Supplementary Information for:

Influence of Nutrients and Currents on the Genomic Composition of Microbes across an Upwelling Mosaic

Lisa Zeigler Allen[1,2], Eric E. Allen[2,3], Jonathan H. Badger[1], John P. McCrow[1], Ian T. Paulsen[4], Liam D. H. Elbourne[4], Mathangi Thiagarajan[5], Doug B. Rusch[5], Kenneth H. Nealson[1], Shannon J. Williamson[1], J. Craig Venter[1], Andrew E. Allen[1§]

[1.] Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, CA, USA
[2.] Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA
[3.] Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA
[4.] Macquarie University, Sydney, NSW, Australia
[5.] Informatics, J. Craig Venter Institute, Rockville, MD, USA

Keywords: Marine / Metagenomics / Upwelling / California Current

Running Title: Microbial Metagenomics across a Southern California Current Upwelling Mosaic

Subject Category: Integrated genomics and post-genomics approaches in microbial ecology

§Corresponding author: Andrew Allen, aallen@jcvi.org

Supplemental Methods:

Nutrient and oceanographic data

Metadata was collected on CalCOFI cruise 0707 using a variety of oceanographic sampling devices, for details see SIO Ref. 07-08. Briefly, a Seabird Electronics, Inc., Conductivity-Temperature-Depth (CTD) instrument (911) equipped with a 24, 10-liter rosette was used to collect seawater. Seawater temperature measurements were taken from the CTD prior to the bottle trip and salinity was analyzed using a Guildline model 8410 Portasal salinometer. Nutrient samples for dissolved silicate, phosphate, nitrate and nitrite were obtained using methods similar to Gordon *et al*., 1993. Water samples for assessing chlorophyll*a* and phaeopigments were filtered onto Whatman GF/F filters and the pigments extracted using methods similar to (Venrick, *et al.* 1984). Pigment concentrations were determined from fluorescence readings before and after acidification with a Turner Designs Fluorometer Model 10-AU-005-CE.

Library Construction and Sequencing

Environmental DNA (eDNA) was isolated from the impact filters and library construction was performed according to Rusch, *et al*. 2007. Briefly, shipboard frozen filters contained in the following storage buffer: 1X Tris-EDTA (TE) buffer (Invitrogen), 50mM EGTA and 50mM EDTA, were thawed to room temperature and aseptically cut for DNA extraction. Filter fragments were placed in 50ml conical tubes and lysis was performed as previously described. DNA was ultimately purified using a phenol/chlorofom extraction method followed by ethanol precipitation. For library construction using Sanger sequencing, eDNA was randomly sheared using nebulization and polished through BAL31 nuclease and T4 DNA polymerase reactions. Fragmented DNA was size selected via agarose gel electrophoresis. BstX1 adapters were ligated onto the inserts prior to ligation into the BstX1 linearized pBR322c2T plasmid vector used for shotgun library preparation. Libraries were then transformed and plated for isolation of colonies as sequencing templates.

Di-deoxy sequencing was used to generate paired-end sequence reads from plasmid templates. For 454 GS FLX Titanium sequencing eDNA was sheared and adaptors ligated. Subsequently, AMPure bead purification was followed by emulsion PCR (emPCR) and sequenced on the Roche 454 Sequencing platform. All sites, except GS260 had enough eDNA to perform Sanger and 454 sequencing from the 0.1μm eDNA.

Accession numbers

The 454 metagenomic datasets were submitted to GenBank with the following accession numbers: GS257 MIDpool-BF-01-669_FRLB9F301.GSMIDSMID11 (0.8um): SRA036240, GS257MIDpool-BF-01-669_FRLB9F301.GSMIDSMID12 (3.0um): SRA036241, GS258-BF-A3B-01-685_GAA1ATG01_EL1.A3B_AGACGCACTC (0.1um): SRA036170, GS258MIDpool-4F-01-593_FQRL34J01.GSMIDSMID5 (0.8um): SRA036235, GS258MIDpool-4F-01-593_FQRL34J01.GSMIDSMID6

(3.0um): SRA036236, GS259-BF-A4B-01-755_GAKKUUN01_EL1.A4B_AGCACTGTAG (0.1um): SRA036171, GS259Pool-4F-01-675_FVQZ7HG01.GSMIDSMID1 (0.8um): SRA036246, GS259Pool-4F-01-675_FVQZ7HG01.GSMIDSMID2 (3.0um): SRA036247, GS260MID70P8-BF-01-580_FVQ8UJY01.GSMIDSMID7 (0.8um): SRA036244, GS260MID83P0UM-BF-01-571_FVQ8UJY02.GSMIDSMID8 (3.0um): SRA036245, GS262-BF-A5B-01-733_GAKKUUN01_EL1.A5B_ATCAGACACG (0.1um): SRA036173, GS262MIDPool-BF-02-616_FR1ZSJO01.GSMIDSMID3 (0.8um): SRA036238, GS262MIDPool-BF-02-616_FR1ZSJO01.GSMIDSMID4 (3.0um): SRA036239, GS263-BF-A6B-01-757_GAKKUUN02_EL1.A6B_ATATCGCGAG (0.1um): SRA036237, GS263MIDPool-BF-01-630_FSP4E4R02.GSMIDSMID5 (0.8um): SRA036242, GS263MIDPool-BF-01-630_FSP4E4R02.GSMIDSMID6 (3.0um): SRA036243, GS264-BF-A7B-01-765_GAKKUUN02_EL1.A7B_CGTGTCTCTA (0.1um): SRA036174, GS264MIDpool-BF-02-689_FVQZ7HG02.GSMIDSMID7 (0.8um): SRA036248, GS264MIDpool-BF-02-689_FVQZ7HG02.GSMIDSMID8 (3.0um): SRA036249.

Removal of 454-sequencing artifacts and protein prediction

To remove artificial replicates in 454-Titanium generated sequences the approach reported by Gomez-Alvarez, *et al*., 2009 was used with the following parameters: sequences sharing >90% nucleotide identity beginning with the same 3 nucleotides were removed.

During our analysis of 454 data it became apparent that many genes were being truncated, mis-annotated, annotated multiple times (as two separate ORF's) or completely missed due to a high rate of frameshift error associated with 454 XLR sequencers. To improve the counts and accuracy of our annotation we implemented a pipeline to correct suspected frameshift related errors on individual sequencing reads. The pipeline can be summarized by the following steps:

1) Identify all stop to stop open reading frames greater than 90 bp where the end of a read/contig was treated as a stop
2) Each stop to stop ORF was converted into a set of overlapping 20 amino acid words with 5 amino acids of overlap with the preceding word
3) The 20 aa words were rapidly aligned to the nraa database using cd-hit-2d with the following non-default parameters: -G 0 -aS 0.9 -c 0.7 –n 5 –g 1
4) The nraa peptides identified by cd-hit-2d were used to construct a mini-database of informative peptides for a more exhaustive frameshift tolerant BLASTX search
5) Metagenomic reads/assemblies were searched against the mini-nraa database using NCBI BLASTX V2.2.23 with the following non-default parameters: -F "m L" -U T -w 11 -e 1e-5 -b 1 -v 5
6) The top non-overlapping BLASTX hits with or without frameshifts were used to identify predicted peptides with sequence similarity based evidence. If the best BLASTX hit contains a frameshift then a

new peptide will supersede and replace any overlapping original peptides thus correcting genes presumably miscalled due to sequencing error. The DNA sequence is not corrected but a record of the frameshift and its location are associated with the newly predicted peptide. If the predicted peptide does not contain a frameshift the new prediction will be kept if only if there was no original gene call covering that portion of the read thus allowing similarity based methods to identify genes potentially missed by the gene finder due to frameshifts. If an original gene call is superseded by these methods it is eliminated from the set of called genes and replaced by the frameshift corrected peptide.

7) Peptides predicted by sequence similarity were then masked from the reads with N's

8) To identify unknown potentially frameshifted genes we employed FragGeneScan V1.13 (Rho et al., 2010; 10.1093/nar/gkq747). While much faster than the homology based approach outlined above, FragGeneScan has a tendency to fuse genes that are present in the same operon and can frequently mis-identify the site of the frameshift by +/- 10 amino acids. With such issues in mind, FragGeneScan is the best alternative for identifying frameshifted genes where sequencing similarity was not an option.

9) After frameshift correction there exist three distinct non-overlapping sets of predicted peptides: open reading frames without frameshifts, frameshifted peptides corrected sequence similarity, and frameshifted peptides identified by FragGeneScan.

The frameshifting rate varies between datasets and is dependent to a degree on the %GC of the sample and the size of the filter that was sequenced. For a typical microbial metagenomic dataset we have documented a modest increase in the number of genes predicted using frameshifted sequence similarity while identifying a large number of new peptides using FragGeneScan (**Table S1**) along with an increase in the average ORF length as ORF's are merged or extended (data not shown). The larger filters (0.8 and 3.0 μm) contain a greater proportion of uncharacterized organisms including eukaryotic genomes and this significantly decreases the ability to correct genes by sequence similarity.

Calibration of sequence similarity based frameshift correction was performed using paired Sanger and 454 sequenced samples not analyzed in this paper. Using reference genomes that were abundantly represented by high identity reads we were able to measure the rate of frameshifts in either the Sanger or 454 datasets based on their position within a read (**Figure S12**). Based on these frequencies there is only a 9.29% chance of finding a 450 bp 454 read without a frameshift. Indeed many reads have multiple frameshifts. The impact these frameshifts have on a gene calling and functional annotation varies both with the gene and with the mechanism by which they are annotated so that not all gene families are impacted evenly.

Protein prediction of Sanger generated sequences involved three steps to identify the most likely protein sequence: (i) Six-frame translation with a 90bp minimum size cutoff; (ii) MetaGeneAnnotator (Noguchi, *et al.* 2008) to identify ORFs >90bp; (iii) ORFs found in step 'ii', are used to identify ORFs in step 'i' and if the ORF in step 'i' extends beyond that found in step 'ii' then that sequence is used as the longest possible sequence.

Bacterial genome estimation and phylogenetic placement of marker peptides

A set of core, mostly single-copy bacterial marker proteins, found in 880 fully sequenced reference genomes, were used to estimate bacterial genome equivalents and for normalization of metagenomic samples. AMPHORA (Wu, *et al.* 2008), a set of hidden Markov models (HMMs), was used to detect the following 31 marker genes: *dnaG, frr, infC, nusA, pgk, pyrG, rplA, rplB, rplC, rplD, rplE, rplF, rplK, rplL, rplM, rplN, rplP, rplS, rplT, rpmA, rpoB, rpsB, rpsC, rpsE, rpsI, rpsJ, rpsK, rpsM. rpsS, smpB, tsf*. These data were then used to extrapolate a predicted genomic size. Although Raes et al. (2007) do not use the more recently published AMPHORA HMMs, an approach similar to (Raes, *et al.* 2007) was taken to assess the genome equivalents sequenced in each sample. Very similar to Raes et al. 2007, we use marker gene density to predict effective genome size and report genome equivalents as the sum of weighted counts for each marker. Whereas Raes et al. 2007 use marker gene length to calibrate marker gene counts, independently for each marker gene, we calibrate marker gene counts with coefficients derived from an additive model. The marker coefficients were derived from a linear model built to estimate average genome number from hits to each of the marker HMMs (identified with AMPHORA HMMs) using simulated metagenomic training data generated by METASIM (Richter, *et al.* 2008). Coefficients for each marker were determined using an additive model of all 31 markers on one linear regression; the formula used was $y=a1 \times 1 + a2x2 + \ldots + anxn$. As with Raes et al. 2007, the coefficients are generally inversely proportional to the length of the marker, as a larger marker would generate more hits. Along these lines, the coefficient is also generally smaller for markers that have 2 copies per genome (**Table S8**), since more hits to these markers does not necessarily mean more genomes. Of note, an additive model (as opposed to an independent model based on independent estimates for genome equivalents for each marker) was selected as it allowed for dependencies of each marker to be captured. A linear model was constructed and evaluated , but deemed less accurate and was not used for the data presented here. Coefficients derived for the additive model do not follow marker length as closely, as in the linear model; some are negative and/or less informative about the number of genomes (lower coefficient) irrespective of length. Genome equivalent data (counts) were then used to predict the average number of peptides per genome (based on the ratio of total bacterial peptides to estimated genome equivalents (**Figure S2c**)) and subsequently genome size. A model to extrapolate genome size from estimates of the number of peptides per genome was chosen that best fit simulated data from bacterial genomes ranging in size from 1-4Mbp, as well as random mixed samples of genomes of the same average range.

