Supplementary Information for

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

Supplementary information contains:

Supplementary Text
Supplementary Methods
Supplementary Figures S1-15
Supplementary Table_S1_F.xlsx
Additional supporting text files, including scripts, supporting text files and fasta files, and SAG genome data, are available on the Giovannoni Lab website at:
http://giovannonilab.science.oregonstate.edu/publications

Supplementary Text

Basic comparative genomics results

Table S1 details the results of our comparative genomics analysis using categorization of the 1764 orthologous clusters (OCs) found in the SAGs according to other SAR11 membership: those unique to a given SAG (unique), those shared by 2-4 SAGs but not by any other SAR11 genomes (shared Ic), those shared by all isolates and 1-3 SAGs (possible core), and those shared in all SAGs and all isolates (true core). All other distributions were not categorized. Each gene categorization and OC membership is cataloged in Table S1 and displayed in order along each scaffold in Figure S15. The incomplete nature of the SAGs makes it impossible to rule out the presence of an ortholog in the source cell and given the high percentage of conserved core genes among distantly related pure culture strains (Grote *et al.*, 2012), possible core orthologs were so termed because their presence in all isolate genomes and at least one SAG made these the most likely candidates for missing orthologs in the remaining SAGs. There were 172 true core (10%) and 480 possible core (27%) OCs. The 544 SAG-specific clusters contained 400 unique (23%) and 144 shared Ic (8%) clusters, and 568 OCs (32%) had some other shared distribution among SAGs and pure-culture genomes.

Negative data with regards to previously reported deep-ocean characteristics

Many characteristics previously associated with deep ocean microorganisms, whether obtained from cultivated strains (Lauro and Bartlett, 2008, Nagata *et al.*, 2010, Simonato *et al.*, 2006)(and references therein) or from cultivation-independent analyses (Konstantinidis *et al.*, 2009), were either not observed in the subclade Ic genomes, or were present, but shared with surface genomes. We examined amino acid substitution patterns among orthologs shared between the SAGs and the surface genomes (Konstantinidis *et al.*, 2009) (Figs. 6, S10). While the overall magnitude of the substitution differences was small, the general trend was for relative increases in cysteine, isoleucine, lysine, asparagine, arginine and tryptophan in the predicted subclade Ic protein sequences at the expense of alanine, aspartatic acid, glutamic acid, methionine, glutamine, threonine and valine. These trends mostly held true when comparing amino acid substitutions with biochemically similar or different residues (+ and - results in blast output, respectively). The exceptions were in the cases of cysteine and asparagine, which showed relative increases only when examining different and similar amino acid substitutions, respectively (Fig. S12). Increased arginine and tryptophan residues were reported in a 4000 m sample from station ALOHA (Konstantinidis *et al.*, 2009), yet many of the other depth-associated trends observed in that whole-community metagenome study were opposite our

more targeted SAR11 taxon analysis, for example, histidine and alanine were more frequently substituted in the surface strains compared to the SAGs, and the reverse was true for isoleucine and lysine.

We identified no unique genes for polyunsaturated fatty acid synthesis, no unique duplications in ATPases or cytochrome c oxidases, and no unique genes for complex polymer utilization, such as pullulan, chitin, or cellulose. AAA240-E13 had a unique predicted xylanase/chitin deacetylase, however an OC with the same annotation was found in the isolate genomes (Table S1). In spite of their isolation from the ALOHA OMZ and considerable recruitment of metagenomic sequences from the heart of the ETSP OMZ at 200 m, the SAGs did not encode a high affinity *cbb*$_3$-type cytochrome c oxidase, nor any genes for alternative terminal electron accepting processes, such as nitrate reductases. The SAG 16S rRNA genes didn't have insertions associated with some piezophilic strains (Lauro *et al.*, 2007) (Fig. S16). There were no unique *tonB*-type transporters, or *ompH/L* genes, and none of the SAGs had genes for the Strickland reaction other than thioredoxin, thioredoxin reductase, and a translation elongation factor (these three genes were not unique to the SAGs). The SAGs had putative cold shock protein genes and copies of the *groE/L* genes, but these were not unique to the Ic clade. AAA240-E13 and AAA288-E13 had a unique *pulJ* type II secretion component, but many of the surface strains had pillin and type II secretion genes. AAA288-N07 contained an intact aerobic carbon monoxide dehydrogenase (CODH) gene cluster, but this was shared with HIMB5 and HIMB114. Unlike the SAR324 and Arctic96-BD19 SAGs reported by (Swan *et al.*, 2011), none of the SAR11 SAGs contained unique autotrophy pathways. Finally, unlike the results of (Konstantinidis *et al.*, 2009), the SAGs showed no significant increases transposable elements compared to the surface SAR11s (Table S1).

Additional observations based on relative abundance of genes in metagenomic datasets

Qualitative assessment of additional clusters showed several additional genes of interest that, although were not statistically more abundant in deep samples compared to surface samples, nevertheless had considerable recruitment at depth compared to surface samples (Table S1). Among the additional genes highly abundant at depth that were also conserved among SAR11 genomes outside of the Ic subclade were predicted ABC-type amino acid transporters, TRAP transporters, glutamate and glutamine synthases, FeS cluster assembly genes, *sec* and twin arginine protein translocases, the *coxL* CODH subunit, subunits of the hetertetrameric sarcosine oxidase (Yilmaz *et al.*, 2011), thioredoxin reductase, and ferrodoxin (Table S1). The unique xylanase/chitin deacetylase in AAA240-E13 (2236673789) was more abundant in deep water

samples, however, the copy of this gene shared with other SAGs and the isolate genomes (2236673495) was also among the most abundant genes in surface samples (Table S1). Ribosomal proteins and respiratory proteins such as cytochromes, NADPH:quinone oxidoreductase, and F_0F_1-type ATP synthases were also highly abundant in deep samples, usually with much lower abundance in surface samples, indicating that these highly conserved SAR11 genes probably have enough nucleotide divergence to serve as clade-specific markers. There were also a number of highly conserved (possible or true core) hypothetical genes that showed high abundance with depth.

Additional phage information

The SAGs shared genes with other SAR11 genomes that had at least partial homology to Pelagiphage genes. Whole genome BLAST of the four recently sequenced Pelagiphage genomes (Zhao *et al.*, 2013) against the SAGs indicated significant hits > 100bp from two of the Pelagiphage, HTVC008M and HTVC019P, to true core and possible core genes such as NAD synthetase, peroxiredoxin, and ribonuclotide reductase, as well as flexible genome genes such as that encoding a small heat shock protein, ribosomal protein S21, and hypothetical proteins (Table S1).

Supplementary methods

Single-cell separation/ MDA

Water samples for single cell analyses were collected from the mesopelagic (770m) using Niskin bottles during Hawaii Ocean Time-series (HOT) Cruise 215, station ALOHA (22°45′N, 158°00′W; 9 September 2009). Replicate, 1 mL aliquots of water were cryopreserved with 6% glycine betaine (Sigma) and stored at −80ºC (Cleland *et al.*, 2004).

Prior to cell sorting, samples with prokaryote cell abundances above $5x10^5$ mL^{-1} were diluted 10x with filter-sterilized field samples and pre-screened through a 70 µm mesh-size cell strainer (BD). For heterotrophic prokaryote detection, diluted subsamples (1-3 mL) were incubated for 10-120 min with SYTO-9 DNA stain (5 µM; Invitrogen). Cell sorting was performed with a MoFlo™ (Beckman Coulter) flow cytometer using a 488 nm argon laser for excitation, a 70 µm nozzle orifice and a CyClone™ robotic arm for droplet deposition into microplates. The cytometer was triggered on side scatter. The “purify 1 drop” mode was used for maximal sort purity. Prokaryote cells were separated from eukaryotes, viruses, and detritus based on SYTO-9 fluorescence (proxy to nucleic acid content) and light side scatter (proxy to particle size)(del Giorgio *et al.*, 1996). *Synechococcus* cells were excluded, based on their autofluorescence

signal. Target cells were deposited into 384-well plates containing 600 nL per well of 1x TE and stored at −80ºC until further processing. Of the 384 wells, 315 were dedicated for single cells, 66 were used as negative controls (no droplet deposited) and 3 received 10 cells each (positive controls).

Cells were lysed and their DNA was denatured using cold KOH (Raghunathan *et al.*, 2005). Genomic DNA from the lysed cells was amplified using multiple displacement amplification (MDA) (Dean *et al.*, 2002, Raghunathan *et al.*, 2005) in 10 µL final volume. The MDA reactions contained 2 U/µL Repliphi polymerase (Epicentre), 1x reaction buffer (Epicentre), 0.4 mM each dNTP (Epicentre), 2 mM DTT (Epicentre), 50 mM phosphorylated random hexamers (IDT) and 1 µM SYTO-9 (Invitrogen) (all final concentration). The MDA reactions were run at 30°C for 12-16 h, then inactivated by a 15 min incubation at 65°C. Amplified genomic DNA was stored at -80°C until further processing. We refer to the MDA products originating from individual cells as single amplified genomes (SAGs).

Prior to cell sorting, the instrument and the workspace were decontaminated for DNA as previously described (Stepanauskas and Sieracki, 2007). High molecular weight DNA contaminants were removed from all MDA reagents by a UV treatment in Stratalinker (Stratagene). An empirical optimization of the UV exposure was performed to remove all detectable contaminants without inactivating the reaction. Cell sorting and MDA setup were performed in a HEPA-filtered environment. As a quality control, the kinetics of all MDA reactions was monitored by measuring the SYTO-9 fluorescence using FLUOstar Omega (BMG). The critical point (Cp) was determined for each MDA reaction as the time required to produce half of the maximal fluorescence. The Cp is inversely correlated to the amount of DNA template (Zhang *et al.*, 2006). The Cp values were significantly lower in 1-cell wells compared to 0-cell wells in all microplates for which MDA kinetics was monitored ($p<0.001$; Wilcoxon Two Sample Test).

MDA products were diluted 50-fold in TE buffer and 500 nL aliquots of diluted MDA product served as the template DNA in 5 µL final volume real-time PCR screens. All PCR reactions were performed using LightCycler 480 SYBR Green I Master Mix (Roche) and the Roche LightCycler® 480 II real-time thermal cycler. PCR amplification of SSU rRNA and metabolic genes from SAGs was done using primers and conditions listed in Table S1. Forward (5′−GTAAAACGACGGCCAGT−3′) and reverse (5′−CAGGAAACAGCTATGACC−3′) M13 sequencing primers were added to the 5′ ends of each target primer pair to aid direct sequencing of PCR products. All PCR reactions were run for 40 cycles at the appropriate annealing temperature, followed by melting curve analysis performed as follows: 95°C for 5 s,

52°C for 1 min, and a continuous temperature ramp (0.11°C/s) from 52 to 97°C. Real-time PCR kinetics and amplicon melting curves served as proxies for detecting SAGs positive for target genes. New, 20 µL PCR reactions were set up for all PCR-positive SAGs and amplicons were sequenced from both ends using Sanger technology by Beckman Coulter Genomics.

Single cell sorting, whole genome amplification, real-time PCR screens and PCR product sequence analyses were performed at the Bigelow Single Cell Genomics Center (www.bigelow.org/scgc).

Selection based on MDA reaction kinetics

Four SAGs in the Ic subclade were picked to cover the total depth of branching observed in the 16S phylogeny, and selected based on MDA reaction kinetics to improve the possibility of more complete genome recovery. The kinetics of all MDA reactions was monitored by measuring the SYTO-9 fluorescence using a FLUOstar Omega (BMG) plate reader. The MDA critical point (Cp) was determined for each reaction as the time required to produce half of the maximal fluorescence. The Cp value is inversely correlated to the amount of DNA template (Zhang *et al.*, 2006).

Sequencing/Assembly/Annotation

Draft genomes were generated at the DOE Joint genome Institute (JGI) using the Illumina technology (Béjà *et al.*, 2002). Illumina standard shotgun libraries were constructed and sequenced using the Illumina HiSeq 2000 platform. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov. All raw Illumina sequence data was passed through DUK, a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts (unpublished). The following steps were then performed for assembly: 1) filtered Illumina reads were assembled using Velvet v. 1.1.04 (Stein *et al.*, 1996), 2) 1–3 kbp simulated paired end reads were created from Velvet contigs using wgsim (http://github.com/lh3/wgsim), 3) Illumina reads were assembled with simulated read pairs using Allpaths–LG v. r41043 (Hsiao *et al.*, 2005). Parameters for assembly steps were: 1) Velvet: 63 -shortPaired and velvetg: -very clean yes -export -Filtered yes -min contig lgth 500 -scaffolding no -cov cutoff 10, 2) wgsim: -e 0 -1 100 -2 100 -r 0 -R 0 -X 0, 3) Allpaths: -LG PrepareAllpathsInputs: PHRED 64=1 PLOIDY=1 FRAG COVERAGE=125 JUMP COVERAGE=25 LONG JUMP COV=50, RunAllpathsLG: THREADS=8 RUN=std shredpairs TARGETS=standard VAPI WARN ONLY=True OVERWRITE=True.

Each raw sequence data set was screened against all finished bacterial and archaeal genome sequences (downloaded from NCBI) and the human genome to identify potential contamination in the sample. Reads were mapped against reference genomes with bwa version 0.5.9 (Philippot, 2002) using default parameters (96% identity threshold). None of the libraries showed significant contamination. Additionally, gene sequences of the final assemblies were compared against the GenBank nr database by BLASTX and taxonomically classified using MEGAN (Beaumont *et al.*, 2002). Genes were identified using Prodigal (Hyatt *et al.*, 2010). The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database (nr), UniProt, TIGRFam, Pfam, KEGG, COG, and InterPro databases. The tRNAScan-SE tool (Hacker and Kaper, 2000) was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA (Pruesse *et al.*, 2007). Other non–coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching genomes for the corresponding Rfam profiles using INFERNAL (Makarova *et al.*, 1999). Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) (Markowitz *et al.*, 2008) platform developed by the Joint Genome Institute, Walnut Creek, CA, USA (http://img.jgi.doe.gov).

