 Carlson CA, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K (2009). Seasonal  
468 dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern  
469 Sargasso Sea. *ISME J* **3**: 283-295.  
470
- 471 Castresana J (2000). Selection of conserved blocks from multiple alignments for their use in  
472 phylogenetic analysis. *Mol Biol Evol* **17**: 540-552.  
473
- 474 Chou HH, Holmes MH (2001). DNA sequence quality trimming and vector removal.  
475 *Bioinformatics* **17**: 1093-1104.  
476
- 477 Coleman ML (2006). Genomic Islands and the Ecology and Evolution of Prochlorococcus.  
478 *Science* **311**: 1768-1770.  
479
- 480 DeLong EF, Yayanos AA (1985). Adaptation of the membrane lipids of a deep-sea bacterium to  
481 changes in hydrostatic pressure. *Science* **228**: 1101-1103.  
482
- 483 DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community  
484 genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496.  
485
- 486 Eddy SR (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.  
487

488 Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high  
489 throughput. *Nucleic Acids Res* **32**: 1792-1797.  
490  
491 Eloe EA, Fadrosch DW, Novotny M, Zeigler Allen L, Kim M, Lombardo M-J *et al.* (2011a). Going  
492 Deeper: Metagenome of a Hadopelagic Microbial Community. *PLOS ONE* **6**: e20388.  
493  
494 Eloe EA, Shulse CN, Fadrosch DW, Williamson SJ, Allen EE, Bartlett DH (2011b). Compositional  
495 differences in particle - associated and free - living microbial assemblages from an extreme  
496 deep - ocean environment. *Environ Microbiol Rep* **3**: 449-458.  
497  
498 Field K, Gordon D, Wright T, Rappe M, Urbach E, Vergin K *et al.* (1997). Diversity and depth-  
499 specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl Environ*  
500 *Microbiol* **63**: 63-70.  
501  
502 Fu L, Niu B, Zhu Z, Wu S, Li W (2012). CD-HIT: accelerated for clustering the next-generation  
503 sequencing data. *Bioinformatics* **28**: 3150-3152.  
504  
505 Giovannoni SJ, Rappe MS, Vergin KL, Adair NL (1996). 16S rRNA genes reveal stratified open  
506 ocean bacterioplankton populations related to the Green Non-Sulfur bacteria. *P Natl Acad Sci*  
507 *USA* **93**: 7979-7984.  
508  
509 Giovannoni SJ, Vergin KL (2012). Seasonality in ocean microbial communities. *Science* **335**:  
510 671-676.  
511  
512 Gordon DA, Giovannoni SJ (1996). Detection of stratified microbial populations related to  
513 Chlorobium and Fibrobacter species in the Atlantic and Pacific oceans. *Appl Environ Microbiol*  
514 **62**: 1171-1177.  
515  
516 Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ *et al.* (2012). Streamlining  
517 and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. *mBio*  
518 **3**: e00252-00212.  
519  
520 Haft DH, Selengut J, Mongodin EF, Nelson KE (2005). A guild of 45 CRISPR-associated (Cas)  
521 protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLOS*  
522 *Comput Biol* **1**: e60.  
523  
524 Hay NA, Tipper DJ, Gygi D, Hughes C (1999). A novel membrane protein influencing cell shape  
525 and multicellular swarming of *Proteus mirabilis*. *J Bacteriol* **181**: 2008-2016.  
526  
527 Ivars-Martínez E, Martín-Cuadrado A-B, D'Auria G, Mira A, Ferriera S, Johnson J *et al.* (2008).  
528 Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas*  
529 *macleodii* suggests alternative lifestyles associated with different kinds of particulate organic  
530 matter. *ISME J* **2**: 1194-1212.  
531  
532 Kappler U, Bennett B, Rethmeier J, Schwarz G, Deutzmann R, McEwan AG *et al.* (2000).  
533 Sulfite: Cytochrome c Oxidoreductase from *Thiobacillus novellus*. *J Biol Chem* **275**: 13202-  
534 13212.  
535