An additional distinction of our approach, relative to Raes et al. (2007), is that the initial AMPHORA HMM searches were performed only on peptides identified phylogenetically (with APIS) as members of Kingdom Bacteria. Subsequently the density of genome equivalents was estimated as the ratio of total bacterial peptides, also predicted peptides determined phylogenetically to be Kingdom Bacteria, to bacterial genome equivalents. Constraint of genome equivalent density and total bacterial peptide estimates to only bacterial

peptides introduces the obvious disadvantage of not counting true bacterial peptides that are either not phylogenetically affiliated with Bacteria or phylogenetically unclassifiable because they are novel. However, modeling of simulated metagenomic data indicated that restriction of genome counts to bacterial peptides facilitated relatively accurate estimation of bacterial genome equivalents from multi-lineage communities; such as those found on the larger size class filters analyzed in this study. Therefore restriction of these analyses to Kingdom Bacteria peptides allowed for comparative analyses of bacterial genomic properties between communites of bacteria captured on different size class filters. Also, it should be noted that even in the smallest size class (0.1 – 0.8 μM) we generally find that around 10-20% of the predicted peptides are phylogenetically affiliated with either Eukarya or viruses. According to our metagneomic data simulations, inclusion of such peptides in bacterial genome equivalent density estimates is significantly more problematic than restriction of analyses to peptides identified as Kingdom Bacteria.

Core bacterial marker peptides from metagenomic data were phylogenetically placed onto a static reference species tree that was produced with PHYML (Guindon, *et al.* 2003) using a concatenated sequence of the 31 marker genes from each reference genome. A maximum likelihood (ML) method of phylogenetic placement was used similar to (von Mering, *et al.* 2007), in which a set of candidate nodes was chosen for each sequence and the ML position selected. Rather than use all possible internal nodes as candidates, a preliminary screening was used to find a subset of internal nodes that fit prebuilt sequence similarity models from each potential subtree. This process uses position weight matrices describing the peptide sequence distributions at every position of each marker at each node in the reference species tree. The distribution of metagenomic peptides in a sample across the reference tree was then visualized as a density map on a circular tree using 2-D kernel density estimation with the kde2d function in the MASS library in R (Venables, *et al.* 2002). Trees were drawn using an in-house Perl script to generate SVG output from the reference tree and taxonomic node abundances for each sample.

SAR11 identification and phylogenetic analyses

16S rRNA gene sequences in all metagenomics datasets were identified using the JCVI metagenomic annotation pipeline and further analyzed using an in-house version of the small subunit RNA Taxonomy Alignment Pipeline (STAP) (Wu, *et al.* 2008) that was modified to search against the Silva Database (Pruesse, *et al.* 2007). STAP automatically generates and curates multiple alignments and builds a phylogenetic tree that is used to for taxonomic assignment of each 16S rRNA query sequence. Sequences that were binned into the SAR11 cluster at the family level were further verified by BLASTN against the GenBank nt database to identify the top 3 hits. If a SAR11-like hit was in the top 3, the query sequence was used in subsequent phylogenetic analyses. This resulted in 162 total SAR11-like sequences across all sites and size classes. Multiple sequence alignments containing reference and query sequences were generated using MUSCLE (Edgar 2004) followed by gap removal using

Gblocks (Castresana 2000). Sequences were full length with a mean of 1417.1bp. Construction of a mid-point rooted maximum likelihood tree was performed using Phyml (GTR substitution model) containing only the reference sequences. Selected references from each subgroup were taken from previous 16S phylogenetic reports (Morris, *et al.* 2005; Stingl, *et al.* 2007). These were used as input into pplacer (Matsen, *et al.* 2010), which enabled phylogenetic placement onto the reference tree.

Assembly

Metagenomic sequence reads were assembled using the Celera (Sanger) (Rusch, *et al.* 2007) or Newbler (454-Titanium) assemblers using an 86% identity cutoff. Four assemblies were used in the analyses: (i) Sanger sequence reads from the 0.1-0.8μm size class for all sites, (ii) 454-Ttianium sequence reads from all three size fractions (0.1-0.1μm, 0.8-3.0μm, and 3.0-200μm), (iii) Sanger sequence reads from GS260 0.1-0.8μm size class, and (iv) 454-Titanuium sequence reads from GS260 0.8-3.0μm size class. Predicted proteins, from unassembled metagenomic reads, were mapped to scaffolds or contigs and their phylogenetic classification from APIS used to assign scaffold/contig taxonomy. Scaffolds/contigs were considered Planctomycete or SAR11 if $\geq$33% of the predicted ORFs in the scaffold/contig were classified as one of these taxa. To identify the contribution of site-specific reads to the aggregate scaffolds/contigs, the abundance of sequence reads from each site and size class was calculated for all scaffolds/contigs. This was performed by taking the presence or absence of a read from each scaffold from each site and computing the Hamming distance between vectors for each site. The distances were reduced to 2-dimensions by Multidimensional Scaling (MDS) using the cmdscale function in R, and plotted.

Statistical analyses

All statistical analyses were performed using the R statistical program (Team 2008) (http://www.gnu.org/software/r/R.html). Principal Component Analysis (PCA) was conducted on the oceanographic metadata and amino acid frequency such that each was a variable considered and the first two principles were plotted. Multiple dimensional scaling (MDS) was performed using distance matrices of specific data.

Supplementary Figure Legends:

Figure S1: Principal component analysis of DNA sequence characteristics from the 0.1μm filter: A) tri-nucleotide composition and B) amino acid frequency.

Figure S2: Genome equivalent and genome size estimation for 454 sequences. A) genome equivalents and B) genome size each plotted with sites on the x-axis binned by size fraction: 0.1μm (black), 0.8μm (orange), 3.0μm (purple). C) Table of data used and obtained in genome equivalent calculations.

Figure S3: Taxonomic classification via APIS (Automated Phylogenetic Inference System) of the 454-Titanium generated sequences. The abundance of each taxonomic category was calculated and the difference from the mean for each site was plotted.

Figure S5: Viral sequence diversity comparison of three filters. Sequence similarity cutoff, y-axis and the number of clusters per viral peptides, x-axis. Sites binned by size fraction: 0.1μm (black), 0.8μm (orange) and 3.0μm (purple).

Figure S6: Heatmap of eukaryotic taxonomic composition of sequences from the 0.8-200µm size classes.

Figure S7: Satellite chlorophyll*a* data of station 83.80 (GS260). Each map is an 8 day composite A) before sampling, B) during sampling, and C) after sampling.

Figure S8: MDS plot of CCP sequences compared to reference genomes. E-values from BLASTP against reference Planctomycete genomes were used to generate distance matrix. Abbreviations are as follows: PM = *P. maris*, RB = *R. baltica*, PL = *P. liminophilus*, GO = *G. obscuriglobus*, BM = *B. marina*, CCP = California Current Planctomycete.

Figure S9: MDS analysis of sample contribution to assembled sequences. All scaffolds A) 0.1μm Sanger sequences and contigs B) 0.1, 0.8 and 3.0μm 454 sequences. *Pelagibacter* sp. binned scaffolds/contigs from C) 0.1μm Sanger sequences and D) 0.1, 0.8 and 3.0μm 454 sequences; black circle indicates large size class libraries (0.8-200μm) with the exception of GS258 0.1μm. Sites were color-coded based on oceanographic metadata: oligotrophic (blue circle), aged-upwelled (green triangle), upwelled (red square).

Figure S10: Glutamine synthetase II sequences identified as Mimivirus-like. A-G are NJ phylogenetic trees eac containing a single environmental (metagenomic) sequence (red) and appropriate reference sequences s. H) Phylogenetic tree derived from pplacer showing all seven environmental (metagenomic) sequences (purple) in the context of reference sequences (*Mimivirus reference sequence).

Figure S11: Transporter proteins binned by substrate versus genome size for each site and size class using the 454 generated sequence data. Each data point represents a discrete size class at a particular site. Estimates for transporters per genome are given on the y-axis and estimates for genome size are given on the x-axis. Colors denote size fraction: 0.1μm (black), 0.8μm (orange), 3.0μm (purple). Shapes indicate operationally defined groups based on oceanographic context: oligotrophic (circle), aged-upwelled (triangle), and upwelled (square). A linear model was fit to the data for each substrate where substrate/genome for each site was the response vector and genome size the linear predictor for the response. The relationship between genome-size normalized abundance for transporters for a particular substrate (y-axis) and estimated genome size (x-axis) provides an indication of how particular transporters scale with genome size or are enriched or depleted in bacterial communities with larger or smaller average genome sizes.

Figure S12: Detection of frameshifts in 454-derived metagenomic sequences. A.) Methodology for identification and correction of frameshifts. B.) Rates of error induced frameshifts and stop codons determined on 454 reads based on comparison to closely related reference genomes.

Supplementary References

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, Molecular biology and evolution, 17(4), 540.

Edgar, RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res, 32(5), 1792.

Guindon, S and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Systematic biology, 52(5), 696.

Matsen, FA, Kodner, RB,Armbrust, EV. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree, BMC bioinformatics, 11, 538.

Morris, RM, Vergin, KL, Cho, JC, Rappe, MS, Carlson, CA,Giovannoni, SJ. (2005). Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site, Limnol. Oceanogr., 50(5), 1687.

Noguchi, H, Taniguchi, T,Itoh, T. (2008). MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes, DNA Res, 15(6), 387.

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, Nucleic Acids Res, 35(21), 7188.

Raes, J, Korbel, JO, Lercher, MJ, von Mering, C,Bork, P. (2007). Prediction of effective genome size in metagenomic samples, Genome biology, 8(1), R10.

Richter, DC, Ott, F, Auch, AF, Schmid, R,Huson, DH. (2008). MetaSim: a sequencing simulator for genomics and metagenomics, PloS one, 3(10), e3373.

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 biology, 5(3), e77.

Stingl, U, Tripp, HJ,Giovannoni, SJ. (2007). Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site, The ISME journal, 1(4), 361.

Team, RDC, (R Foundation for Statistical Computing, Vienna, Austria, 2008).

Venables, WN and Ripley, BD, *Modern applied statistics with S*, 4th ed. (Springer, 2002).

Venrick, EL and Hayward, TL. (1984). Determining Chlorophyll on the 1984 CalCOFI surveys., California Cooperative Oceanic Fisheries Investigations, Data Report XXV, 74.

von Mering, C, Hugenholtz, P, Raes, J, Tringe, SG, Doerks, T, Jensen, LJ *et al.* (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments, Science (New York, N.Y, 315(5815), 1126.

Wu, D, Hartman, A, Ward, N,Eisen, JA. (2008). An automated phylogenetic tree-based small subunit rRNA taxonomy and alignment pipeline (STAP), PloS one, 3(7), e2566.

Wu, M and Eisen, JA. (2008). A simple, fast, and accurate method of phylogenomic inference, Genome biology, 9(10), R151.

Supplementary Tables:

Library	Size Fraction	Sequencing Reads	Original Predicted Proteins	Proteins with Frameshifts	Proteins without Frameshifts	Fraggenescan Predictions	Homology based Predictions	Final Proteins Predicted
GS257	0p1	271025	188910	116441	72469	75580	130973	279022
GS257	0p8	378864	187948	125422	62526	213027	67183	342736
GS257	3p0	272795	181001	114673	66328	119562	91345	277235
GS258	0p1	272039	214490	107541	106949	66412	116840	290201
GS258	0p8	346162	192413	117482	74931	185489	57302	317722
GS258	3p0	270789	152102	92180	59922	147145	42865	249932
GS259	0p1	320468	124202	67176	57026	91672	112469	261167
GS259	0p8	240049	156121	99106	57015	124740	58131	239886
GS259	3p0	307215	133926	87130	46796	190399	31447	268642
GS260	0p8	488373	158719	101807	56912	277256	52470	386638
GS260	3p0	516656	107552	71883	35669	323430	13975	373074
GS262	0p1	184469	74716	40750	33966	52746	66640	153352
GS262	0p8	310918	193672	112059	81613	138544	80492	300649
GS262	3p0	301517	144367	88372	55995	159241	45148	260384
GS263	0p1	253153	146189	83493	62696	69556	105630	237882
GS263	0p8	267507	171412	95140	76272	112863	68624	257759
GS263	3p0	269943	122883	68758	54125	140596	28570	223291
GS264	0p1	313160	176135	104155	71980	74754	149254	295988
GS264	0p8	206259	158733	98571	60162	67524	90621	218307
GS264	3p0	247192	133890	80314	53576	115392	49243	218211

Table S1: Protein prediction statistics using Frameshift correction pipeline. The number of proteins predicted using standard metagene ORF calling are reported in the 'original predicted proteins' column. These predicted proteins are further partitioned, in the next two columns, into the total number of proteins with and without frameshifts. The final predicted protein set (final column) is comprised of proteins without frameshifts in addition to those obtained through the FragGeneScan and homology based frameshift correction pipeline.

