Classification:

Biological Sciences; Ecological sciences

Title:

Metagenome-assembled genomes uncover a global brackish microbiome

Authors:

Luisa W. Hugerth[1], John Larsson[2], Johannes Alneberg[1], Markus V. Lindh[2], Catherine Legrand[2], Jarone Pinhassi[2], Anders F. Andersson[1]

1. KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Division of Gene Technology
   Tomtebodavägen 23A
   SE-17165 Solna
   SWEDEN
2. Centre for Ecology and Evolution in Microbial model Systems - EEMiS, Linnaeus University
   Barlastgatan 11
   SE-39182, Kalmar
   SWEDEN

Corresponding author

Anders F. Andersson

Tomtebodavägen 23A
SE-17165 Solna
SWEDEN

+46 73-9838962

anders.andersson@scilifelab.se

Keywords:

Metagenomics; Brackish water; Marine biology; Microbial ecology

Abstract

Microbes are main drivers of biogeochemical cycles in oceans and lakes. Yet an understanding of the regulation of such processes is hampered by limited genome-context insight into the metabolic potential of bacterial populations. Here we explored an automatic binning approach to reconstruct representative bacterioplankton genomes from metagenomic samples across a time-series in the Baltic Sea. The 30 unique genomes assembled represent novel species within typical marine and freshwater clades. Analysis of the first genomes for abundant lineages entirely lacking reference genomes, such as OM182, acIV and LD19, uncovered divergent ecological adaptations. While phylogenetic patterns in the seasonal succession of the investigated genomes were evident, closely related genomes sometimes displayed distinct seasonal patterns, that could to some extent be explained by gene content. Signs of streamlining were evident in most genomes; and genome sizes correlated with abundance variation across filter size fractions. Comparisons of 86 aquatic metagenomes against the assembled genomes revealed significant fragment recruitment from brackish waters in North America, but little from lakes or oceans, suggesting the existence of a global brackish microbiome. Current estimates of evolutionary rates imply brackish bacteria diverged from freshwater and marine relatives over 100,000 years ago, long before the Baltic Sea was formed (8000 ya) - markedly contrasting the evolutionary history of Baltic Sea macro-organisms, which are locally adapted populations of nearby meta-populations. We have thus demonstrated how metagenome-assembled genomes enable an integrated analysis of ecological patterns, functional potential and evolutionary history of several relevant genomes at a time in natural communities.

Significance statement

Prokaryotes are main drivers of biogeochemical cycles in oceans and lakes. However, a majority of the most abundant populations are still uncharacterized due to difficulties in retrieving cultures and genome information. Here, we used isolation-free metagenomic sequencing and automatic binning to reconstruct genomes from abundant prokaryotic populations of the Baltic Sea, one of world's largest brackish environments. Results showed genome features influencing functional potential and seasonal progression of 30 strains of Bacteria and Archaea. Comparisons to global metagenomic data placed these populations in biogeographic and evolutionary contexts, revealing brackish prokaryotes have a worldwide distribution, having diverged from marine and freshwater representatives long before current brackish water bodies formed. This suggests a global reservoir of brackish lineages that colonise new environments over geological time.

Introduction

Microorganisms in aquatic environments play a crucial role in determining global fluxes of energy and turnover of elements essential to life. To understand these processes through comprehensive analyses of microbial ecology, evolution and metabolism, sequenced reference genomes of representative native prokaryotes are crucial. If these are obtained from isolates, the encoded information can be complemented by phenotypic assays and ecophysiological response experiments to provide insights into factors that regulate the activity of populations of prokaryotes in particular biogeochemical processes. However, obtaining and characterising new pure cultures is invariably a slow process, even with recent advances in high-throughput dilution culturing approaches (1) . Most notoriously, the highly-abundant slow-growing oligotrophic lineages typical of pelagic environments (2, 3) remain severely underrepresented in current culture collections (Yooseph et al).