536 Kappler U, Davenport K, Beatson S, Lucas S, Lapidus A, Copeland A *et al.* (2012). Complete  
537 genome sequence of the facultatively chemolithoautotrophic and methylotrophic alpha  
538 Proteobacterium *Starkeya novella* type strain (ATCC 8093(T)). *Stand Genomic Sci* **7**: 44-58.  
539  
540 Karner MB, DeLong EF, Karl DM (2001). Archaeal dominance in the mesopelagic zone of the  
541 Pacific Ocean. *Nature* **409**: 507-510.  
542  
543 King GM, Smith CB, Tolar B, Hollibaugh JT (2013). Analysis of composition and structure of  
544 coastal to mesopelagic bacterioplankton communities in the northern gulf of Mexico. *Front*  
545 *Microbiol* **3**: 438.  
546  
547 Konstantinidis KT, Tiedje JM (2007). Prokaryotic taxonomy and phylogeny in the genomic era:  
548 advancements and challenges ahead. *Curr Opin Microbiol* **10**: 504-509.  
549  
550 Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009). Comparative Metagenomic Analysis of  
551 a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North  
552 Pacific Subtropical Gyre. *Appl Environ Microbiol* **75**: 5345-5355.  
553  
554 Lasken RS (2013). Single-cell sequencing in its prime. *Nat Biotechnol* **31**: 211-212.  
555  
556 Lauro FM, Bartlett DH (2008). Prokaryotic lifestyles in deep sea habitats. *Extremophiles* **12**: 15-  
557 25.  
558  
559 Madigan MT, Martinko JM, Parker J (2000). *Brock Biology of Microorganisms*, 9th edn.  
560 Prentice-Hall.  
561  
562 Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P *et al.* (2011).  
563 Evolution and classification of the CRISPR/Cas systems. *Nat Rev Micro* **9**: 467-477.  
564  
565 Martin-Cuadrado A-B, López-García P, Alba J-C, Moreira D, Monticelli L, Strittmatter A *et al.*  
566 (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLOS ONE* **2**:  
567 e914.  
568  
569 Morris R, Rappé M, Connon S, Vergin K, Siebold WA, Carlson CA *et al.* (2002). SAR11 clade  
570 dominates ocean surface bacterioplankton communities. *Nature* **420**: 806-810.  
571  
572 Morris RM, Longnecker K, Giovannoni SJ (2006). *Pirellula* and OM43 are among the dominant  
573 lineages identified in an Oregon coast diatom bloom. *Environ Microbiol* **8**: 1361-1370.  
574  
575 Morris RM, Frazar CD, Carlson CA (2012). Basin-scale patterns in the abundance of SAR11  
576 subclades, marine Actinobacteria (OM1), members of the Roseobacter clade and OCS116 in  
577 the South Atlantic. *Environ Microbiol* **14**: 1133-1144.  
578  
579 Nagata T, Tamburini C, Arístegui J, Baltar F, Bochkansky AB, Fonda-Umani S *et al.* (2010).  
580 Emerging concepts on microbial processes in the bathypelagic ocean – ecology,  
581 biogeochemistry, and genomics. *Deep-Sea Res II* **57**: 1519-1536.  
582  
583 Pester M, Schleper C, Wagner M (2011). The Thaumarchaeota: an emerging view of their  
584 phylogeny and ecophysiology. *Curr Opin Microbiol* **14**: 300-306.  
585

586 Quaiser A, Zivanovic Y, Moreira D, López-García P (2010). Comparative metagenomics of  
587 bathypelagic plankton and bottom sediment from the Sea of Marmara. *ISME J* **5**: 285-304.  
588  
589 Reinthaler T, van Aken HM, Herndl GJ (2010). Major contribution of autotrophy to microbial  
590 carbon cycling in the deep North Atlantic's interior. *Deep-Sea Res II* **57**: 1572-1580.  
591  
592 Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F *et al.* (2013). Insights  
593 into the phylogeny and coding potential of microbial dark matter. *Nature*.  
594  
595 Robbertse B, Yoder RJ, Boyd A, Reeves J, Spatafora JW (2011). Hal: an Automated Pipeline  
596 for Phylogenetic Analyses of Genomic Data. *PLOS Currents Tree of Life* **3**: RRN1213.  
597  
598 Robinson C, Steinberg DK, Anderson TR, Arístegui J, Carlson CA, Frost JR *et al.* (2010).  
599 Mesopelagic zone ecology and biogeochemistry—a synthesis. *Deep-Sea Res II* **57**: 1504-1518.  
600  
601 Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F  
602 *et al.* (2009). Explaining microbial population genomics through phage predation. *Nat Rev*  
603 *Microbiol* **7**: 828-836.  
604  
605 Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007a). The  
606 Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical  
607 Pacific. *PLOS Biol* **5**: e77.  
608  
609 Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007b). The  
610 Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical  
611 Pacific. *Plos Biol* **5**: e77.  
612  
613 Schattenhofer M, Fuchs BM, Amann R, Zubkov MV, Tarran GA, Pernthaler J (2009). Latitudinal  
614 distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environ Microbiol* **11**:  
615 2078-2093.  
616  
617 Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ (2010). The presence of the  
618 glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. *Environ*  
619 *Microbiol* **12**: 490-500.  
620  
621 Shi Y, Tyson GW, Eppley JM, DeLong EF (2011). Integrated metatranscriptomic and  
622 metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J* **5**: 999-  
623 1013.  
624  
625 Simonato F, Campanaro S, Lauro FM, Vezzi A, D'Angelo M, Vitulo N *et al.* (2006).  
626 Piezophilic adaptation: a genomic point of view. *J Biotechnol* **126**: 11-25.  
627  
628 Smedile F, Messina E, La Cono V, Tsoy O, Monticelli LS, Borghini M *et al.* (2013). Metagenomic  
629 analysis of hadopelagic microbial assemblages thriving at the deepest part of Mediterranean  
630 Sea, Matapan-Vavilov Deep. *Environ Microbiol* **15**: 167-182.  
631  
632 Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with  
633 thousands of taxa and mixed models. *Bioinformatics* **22**: 2688-2690.  
634  
635 Stamatakis A, Hoover P, Rougemont J (2008). A Rapid Bootstrap Algorithm for the RAxML Web  
636 Servers. *Syst Biol* **57**: 758-771.