			Sanger				454-Titanium			
JCVI ID	Filter Size (µm)	CalCOFI ID	Reads	Peptides	APIS Trees	% Trees	Reads	Peptides *	APIS Trees	% Trees
GS257	0.1	87.40	46621	74568	40290	54.00	271025	279022	145948	52.31%
GS258	0.1	87.80	46760	73225	40628	55.50	272039	290201	156852	54.05%
GS259	0.1	83.110	47173	77206	42374	55.40	320468	261167	97215	37.22%
GS260	0.1	83.80	46325	52870	22758	43.00	—	—	—	—
GS262	0.1	80.90	45964	64383	43035	57.90	184469	153352	60211	39.26%
GS263	0.1	77.60	46922	64383	36236	56.30	253153	237882	103908	43.68%
GS264	0.1	77.49	41199	73829	43563	59.00	313160	295988	144964	48.98%
GS257	0.8	87.40	—	—	—	—	378864	342736	80938	23.62%
GS258	0.8	87.80	—	—	—	—	346162	317722	73639	23.18%
GS259	0.8	83.110	—	—	—	—	240049	239886	113758	47.42%
GS260	0.8	83.80	—	—	—	—	488378	386638	57131	14.78%
GS262	0.8	80.90	—	—	—	—	310918	219036	78236	35.72%
GS263	0.8	77.60	—	—	—	—	267507	257759	43677	16.94%
GS264	0.8	77.49	—	—	—	—	206259	218307	61896	28.35%
GS257	3	87.40	—	—	—	—	272795	277235	11953	4.31%
GS258	3	87.80	—	—	—	—	270789	249932	68448	27.39%
GS259	3	83.110	—	—	—	—	307215	268642	39128	14.57%
GS260	3	83.80	—	—	—	—	516684	373074	85751	22.98%
GS262	3	80.90	—	—	—	—	301517	204389	41953	20.53%
GS263	3	77.60	—	—	—	—	269943	223291	111546	49.96%
GS264	3	77.49	—	—	—	—	247192	218211	64945	29.76%
Total or Average Percentage:			**320,964**	**480,464**	**268,884**	**54.40%**	**4,424,272**	**5,314,470**	**1,642,097**	**31.75%**

*Frameshift corrected

285

Table S2: Sequencing statistics for each sample.

A.) Taxa abundances	Abundance					
	0.1µm			0.8 and 3.0µm		
Taxonomy Lineage	**oligotrophic**	**aged upwelled**	**upwelled**	**oligotrophic**	**aged upwelled**	**upwelled**
Bacteria;Actinobacteria	0.032	0.045	0.033	0.008	0.008	0.009
Bacteria;Bacteroidetes/Chlorobigroup	0.188	0.316	0.366	0.282	0.255	0.344
Bacteria;Chlamydiae/Verrucomicrobia group	0.025	0.014	0.013	0.074	0.014	0.038
Bacteria;Cyanobacteria	0.003	0	0.002	0.016	0.026	0.026
Bacteria;Cyanobacteria;Chroococcales; Synechococcus	0.001	0.002	0.002	0.055	0.016	0.028
Bacteria;Cyanobacteria;Prochlorales; Prochlorococcaceae;Prochlorococcus	0.011	0.001	0.001	0.139	0.003	0.004
Bacteria;Firmicutes	0.021	0.011	0.014	0.015	0.022	0.018
Bacteria;Planctomycetes	0.004	0.023	0.005	0.006	0.264	0.004
Bacteria;Proteobacteria	0.003	0	0.001	0.001	0.001	0.001
Bacteria;Proteobacteria;Alphaproteobacteria	0.098	0.075	0.092	0.04	0.06	0.083
Bacteria;Proteobacteria;Alphaproteobacteria; Rhodobacterales	0.088	0.143	0.116	0.072	0.09	0.147
Bacteria;Proteobacteria;Alphaproteobacteria; Rickettsiales;SAR11cluster	0.232	0.035	0.062	0.043	0.034	0.02
Bacteria;Proteobacteria;Betaproteobacteria	0.025	0.025	0.027	0.011	0.012	0.012
Bacteria;Proteobacteria;Deltaproteobacteria	0.012	0.004	0.009	0.01	0.009	0.013
Bacteria;Proteobacteria;Epsilonproteobacteria	0.009	0.005	0.006	0.004	0.005	0.005
Bacteria;Proteobacteria;Gammaproteobacteria	0.199	0.273	0.223	0.184	0.145	0.213
Bacteria;Spirochaetes	0.007	0.003	0.004	0.007	0.005	0.007
Bacteria;Thermotogae	0.003	0.001	0.001	0.003	0.002	0.001
Other	0.025	0.009	0.013	0.019	0.02	0.018
Unknown Bacteria	0.013	0.012	0.01	0.013	0.008	0.011

290

B.) Statistical significance of abundance shifts across size class and environments

	Statistical significance of abundance change across category									
	Size Class		Environment (all size classes)		Environment (0.1µm)		Environment (0.8µm)		Environment (3.0µm)	
Taxonomy Lineage	**pvalue**	**sig.**	**pvalue**	**sig.**	**pvalue**	**sig.**	**pvalue**	**sig.**	**pvalue**	**sig.**
Bacteria;Actinobacteria	2.02E-115	1	1.82E-86	1	1.96E-15	1	0.00E+00	1	0.04	0
Bacteria;Bacteroidetes/Chlorobigroup	2.85E-253	1	1.12E-03	1	5.26E-120	0	8.52E-21	1	0.01	0
Bacteria;Chlamydiae/Verrucomicrobia group	5.67E-26	1	1.59E-68	1	1.35E-02	1	1.70E-05	1	0.27	1
Bacteria;Cyanobacteria	1.86E-22	1	3.40E-02	0	9.90E-07	0	5.01E-127	*1	0.03	0
Bacteria;Cyanobacteria;Chroococcales; Synechococcus	4.07E-81	1	3.58E-03	0	5.44E-02	1	7.13E-01	1	0.08	0
Bacteria;Cyanobacteria;Prochlorales; Prochlorococcaceae;Prochlorococcus	5.44E-68	1	3.01E-23	1	8.44E-72	1	7.10E-144	1	0.00	1
Bacteria;Firmicutes	2.24E-18	1	1.98E-01	0	9.92E-02	0	2.32E-09	1	0.66	1
Bacteria;Planctomycetes	1.93E-46	1	9.08E-15	1	1.51E-03	0	7.46E-11	1	0.66	1
Bacteria;Proteobacteria	6.26E-05	1	4.88E-01	0	4.28E-02	0	5.88E-05	0	0.92	0
Bacteria;Proteobacteria;Alphaproteobacteria	8.19E-183	1	3.70E-135	1	5.65E-06	1	3.03E-31	1	0.00	1
Bacteria;Proteobacteria;Alphaproteobacteria; Rhodobacterales	5.62E-105	1	3.44E-01	0	2.01E-02	1	2.38E-06	1	0.00	0
Bacteria;Proteobacteria;Alphaproteobacteria; Rickettsiales;SAR11cluster	0.00E+00	1	5.31E-284	1	4.94E-01	1	3.66E-19	1	0.00	0
Bacteria;Proteobacteria;Betaproteobacteria	1.88E-71	1	1.19E-03	1	2.44E-06	1	2.66E-59	*1	0.10	0
Bacteria;Proteobacteria;Deltaproteobacteria	9.45E-10	1	2.03E-01	0	5.87E-02	0	7.04E-03	0	0.52	0
Bacteria;Proteobacteria;Epsilonproteobacteria	2.84E-15	1	8.94E-02	0	1.10E-01	0	4.26E-02	*1	0.36	0
Bacteria;Proteobacteria;Gammaproteobacteria	7.05E-249	1	3.27E-11	1	3.24E-12	1	7.43E-122	1	0.00	0
Bacteria;Spirochaetes	3.22E-04	1	1.38E-01	0	4.19E-07	0	9.49E-07	0	0.00	0
Bacteria;Thermotogae	1.34E-03	1	6.31E-04	1	3.93E-07	1	1.29E-99	1	0.00	1
Other	2.52E-19	1	1.61E-19	1	1.13E-03	1	1.75E-13	1	0.44	1
Unknown Bacteria	1.75E-11	1	3.62E-01	0	3.28E-01	0	2.08E-06	1	0.59	0

Table S3: Sequence abundance of ecologically relevant organisms taxonomically classified using bacterial core HMMs followed by mapping to a reference tree. A.) Taxa abundances were calculated by total peptides mapped to nodes (taxa) indicated divided by the total number of sequences for each habitat (column header). B.) Statistical significance of changes between size class (0.1 and >0.8um) and environment (oligotrophic, aged-upwelled, upwelled) were calculated using chi-square goodness of fit analyses. Observed values equal the raw counts for each taxa for each category evaluated (habitat or size class) and expected values equal the number of proteins per category evaluated in a given taxa normalized by the contribution of that category to the total (e..g, the number one would expect for each taxa according only to level of sampling differences between size classes and habitats). The p-value was obtained using the right-tailed probability of the chi-squared distribution. The significance of the p-value is given as a 1 (significant) or 0 (not significant) using a Bonferonni correction. Colors indicate whether the taxa abundance for significant values are greater in which size class (0.1 (black) or >0.8um (purple)) or environment (oligotrophic (blue), aged-upwelled (green), upwelled (red). *indicate where abundances were equal in the aged-upwelled and upwelled. Non-significant values are grey.

Taxonomy	**0.1**µm	**0.8**µm	**3.0**µm
Archaea	0.00707	0.007292	0.007055
Bacteria	0.902038	0.750075	0.766224
Eukaryota	0.030145	0.145655	0.134822
Mixed	0.041351	0.043445	0.043695
Unassigned	0.877049	1.779519	1.950267
Viruses	0.019396	0.053533	0.048205

Table S5: Phylogenetic profiling of ORFs using APIS. Kingdom level taxonomic abundances are shown, indicating a rise in viral sequences in the larger size classes (0.8 and 3.0µm derived sequences).

310

Taxonomy	% Taxa
Unassigned	62.50%
Other Bacteria*	1.39%
Proteobacteria	9.72%
Mixed	6.67%
Bacteroidetes	6.39%
Verrucomicrobia	4.72%
Lentisphaerae	2.22%
Cyanobacteria	1.67%
Firmicutes	1.67%
Acidobacteria	0.83%
Gemmatimonadetes	0.56%
Chlorobi	0.28%
Actinobacteria	0.28%
Ascomycota	0.28%
Bacillariophyta	0.28%
Euryarchaeota	0.28%

Table S6: Planctomycete pangenome taxonomic assignment of predicted ORFs. Percent is given as the amount of taxonomic group per total proteins designated as belonging to pangenome.