Orthology with other SAR11s

SAG orthologs were determined using the automated phylogenomics pipeline Hal (Robbertse *et al.*, 2011) combined with several curation steps. Protein-coding fastas for 8 pure-culture SAR11 genomes were obtained from IMG (https://img.jgi.doe.gov). Initial orthologous clusters were established by all vs. all BLASTP and subsequent Markov clustering (MCL) at the inflation parameter (*I*) 1.5, consistently with our previous work (Grote *et al.*, 2012). These clusters were then filtered for genes with ≥ 30% difference to the median gene length of the cluster, similarly to (Grote *et al.*, 2012), using cluster_size_filter.pl iteratively until clusters no longer split. Because of the fragmentary nature of the SAGs, genes from these genomes were not subjected to the same size filter, as it has been shown that highly incomplete genomes contain large numbers of partial open reading frames (ORFs) (Klassen and Currie, 2012). A separate evaluation of orthologs, based on synteny, was conducted using the methods of (Yelton *et al.*, 2011). Syntenic genes from all pairwise comparisons were coalesced into clusters using synteny_clustering.py, and these clusters were compared with those from the Hal pipeline using cluster_comparison.py. In cases where the syntenic clusters contained additional genes to those in the Hal clusters, the cluster assignment based on synteny was used. Hypothetical

proteins associated with extraneous, unused scaffolds from HTCC1002 (scaffolds 3-5) (Grote *et al.*, 2012) were manually removed. Finally, orthologous clusters for single-copy housekeeping genes were manually inspected to ensure no erroneous mis-assignment during these automated steps. As part of the Hal pipeline, cluster data is output in the form of a heatmap. This heatmap was manually curated to reflect the changes to the original clustering based on our filters. This heat map can be queried to establish numbers of shared orthologs between different genomes. Table S1 and Figure S15 display five gene categories: 1) orthologs unique to an individual SAG (unique), 2) orthologs shared among SAGs but not other SAR11 strains (shared Ic), 3) orthologs shared among all pure culture SAR11 genomes and between one and three SAGs (potential core), 4) orthologs shared among all SAR11 pure culture genomes and all SAGs (true core), 5) everything else. Note that many clusters contain more than one gene from a given genome. These may be considered paralogs for the purposes of this analysis, the same as in (Grote *et al.*, 2012), and paralogs were ignored for the purposes of counting shared orthologs across clusters. However, caution should be used in interpreting duplicate genes within clusters for two reasons: 1) due to the incomplete nature of the SAGs and the potential abundance of partial or split ORFs across multiple scaffolds, duplicate genes from SAGs in a given cluster may not reflect “real” duplicates, and 2) given the clustering is the result of a partially automated pipeline and not hand-curated for each genome, the possibility remains that clusters may be artificially concatenated due to the inflation parameter selected for Markov Clustering.

Post-assembly SAG scaffold QC

Amplification of genomes from single cells increases the risk of contaminant DNA being included in the downstream genome sequencing and subsequent assembly compared to traditional sequencing of microbial genomes using billions-trillions of chromosomal copies. Although pre- and post-amplification QC has very high success with decreasing such contamination, sequenced SAG scaffolds were evaluated for possible cases of contamination using BLASTN, tetramer analysis, and manual inspection all scaffolds smaller than 1000bp for regions of low predicted ORF density. BLASTN of each scaffold was performed against the nr database with default settings. Scaffolds were flagged for further evaluation if they contained BLASTN hits with > 500bp alignment length to non-alphaproteobacteria, or if all BLASTN hits were to non-alphaproteobacteria. Tetramer analysis of each scaffold was carried out using a sliding window of 2000bp and a 200bp step. Principal Components Analysis (PCA) was carried out on the results to identify outlying scaffolds/regions of scaffolds, and these were flagged for

further evaluation. Those scaffolds flagged from the BLASTN and tetramer analyses were manually inspected for gene conservation among the other SAR11 strains. Those containing genes conserved in other strains were ruled out as contaminants. Finally all scaffolds smaller than 1000bp were manually inspected for predicted ORF density. The streamlined nature of SAR11 genomes results in very small non-coding regions throughout the genome, so scaffolds were visually screened based on coding density. All flagged scaffolds resulting from these analyses are identified in Table S1. However, it should be noted that a scaffold with BLASTN hits to non-alphaproteobacteria or outlying tetramer scores is not sufficient to identify contamination as these may simply reflect hypervariable regions within the SAG and/or instances of horizontal gene transfer. Therefore, scaffolds flagged as possible contaminants are to be treated with caution.

Circos plot

Circos plots were generated using Circos version 0.62-1. Scaffolds for each SAG were placed in order of size from largest to smallest. IMG annotations and OC designations were used as the base categorizations for the plot. GC content for each gene is represented as displayed in IMG. Links files were produced using a custom python script to organize segment duplications by OC designations using previously created tab-delimited files. The python script, Circos configuration files, and input data are available upon request.

Estimated genome completion

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 genes found in the SAGs (Table S1).

AAI vs. 16S identity

Average amino acid identity was calculated with the scripts/methods of (Yelton *et al.*, 2011). Pairwise 16S rRNA gene identity was calculated with megablast using default settings. All data is available in Table S1. The heatplot of AAI vs. 16S was made in R.

COG distribution

The barplot (Fig. 4) and boxplot (Fig. S9) of the COG distribution among SAR11 genomes was built using data supplied by IMG (Table S1) with R.

Amino acid distribution

Amino acid substitution pattern analysis was performed as described in Konstantinidis et al. 2009 (Konstantinidis *et al.*, 2009). Briefly, the fasta file for each HAL cluster was used as a BLASTP query against a database containing non-redundant proteins from 8 surface SAR11 strains and 4 SAGs from 770m at ALOHA. Alignments with an expect value of less than 1×10^{-20} were discarded. For remaining alignments, the number of similar amino acid substitutions, different amino acid substitutions and gaps was calculated between all surface strain proteins and SAG proteins. The fold-change in abundance of each amino acid between surface strain proteins and SAG proteins was calculated. All analyses are available in the IPython notebook 'Calculating amino acid changes.ipynb'.

Intergenic spacer analysis

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 with histograms and box plots, and R was also used to run the Wilcoxon rank sum analysis:

```
> wilcox.test(hist$surface,histr$Ic,paired=FALSE,conf.int=TRUE,conf.level=0.95)
```

All analyses only included non-zero intergenic regions calculated by IMG.

Transposable elements

Transposable elements were assessed using the sequences collected by Brian Haas of the Broad Institute for the program TransposonPSI (http://transposonpsi.sourceforge.net). Sequences from this library (1537) were searched against a database of SAR11 genomes (see metagenomic reciprocal best blast, below) using tblastn on default settings.

Deep water adaptation genes

Genes identified as potential deep ocean adaptations were searched for by inspection of the SAG annotations, inspection of KEGG pathways on IMG, and BLASTP of homologs for the various features against the SAR11 genomes. The fasta file containing these homologs is available in Supplemental Information. The presence of 16S rRNA gene insertions was tested by aligning all SAR11 16S sequences with those of known piezophiles carrying the insertions (Lauro *et al.*, 2007) using MUSCLE (Edgar, 2004), and manually inspecting the alignment.

CRISPR region analysis and cas gene search

CRISPRs are annotated by IMG using the CRISPR recognition tool and PILER (see http://img.jgi.doe.gov for more details). The designated AAA240-E13 CRISPR region was further analyzed by examining reciprocal best BLAST hits (see "Metagenomics," below) to those coordinates (scaffold 14, bases 24207-26268). Sequences were identified as having recruited to the corresponding coordinates of the single-scaffold AA240-E13 genome (Table S1) by use of metagenome_index.pl. To build the local recruitment plot in Figure 7, these sequences were aligned to the intergenic/CRISPR region with BLASTN with low-complexity filtering disabled. Additionally, the entire region was searched against the IMG v400 custom protein database with BLASTX on default settings to identify potential protein-coding regions. HMMs developed for a wide range of cas protein families were developed by Haft *et al.* (Haft *et al.*, 2005) and Makarova *et al.* (Makarova *et al.*, 2011), and made available through TIGRFAM (http://www.jcvi.org/cgi-bin/tigrfams/index.cgi). Each of the 46 HMMs from *Haft et al.* Table 1 and the 32 additional HMMs from Makarova *et al.* Table S4 were searched against the protein-coding sequences from all four SAGs using hmmsearch (part of the HMMER3 package (Eddy, 2011)) on default settings. Only 8 HMMs returned any hits to the SAGs. Of those, only three had hits in AAA240-E13. TIGR00372 had hits to hypothetical proteins in orthologous cluster 15001317. TIGR01587 had hits to type II DNA helicases conserved in all sequenced SAR11 genomes. TIGR03117 had hits to a RecG-like helicase, also conserved in all SAR11 genomes. The remaining HMMs with hits in SAR11 had low scores and small alignment regions. All positive search results are provided with our additional files.

Phylogenetics

SAG 16S rRNA gene sequences were combined with those of SAR11 strains in previously established subclades (Grote *et al.*, 2012, Morris *et al.*, 2005, Vergin *et al.*, 2013) (SAG_reference.fasta). Sequences were aligned with MUSCLE (Edgar, 2004) using default settings and the tree was created with RAxML (Stamatakis *et al.*, 2008). Thirty seven alphaproteobacterial and two outgroup 16S rRNA sequences were aligned with the subclade Ic clone sequence from (Vergin *et al.*, 2013) and those of the sequenced Ic SAGs. AAA240-E13 had two partial 16S rRNA gene fragments after assembly, and the larger of the two was used. Alignments were performed with MUSCLE using default settings (Edgar, 2004), poorly aligned positions were removed with Gblocks (Castresana, 2000) using the following settings:

–b1=(seqs/2)+1 –b2=(seqs/2)+1 –b3=seqs/2 –b4=2 –b5=h

The final tree was computed using RAxML (Stamatakis, 2006) using the following settings:

raxmlHPC-PTHREADS -x 1234 -T 4 -f a -m GTRGAMMA -\# 1000

The phylogenomic tree of SAR11 strains was calculated as part of the Hal pipeline (Robbertse *et al.*, 2011, Thrash *et al.*, 2011), in this case, using liberal Gblocks settings, maximum-likelihood (RAxML), and 10% allowed missing data, which resulted in a concatenated alignment of 322 protein-coding genes (84,355 sites).

Because of the small length of the proteorhodopsin gene and difficulty with the alignments using MUSCLE (data not shown), the combined, iterative alignment/phylogeny program HandAlign was utilized for the proteorhodopsin phylogeny (Westesson *et al.*, 2012). Homologs were selected based on a BLASTP of the proteorhodopsin orthologs (cluster 1500536) in SAR11 against the IMG v350 protein gene database using default settings. Large gene homologs were removed using the cluster_size_filter.pl script:

$ perl ~/scripts/cluster_size_filter.pl rhodopsin.tab 70 130

Eukaryotic rhodopsin homologs were removed manually as well as those from HTCC2255/HTCC2999 as these constituted potential contaminants. The tree was calculated using the following commands:

$ handalign clean_rhodopsins3.faa --hmmoc-root hmmoc -l 255 -ts 2000 -af clean_rhodopsins3.trace.stk -ub > clean_rhodopsins3.MAP.stk

$ stocktree.pl clean_rhodopsins3.MAP.stk > clean_rhodopsins3.MAP.newick

Based on the results of (Grote *et al.*, 2012), we knew distinction of the sarcosine dehydrogenase (sardh) and dimethylglycine dehydrogenase (dmgdh) from within the same OC (cluster 150013) required phylogenetic analysis. This was done using the bash script protPipeline3, which automates a phylogenetic pipeline including alignment with MUSCLE (Edgar, 2004), Gblocks editing of poorly aligned sites (Castresana, 2000), ProtTest amino acid substitution modeling (Abascal *et al.*, 2005), and maximum-likelihood tree construction using RAxML (Stamatakis, 2006). The clusters formed based on the tree were designated 1500013.f.ok (dmgdh), 150013.f1.ok, 1500013.f2.ok, and 1500013.f3.ok (sardh).

All unaligned fasta files for each of the single gene trees and the super alignment and model file for the concatenated protein tree provided in Supplemental Information. All tree labels were manipulated using the Newick utilities (Junier and Zdobnov, 2010).

Metagenomics

Seven samples were collected at the BATS site (31° 40’ N, 64° 10’ W) during August 19, 2002 (0, 40, 80, 120, 160, 200, and 250 m). To collect the microbial biomass, ~ 70 L of seawater was filtered through 0.2-µm polyethersulfone membranes (Supor, Pall, East Hills, NY, USA) using McLane Research Laboratories, Inc. (East Falmouth, MA) in situ water transfer systems and

membranes were stored at -80°C until further processing. Nucleic acids were extracted and purified as described (Morris *et al.*, 2005). Library construction and shotgun sequencing were performed using a first-generation GS-FLX protocol (454 Life Sciences), yielding average read lengths of 226 bp. BATS metagenomes and corresponding geochemical metadata are available on CAMERA (http://camera.calit2.net) under the project ID CAM_PROJ_BATS. Data was also analyzed from 454 metagenomic sequences collected from Station ALOHA (Shi *et al.*, 2011), the ETSP OMZ (Stewart *et al.*, 2012), Puerto Rico Trench (Eloe *et al.*, 2011a), the Sea of Marmara (Quaiser *et al.*, 2010), and the Matapan-Vavilov Deep in the Mediterranean Sea (Smedile *et al.*, 2013). All 454 metagenomic datasets were quality trimmed using lucy v1.2 (http://sourceforge.net/projects/lucy/) using a minimum good sequence length of 100bp and default error settings. Trimmed reads were dereplicated using CD-HIT-454 v4.6.1 (https://code.google.com/p/cdhit/) with the following parameters: (-c 1.00 –n 8). Sanger sequences from ALOHA (Konstantinidis *et al.*, 2009), and the GOS (Rusch *et al.*, 2007, Venter *et al.*, 2004), were also subjected to the reciprocal best blast protocol, below, but without normalization due to their difference in read size. GOS sequences were first separated according to sample, and individual databases were created for each; samples with FILTER_MIN > 0.22 µm were excluded, as well as a single sample MOVE858 where FILTER_MAX was 0.22 µm. GOS surface sequences from Antarctica and the Southern Ocean (Brown *et al.*, 2012) (Table S1) were dereplicated as above, but without quality trimming (no fastq files were available). For these samples, all available size fractions were combined for each sample. These were also excluded from gene centric normalization.

Comparative recruitment of metagenomic sequences was completed using a reciprocal best BLAST (rbb). A total of 12 SAR11 genomes were included in the rbb- eight pure culture genomes (HTCC1062, HTCC1002, HTCC9565, HTCC7211, HIMB5, HIMB114, IMCC9063, HIMB59), and the four SAR11 SAGs. All incomplete SAR11 genome sequences were artificially concatenated in the order presented by IMG, with a string of nine “N” at each scaffold break for later identification of boundaries. The concatenated SAR11 genome sequences were searched against each metagenome database with BLASTN on default settings. All hits to SAR11 genomes were then searched against the entire IMG database (v400), which had been customized to contain only these 12 SAR11 genome sequences (others have been completed and deposited since this analysis began) using BLASTN on default settings except –max_target_seqs, set to 100000000, to allow for inclusion of all possible hits. Amino acid and nucleotide versions of this database, IMG v400 custom, were used in the average genome size and relative abundance calculations, respectively, below. Best hits from the second BLASTN

(reciprocal best hits- rbh) to the 12 SAR11 sequences were plotted with fragment_recruiter.r. Localization of each rbh to each gene in the SAR11 genomes was completed using the script metagenome_index.pl. Gene coordinates were established for each genome fasta using the gene coordinate workflow available in Supplemental Information. The gene coordinates for the concatenated SAGs is recorded in Table S1 under the tabs Start Coord (single) and End Coord (single). For relative abundance estimates of SAR11 in a given dataset, each set of 454 metagenomic sequences by sample was searched against the entire IMG v400 custom genome database using megablast with default settings. Best hits were then counted as either SAR11-like or non SAR11-like, based on whether they recruited to any of the 12 SAR11 genomes or not, respectively.