637  
638 Stepanauskas R (2012). Single cell genomics: an individual look at microbes. *Curr Opin*  
639 *Microbiol*: 1-8.  
640  
641 Stewart FJ, Ulloa O, DeLong EF (2012). Microbial metatranscriptomics in a permanent marine  
642 oxygen minimum zone. *Environ Microbiol* **14**: 23-40.  
643  
644 Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D *et al.* (2011). Potential  
645 for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. *Science*  
646 **333**: 1296-1300.  
647  
648 Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ *et al.* (2011). Phylogenomic  
649 evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci Rep* **1**: 1-9.  
650  
651 Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA *et al.* (2009).  
652 Seasonality and vertical structure of microbial communities in an ocean gyre. *ISME J* **3**: 1148-  
653 1163.  
654  
655 Tripp HJ, Kitner JB, Schwalbach MS, Dacey JWH, Wilhelm LJ, Giovannoni SJ (2008). SAR11  
656 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452**: 741-744.  
657  
658 Tully BJ, Nelson WC, Heidelberg JF (2011). Metagenomic analysis of a complex marine  
659 planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* **14**: 254-267.  
660  
661 Varela MM, Van Aken HM, Herndl GJ (2008). Abundance and activity of Chloroflexi - type  
662 SAR202 bacterioplankton in the meso - and bathypelagic waters of the (sub) tropical Atlantic.  
663 *Environ Microbiol* **10**: 1903-1911.  
664  
665 Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004).  
666 Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.  
667  
668 Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappé MS, Giovannoni SJ (2007). High  
669 intraspecific recombination rate in a native population of *Candidatus Pelagibacter ubique*  
670 (SAR11). *Environ Microbiol* **9**: 2430-2440.  
671  
672 Vergin KL, Beszteri B, Monier A, Thrash JC, Temperton B, Treusch AH *et al.* (2013). High-  
673 resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by  
674 phylogenetic placement of pyrosequences. *ISME J*: 1-11.  
675  
676 Viklund J, Ettema TJG, Andersson SGE (2012). Independent Genome Reduction and  
677 Phylogenetic Reclassification of the Oceanic SAR11 Clade. *Mol Biol Evol* **29**: 599-615.  
678  
679 Vos M, Didelot X (2009). A comparison of homologous recombination rates in bacteria and  
680 archaea. *ISME J* **3**: 199-208.  
681  
682 Wang F, Wang J, Jian H, Zhang B, Li S, Wang F *et al.* (2008). Environmental Adaptation:  
683 Genomic Analysis of the Piezotolerant and Psychrotolerant Deep-Sea Iron Reducing Bacterium  
684 *Shewanella piezotolerans* WP3. *PLOS ONE* **3**: e1937.  
685

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

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

748

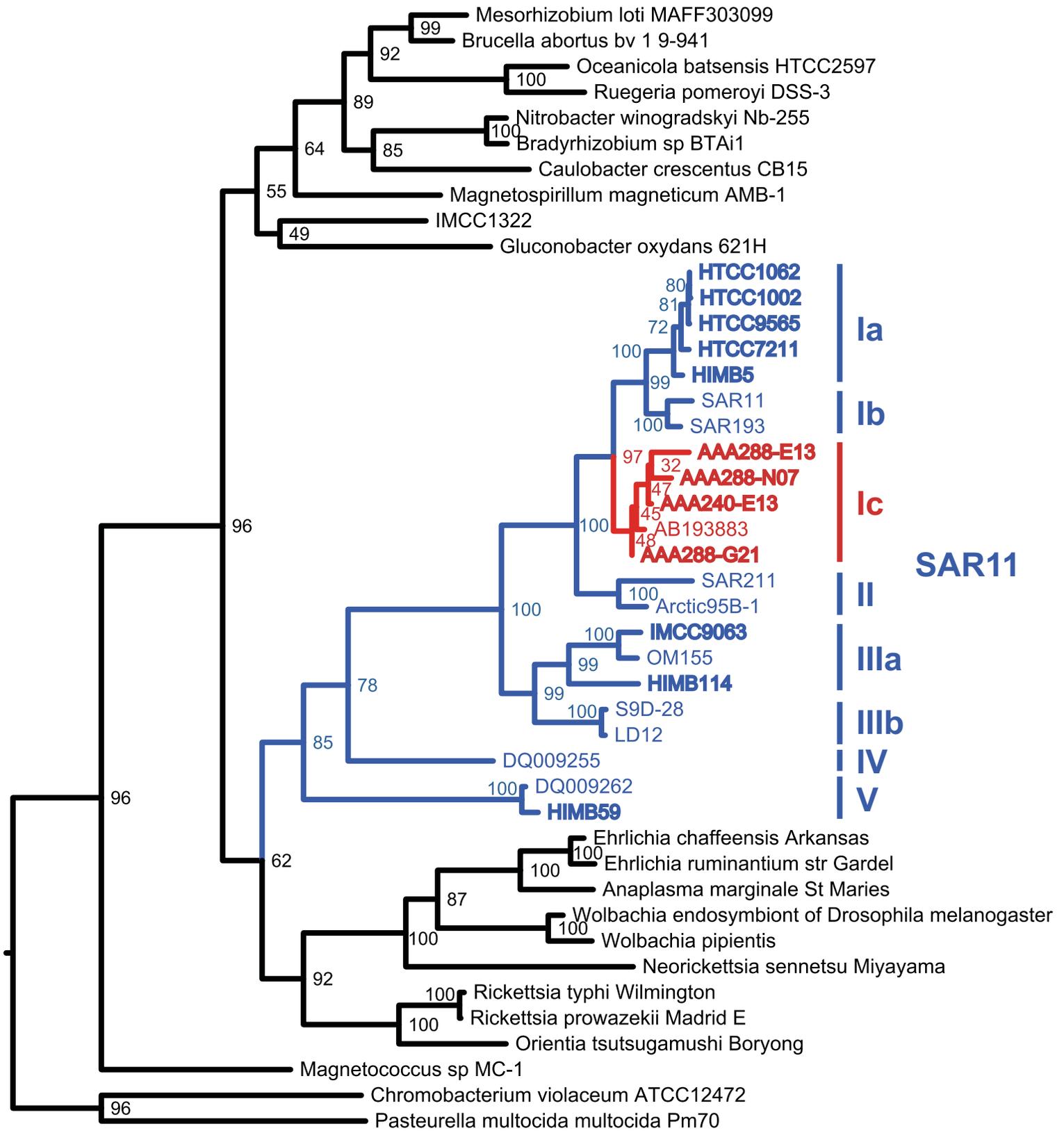
749 Figure 6. Fold-change in amino acid substitutions between the SAGs and the surface genomes.  
750 Pair-wise substitutions were quantified based on BLAST alignments of homologs between  
751 surface genomes and SAGs. X- unknown codons.