315

				Within Subgroups				Outgroups					
Site	Size Class	# of sar11	% of total sar11	1a	1b	2	3/4	1a	1b	1	2	1/2	3/4
GS257	0p1	49	28.00	18	5	10	4	8	-	-	1	1	2
GS258	0p1	12	6.86	3	-	4	2	-	1	-	1	-	1
GS259	0p1	56	32.00	10	12	6	-	8	1	2	9	2	6
GS262	0p1	28	16.00	4	2	7	4	3	3	1	3	-	1
GS263	0p1	8	4.57	2	-	5	1	-	-	-	-	-	-
GS264	0p1	15	8.57	3	-	7	2	1	-	-	-	-	2
GS257	0p8	1	0.57	-	-	1	-	-	-	-	-	-	-
GS259	0p8	1	0.57	-	-	1	-	-	-	-	-	-	-
GS257	3p0	2	1.14	1	-	1	-	-	-	-	-	-	-
GS259	3p0	1	0.57	1	-	-	-	-	-	-	-	-	-
GS263	3p0	1	0.57	1	-	-	-	-	-	-	-	-	-
GS264	3p0	1	0.57	1	-	-	-	-	-	-	-	-	-

Table S7: SAR11 subgroup identification from fragment 16S rRNA gene sequences found in 454-Titanium generated reads. SAR11 sequences were identified via a JCVI version of STAP, further confirmed by BLASTP against nr. Classification into subgroups was from analysis of placement on reference tree, a (-) indicates no representatives for that specific group were identified. Outgroups denote sequences that were not clading with a reference.

Marker	Length (aa)	Linear Coefficient
dnaG	613.1	0.016009
frr	184.7	0.057761
infC	174.4	0.090331
nusA	454.1	0.021356
pgk	401	0.041711
pyrG	546.4	0.011538
rplA	232.4	0.086567
rplB	276.6	-0.030775
rplC	217.1	0.063662
rplD	208.4	-0.00067
rplE	181.3	0.051631
rplF	178.7	0.067095
rplK	143	0.045354
rplL	124.8	0.032811
rplM	147.7	0.00696
rplN	122.6	-0.01531
rplP	139.7	-0.061306
rplS	123.2	-0.003455
rplT	120.3	-0.046662
rpmA	88.5	-0.026573
rpoB	1308.4	0.072564
rpsB	262.3	0.045404
rpsC	237	0.009766
rpsE	177.5	0.05338
rpsI	137.1	-0.013871
rpsJ	103.9	-0.010422
rpsK	130	-0.035316
rpsM	121.9	-0.013983
rpsS	92.4	0.016069
smpB	157.3	0.027191
tsf	286.5	0.061926

Table S8: Markers used in genome equivalent and size estimations. Marker length and coefficients for the additive model are given. Coefficients were determined from a simulated metagenomic dataset using reference genomes.

Figure S1

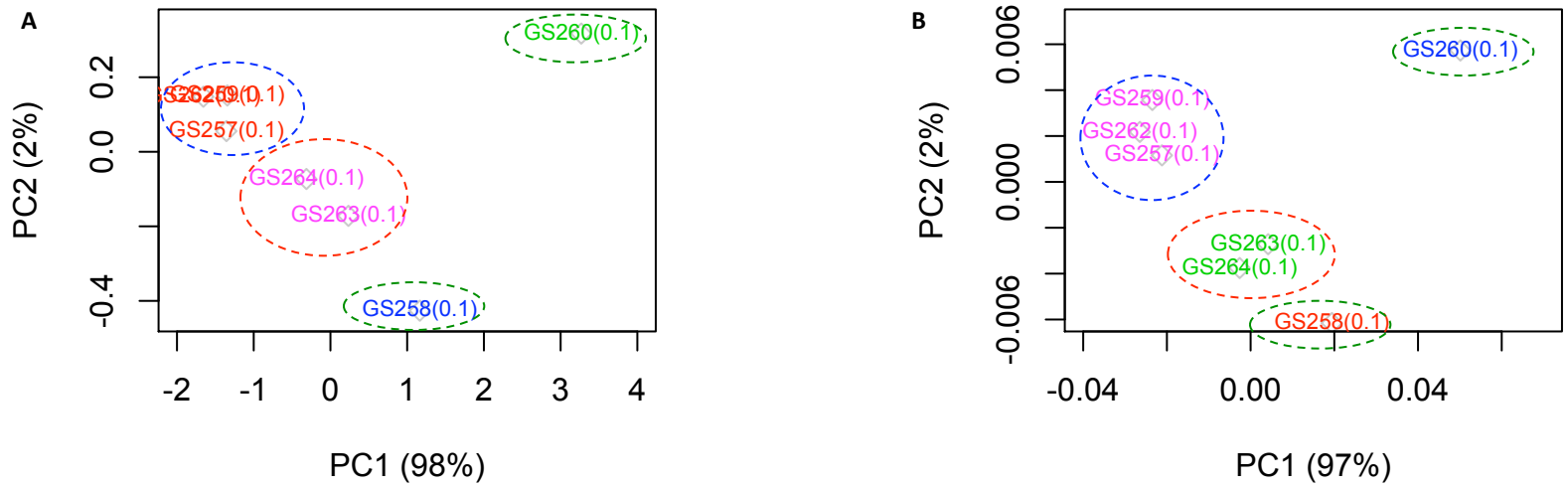
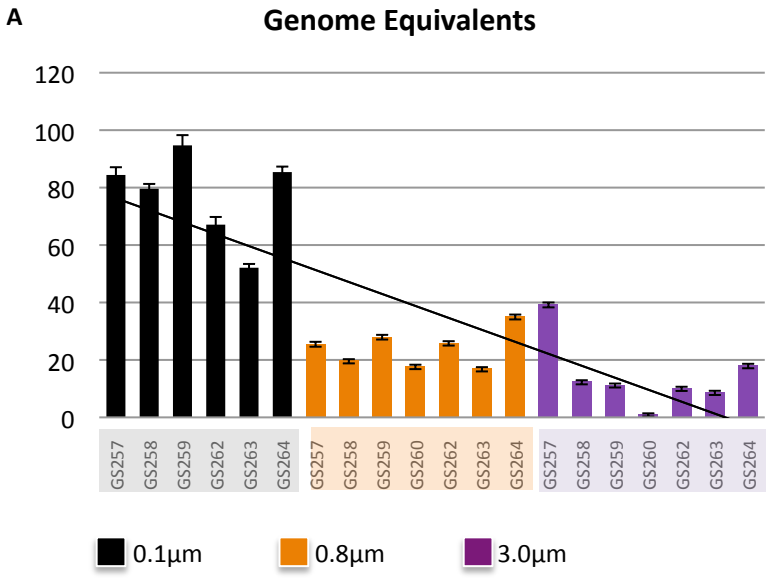
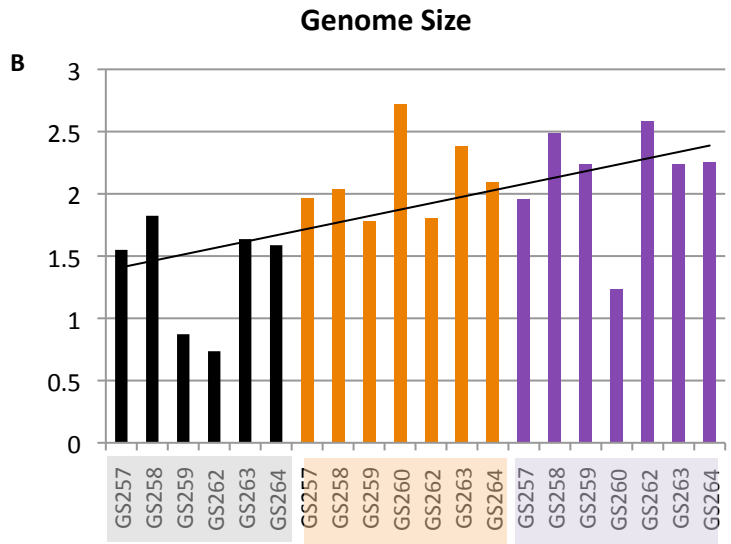

A
PC2 (2%)
0.2
0.0
-0.4
GS260(0.1)
GS259(0.1)
GS257(0.1)
GS264(0.1)
GS263(0.1)
GS258(0.1)
-2
-1
0
1
2
3
4
PC1 (98%)
B
PC2 (2%)
0.006
0.000
-0.006
GS260(0.1)
GS259(0.1)
GS262(0.1)
GS257(0.1)
GS263(0.1)
GS264(0.1)
GS258(0.1)
-0.04
0.00
0.04
PC1 (97%)

Figure S2

A



B



C

Library	Size Fraction	Genome Equivalents	SD	StdErr	bacterial peps	bacterial peps/ genome	Genome Size
454-Titanium							
GS257	0p1	84.4	28.3	2.70	128796	1526.02	1.55
GS258	0p1	79.6	17.6	1.68	140769	1768.45	1.82
GS259	0p1	94.7	37.5	3.57	87734	926.44	0.87
GS262	0p1	67.1	28.1	2.68	54049	805.50	0.74
GS263	0p1	52.1	13.8	1.31	83489	1602.48	1.64
GS264	0p1	85.4	20.1	1.91	133150	1559.13	1.59
GS257	0p8	25.6	7.4	0.70	48503	1894.65	1.97
GS258	0p8	19.8	4.8	0.46	38685	1953.79	2.03
GS259	0p8	28.1	6.6	0.63	48544	1727.54	1.78
GS260	0p8	17.8	5.6	0.53	45566	2559.89	2.72
GS262	0p8	26	5.5	0.52	45434	1747.46	1.80
GS263	0p8	17	5.1	0.49	38514	2265.53	2.39
GS264	0p8	35.1	7.3	0.70	70471	2007.72	2.09
GS257	3p0	39.3	7.4	0.70	74127	1886.18	1.96
GS258	3p0	12.5	4.5	0.43	29405	2352.40	2.48
GS259	3p0	11.4	4	0.38	24363	2137.11	2.24
GS260	3p0	1.3	1.1	0.10	1617	1243.85	1.23
GS262	3p0	10.2	4.8	0.46	24908	2441.96	2.58
GS263	3p0	8.8	4.8	0.46	18806	2137.05	2.24
GS264	3p0	18.1	5.8	0.55	38936	2151.16	2.26
Sanger							
GS257	0p1	23.5	9.6	3.63	37170	1581.70	1.61
GS258	0p1	15.2	5.8	2.19	37561	2471.12	2.62
GS259	0p1	28.4	6.3	2.38	40080	1411.27	1.42
GS260	0p1	9.4	4.2	1.59	23006	2447.45	2.59
GS262	0p1	27.5	8.4	3.17	40413	1469.56	1.49
GS263	0p1	16.3	5.9	2.23	32938	2020.74	2.11
GS264	0p1	18.5	6.8	2.57	40433	2185.57	2.30