Investigating enrichment of SAR11 gene clusters at depth in metagenomic samples

A complete description of methods can be found in the supplementary file 'Statistical_analyses.pdf'. This document was generated with a runnable Sweave script, which is available on the Giovannoni GitHub repository (git@github.com:giovannoni-lab/bathytype.git). Briefly, metagenomes were classified as 'surface' (<200m) or 'deep' (≥200m) and the abundance of gene clusters as determined by reciprocal best-BLAST (see above) in each dataset was calculated. To identify statistically significant differences in gene cluster abundance between deep and surface samples, the R package DESeq (Anders & Huber, 2010) was used to correctly account for over-dispersion typically associated with gene abundances in high throughput sequencing studies. To validate the *a posteriori* classification of 'deep' and 'surface' samples, blind estimation of dispersion followed by clustering analysis of samples was performed. This analysis showed the MATA, MARM, PRT and CHOMZ.800 samples did not cluster correctly with other 'deep' samples. Therefore, DESeq analysis was performed both against the complete dataset and a subset with samples MATA, MARM, PRT and CHOMZ.800 removed.

Investigating SAR11 Ic subclade abundance as a function of temperature

A complete description of methods can be found in the supplementary file 'Statistical_analyses.pdf'. This document was generated with a runnable Sweave script, which is available on the Giovannoni GitHub repository (git@github.com:giovannoni-lab/bathytype.git). To determine if increased SAR11 Ic abundance at depth was a result of an adaptation to lower temperatures, SAR11 Ic abundance and total SAR11 abundance was calculated in 91 GOS samples by reciprocal best-BLAST (see above). Negative binomial regression GLM analysis

was then applied to the data, using the total SAR11 counts as an offset. As the data consisted of different sequencing types (Sanger and 454), the sequencing type was included as a factor in the model. Model terms were sequentially dropped to test each term for significance. Sequencing type was shown to not be a significant regression factor and so was dropped from the model. The resultant model was then compared to the null model and the Nagelkerke coefficient of determination (Nagelkerke, 1991) was calculated.

References

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

Beaumont HJE, Hommes NG, Sayavedra-Soto LA, Arp DJ, Arciero DM, Hooper AB *et al.* (2002). Nitrite reductase of *Nitrosomonas europaea* is not essential for production of gaseous nitrogen oxides and confers tolerance to nitrite. *J Bacteriol* **184:** 2557–2560.

Béjà O, Koonin EV, Aravind L, Taylor LT, Seitz H, Stein JL *et al.* (2002). Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl Environ Microb* **68:** 335–345.

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.

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

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

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P *et al.* (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* **99:** 5261-5266.

del Giorgio PA, Bird DF, Prairie YT, Planas D (1996). Flow cytometric determination of bacterial abundance in lake plankton with the green nucleic acid stain SYTO 13. *Limnol Oceanogr* **41**.

Dobbek H, Gremer L, Meyer O, Huber R (1999). Crystal structure and mechanism of CO dehydrogenase, a molybdo iron-sulfur flavoprotein containing S-selanylcysteine. *Proc Natl Acad Sci USA* **96:** 8884-8889.

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, Malfatti F, Gutierrez J, Hardy K, Schmidt WE, Pogliano K *et al.* (2011b). Isolation and characterization of a psychropiezophilic alphaproteobacterium. *Appl Environ Microbiol* **77:** 8145-8153.

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.

Hacker J, Kaper JB (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54:** 641–679.

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.

Hsiao WWL, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FSL (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLOS Genet* **1:** e62.

Hyatt D, Chen G-L, LoCascio P, Land M, Larimer F, Hauser L (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11:** 119.

Junier T, Zdobnov EM (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26:** 1669-1670.

Klassen JL, Currie CR (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* **13:** 14.

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.

Lauro FM, Chastain RA, Blankenship LE, Yayanos AA, Bartlett DH (2007). The unique 16S rRNA genes of piezophiles reflect both phylogeny and adaptation. *Appl Environ Microbiol* **73:** 838-845.

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

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI *et al.* (1999). Comparative genomics of the *Archaea* (*Euryarchaeota*): Evolution of conserved protein families, the stable core, and the variable shell. *Gen Res* **9:** 608–628.

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.

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D *et al.* (2008). IMG/M: a data management and analysis system for metagenomes. *Nucl Acids Res* **36:** D534-538.

Morris R, Vergin K, Cho J, Rappe M, Carlson C, Giovannoni S (2005). Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. *Limnol Oceanogr* **50:** 1687-1696.

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.

Nagelkerke NJD (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78:** 691-692.

Philippot L (2002). Denitrifying genes in bacterial and Archaeal genomes. *BBA-Gene Struct Expr* **1577:** 355–376.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucl Acids Res* **35:** 7188–7196.

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.

Raghunathan A, Ferguson Jr HR, Bornarth CJ, Song W, Driscoll M, Lasken RS (2005). Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* **71:** 3342-3347.

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

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

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.

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996). Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178:** 591–599.

Stepanauskas R, Sieracki ME (2007). Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* **104:** 9052–9057.

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.

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, 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.

Westesson O, Barquist L, Holmes I (2012). HandAlign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. *Bioinformatics* **28:** 1170-1171.

Yelton AP, Thomas BC, Simmons SL, Wilmes P, Zemla A, Thelen MP *et al.* (2011). A Semi-Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and Poorly Annotated Bacterial and Archaeal Genes. *PLOS Comput Biol* **7:** e1002230.

Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G *et al.* (2011). The genomic standards consortium: bringing standards to life for microbial ecology. *ISME J* **5:** 1565-1567.

Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW *et al.* (2006). Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24:** 680-686.

Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC *et al.* (2013). Abundant SAR11 viruses in the ocean. *Nature* **494:** 357-360.

Supplementary Figures

Figure S1. Maximum-likelihood tree of 16S rRNA genes for the SAR11-type SAGs used for initial selection of SAGs for genome sequencing. Outgroup is *R. prowazekii*; bootstrap values (n=100) are indicated at the nodes; scale bar indicates 0.3 changes per position.

Figure S2. Recruitment plots of reciprocal best hit metagenomic sequences for Station ALOHA, by genome. Colors indicate depth. Y-axes: % identity; x-axes, genome length.

Figure S3. Recruitment plots of reciprocal best hit metagenomic sequences for the Eastern Tropical South Pacific Oxygen Minimum Zone, by genome. Colors indicate depth. Y-axes: % identity; x-axes, genome length.

Figure S4. Recruitment plots of reciprocal best hit metagenomic sequences for BATS, by genome. Colors indicate depth. Y-axes: % identity; x-axes, genome length.

Figure S5. Recruitment plots of reciprocal best hit metagenomic sequences for the Puerto Rico Trench, by genome. Y-axes: % identity; x-axes, genome length.

Figure S6. Recruitment plots of reciprocal best hit metagenomic sequences for the Sea of Marmara, by genome. Y-axes: % identity; x-axes, genome length.

Figure S7. Recruitment plots of reciprocal best hit metagenomic sequences for the Matapan-Vavilov Deep, by genome. Y-axes: % identity; x-axes, genome length.

Figure S8. Recruitment plots of reciprocal best hit Sanger sequenced metagenomic sequences for Station ALOHA at 4000 m, by genome. Y-axes: % identity; x-axes, genome length.

Figure S9. Boxplot of COG distributions between surface (red) and subclade Ic genomes (“sag”, blue). Boxes indicate the upper and lower quartile, with the median as a black bar, error bars indicate adjacent values, and open circles indicate outside values. Orange boxes highlight COG categories with non-overlapping distributions.

Figure S10. Relative abundance of amino acids based on their classification as similar or different amino acid substitutions.

Figure S11. Analysis of intergenic space. Distributions of intergenic spaces are plotted in histograms for A) all surface genomes and B) all Ic genomes, with an x-axis limit of 1000. Box plots of the same data, C) including outliers and D) excluding outliers to more easily view the median values. Wilcoxon rank sum results are reported in C.

Figure S12. Putative partial pathway for purine degradation in subclade Ic (taken from KEGG). Enzymes for each transition are labeled, along with the corresponding OC from this analysis, and colored according to the key. The uncolored S-allantoin amidohydrolase indicates no matching annotations in any of the genomes. Gene organization on AAA240-E13 is depicted in the upper right. Note that the xanthine dehydrogenase clustered with the aerobic carbon monoxide dehydrogenase large subunit genes (Table S1), but we expect this was because they are part of a larger molybdenum hydroxylase family (Dobbek *et al.*, 1999), and not true orthologs in this case. Although present in AAA288-E13 and AAA288-N07, the adenosine deaminase was missing in AAA240-E13 and therefore not included in the gene neighborhoods.

Figure S13. Relative abundance of orthologous clusters (OCs) by depth and location according to number of reciprocal best blast hits in that sample. Each heat plot contains all OCs found in the pan-genome (x-axis) of the 12 genomes analyzed (y-axis). Colors (blue to white to red) indicate increasing relative abundance, according to each plot-specific scale. Relative abundance of OCs for each genome was used to calculate Bray-Curtis dissimilarities and hierarchically clustered using unweighted pair group method with arithmetic mean (UPGMA). All descendent links below a cluster node k share a color if k is the first node below the cut threshold of 0.7 × maximum cluster difference (default matplotlib/MATLAB behavior for dendrograms). Red boxes highlight samples where the sublclade Ic SAGs grouped together hierarchically based on gene recruitment patterns. ALOHA- Station ALOHA, ESTP OMZ- Eastern Subtropical Pacific Oxygen Minimum Zone, BATS- Bermuda Atlantic Time-series Study site, MARM- Sea of Marmara, PRT- Puerto Rico Trench, MATA- Matapan-Vavilov Deep.

Figure S14. Bayesian phylogenetic analysis of proteorhodopsin sequences using HandAlign. SAR11 proteorhodopsins are highlighted in red, along with subclade designations. Scale bar indicates changes per position.

Figure S15. Circos plot of SAG genomes, arranged by scaffold length in descending order. Outer ring: GC content. Center ring: scaffolds proportionate to size. Inner ring: predicted genes colored according to OC category: blue- true core; green- potential core; orange- shared Ic; red- unique Ic; black- rRNA/tRNA; uncolored- remaining distributions. Additionally, shared Ic clusters are connected by lines of the same color.

Figure S16. MUSCLE alignment of the 16S rRNA gene for genome-sequenced SAR11 strains, including the SAGs, as well as sequences with known insertions from (Eloe *et al.*, 2011b, Lauro *et al.*, 2007) for comparison.

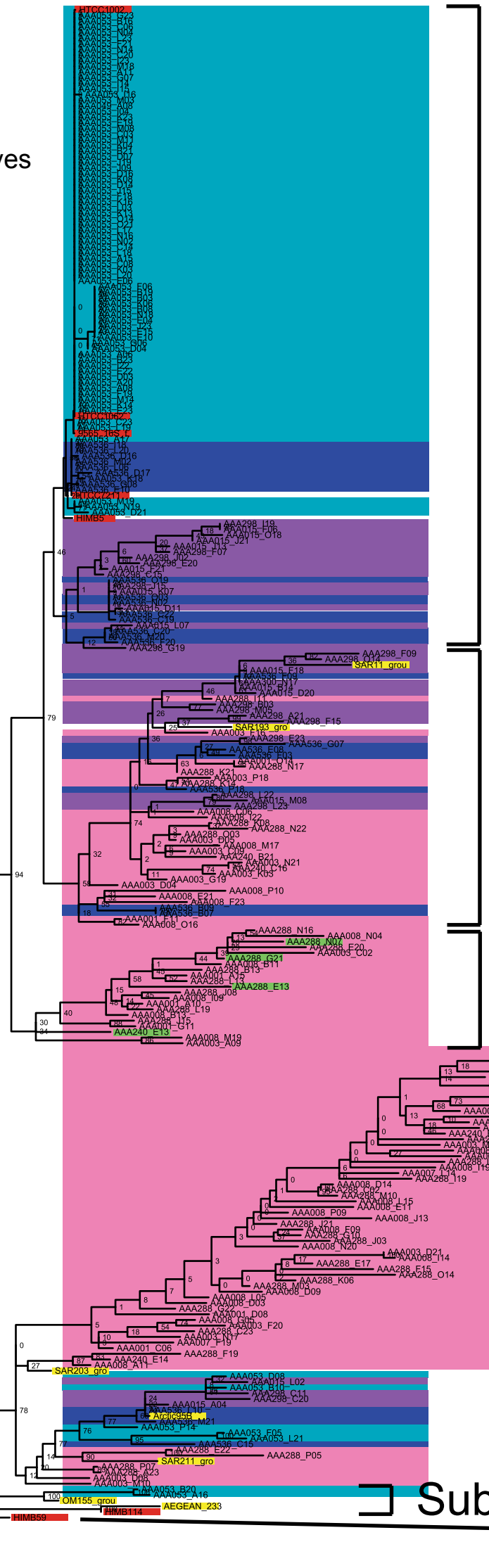
Supplementary Table

Table S1. Excel spreadsheet (Table_S1_F.xlsx) of the comparative genomics and metagenomics data for the SAR11 subclade Ic SAGs.