Metagenomics offers a potential shortcut to much of the information obtained from pure culture genome sequencing (Tyson, 2004, Venter, 2004). The last decade's revolution in DNA sequencing throughput and cost has provided researchers with the unprecedented possibility of obtaining sequences corresponding to thousands of genomes at a time without the need for isolation, cultivation or enrichment. However, despite vast amounts of sequence data allowing inferences on global distribution of phylogenetic lineages and metabolic potentials (4–10), the issue of structuring the data into genomes has remained. This is critical, because while individual genes or genome fragments provide useful information on the metabolic potential of a community, in practice most biochemical transformations take place inside a cell, involving sets of genes structured in controlled pathways. Gaining insight into these pathways requires understanding which genes coexist inside individual microbes. Furthermore, reconstructed genomes from naturally abundant microbes serve as references that allow high-quality annotations to be made in environmental sequencing efforts where otherwise a majority of sequences would remain unclassified (5, 6). Single-cell sequencing has emerged as a very powerful approach to obtain coherent data from individual lineages (3, 11–16). This approach allows researchers to select certain targets of interest, based on e.g. cell characteristics or genetic markers, to address specific research questions (11, 16–20). However, single-cell sequencing requires a highly specialized laboratory facility, and single-amplified genomes (SAGs) typically have low genome coverage, due to the small amount of DNA in each cell and associated whole-genome amplification biases (21).

While metagenomics can offer, in principle, unlimited amounts of starting material and little amplification bias, it has until recently been impossible to automatically reconstruct full genomes from the mass of genome fragments (contigs or scaffolds) generated from complex natural communities. Approaches based on sequence composition, e.g. tetranucleotide frequencies, have been successfully used to reconstruct near-complete genomes from metagenomic contigs without the use of reference genomes, but generally can discriminate only down to the genus-level (22, 23). More recently, coverage variation across multiple samples have started to be used, which allows binning down to the species and sometimes strain level (24–27). At the same time as genomes are reconstructed, the abundance distribution of these genomes across the samples are obtained, allowing ecological

inferences. The CONCOCT (Clustering of contigs based on coverage and composition) approach uses a combination of these two data sources and was shown to give high accuracy and recall on both model and real human gut microbial communities (28), but have not yet been applied to aquatic communities.

The Baltic Sea is, in many aspects, one of the most thoroughly studied aquatic ecosystems (29). It presents unique opportunities for obtaining novel understanding of how environmental forcing determines ecosystem structure and function, due to its strong gradients in salinity (North-Southwest), redox (across depths) and organic and nutrient loading (from coasts to center), as well as pronounced seasonal changes in growth conditions. 16S rRNA gene-based studies have revealed prominent shifts in the microbial community composition along these dimensions (30–34). The community composition of surface waters changes gradually along the 2000 km salinity gradient, from mainly freshwater lineages in the low salinity North to mostly marine lineages in the higher salinity South-West, and a mixture in the mesohaline central Baltic Sea (30). The phylogenetic resolution of 16S amplicons, however, does not permit determining whether the prokaryote populations are locally adapted freshwater and marine populations or represent distinct brackish strains. A recent Baltic Sea metagenomic study showed how a shift in genetic functional potential along the salinity gradient paralleled this phylogenetic shift in bacterial community composition (6). However, since genes were not binned into genomes, different sets of distinguishing gene functions could not be assigned to the genomic context of specific taxa. These considerations suggest that analysis approaches that efficiently provide reference genomes would be invaluable for a richer exploration of available and future omics data.

Here, we used metagenome time-series data from a sampling station in the central Baltic Sea to generate metagenome-assembled genomes (MAGs) corresponding to several of the most abundant, and mostly uncultured, lineages in this environment. We use these data to compare functional potentials between phylogenetic lineages and their relation to seasonal succession. By comparing the MAGs with metagenome data from globally distributed sites, we propose that these are specialised brackish populations, that evolved long before the formation of the Baltic Sea, and whose closest relatives are today found in other brackish environments across the globe.