Figure S3

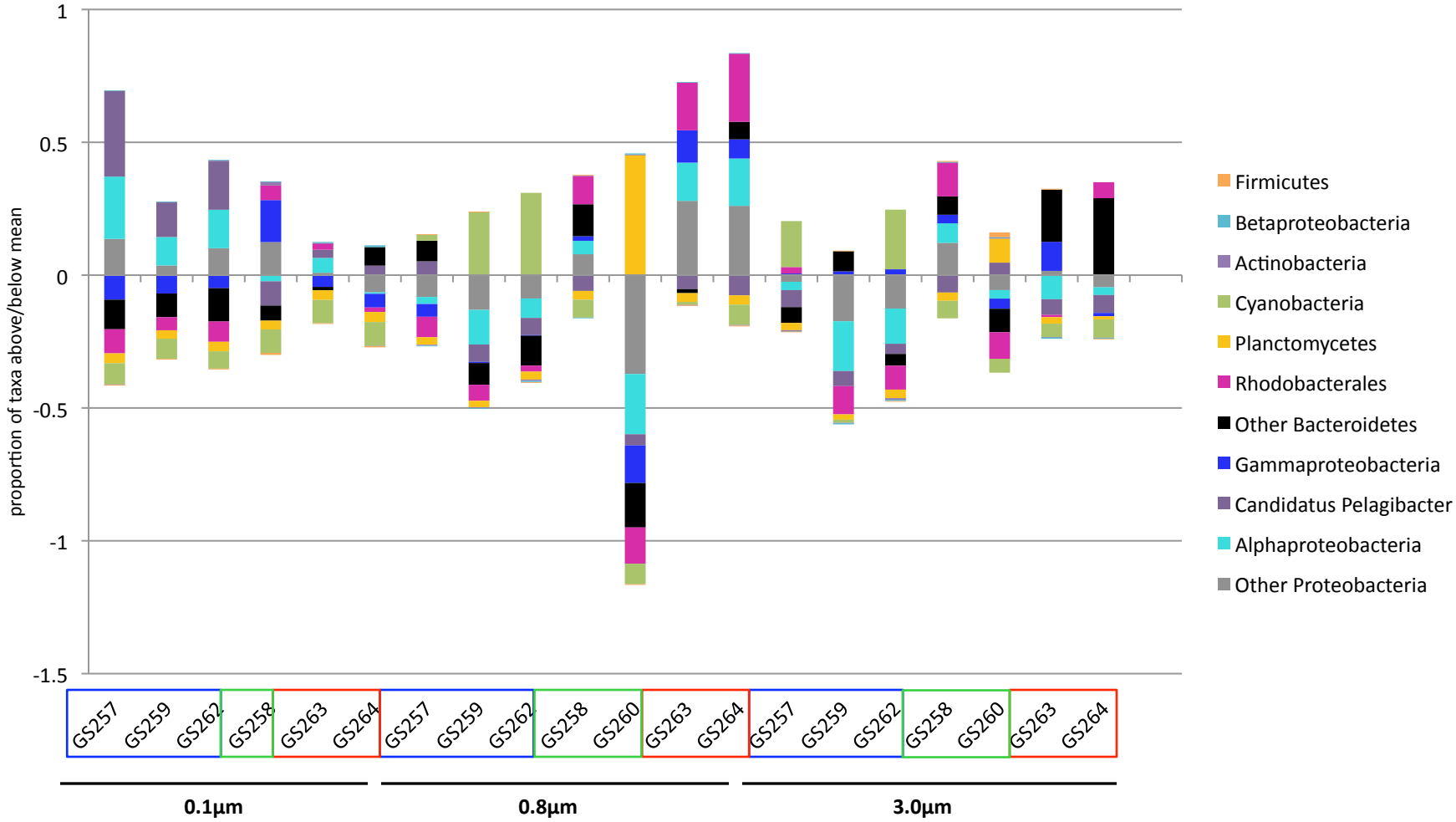

1
0.5
0
-0.5
-1
-1.5
proportion of taxa above/below mean
GS257
GS259
GS262
GS258
GS263
GS264
GS257
GS259
GS262
GS258
GS260
GS263
GS264
GS257
GS259
GS262
GS258
GS260
GS263
GS264
0.1µm
0.8µm
3.0µm
Firmicutes
Betaproteobacteria
Actinobacteria
Cyanobacteria
Planctomycetes
Rhodobacterales
Other Bacteroidetes
Gammaproteobacteria
Candidatus Pelagibacter
Alphaproteobacteria
Other Proteobacteria

Figure S4

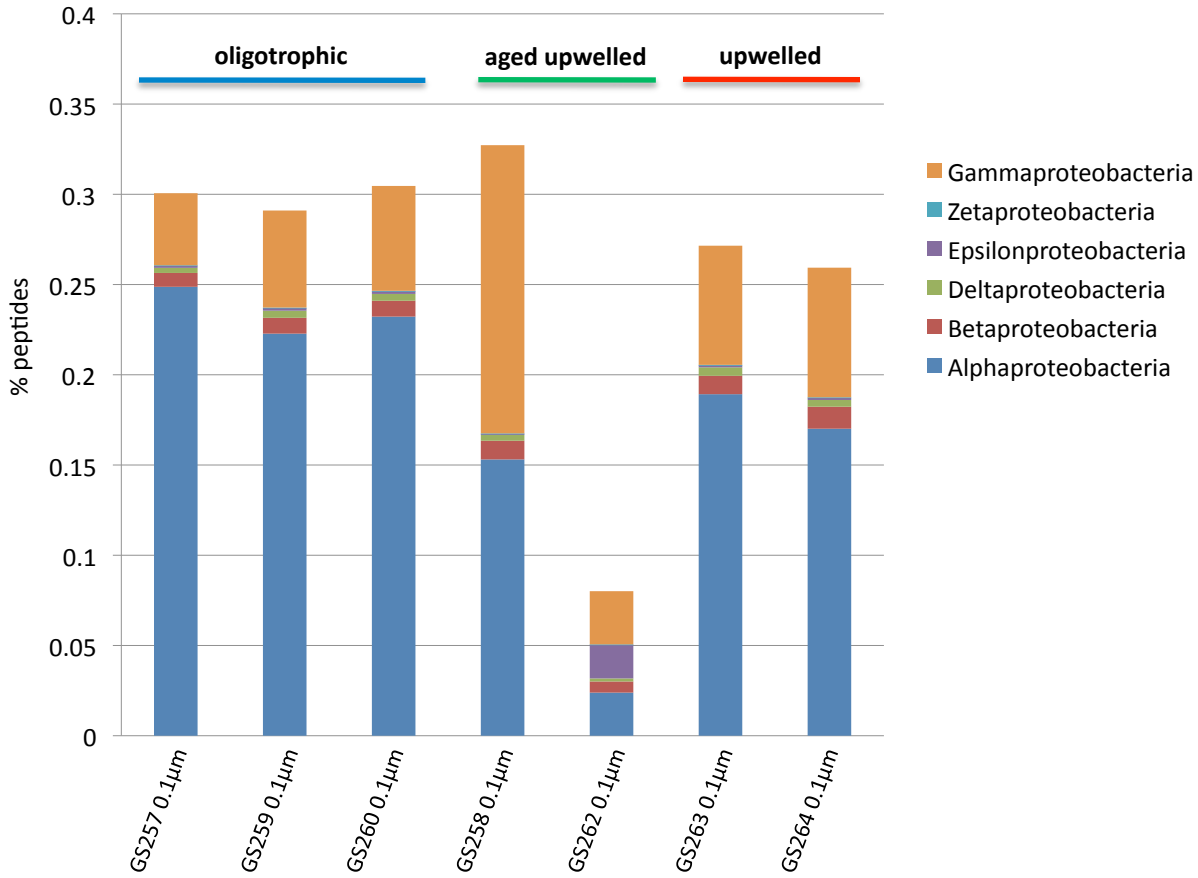

oligotrophic
aged upwelled
upwelled
% peptides
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
GS257 0.1µm
GS259 0.1µm
GS260 0.1µm
GS258 0.1µm
GS262 0.1µm
GS263 0.1µm
GS264 0.1µm
Gammaproteobacteria
Zetaproteobacteria
Epsilonproteobacteria
Deltaproteobacteria
Betaproteobacteria
Alphaproteobacteria

Figure S5

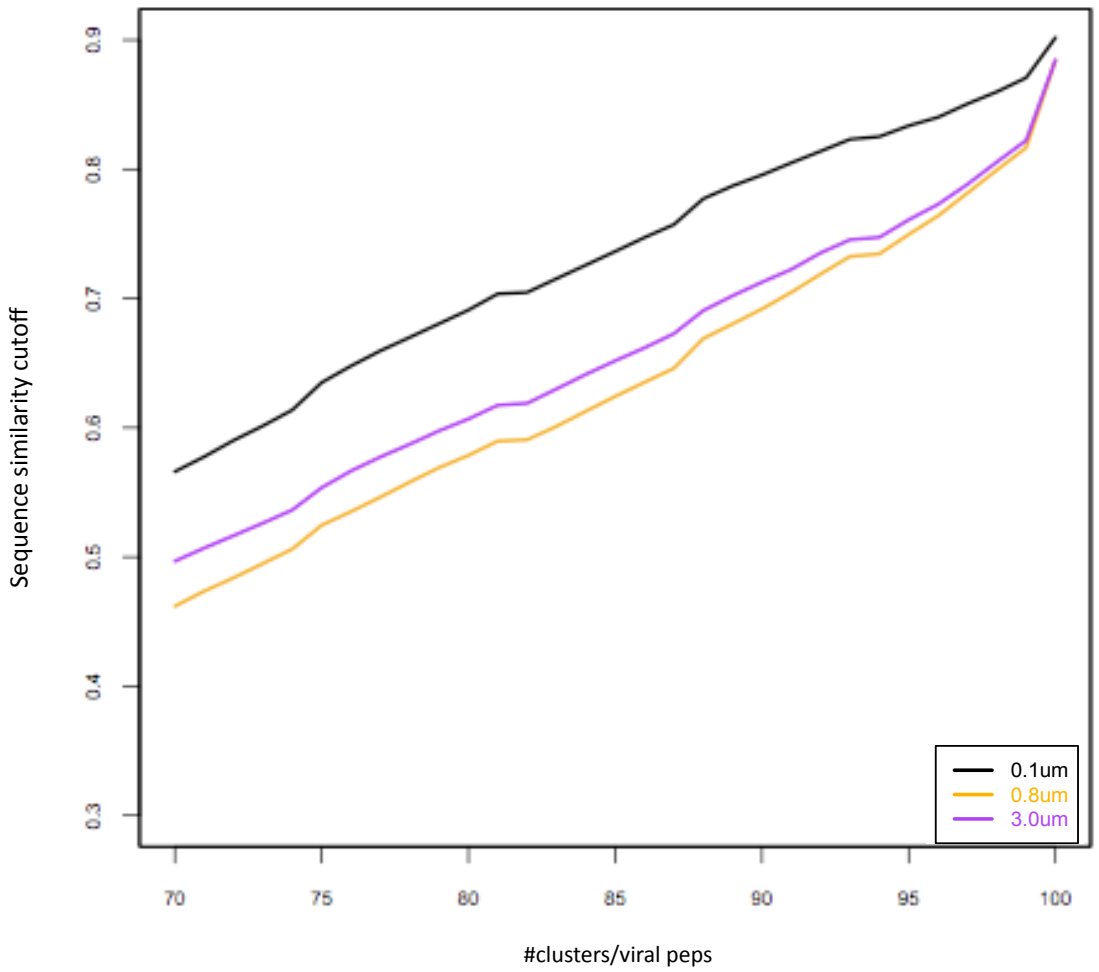

Sequence similarity cutoff
0.3
0.4
0.5
0.6
0.7
0.8
0.9
70
75
80
85
90
95
100
#clusters/viral peps
0.1um
0.8um
3.0um

Figure S6

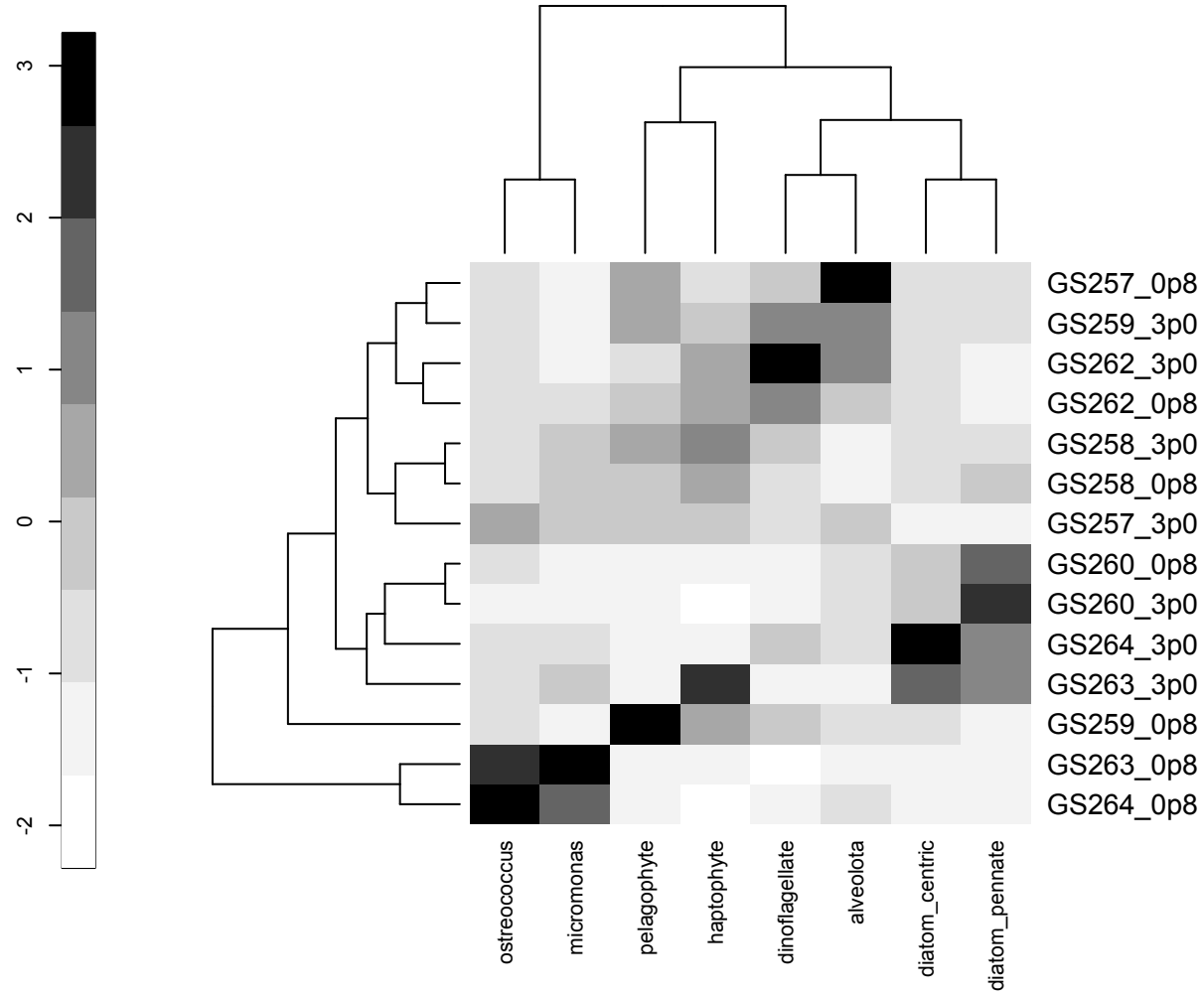

3
2
1
0
-1
-2
GS257_0p8
GS259_3p0
GS262_3p0
GS262_0p8
GS258_3p0
GS258_0p8
GS257_3p0
GS260_0p8
GS260_3p0
GS264_3p0
GS263_3p0
GS259_0p8
GS263_0p8
GS264_0p8
ostreococcus
micromonas
pelagophyte
haptophyte
dinoflagellate
alveolota
diatom_centric
diatom_pennate