Figure S1

ref. sequences w/out cultured representatives
genome-sequenced representatives
SAGs selected for sequencing

Maine
Mediteranean
Surface
Mesopelagic



Subclade Ia

Subclade Ib

“Subclade Ic”

Subclade II

Subclade IIIa

Subclade V

Figure S2

Station ALOHA

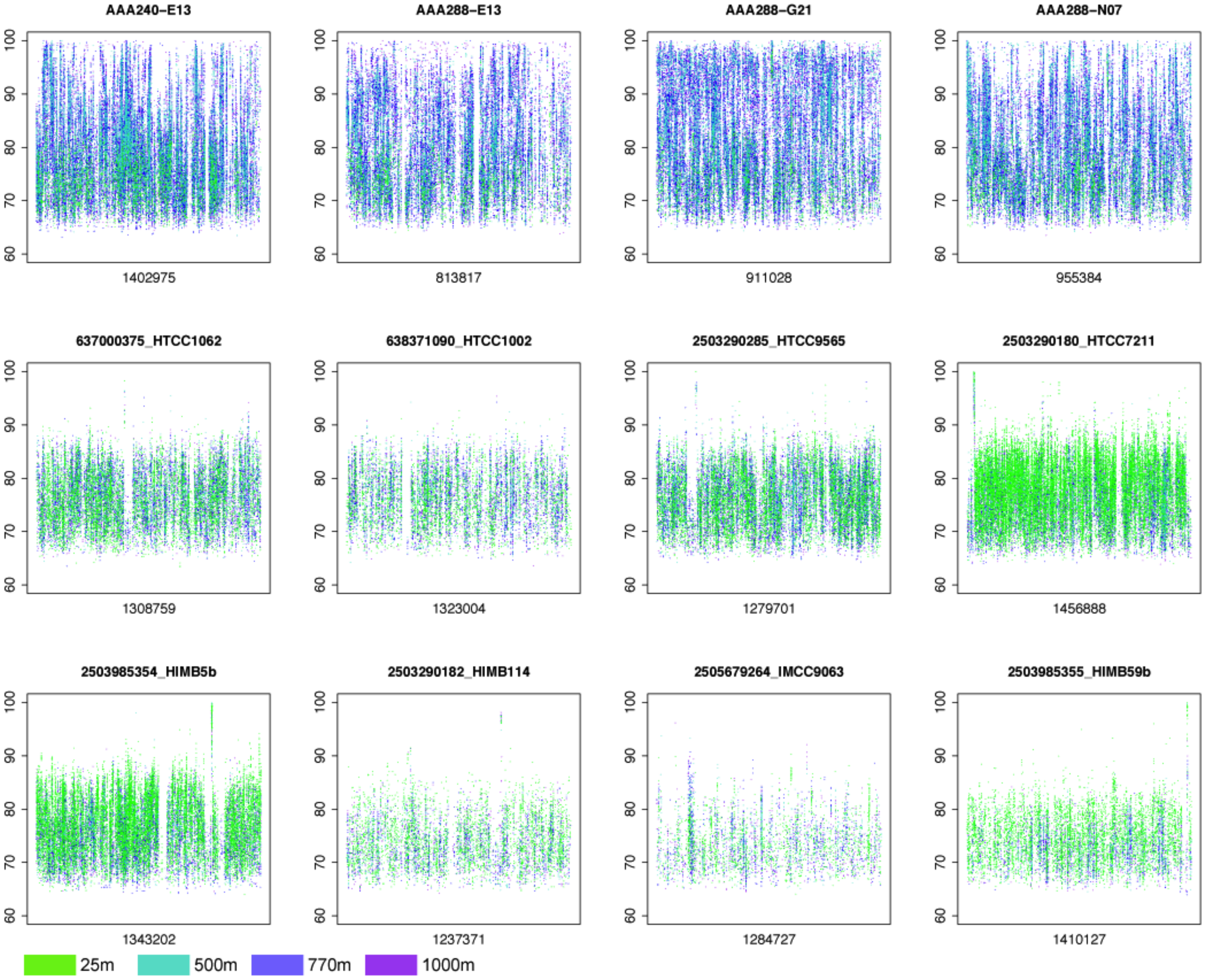


Figure S3

Eastern Tropical South Pacific OMZ

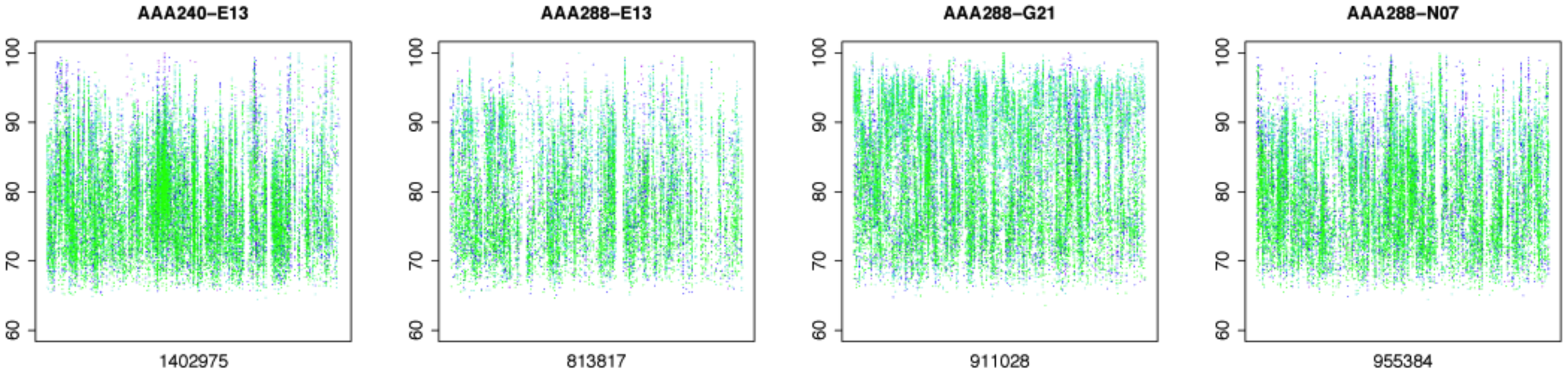


Figure S4

Bermuda Atlantic Time-series Study

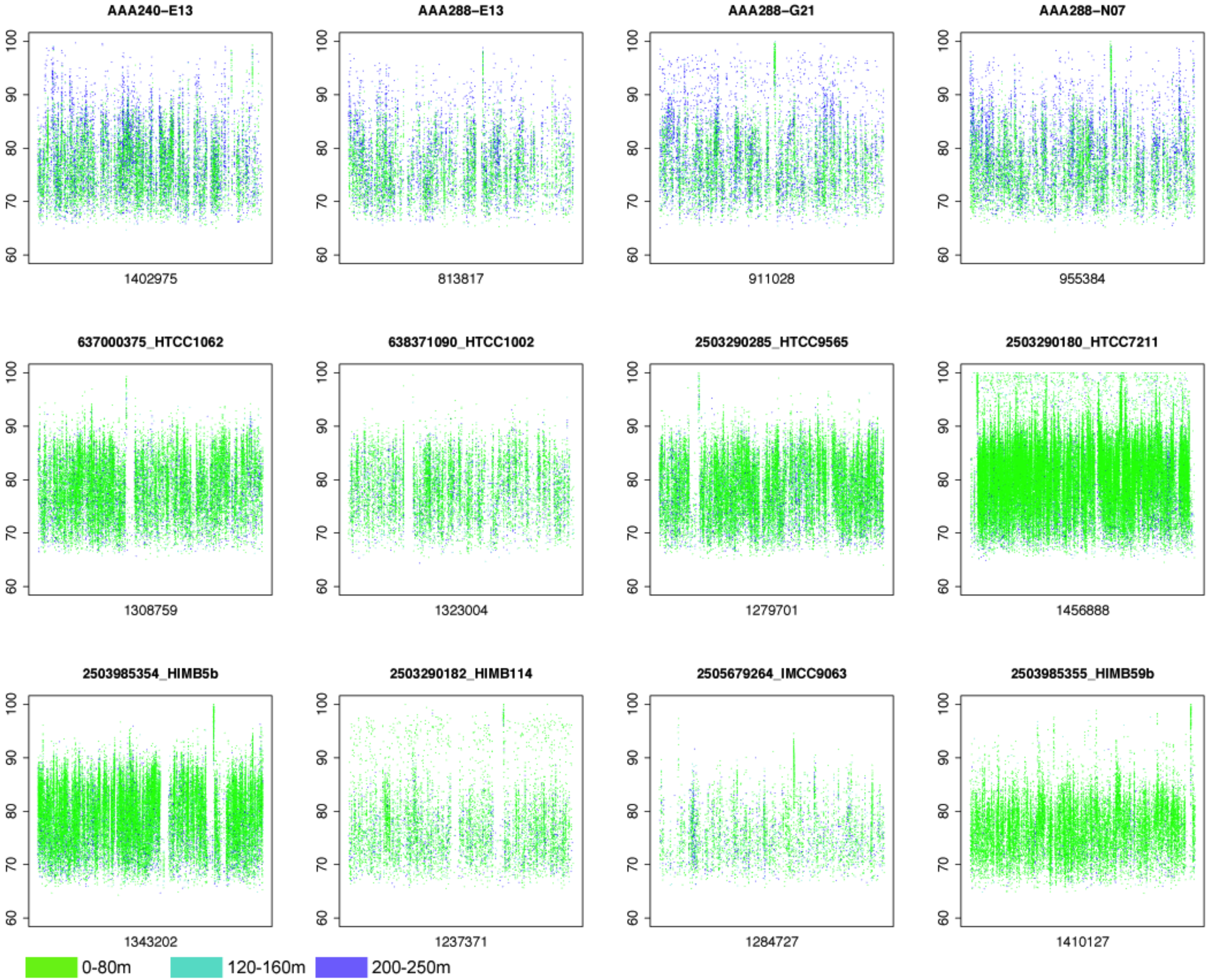


Figure S5

Puerto Rico Trench

Figure S6

Sea of Marmara

Figure S7

Matapan-Vavilov Deep

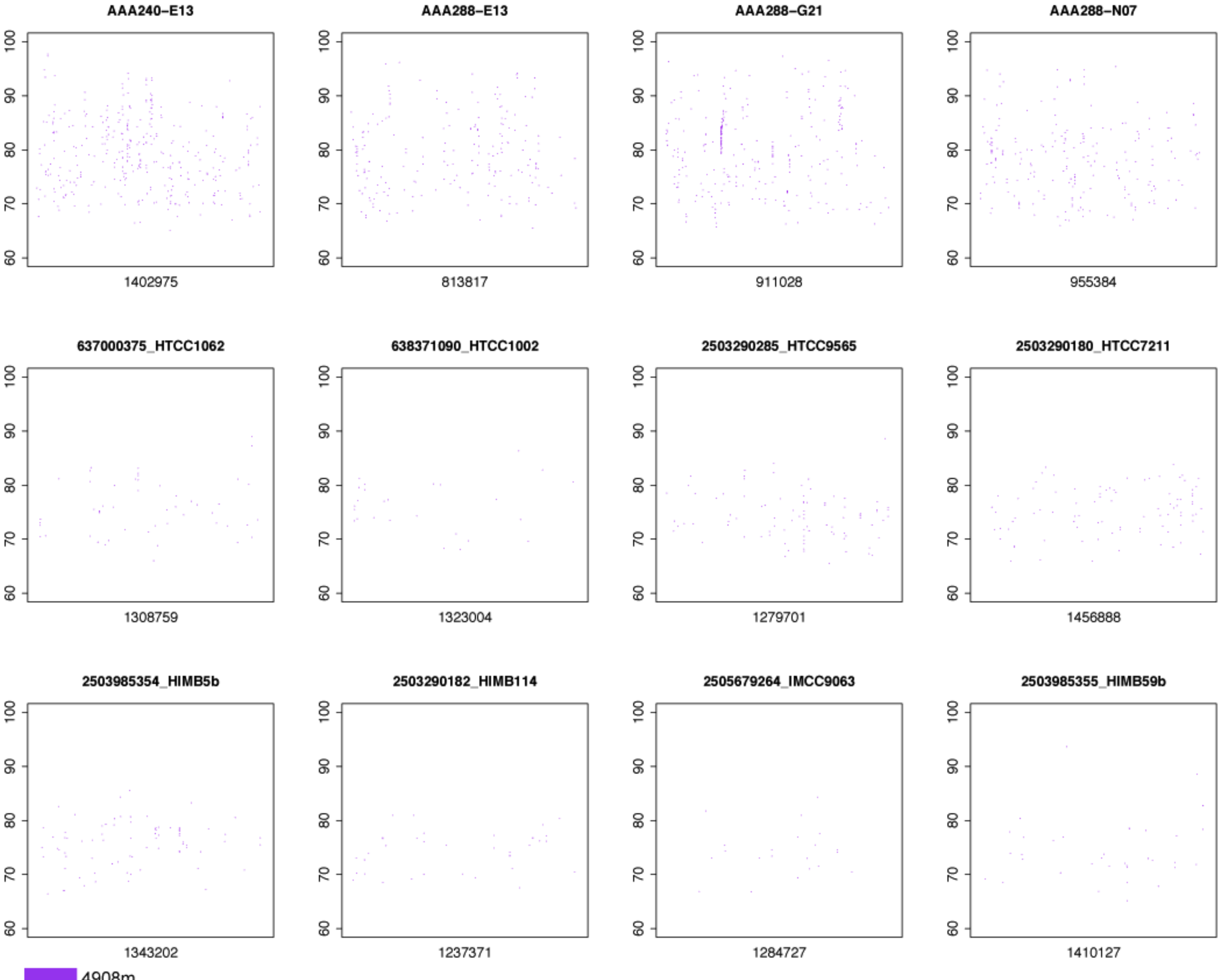


Figure S8

Station ALOHA 4000m Sanger shotgun sequences

Figure S9

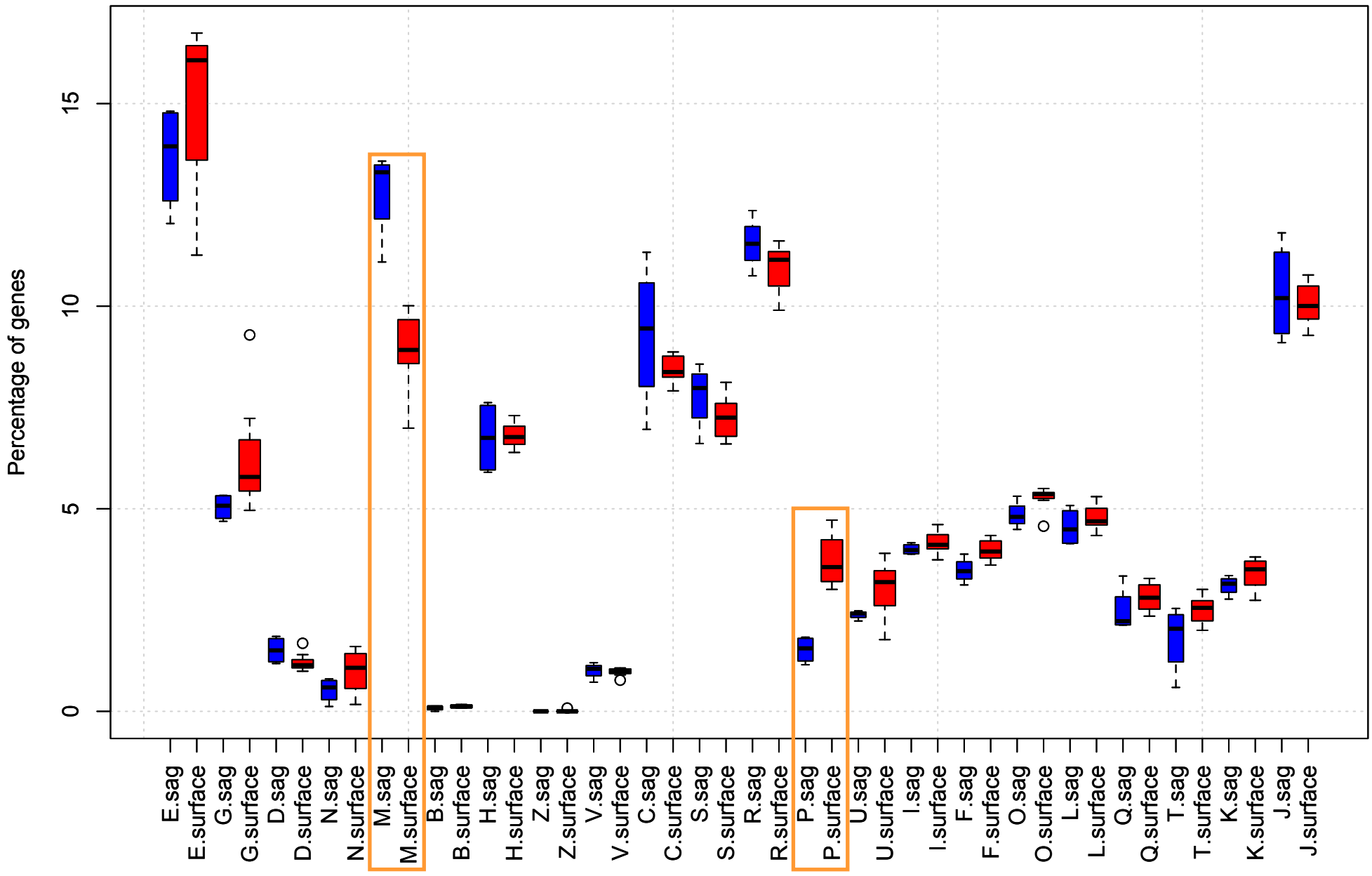
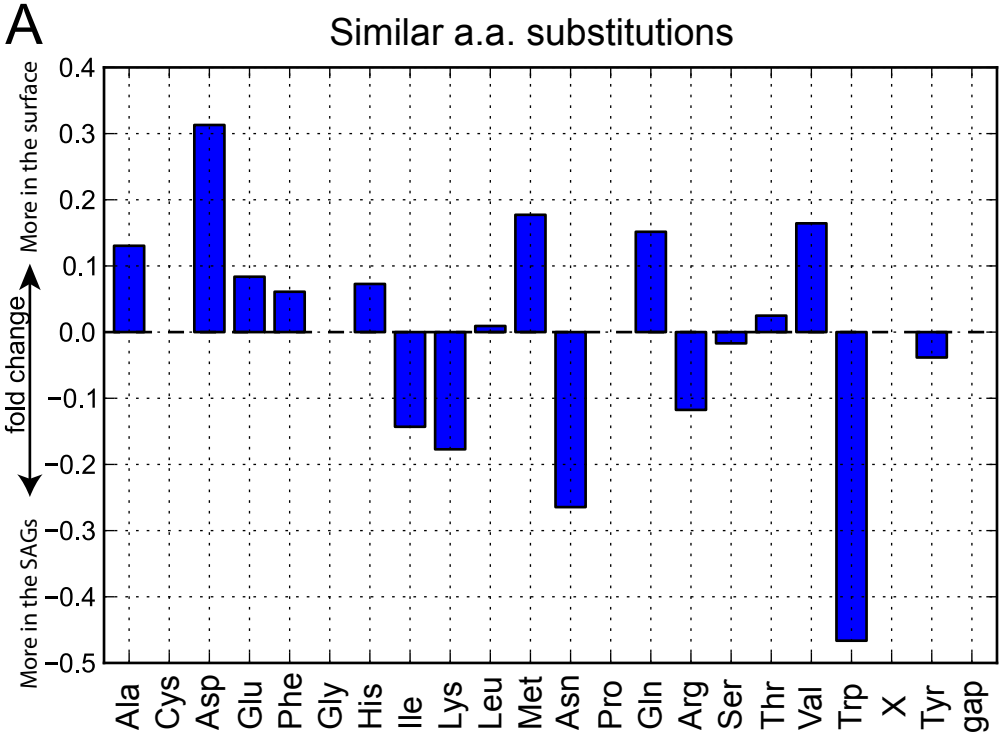