752

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

758

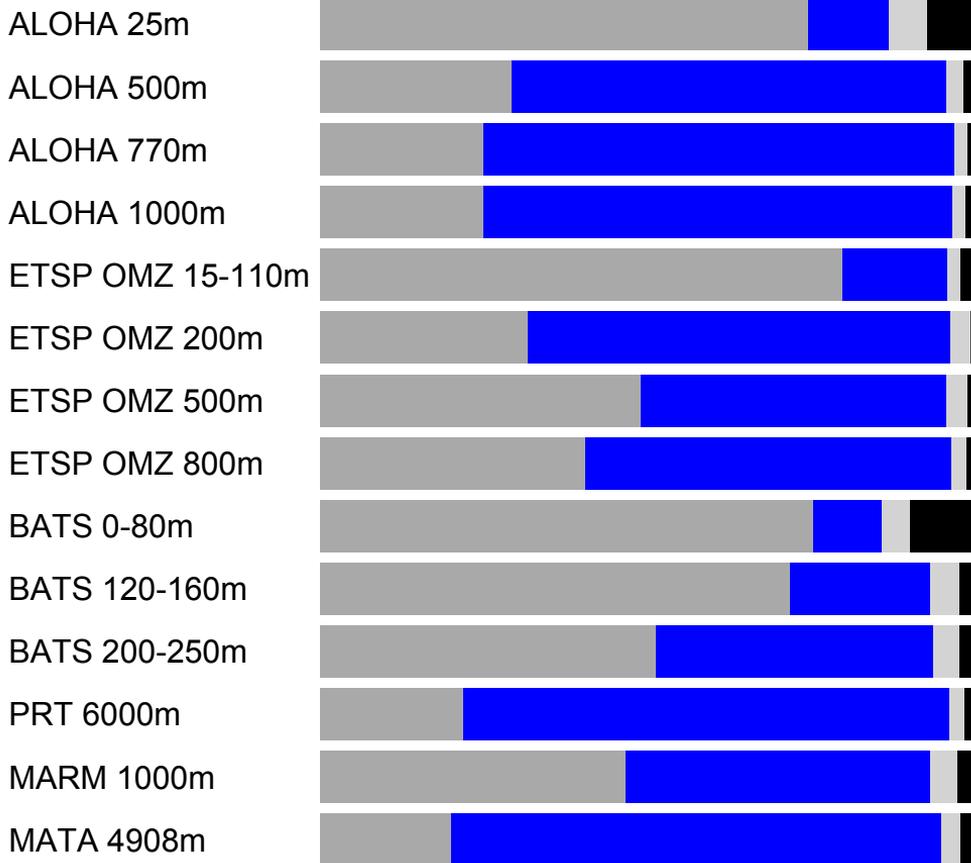
759 Figure 8. Plot of normalized mean vs. log-fold change for surface vs. deep gene clusters.