Figure S7

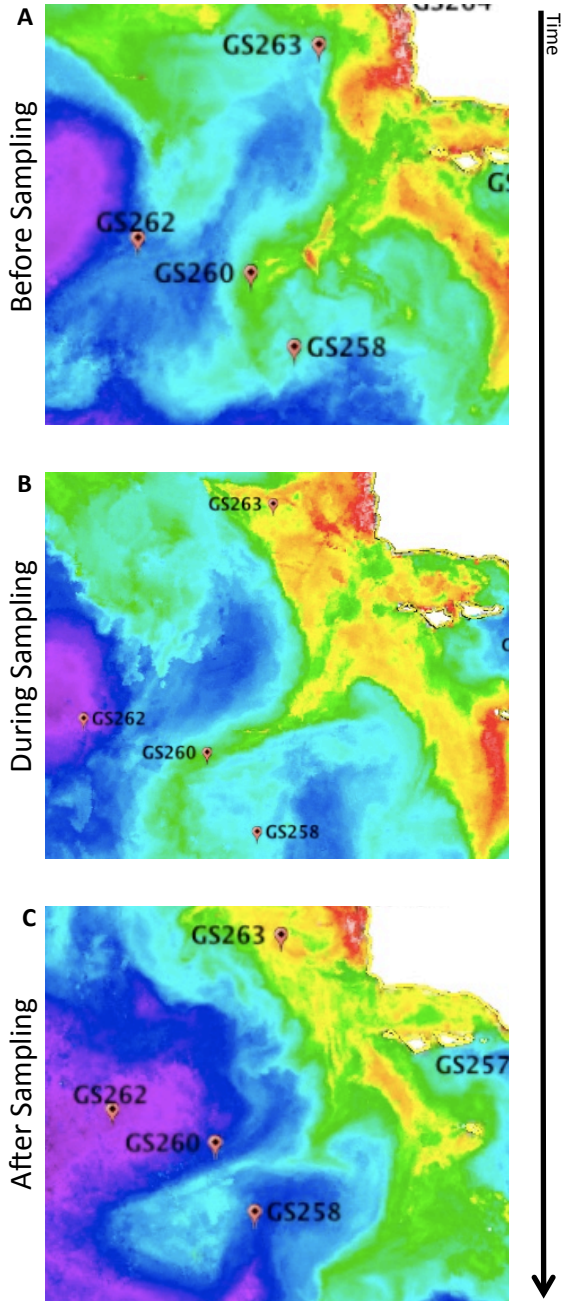

A
Before Sampling
GS263
GS262
GS260
GS258
B
During Sampling
GS263
GS262
GS260
GS258
C
After Sampling
GS263
GS257
GS262
GS260
GS258
Time