Percentage of genes
0
5
10
15
E.sag
E.surface
G.sag
G.surface
D.sag
D.surface
N.sag
N.surface
M.sag
M.surface
B.sag
B.surface
H.sag
H.surface
Z.sag
Z.surface
V.sag
V.surface
C.sag
C.surface
S.sag
S.surface
R.sag
R.surface
P.sag
P.surface
U.sag
U.surface
I.sag
I.surface
F.sag
F.surface
O.sag
O.surface
L.sag
L.surface
Q.sag
Q.surface
T.sag
T.surface
K.sag
K.surface
J.sag
J.surface

Figure S10


A
Similar 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

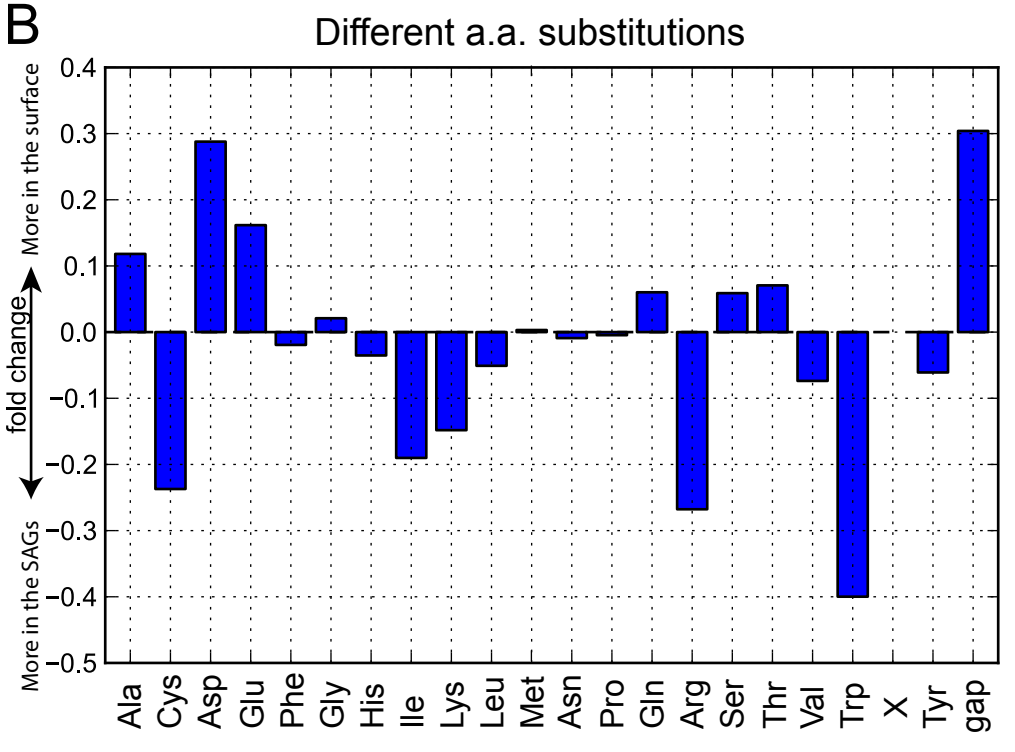
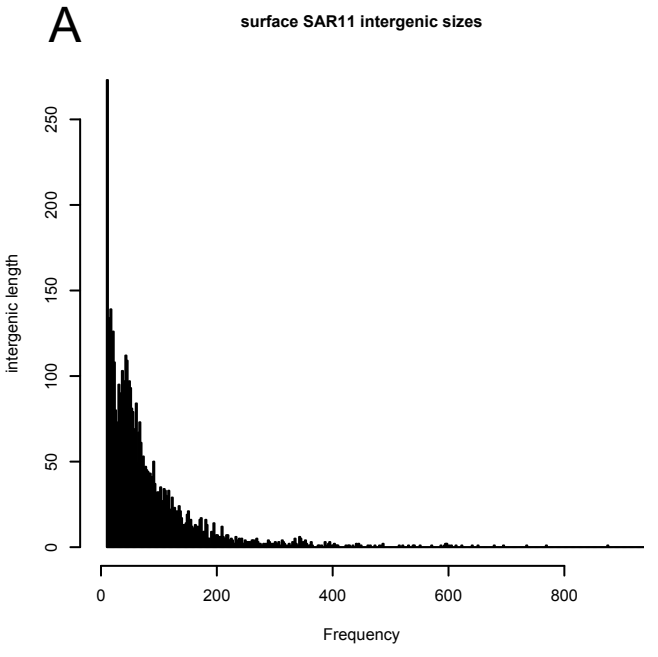
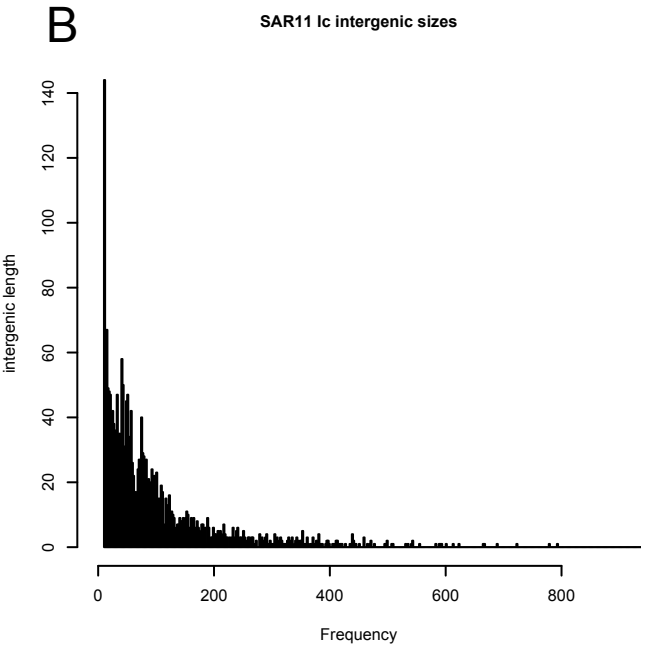
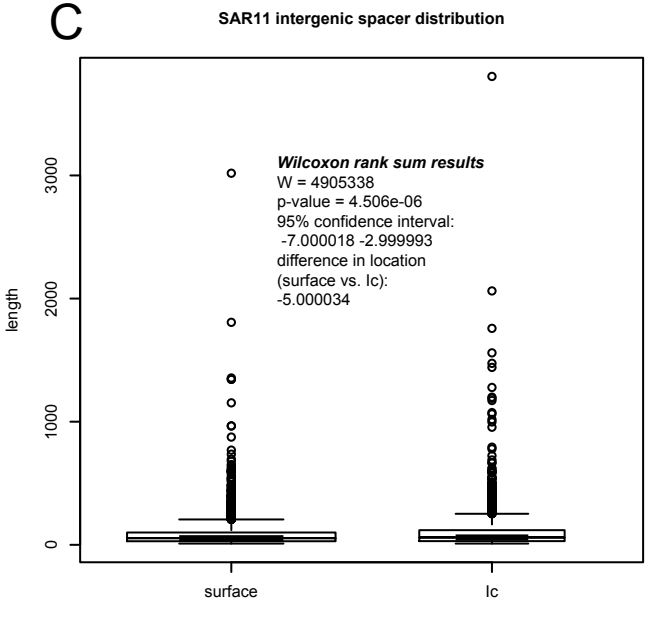

B
Different a.a. substitutions
More in the surface
fold change
More in the SAGs
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
Ala
Cys
Asp
Glu
Phe
Gly
His
Ile
Lys
Leu
Met
Asn
Pro
Gln
Arg
Ser
Thr
Val
Trp
X
Tyr
gap

Figure S11


A
surface SAR11 intergenic sizes
intergenic length
0
50
100
150
200
250
0
200
400
600
800
Frequency


B
SAR11 Ic intergenic sizes
intergenic length
0
20
40
60
80
100
120
140
0
200
400
600
800
Frequency


C
SAR11 intergenic spacer distribution
length
0
1000
2000
3000
Wilcoxon rank sum results
W = 4905338
p-value = 4.506e-06
95% confidence interval:
-7.000018 -2.999993
difference in location
(surface vs. Ic):
-5.000034
surface
Ic

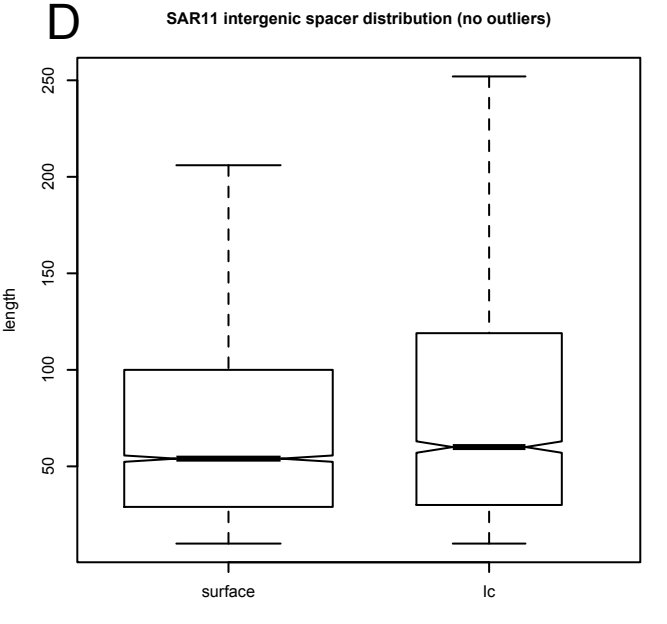

D
SAR11 intergenic spacer distribution (no outliers)
length
50
100
150
200
250
surface
Ic