Results and Discussion

MAG reconstruction

We conducted shotgun metagenomic sequencing on 37 surface water samples collected from March to December in 2012 at the Linnaeus Microbial Observatory, 10 km east of Öland in the central Baltic Sea. On average, 14.5 million read-pairs were assembled from each sample, yielding a total of 1,443,953,143 bp across 4,094,883 contigs. In order to bin contigs into genomes, the CONCOCT software (28) was run on each assembled sample separately, using information on the contigs' coverages across all samples (Fig. S1). Single-copy genes (SCGs) were used to assess completeness and purity of the bins. We defined approved bins as having at least 30 of 36 SCGs present (Supplementary file S1), of which not more than two in multiple copies. This resulted in the identification of 83 genomic

bins, hereafter referred to as metagenome assembled genomes (MAGs). The completeness of these MAGs was further validated by assessing the presence and uniqueness of a set of phylum- and class-specific SCGs (Supplementary file 1). Based on these, we estimate the MAGs to be on average 82.7% complete with 1.1% of bases misassembled or wrongly binned. Some MAGs were estimated to be 100% complete. In comparison, recent single-amplified genome studies of free-living aquatic bacteria have obtained average completeness of 55-68% (3, 14). Importantly, the number of MAGs reconstructed from each sample correlated with the number of reads generated from it and there is no sign of saturation in this trend (Fig. S2), meaning that more genomes can easily be reconstructed by deeper sequencing of the same samples. Every sample with over 20 million reads passing quality control yielded at least 3 approved bins. Further, while only highly complete genomes were selected for this study, other research questions might be adequately addressed with partial genomes, many more of which were generated.

In the original CONCOCT study (28), binning was done on a coassembly of all samples. Here we employed an alternative strategy where binning was run on each sample separately, this way community complexity was minimized and binning accuracy increased. Since this strategy may reconstruct the same genome multiple times over the time-series, the 83 complete MAGs were further clustered based on sequence identity. Thirty distinct clusters (BACL [BAltic CLuster] 1-30) with >99% intra-cluster sequence identity were formed (<70% between-cluster identity; 95% sequence identity is a stringent cut-off for bacterial species definition (35), that included between one and 14 MAGs each (Fig. S3; Table 1 ; Supplementary file 1). Having several MAGs in the same cluster increases the reliability of the analyses performed, especially in the case of results based on the absence of a sequence, such as missing genes.

The genome clusters generated represent environmentally abundant strains, together corresponding to on average 13% of the shotgun reads in each sample (range: 4 - 23%). This shows that the CONCOCT approach successfully reconstructs novel genomes of environmentally relevant bacteria.

Table 1: Overview of clusters, sorted by taxonomy.

Cluster	**Num MAGs**	**Avg bin size (Mb)**	**%Coding**	**%GC**	**Taxonomy**	**% Abundance avg (max)**	**Completenes s avg (max) %**
BACL2	7	1.07	94.3	44.3	Actinobacteria; acI	1.20 (6.47)	78.7 (88.2)
BACL4	6	1.01	94.8	41	Actinobacteria; acI	0.36 (1.16)	80.8 (90.4)
BACL15	2	1.11	94.9	47.1	Actinobacteria; acI	0.37 (1.19)	84.2 (93.4)
BACL6	4	1.55	94.8	51.3	Actinobacteria; acIV	0.34 (2.92)	80.0 (86.0)
BACL17	2	1.45	95.5	52.3	Actinobacteria;	0.21 (1.10)	82.7 (93.4)