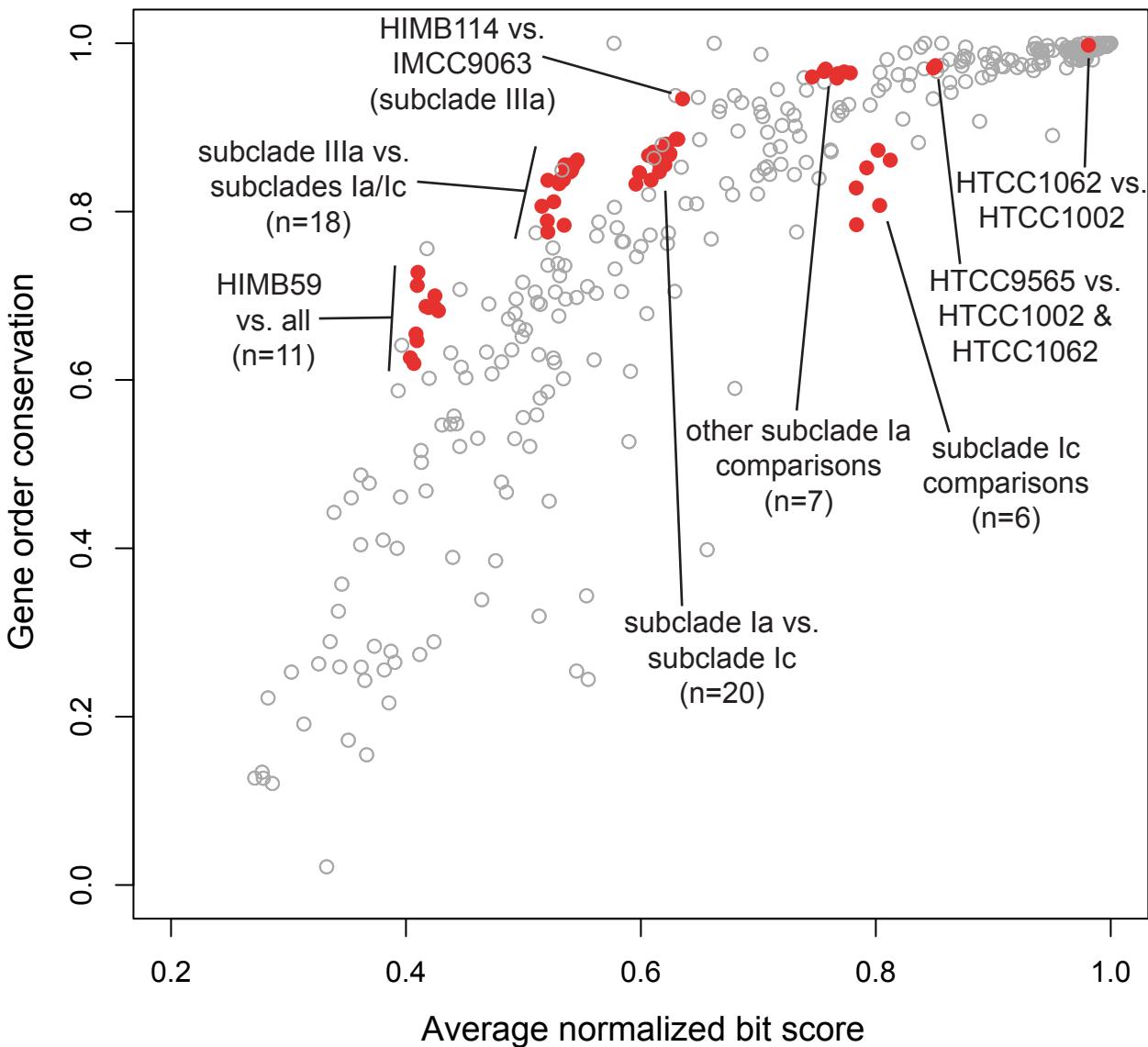


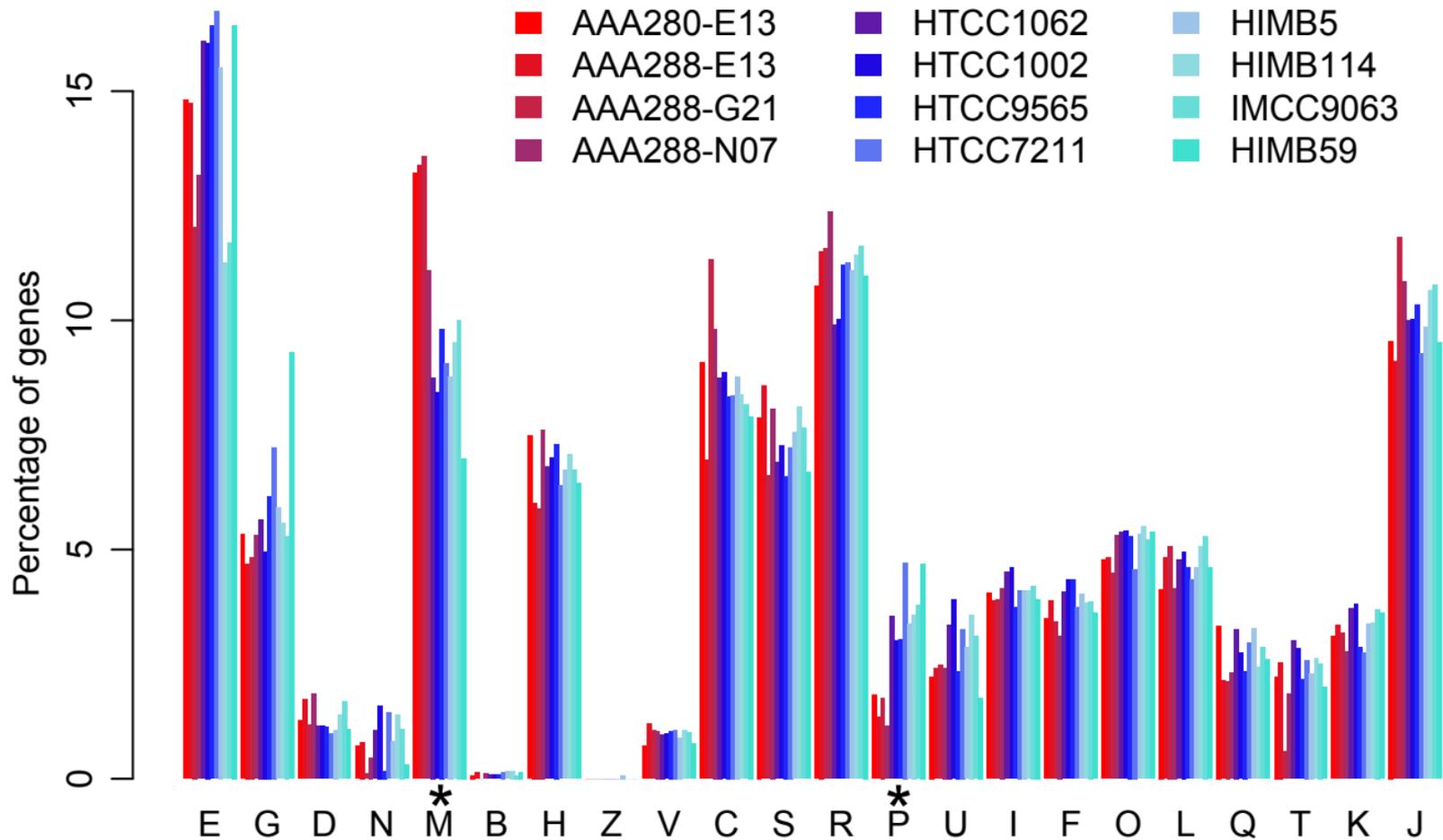
# Normalized aggregate