Figure S12

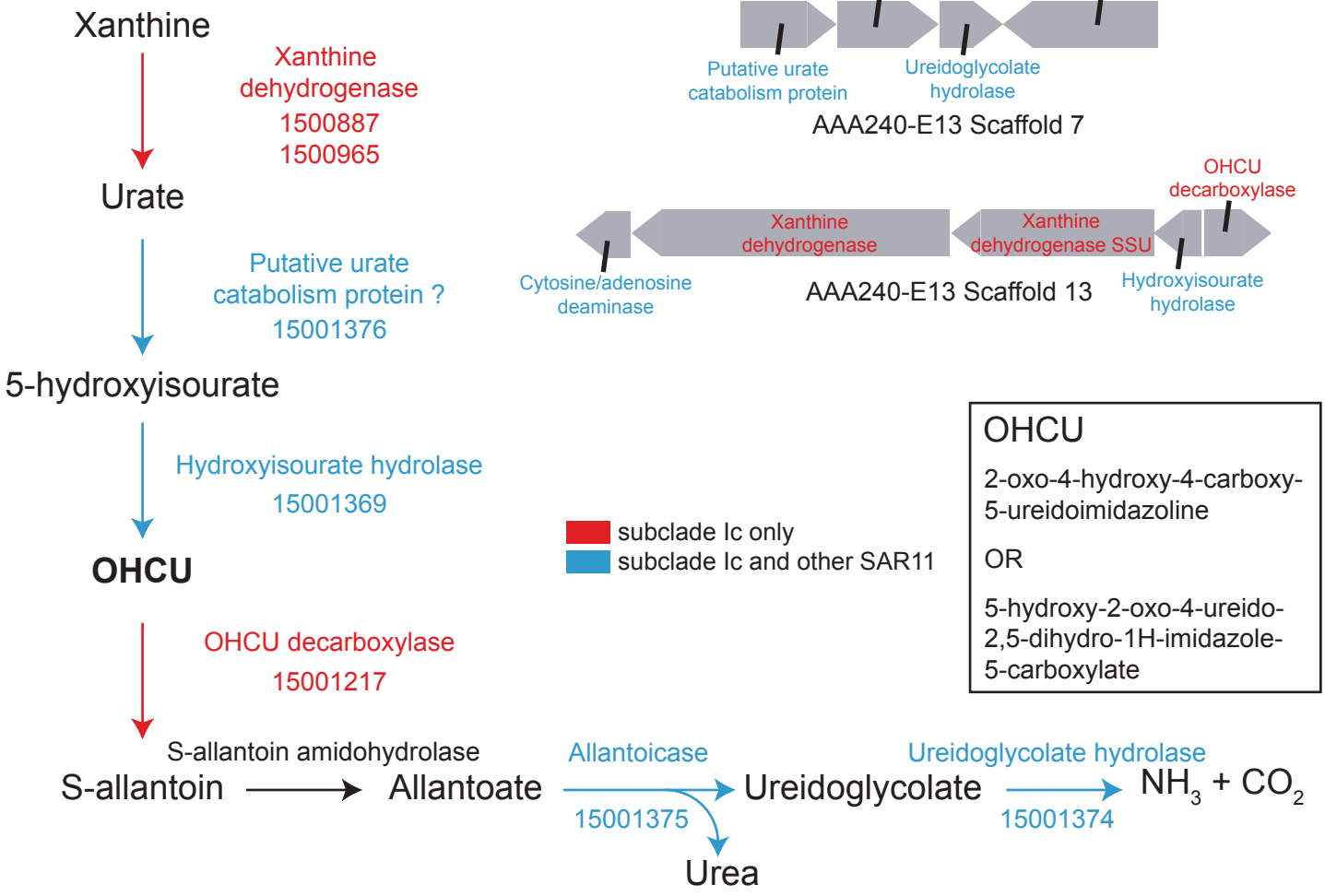

Xanthine
Xanthine dehydrogenase
1500887
1500965
Urate
Putative urate catabolism protein ?
15001376
5-hydroxyisourate
Hydroxyisourate hydrolase
15001369
OHCU
OHCU decarboxylase
15001217
S-allantoin amidohydrolase
S-allantoin
Allantoate
Allantoicase
15001375
Urea
Ureidoglycolate
Ureidoglycolate hydrolase
15001374
NH3 + CO2
Allantoicase
Cytosine/uracil/thiamine/allantoin permeases
Putative urate catabolism protein
Ureidoglycolate hydrolase
AAA240-E13 Scaffold 7
OHCU decarboxylase
Xanthine dehydrogenase
Xanthine dehydrogenase SSU
Cytosine/adenosine deaminase
Hydroxyisourate hydrolase
AAA240-E13 Scaffold 13
subclade Ic only
subclade Ic and other SAR11
OHCU
2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline
OR
5-hydroxy-2-oxo-4-ureido-2,5-dihydro-1H-imidazole-5-carboxylate

Figure S13

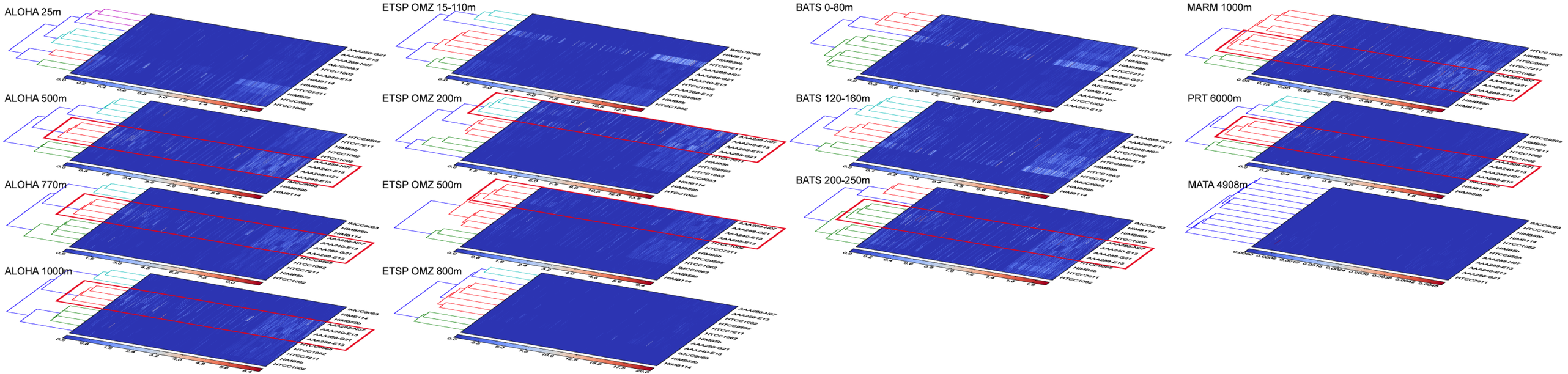

ALOHA 25m
ETSP OMZ 15-110m
BATS 0-80m
MARM 1000m
ALOHA 500m
ETSP OMZ 200m
BATS 120-160m
PRT 6000m
ALOHA 770m
ETSP OMZ 500m
BATS 200-250m
MATA 4908m
ALOHA 1000m
ETSP OMZ 800m