					acIV		
BACL19	1	1.26	95.3	58.1	Actinobacteria; acIV	0.09 (0.57)	59.6
BACL27	1	1.67	94.7	50.4	Actinobacteria; acIV	0.29 (1.73)	69.9
BACL25	1	1.26	93.5	55.7	Actinobacteria; Luna	0.08 (0.62)	77.2
BACL28	1	1.12	94	51.7	Actinobacteria; Luna	0.13 (0.96)	65.4
BACL10	3	2.57	90.2	50.5	Alphaproteobacteria; Rhodobacter	1.22 (10.91)	84.0 (88.7)
BACL5	5	1.08	96.5	30.1	Alphaproteobacteria; SAR11	0.79 (2.92)	80.2 (89.5)
BACL20	1	1.14	95.8	30.9	Alphaproteobacteria; SAR11	0.59 (2.56)	66.9
BACL7	3	1.74	95.4	49	Bacteroidetes; Cryomorphaceae	0.32 (1.40)	99.4 (100)
BACL11	3	1.19	96	32.9	Bacteroidetes; Cryomorphaceae	0.43 (1.70)	75.1 (84.9)
BACL18	2	1.32	94.5	57.5	Bacteroidetes; Cryomorphaceae	0.18 (0.93)	78.6 (85.7)
BACL23	1	1.73	95.3	54.6	Bacteroidetes; Cryomorphaceae	0.15 (1.45)	98.3
BACL8	3	1.74	93.6	38.7	Bacteroidetes; Flavobacteriaceae	0.51 (1.54)	89.4 (98.3)
BACL21	1	1.92	93.1	44.1	Bacteroidetes; Flavobacteriaceae	0.21 (1.09)	97.5
BACL22	1	2.41	91	32.1	Bacteroidetes; Flavobacteriaceae	0.14 (0.98)	91.6
BACL29	1	1.48	95.2	30	Bacteroidetes; Flavobacteriaceae	0.08 (0.49)	88.2
BACL12	2	2.64	93.6	47.2	Bacteroidetes; Sphingobacteriales	0.20 (2.14)	85.3 (93.3)
BACL14	2	1.19	94.2	38.7	Betaproteobacteria	0.42(1.44)	86.5 (87.8)

					; OM43		
BACL30	1	1.81	92.2	63.6	Cyanobacteria; Cyanobium	0.38 (1.26)	79.2
BACL3	6	2.23	90.5	53	Gammaproteobacteria; OM182	0.80 (3.54)	85.8 (90.9)
BACL1	14	1.37	95.2	40	Gammaproteobacteria; SAR86	1.74 (5.23)	85.9 (90.9)
BACL16	2	2.26	91	51.3	Gammaproteobacteria; SAR92	0.37 (2.41)	96.2 (97.2)
BACL26	1	1.91	91	46	Gammaproteobacteria; SAR92	0.13 (0.70)	83.2
BACL13	2	1.01	90.9	31.9	Thaumarchaeota; Nitrosopumilaceae	0.11 (0.88)	70.0 (78.4)
BACL9	3	1.50	93.7	56.1	Verrucomicrobia; LD19	0.36 (1.54)	73.6 (76.4)
BACL24	1	2.98	88.4	53.1	Verrucomicrobia; Opitutaceae	0.08 (2.25)	92.1

Phylogeny and functional potential

The reconstructed genomes belong to Actinobacteria, Bacteroidetes, Cyanobacteria, Verrucomicrobia, Alpha-, Beta and Gammaproteobacteria and Thaumarchaeota (Table 1, Fig. 1; Fig. S4, Supplementary file 2). Phylogenetic reconstruction using concatenated core proteins placed all MAGs consistently within clusters, verifying that the binning was reproducible (Fig. S4). Based on average nucleotide identity, only BACL8 was estimated to have >70% DNA identity with its nearest neighbour in the phylogenetic tree. In this and many other cases, the closest relative was not an isolate, but a SAG, reflecting these methods' ability to recover genomes from abundant, but potentially uncultivable, species.

The broad phylogenetic representation allowed us to compare functional potential between taxonomic groups in this ecosystem. Non-metric multidimensional scaling based on abundance of functional gene categories grouped the MAG clusters according to their phylogeny (Fig. 2; Fig. S5; Supplementary File 3), also confirmed by ANOSIM analysis (Supplementary table 1). Specifically, Alphaproteobacterial clusters encoded a significantly higher proportion of genes in the "Amino acid transport and metabolism" COG category compared to all other clusters (Welch's t-test $p<0.001$). Actinobacteria were significantly enriched in genes in the "Carbohydrate transport and metabolism" COG category ($p=0.04$). Enzymes involved in carboxylate degradation, such as formyl-CoA transferase (EC:2.8.3.16) of the oxalate degradation pathway, were significantly more abundant in

Gammaproteobacteria compared to all other clusters (p=0.019). Carboxylate degradation enzymes were also abundant in Alphaproteobacteria and Bacteroidetes, but significantly lower in proportion among the Actinobacteria (p<0.01). Bacteroidetes and Verrucomicrobia had the largest number of of glycoside hydrolase genes, including xylanases, endochitinases and glycogen phosphorylases (Fig. S6), and thus appear well suited for degradation of polysaccharides such as cellulose, chitin and glycogen, in line with previous findings (13, 36–39).