reciprocal best blast hits (%)

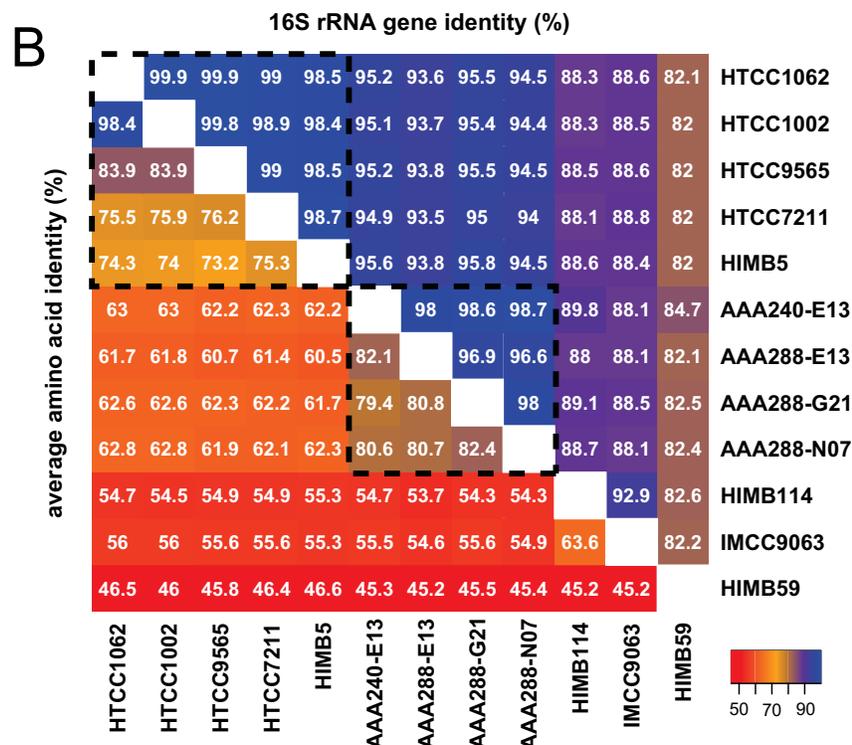
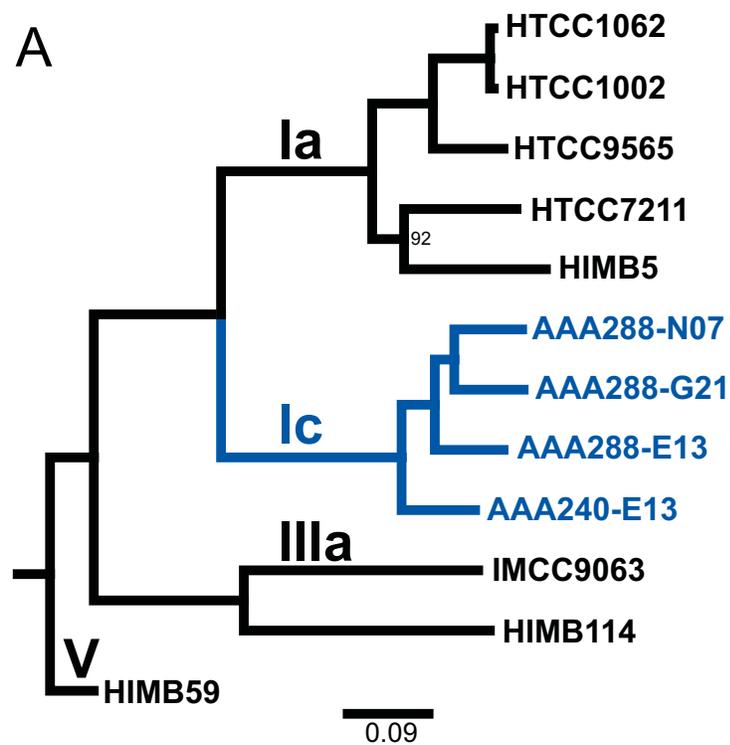