Figure S8

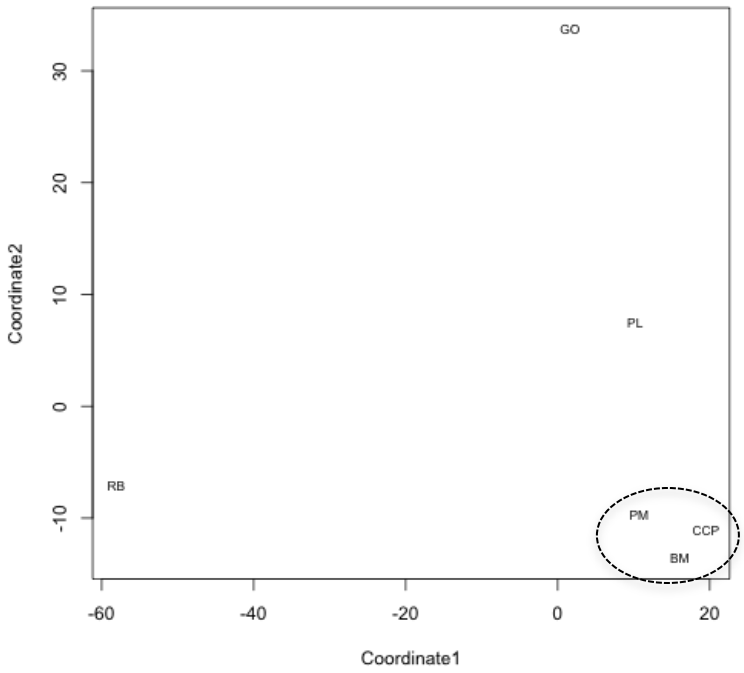

GO
PL
RB
PM
CCP
BM
Coordinate2
30
20
10
0
-10
-60
-40
-20
0
20
Coordinate1

Figure S9

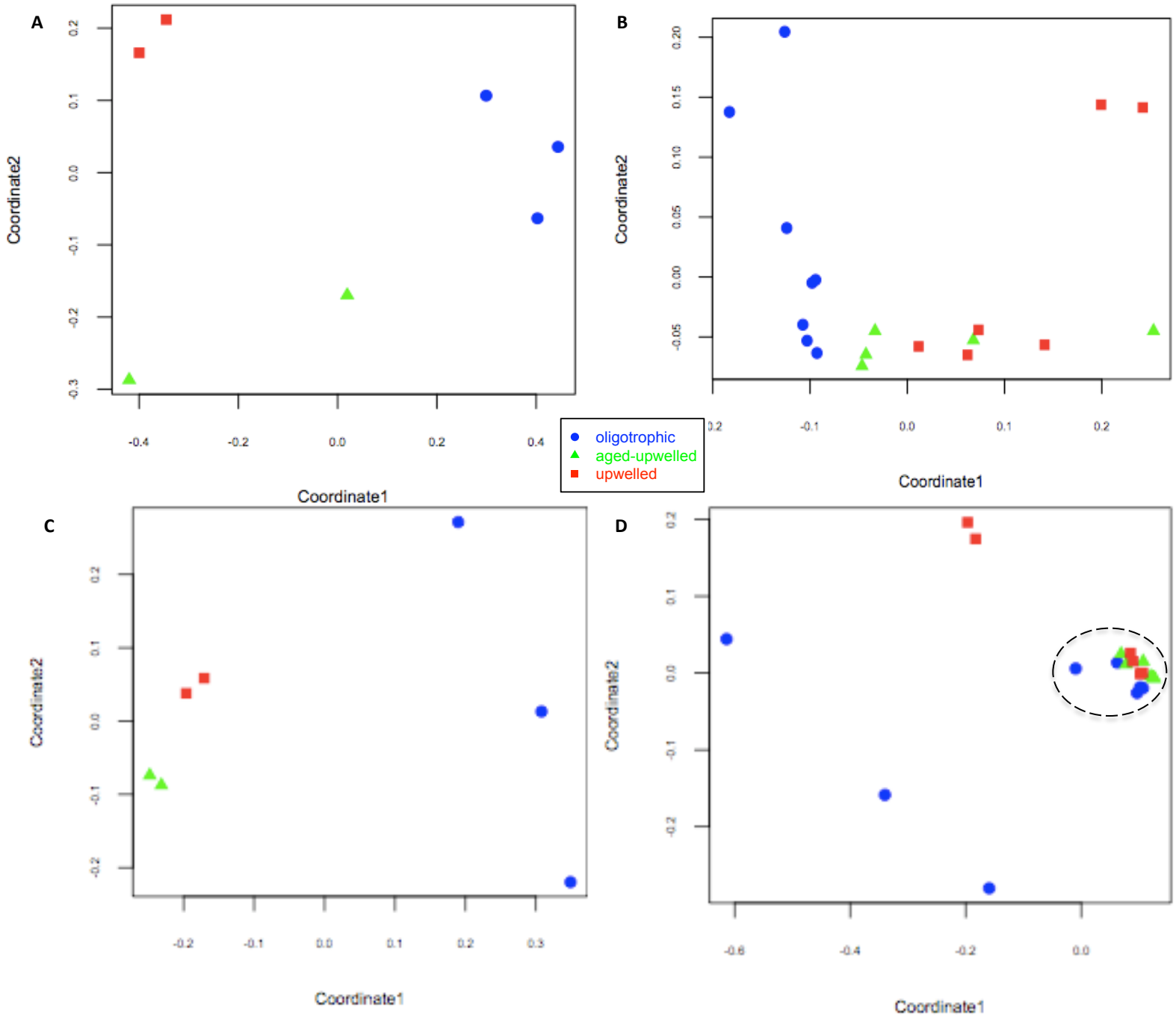
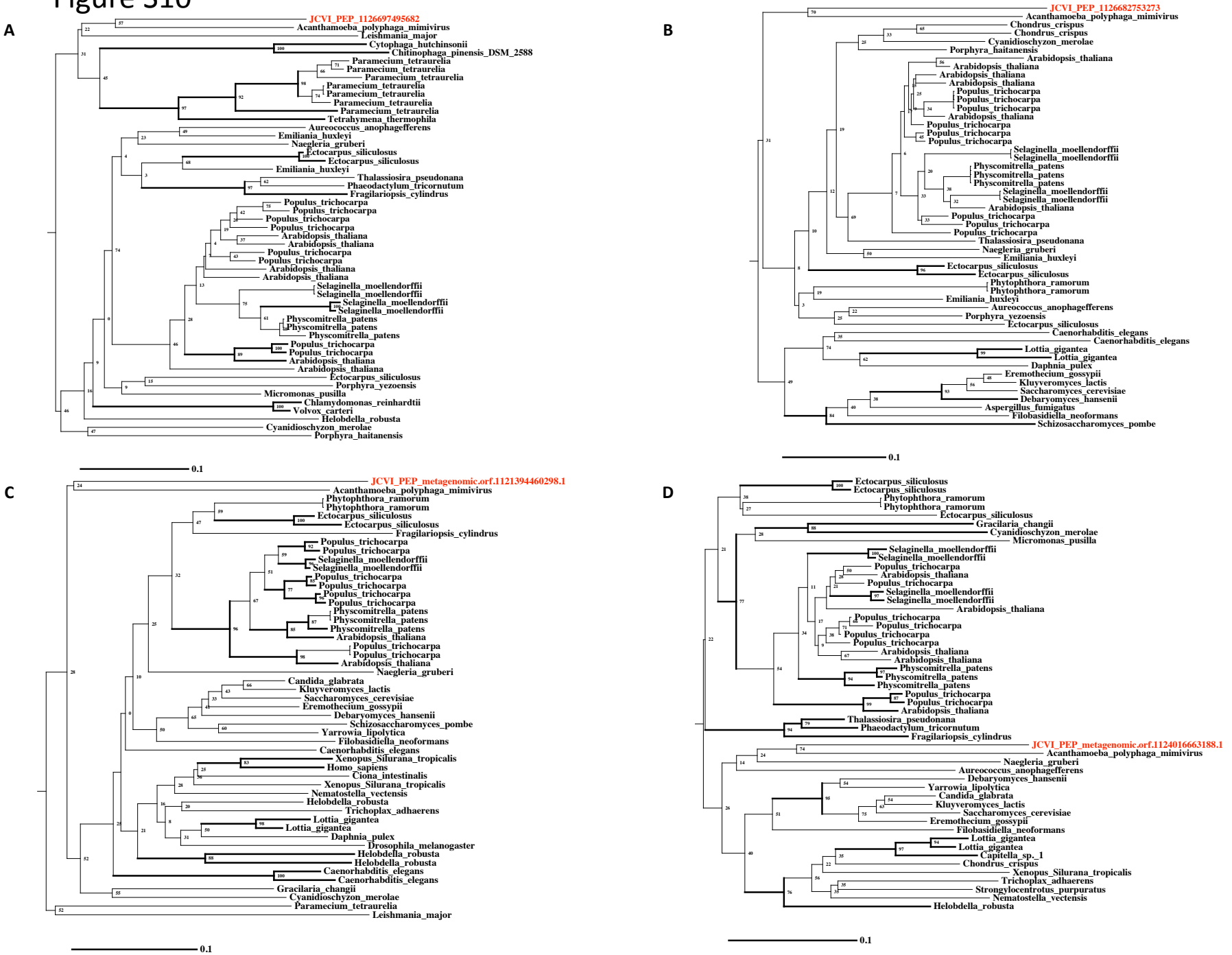
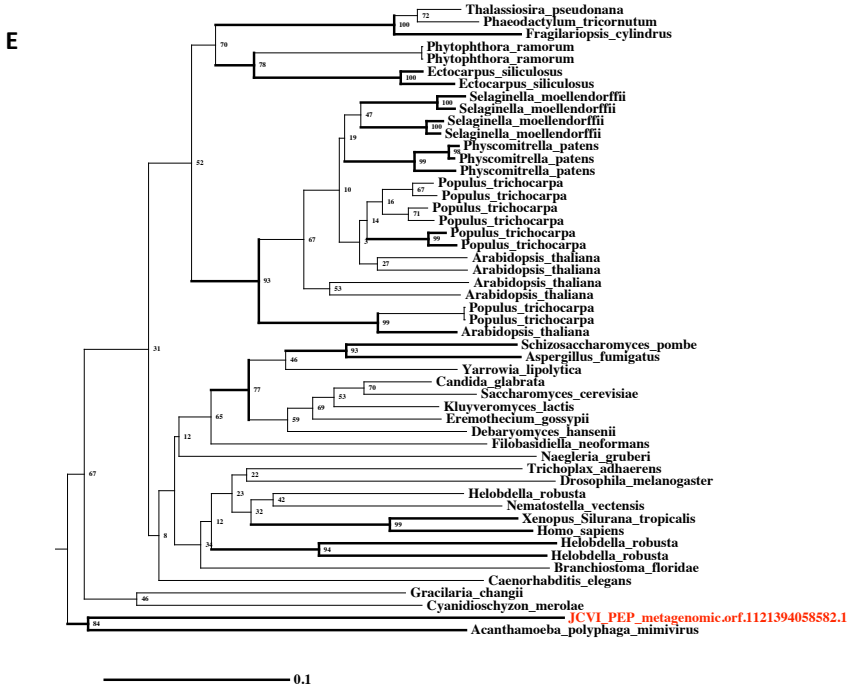
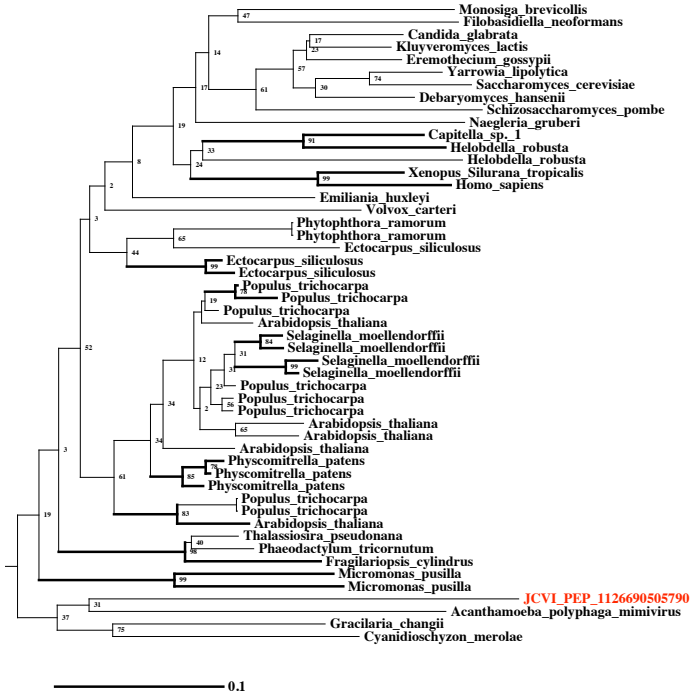
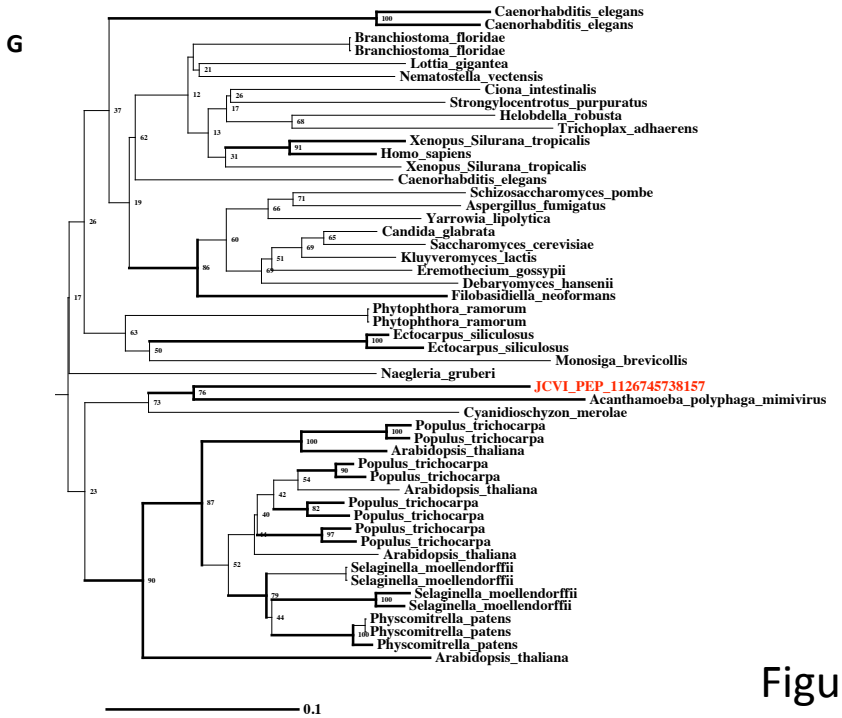

A
B
C
D
Coordinate1
Coordinate2
oligotrophic
aged-upwelled
upwelled

Figure S10


A
JCVI_PEP_1126697495682
Acanthamoeba_polyphaga_mimivirus
Leishmania_major
Cytophaga_hutchinsonii
Chitinophaga_pinensis_DSM_2588
Paramecium_tetraurelia
Paramecium_tetraurelia
Paramecium_tetraurelia
Paramecium_tetraurelia
Paramecium_tetraurelia
Paramecium_tetraurelia
Paramecium_tetraurelia
Tetrahymena_thermophila
Aureococcus_anophagefferens
Emiliania_huxleyi
Naegleria_gruberi
Ectocarpus_siliculosus
Ectocarpus_siliculosus
Emiliania_huxleyi
Thalassiosira_pseudonana
Phaeodactylum_tricornutum
Fragilariopsis_cylindrus
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Arabidopsis_thaliana
Arabidopsis_thaliana
Populus_trichocarpa
Populus_trichocarpa
Arabidopsis_thaliana
Arabidopsis_thaliana
Selaginella_moellendorffii
Selaginella_moellendorffii
Selaginella_moellendorffii
Selaginella_moellendorffii
Physcomitrella_patens
Physcomitrella_patens
Physcomitrella_patens
Populus_trichocarpa
Populus_trichocarpa
Arabidopsis_thaliana
Arabidopsis_thaliana
Ectocarpus_siliculosus
Porphyra_yezoensis
Micromonas_pusilla
Chlamydomonas_reinhardtii
Volvox_carteri
Helobdella_robusta
Cyanidioschyzon_merolae
Porphyra_haitanensis
0.1
B
JCVI_PEP_1126682753273
Acanthamoeba_polyphaga_mimivirus
Chondrus_crispus
Chondrus_crispus
Cyanidioschyzon_merolae
Porphyra_haitanensis
Arabidopsis_thaliana
Arabidopsis_thaliana
Arabidopsis_thaliana
Arabidopsis_thaliana
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Arabidopsis_thaliana
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Selaginella_moellendorffii
Selaginella_moellendorffii
Physcomitrella_patens
Physcomitrella_patens
Physcomitrella_patens
Selaginella_moellendorffii
Selaginella_moellendorffii
Arabidopsis_thaliana
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Thalassiosira_pseudonana
Naegleria_gruberi
Emiliania_huxleyi
Ectocarpus_siliculosus
Ectocarpus_siliculosus
Phytophthora_ramorum
Phytophthora_ramorum
Emiliania_huxleyi
Aureococcus_anophagefferens
Porphyra_yezoensis
Ectocarpus_siliculosus
Caenorhabditis_elegans
Caenorhabditis_elegans
Lottia_gigantea
Lottia_gigantea
Daphnia_pulex
Eremothecium_gossypii
Kluyveromyces_lactis
Saccharomyces_cerevisiae
Debaryomyces_hansenii
Aspergillus_fumigatus
Filobasidiella_neoformans
Schizosaccharomyces_pombe
0.1
C
JCVI_PEP_metagenomic.orf.1121394460298.1
Acanthamoeba_polyphaga_mimivirus
Phytophthora_ramorum
Phytophthora_ramorum
Ectocarpus_siliculosus
Ectocarpus_siliculosus
Fragilariopsis_cylindrus
Populus_trichocarpa
Populus_trichocarpa
Selaginella_moellendorffii
Selaginella_moellendorffii
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Physcomitrella_patens
Physcomitrella_patens
Physcomitrella_patens
Arabidopsis_thaliana
Populus_trichocarpa
Populus_trichocarpa
Arabidopsis_thaliana
Naegleria_gruberi
Candida_glabrata
Kluyveromyces_lactis
Saccharomyces_cerevisiae
Eremothecium_gossypii
Debaryomyces_hansenii
Schizosaccharomyces_pombe
Yarrowia_lipolytica
Filobasidiella_neoformans
Caenorhabditis_elegans
Xenopus_Silurana_tropicalis
Homo_sapiens
Ciona_intestinalis
Xenopus_Silurana_tropicalis
Nematostella_vectensis
Helobdella_robusta
Trichoplax_adhaerens
Lottia_gigantea
Lottia_gigantea
Daphnia_pulex
Drosophila_melanogaster
Helobdella_robusta
Helobdella_robusta
Caenorhabditis_elegans
Caenorhabditis_elegans
Gracilaria_changii
Cyanidioschyzon_merolae
Paramecium_tetraurelia
Leishmania_major
0.1
D
Ectocarpus_siliculosus
Ectocarpus_siliculosus
Phytophthora_ramorum
Phytophthora_ramorum
Ectocarpus_siliculosus
Gracilaria_changii
Cyanidioschyzon_merolae
Micromonas_pusilla
Selaginella_moellendorffii
Selaginella_moellendorffii
Populus_trichocarpa
Arabidopsis_thaliana
Populus_trichocarpa
Selaginella_moellendorffii
Selaginella_moellendorffii
Arabidopsis_thaliana
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Populus_trichocarpa
Arabidopsis_thaliana
Arabidopsis_thaliana
Physcomitrella_patens
Physcomitrella_patens
Physcomitrella_patens
Populus_trichocarpa
Populus_trichocarpa
Arabidopsis_thaliana
Thalassiosira_pseudonana
Phaeodactylum_tricornutum
Fragilariopsis_cylindrus
JCVI_PEP_metagenomic.orf.1124016663188.1
Acanthamoeba_polyphaga_mimivirus
Naegleria_gruberi
Aureococcus_anophagefferens
Debaryomyces_hansenii
Yarrowia_lipolytica
Candida_glabrata
Kluyveromyces_lactis
Saccharomyces_cerevisiae
Eremothecium_gossypii
Filobasidiella_neoformans
Lottia_gigantea
Lottia_gigantea
Capitella_sp._1
Chondrus_crispus
Xenopus_Silurana_tropicalis
Trichoplax_adhaerens
Strongylocentrotus_purpuratus
Nematostella_vectensis
Helobdella_robusta
0.1