Transporter proteins mediate many of the interactions between a cell and its surroundings, thus providing insights into an organism's niche. A detailed analysis of transporter genes in the 30 MAG clusters (Fig. S7; Supplementary File 4) revealed important general patterns, such as a high diversity of genes for amino acid uptake in Actinobacteria and Alphaproteobacteria, a lack of genes for carboxylic acid uptake and a multitude of genes for polyamine uptake in Actinobacteria, and a high diversity of ABC-type sugar transport genes in Actinobacteria and Alphaproteobacteria. The Gammaproteobacteria, Bacteroidetes and Verrucomicrobia encoded a large number of TonB-dependent transporter genes, likely involved in carbohydrate, vitamin and iron chelator uptake in these clades (40). Phosphate uptake systems, such as the high affinity PstS transporter, were highly abundant in the Betaproteobacterial BACL14, while Thaumarchaeon BACL13 had the highest proportion of phosphonate transporter genes, followed by the Cyanobacterium BACL30 and acIV genome clusters. The BACL30 and the BACL13 were sparse in uptake systems for organic molecules in general, consistent with these organisms' photoautotrophic and chemoautotrophic lifestyles, respectively. In addition to the *amoA* ammonia monooxygenase gene, BACL13 encoded urease genes, indicating the capability to also utilise urea for nitrification, as previously observed in Arctic *Nitrosopumilus (41)*. In line with the OM43 clade comprising simple obligate methylotrophs with extremely small genomes (42), betaproteobacterial OM43 cluster (BACL14), which encoded a methanol dehydrogenase gene and genes for formaldehyde assimilation, was also sparse in uptake systems.

All clusters belonging to the Bacteroidetes and Gammaproteobacteria lineages contained the Na+-transporting NADH:ubiquinone oxidoreductase (NQR) enzyme. This enzyme is involved in the oxidative respiration pathway in some bacteria and is similar to the typical H+-transporting ndh NADH dehydrogenase (43)). However, the NQR enzyme exports sodium from the cell and thereby creates a gradient of Na+ ions, in contrast to the proton gradient generated by the ndh enzyme. The use of the NQR enzyme has been shown to be correlated with salinity (increasing Na+ concentrations) in bacterial communities (6). NQR-containing MAG clusters were generally the ones with closest relatives in the marine environment (e.g. Bacteroidetes, see section on biogeography below), while genome clusters more closely related to freshwater bacteria (e.g. Actinobacteria) contained the H+-transporting enzyme. An exception to this were the SAR11 MAGs, which harbored the H+-transporting enzyme despite having predominantly marine relatives. The genomes containing NQR enzymes in our dataset also contained a significantly higher proportion of Na+ symporters and antiporters (for e.g. dicarboxylates, disaccharides and amino-acids), as well as TonB-dependent transporters, compared to the other genomes (Welch's t-test p<0.001). In contrast, ATP-driven ABC-transporters were significantly less abundant in these clusters

($p<0.001$), strongly indicating that these bacteria have reduced their energy requirement by making use of the sodium motive force generated by the NQR enzyme to drive transport processes, a trade-off that has been suggested previously (43). Interestingly, TonB-dependent transporters also require energy derived from charge separation across cellular membranes, in this case in the form of the proton motive force (44). The significant enrichment of TonB transporters in NQR-containing genomes suggests that these proteins may also utilize the sodium motive force.

Novelly sequenced lineages

The MAG approach has previously proven useful for closing gaps in the tree of life by the reconstruction of genomes from uncultivated species (e.g. (45–47). Here we report the first draft genomes for the oligotrophic marine Gammaproteobacteria OM182, and for the typically freshwater Verrucomicrobia subdivision LD19 and the Actinobacteria clade acIV . Annotations for these genomes are found in Supplementary file 3 and Supplementary file 4.