Figure S14

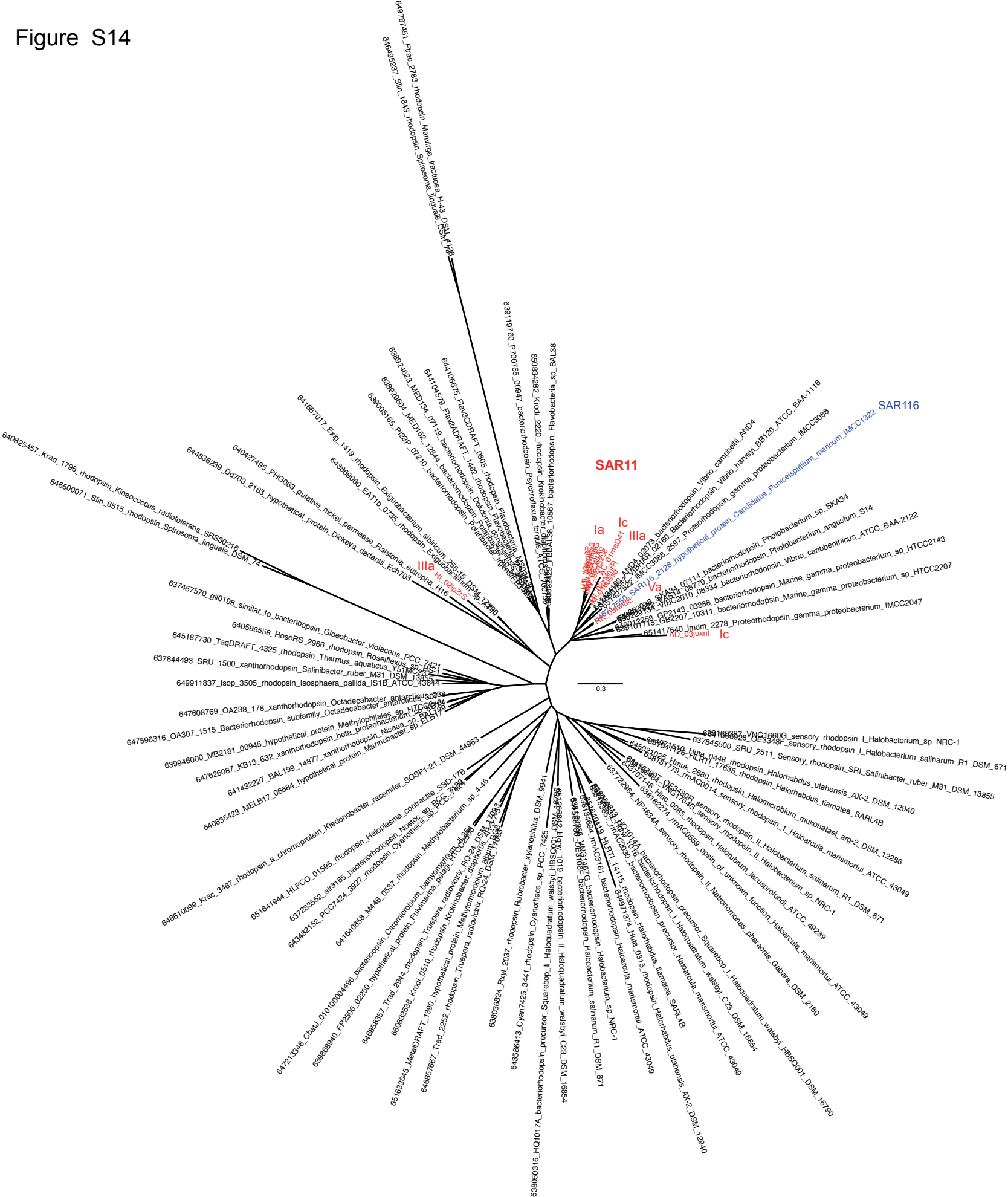


Figure S15

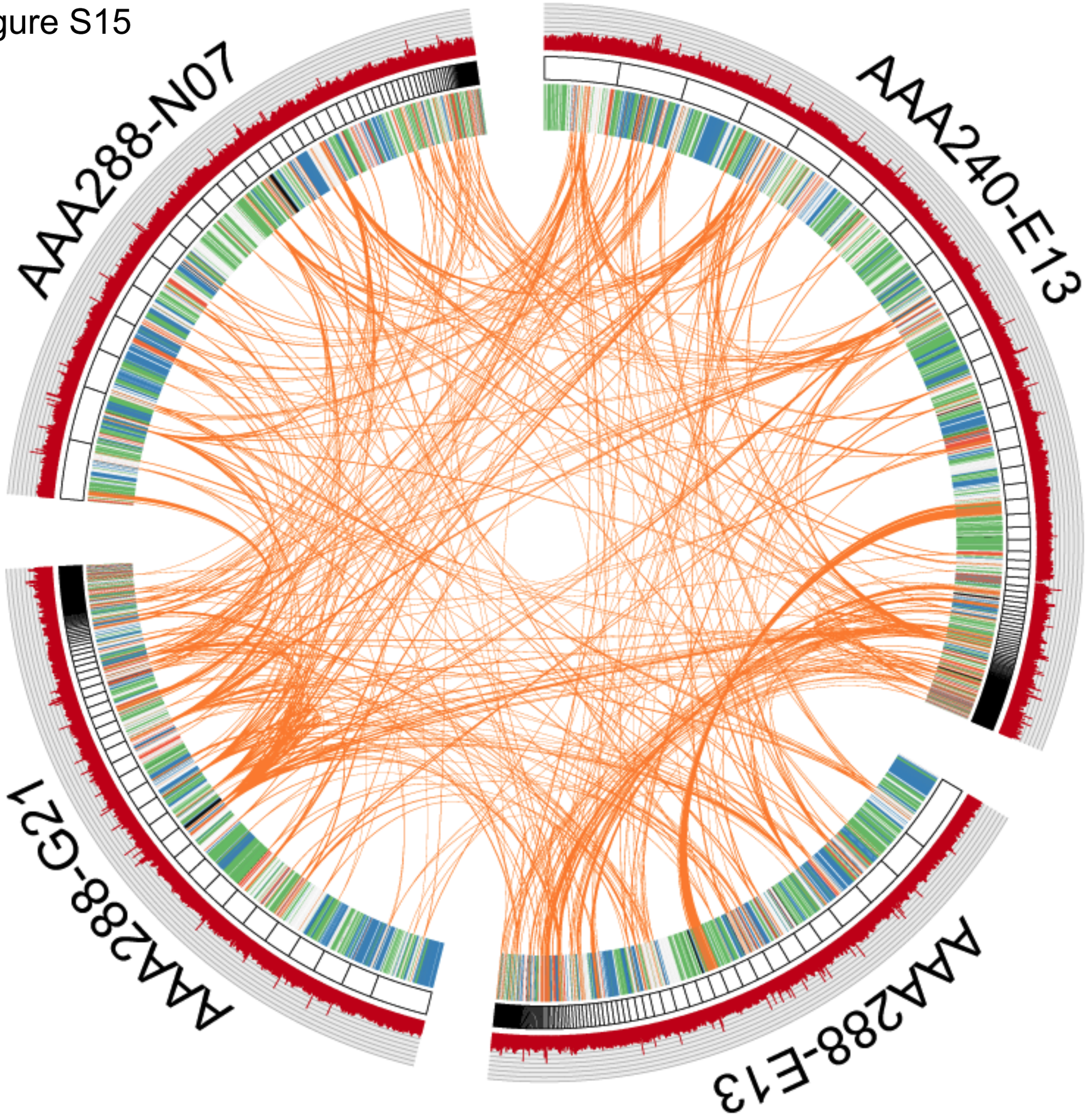
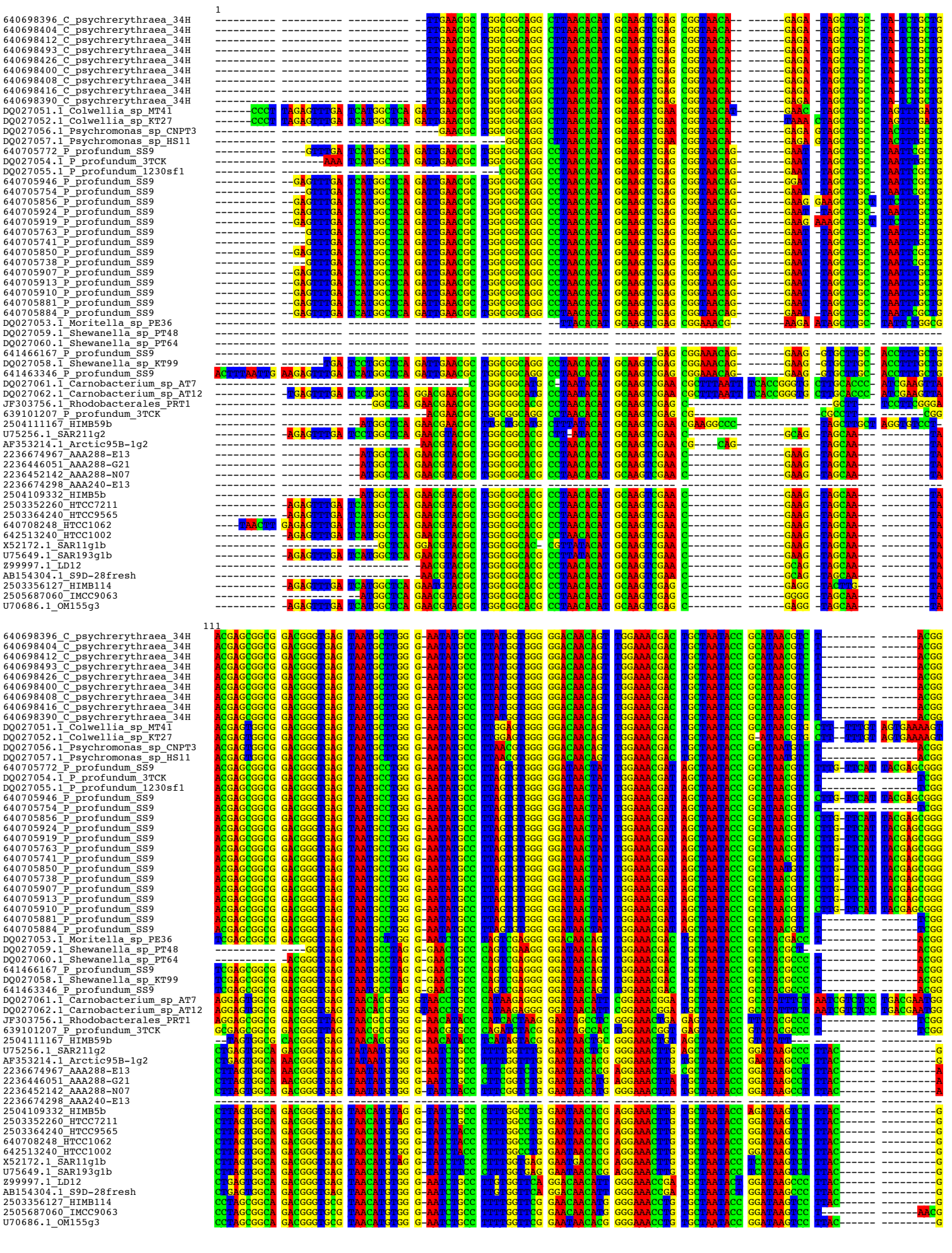
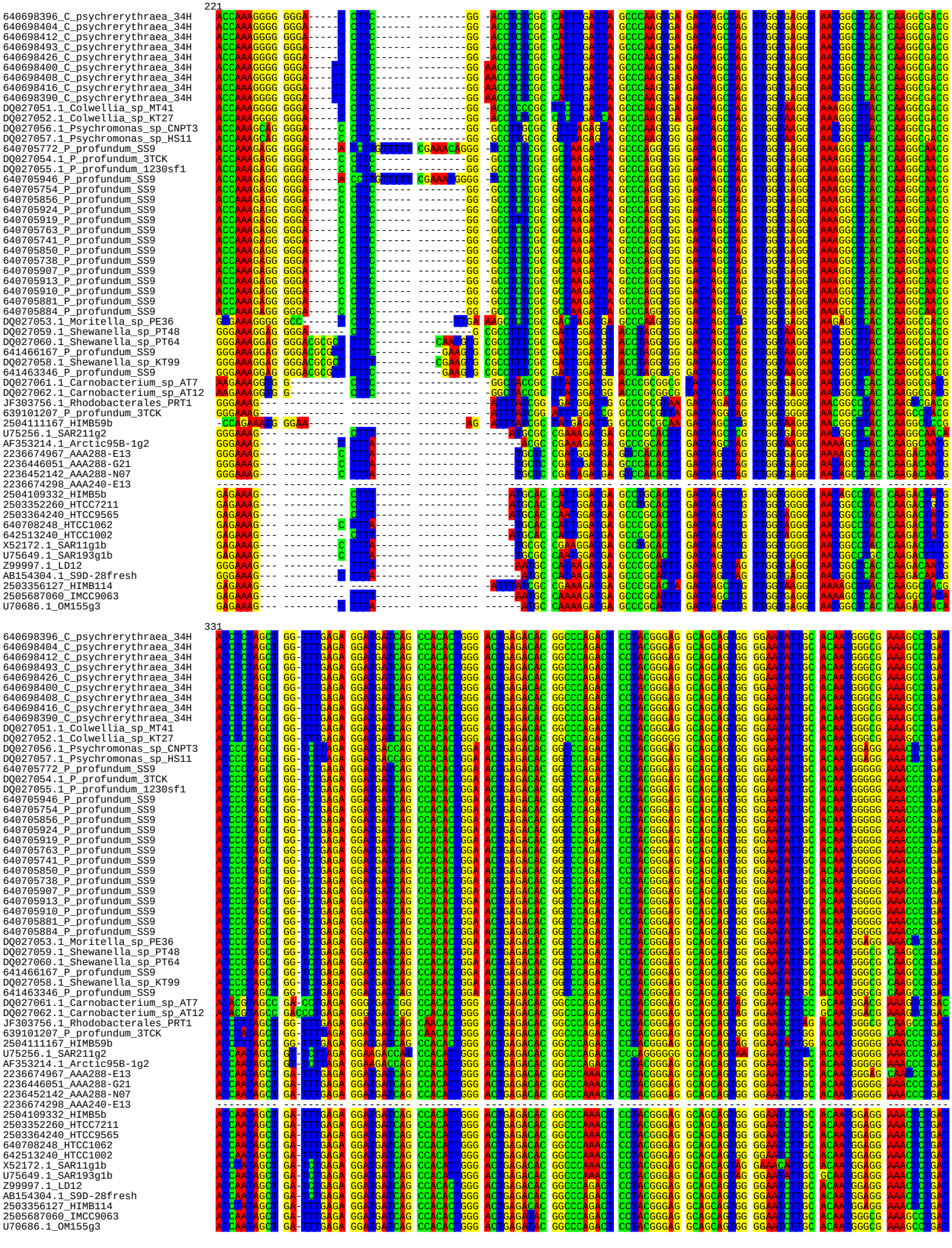
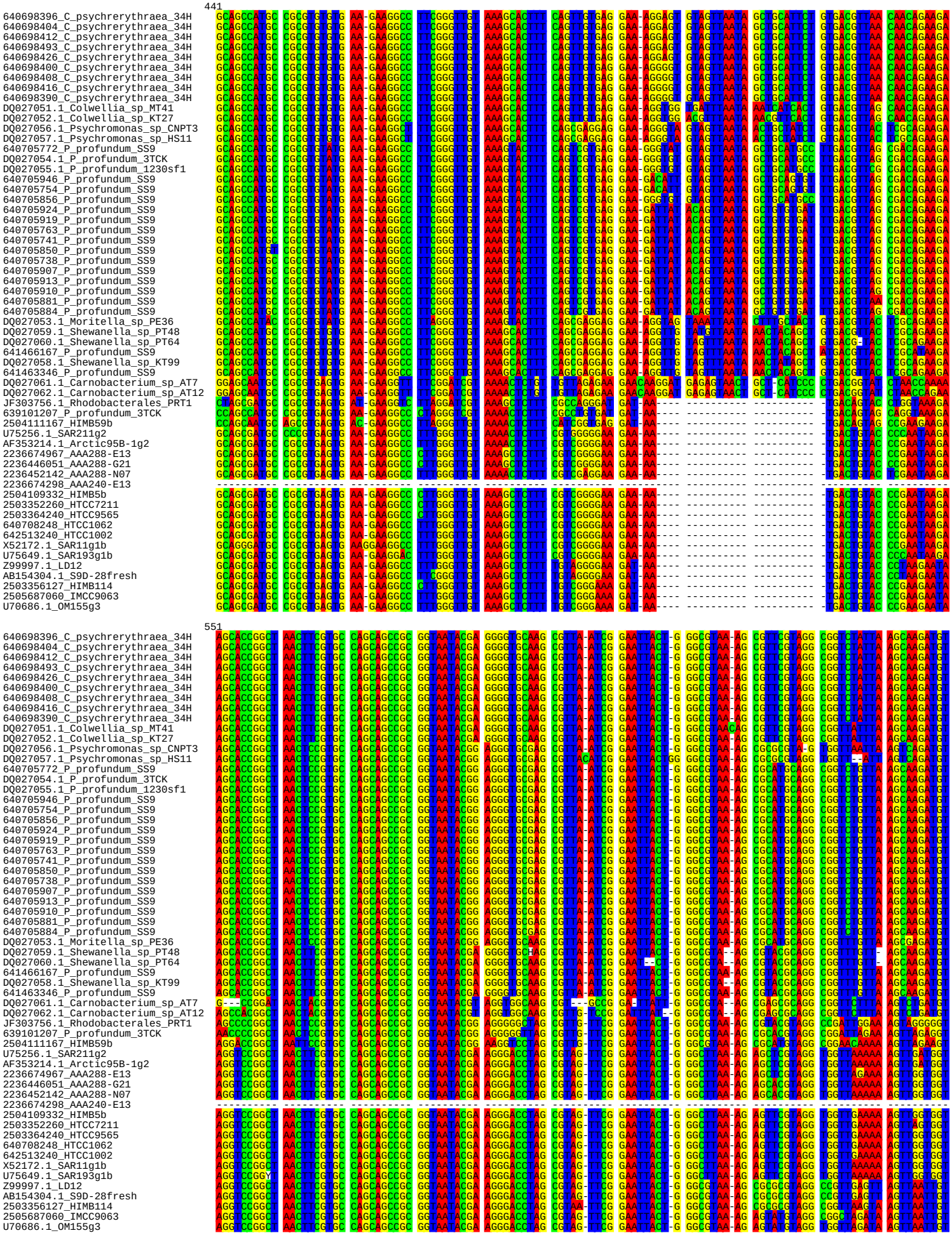
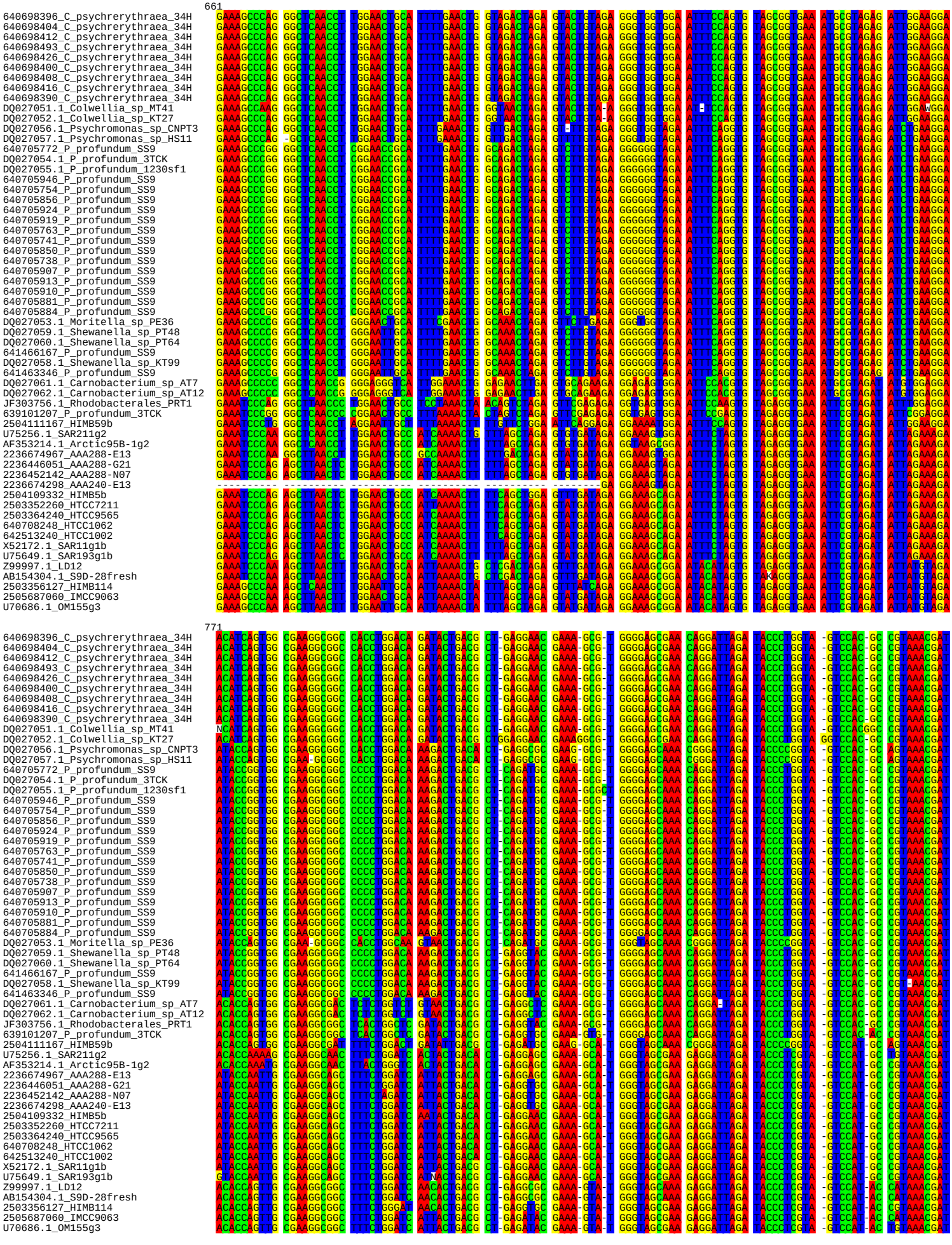