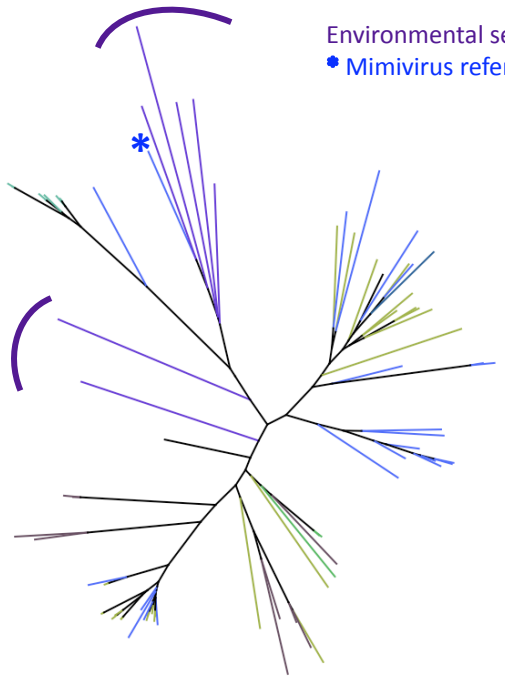


Figure S10 cont.

Figure S11

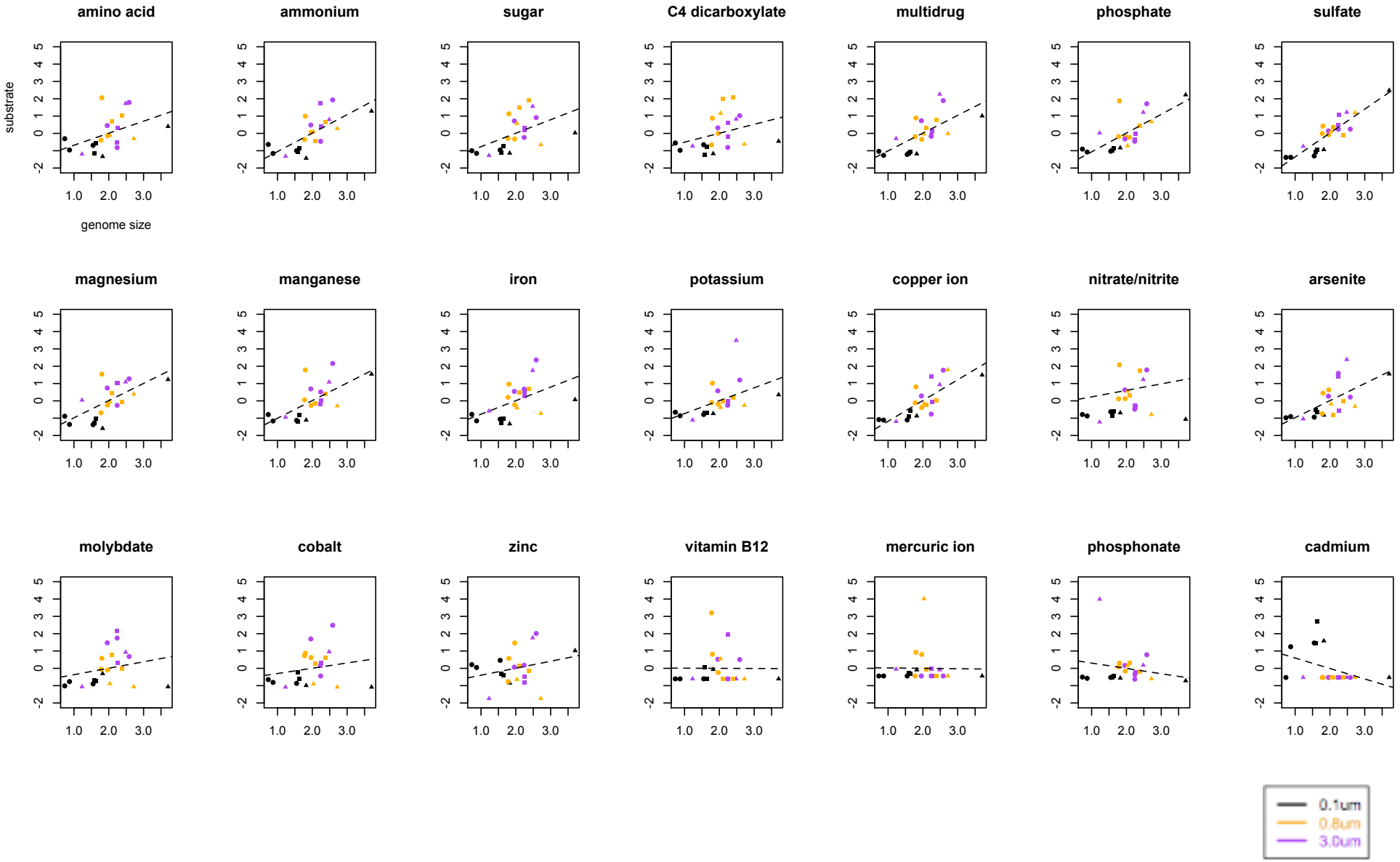
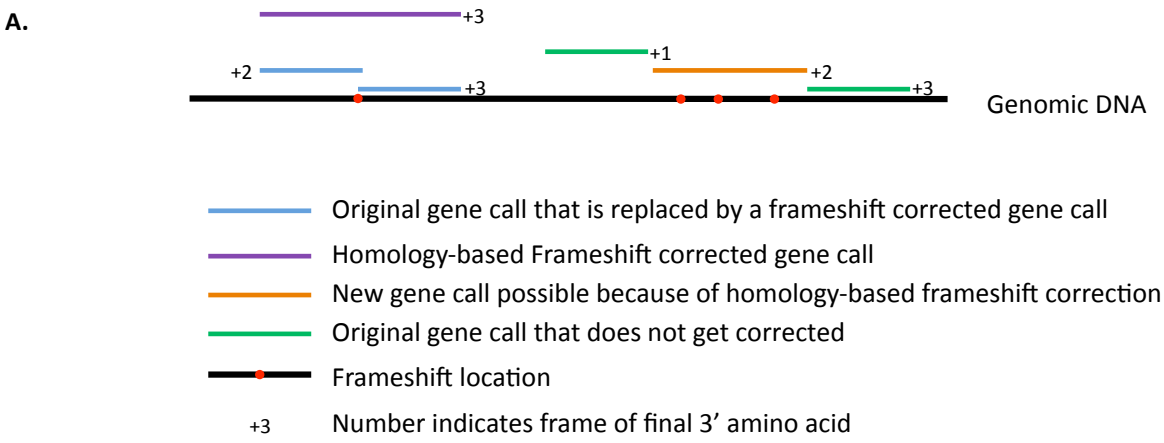
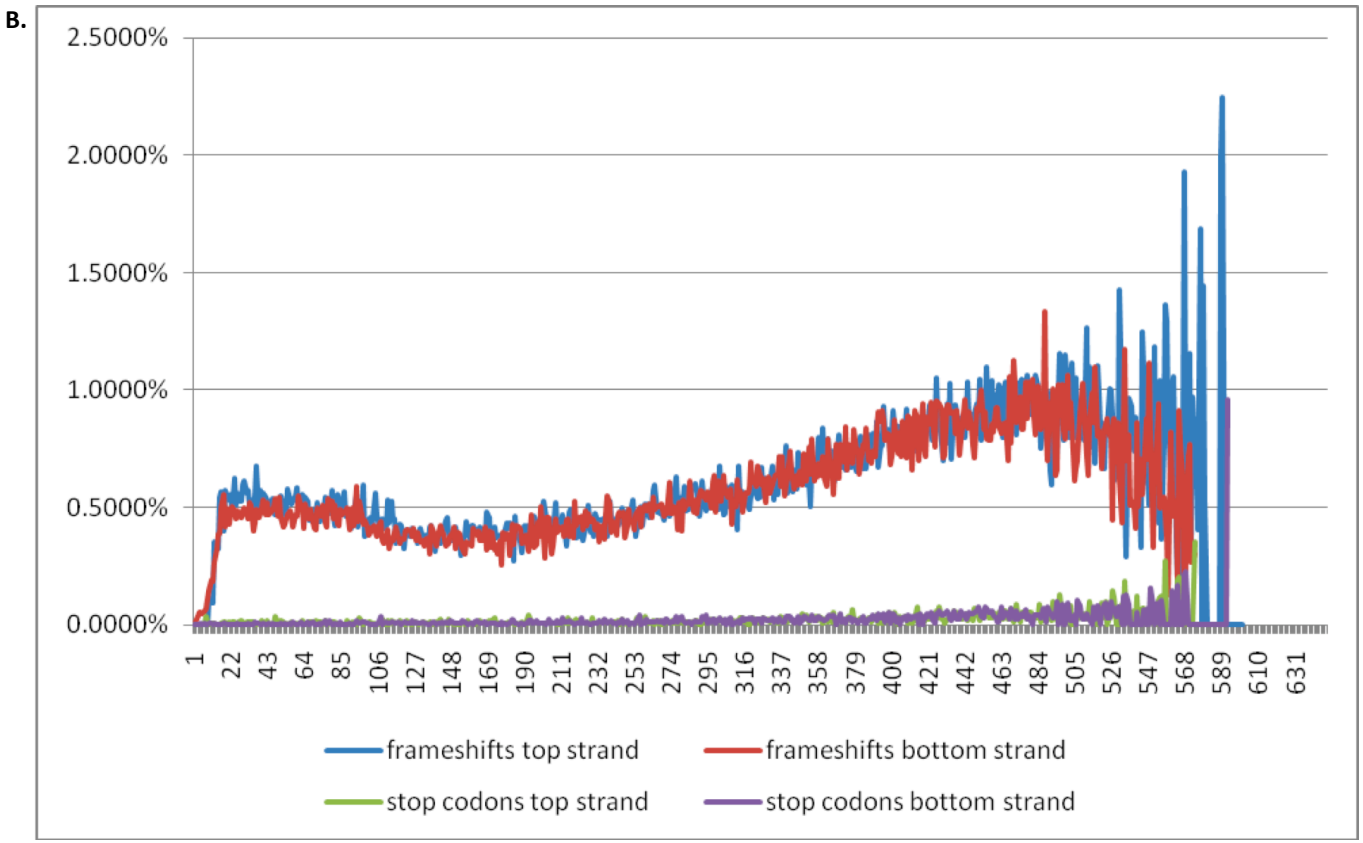

amino acid
ammonium
sugar
C4 dicarboxylate
multidrug
phosphate
sulfate
magnesium
manganese
iron
potassium
copper ion
nitrate/nitrite
arsenite
molybdate
cobalt
zinc
vitamin B12
mercuric ion
phosphonate
cadmium
substrate
genome size
-2 0 1 2 3 4 5
1.0 2.0 3.0
0.1um
0.8um
3.0um

Figure S12


A.
+3
+1
+2
+2
+3
+3
Genomic DNA
Original gene call that is replaced by a frameshift corrected gene call
Homology-based Frameshift corrected gene call
New gene call possible because of homology-based frameshift correction
Original gene call that does not get corrected
Frameshift location
+3
Number indicates frame of final 3' amino acid


B.
2.5000%
2.0000%
1.5000%
1.0000%
0.5000%
0.0000%
1
22
43
64
85
106
127
148
169
190
211
232
253
274
295
316
337
358
379
400
421
442
463
484
505
526
547
568
589
610
631
frameshifts top strand
frameshifts bottom strand
stop codons top strand
stop codons bottom strand