OM182 is a globally abundant gammaproteobacteria which has been grown in enrichment culture, but never sequenced. BACL3 includes a 16S gene 99% identical to that of the OM182 isolate HTCC2188 (48). This MAG cluster shares common features with other Gammaproteobacteria, such as a variety of glycoside hydrolases and carboxylate degradation enzymes. It also encodes the *cysA* sulfate transporter and a complete set of genes for assimilatory sulfate reduction to sulfide and for production of cysteine from sulfide and serine via *cysK* and *cysE*. Genes for sulfite production from both thiosulfate (via *glpE*) and taurine (via *tauD*) are also encoded in the genome, and this is the only MAG cluster to encode the full set of genes for intracellular sulfur oxidation (*dsrCEFH*). BACL3 thus appears remarkably well-suited for metabolising different inorganic and organic sulphur sources, the latter potentially originating from phytoplankton blooms (49), even more so than previously sequenced isolates of oligotrophic marine gammaproteobacteria (50, 51).

Two verrucomicrobial genome MAG clusters were reconstructed. BACL9 MAGs include 16S rRNA genes 99% identical to that of the globally distributed freshwater clade LD19 (52), a subdivision within the Verrucomicrobia first identified in 1998 (53), but still lacking cultured or sequenced representatives. Previous 16S-based analysis placed LD19 as a sister group to a subdivision with acidophilic methanotrophs (54). Accordingly, BACL9 is placed as a sister clade to the acidophilic methanotroph *Methylacidiphilum infernorum* (55) in the genome tree, but does not encode methane monooxygenase genes and thus lacks the capacity for methane oxidation seen in *M. infernorum*. Interestingly, BACL9 contains a set of genes that together allow for production of 2,3-butanediol from pyruvate (via acetolactate and acetoin). Butanediol plays a role in regulating intracellular pH during fermentative anaerobic growth and biofilm formation (56). This is also the only MAG with the genetic capacity to synthesize hopanoid lipids, which have been implicated in enhanced pH tolerance in bacteria by stabilizing cellular membranes (57, 58). This indicates an adaptation to withstanding lowered intracellular pH such as that induced by fermentative growth under anaerobic conditions. Such conditions occur in biofilms (56), and it remains to be shown whether these planktonic bacteria can form biofilms to grow attached to particles in the water column.

BACL 6, 17, 19 and 27 all belong to actinobacterial clade acIV, of the order Acidimicrobiales. Most isolates of the order Acidimicrobiales are acidophilic, and no genomes have been reported for acIV, despite its numeric importance in lake water systems (59). Compared to the other typically freshwater clades acI (BACL 2, 4, 15) and Luna (BACL 25, 28), that instead belong to the order Actinomycetales, acIV MAG clusters had larger genome sizes and contained a significantly lower proportion of genes in the Carbohydrate transport and metabolism COG category ($p<0.01$), in particular ABC-type sugar transporters (Fig. S7). AcIV and acI were also impoverished for phosphotransferase (PTS) genes and amino acid transporters, compared to Luna MAGs. In contrast, acIV MAG clusters contained a significantly higher proportion of genes in the Lipid transport and metabolism COG category ($p=0.02$), and a significantly higher total proportion of enzymes involved in fatty-acid oxidation ($p<0.001$), indicating that these Actinobacteria may use lipids as a carbon source.

The only cyanobacterial genome assembled was BACL30. While it is placed in the phylogenetic tree as a distant neighbour to *Cyanobium gracile*, its 16S gene is only 97% identical with it, the same identity as with *Synechococcus* and *Prochlorococcus*. This genome contains genes for the pigments phycocyanin (PC) and phycoerythrin (PE) and belongs to the Type IIB pigment gene cluster recently identified as being dominant in the Baltic Sea (60). The PC genes *cpcBA* and the intergenic spacer are 100% identical to sequences in the Type IIB pigment clade. Phylogenies of PC and PE subunits as well as 6 ribosomal proteins consistently placed this cyanobacterial MAG within the Type IIB pigment clades and within the clade of picocyanobacteria whose members are abundant in the Baltic Sea, but for which a reference genome has been unavailable (Fig. S8). BACL30 contains the high affinity *pstS* phosphate transporter, but lacks the *phoU* regulatory gene as well as an alkaline phosphatase. In this respect the genome is similar to the coastal strain *Synechococcus* CC9311 (61), likely reflecting adaptation to higher phosphorous loads compared to the open oceans.