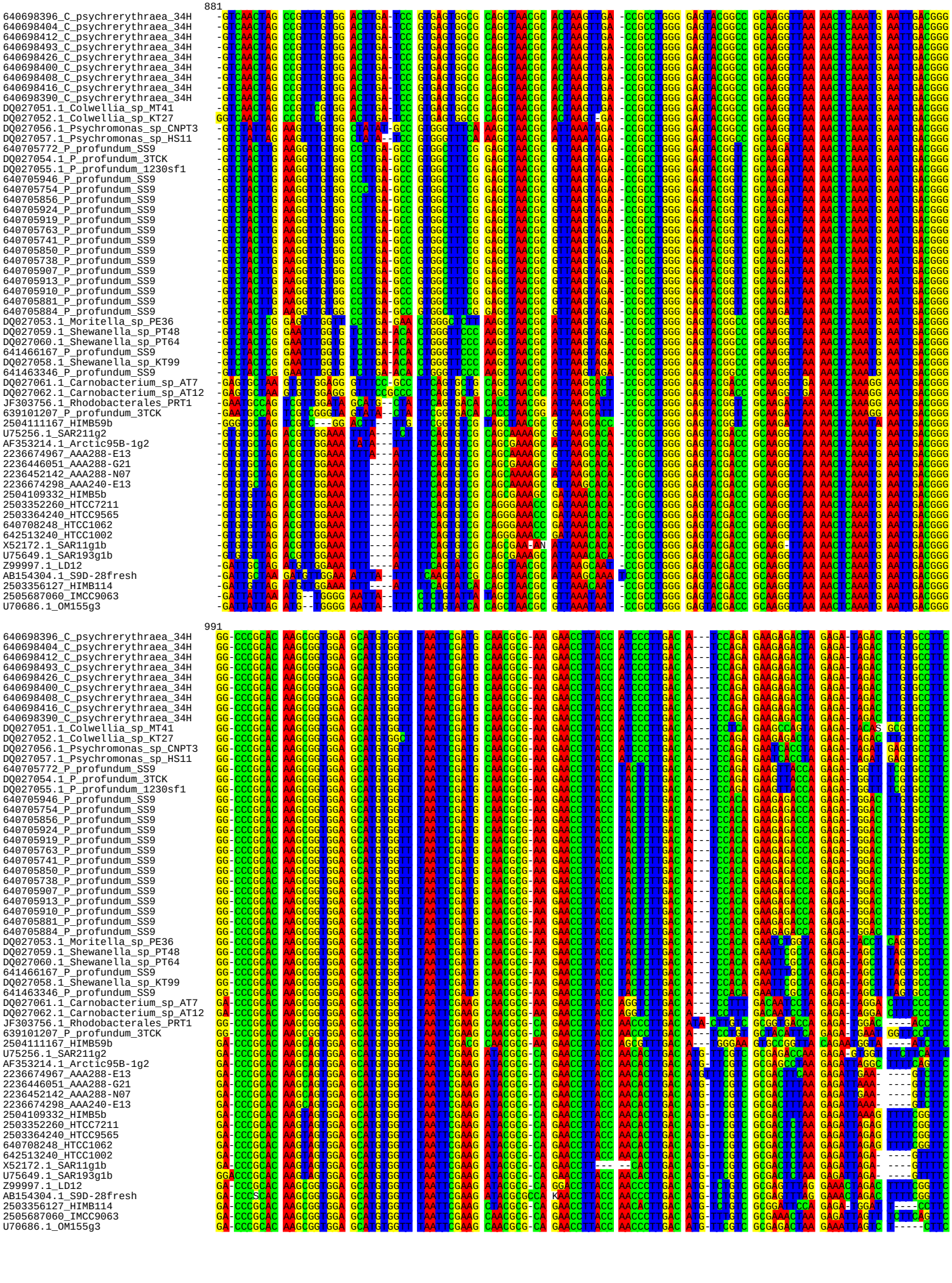
Figure S16

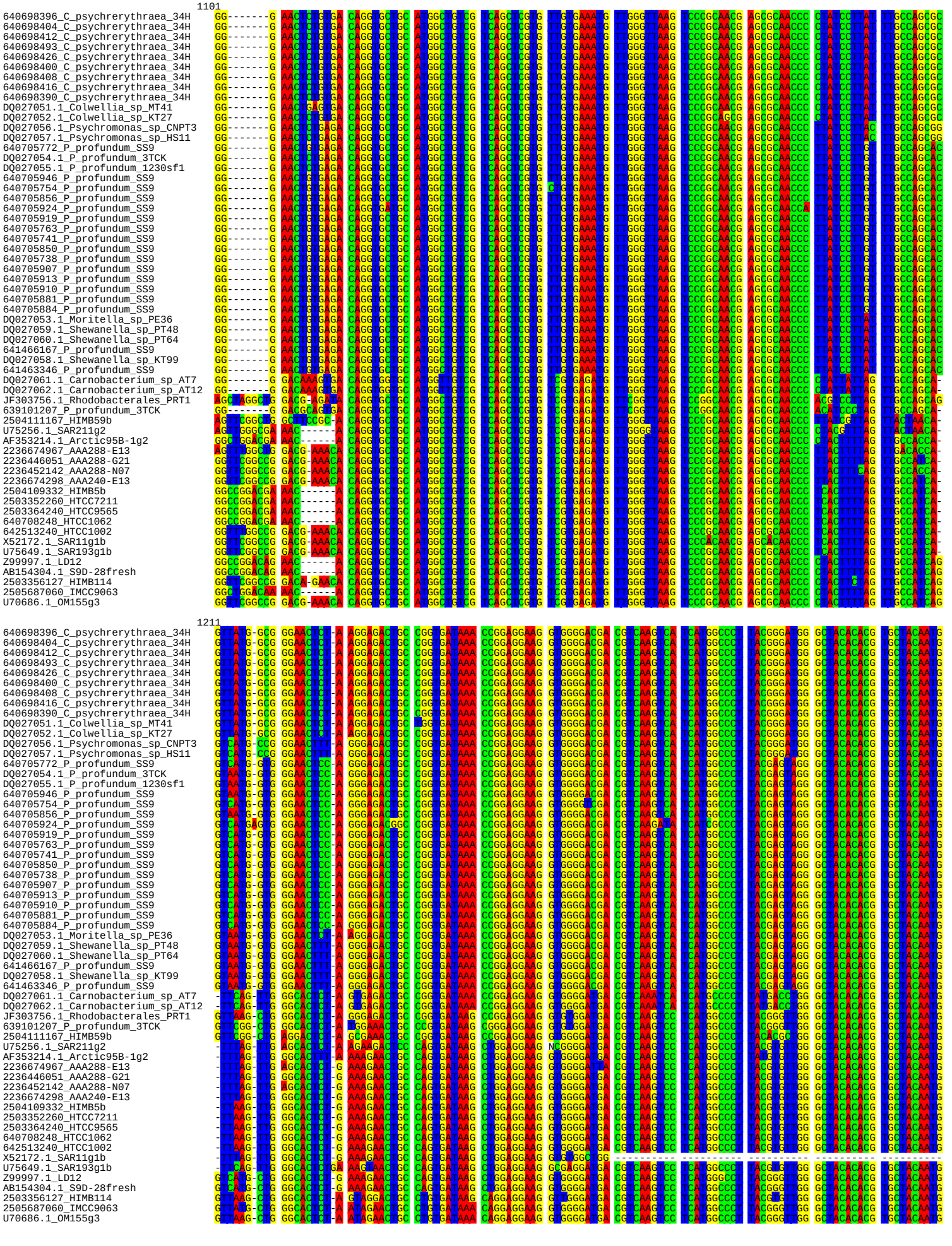

Alignment: /Users/jozenn/Documents/Giovannoni Lab/Projects/SAGs/Results/16S_tree_for_ms/piezo16s_cln_al.fna
Seaview [blocks=10 fontsize=6 LETTER] on Tue Jan 8 09:20:25 2013
1
111
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3

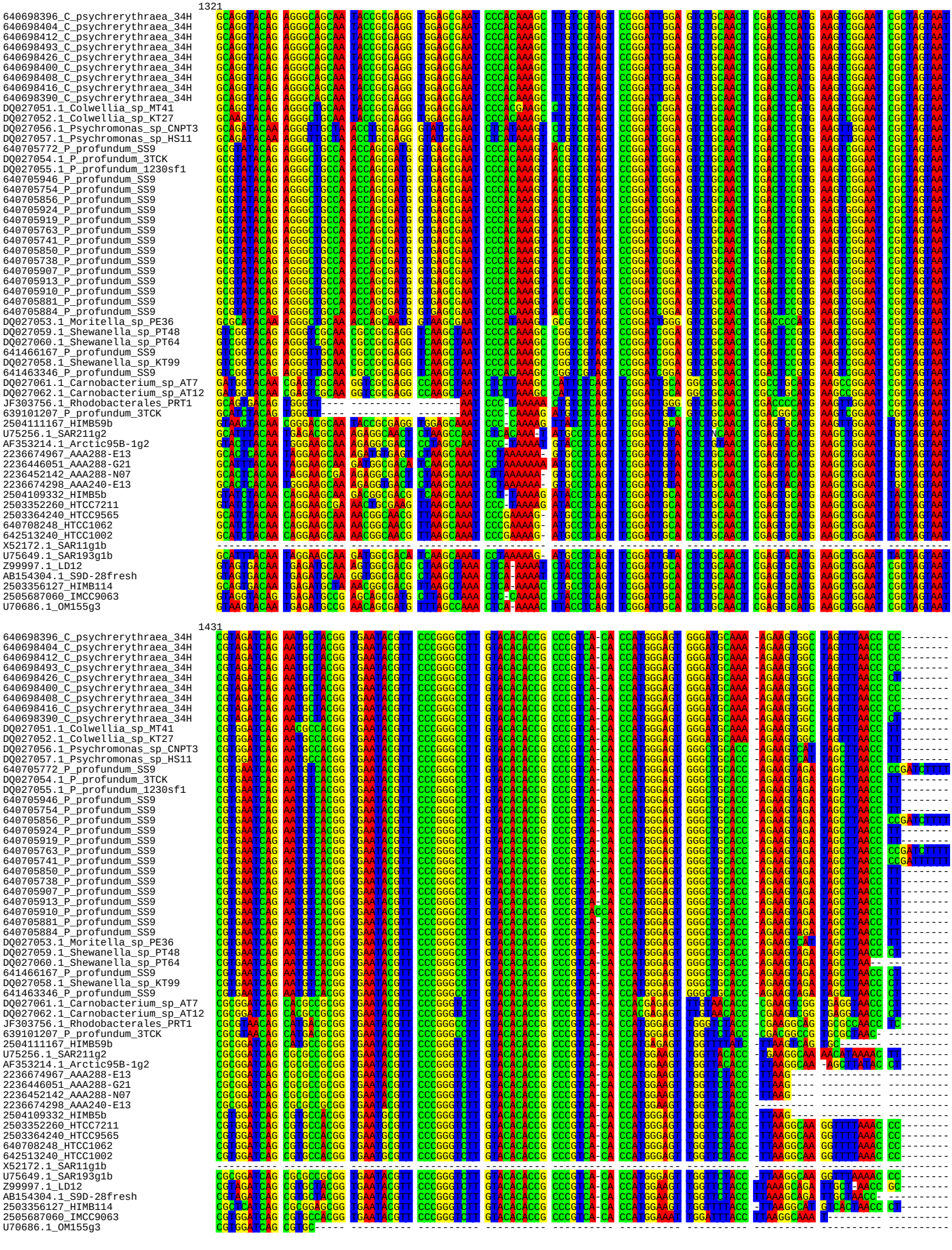

221
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3
331

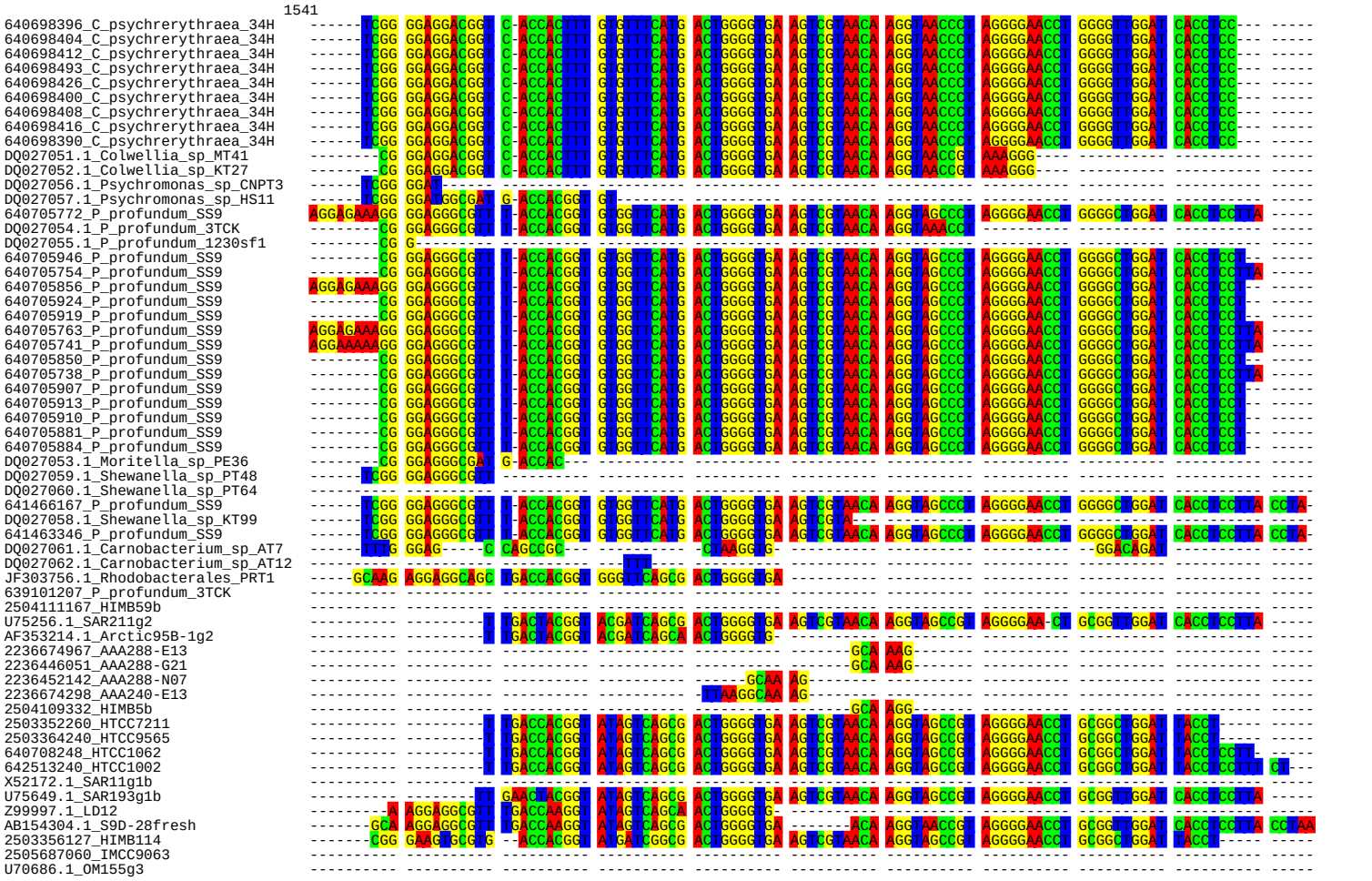

441
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3
551
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3


661
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3
771


881
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3
991


1101
1211
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3


1321
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3
1431


1541
640698396_C_psychrerythraea_34H
640698404_C_psychrerythraea_34H
640698412_C_psychrerythraea_34H
640698493_C_psychrerythraea_34H
640698426_C_psychrerythraea_34H
640698400_C_psychrerythraea_34H
640698408_C_psychrerythraea_34H
640698416_C_psychrerythraea_34H
640698390_C_psychrerythraea_34H
DQ027051.1_Colwellia_sp_MT41
DQ027052.1_Colwellia_sp_KT27
DQ027056.1_Psychromonas_sp_CNPT3
DQ027057.1_Psychromonas_sp_HS11
640705772_P_profundum_SS9
DQ027054.1_P_profundum_3TCK
DQ027055.1_P_profundum_1230sf1
640705946_P_profundum_SS9
640705754_P_profundum_SS9
640705856_P_profundum_SS9
640705924_P_profundum_SS9
640705919_P_profundum_SS9
640705763_P_profundum_SS9
640705741_P_profundum_SS9
640705850_P_profundum_SS9
640705738_P_profundum_SS9
640705907_P_profundum_SS9
640705913_P_profundum_SS9
640705910_P_profundum_SS9
640705881_P_profundum_SS9
640705884_P_profundum_SS9
DQ027053.1_Moritella_sp_PE36
DQ027059.1_Shewanella_sp_PT48
DQ027060.1_Shewanella_sp_PT64
641466167_P_profundum_SS9
DQ027058.1_Shewanella_sp_KT99
641463346_P_profundum_SS9
DQ027061.1_Carnobacterium_sp_AT7
DQ027062.1_Carnobacterium_sp_AT12
JF303756.1_Rhodobacterales_PRT1
639101207_P_profundum_3TCK
2504111167_HIMB59b
U75256.1_SAR211g2
AF353214.1_Arctic95B-1g2
2236674967_AAA288-E13
2236446051_AAA288-G21
2236452142_AAA288-N07
2236674298_AAA240-E13
2504109332_HIMB5b
2503352260_HTCC7211
2503364240_HTCC9565
640708248_HTCC1062
642513240_HTCC1002
X52172.1_SAR11g1b
U75649.1_SAR193g1b
Z99997.1_LD12
AB154304.1_S9D-28fresh
2503356127_HIMB114
2505687060_IMCC9063
U70686.1_OM155g3