Ia
  Ic
  IIIa
  Va

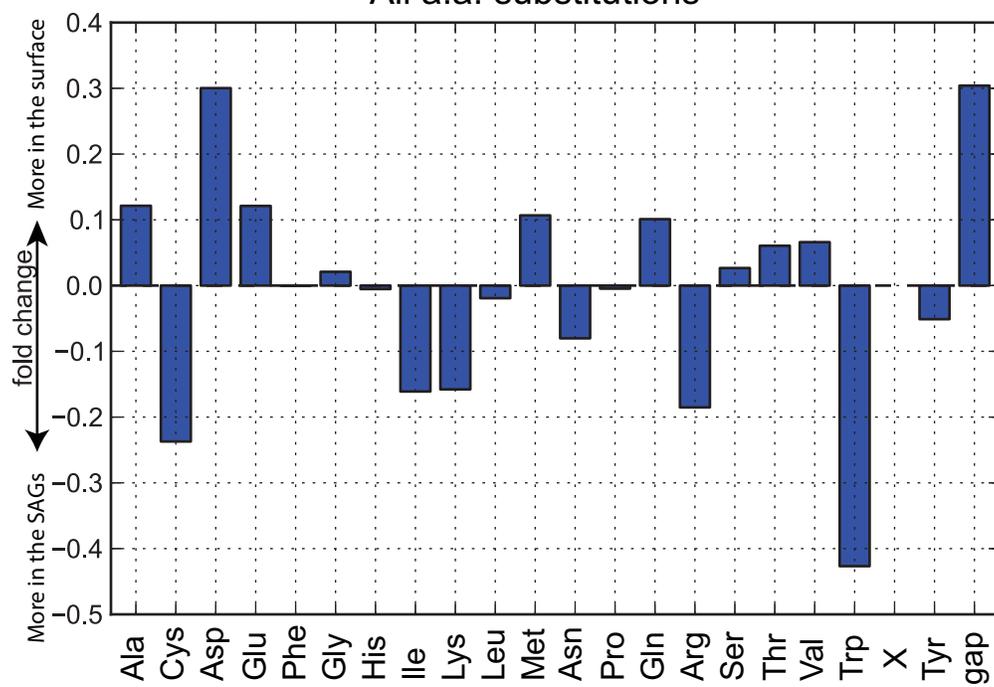




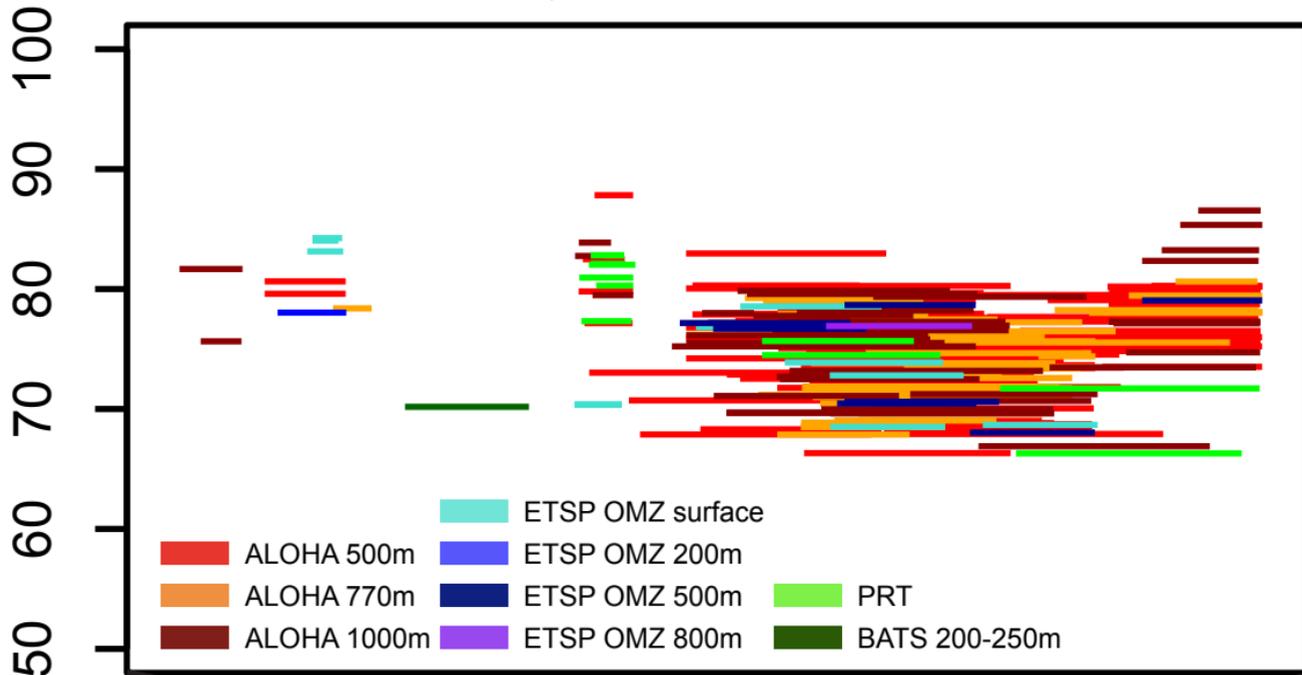




### All a.a. substitutions



# scaffold 14, bases 24207-26268

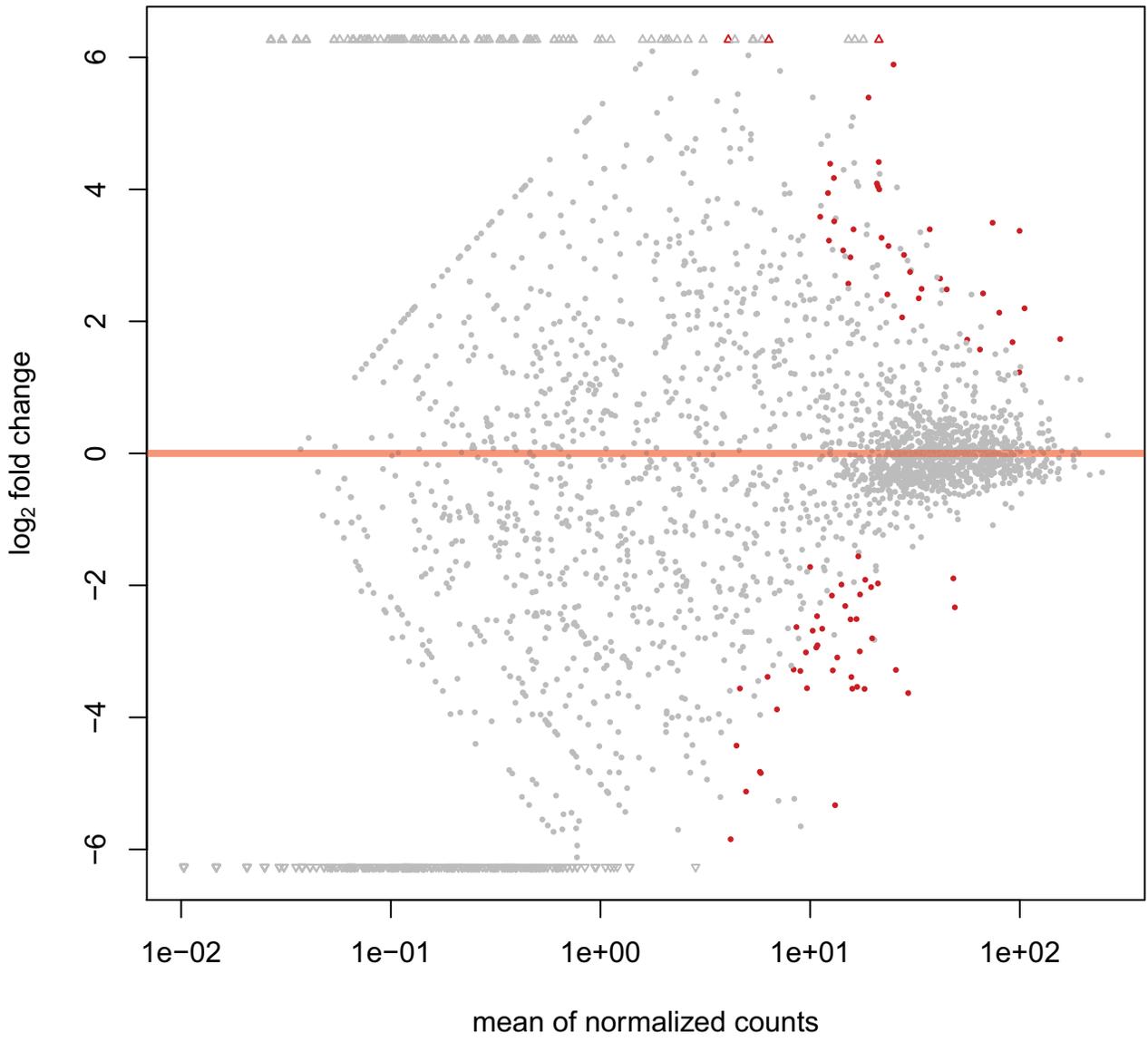


21456

24456

27456





## Tables

Table 1. Subclade Ic SAG genome characteristics

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

<sup>#</sup>Values from (Grote *et al.*, 2012) and IMG, \*actual (not estimated) sizes