Genome streamlining and inferred cell sizes

Oligotrophic bacterioplankton are characterised by streamlined genomes, i.e. small genomes with high coding densities and low numbers of paralogs (62). For the few cultured oligotrophs, such as *Prochlorococcus* (63) and SAR11 (64), this coincides with small cell sizes. The small cells render high surface-to-volume ratios, beneficial for organisms that compete for very-low-concentration nutrients (65). SAG sequencing has shown that genomic streamlining is a widely distributed feature among abundant bacterioplankton (3), contrasting to most cultured marine bacteria. Lauro et al identified genome features for predicting if an organism or community is oligotrophic or copiotrophic (2). Ordination using some of these features (coding density, GC-content and proportion of five COG categories; Swan, 2013) separated our MAG clusters from marine isolate genomes (Fig. 3A). The exceptions were isolates of picocyanobacteria, SAR11 and OM43, that overlapped with our genome MAG clusters, and the SAR92, OM182 and Opitutaceae genome MAG clusters that overlapped with the isolates. Hence, most of the MAGs displayed pronounced signs of streamlining. These features, with

the exception of GC-content, were found to be highly correlated with genome size (Fig. S9), and genome size alone gave equally strong separation (Fig. 3B-C).

Interestingly, several of the Bacteroidetes MAG clusters appear to be streamlined, despite Bacteroidetes being generally described as copiotrophic (62). One of them (BACL11), which represents a novel branch in the Cryomorphaceae (Fig. 1), has a particularly small genome (1.19 Mbp [range 1.16 - 1.21] MAG size, at 75% estimated completeness) with only 4% non-coding DNA. It encodes a smaller number of transporters than the other Bacteroidetes MAG clusters and only one type of glycoside hydrolase. It also has a comparatively low GC-content (33%). However, the Polaribacter MAG cluster (BACL22), which has the largest genome and lowest gene density of the Bacteroidetes genome MAG clusters, has equally low GC-content (32%), as previously observed in planktonic and algae-attached *Polaribacter* isolates (66–73). Since, in general, GC-content correlates very weakly with either genome size or gene density (Fig. S9), this may not be an optimal marker for genome streamlining. Supporting the impression that the MAGs represent small and streamlined genomes, with little metabolic flexibility, most MAG clusters (25 of 30) encode bacteriorhodopsins (PF01036, Supplementary file S3), which allows them to adopt a photoheterotrophic lifestyle when their required substrates for chemoheterotrophy are not available.

By mapping shotgun reads from different filter fractions (0.1 - 0.8, 0.8 - 3.0 and >3.0 µm) from a previous spatial metagenomic survey of the Baltic Sea (Dupont, 2014), we could investigate how our MAG cluster cells were distributed across size fractions. Comparing counts of mapped reads between the 0.8 - 3.0 and 0.1 - 0.8 fractions showed that Bacteroidetes tended to be captured on the 0.8 um filter to a higher extent than Actinobacteria (Fig. 3D). This bias could be driven by Bacteroidetes being, to a higher extent, attached to organic matter particles or phytoplankton (74). However, comparing the >3.0 um with the 0.8 - 3.0 um fraction showed a clear bias only for one of the Bacteroidetes clusters (BACL12; Fig. 3E). This cluster has the largest genomes (2.5 and 2.8 Mbp) of the reconstructed Bacteroidetes and is the only representative of the Sphingobacteriales (Fig. 1). Sphingobacteria have previously been suggested to bind to algal surfaces, explained in part by the presence of glycosyltransferase genes (75). We did not find significantly more glycosyltransferases in BACL12 than in the other Bacteroidetes. Instead it encodes a greater number of genes containing carbohydrate-binding module (CBM) domains than the other clusters (12 vs. 1.4 average of all clusters and 1.3 for the other Bacteroidetes), which may facilitate adhesion to particles or phytoplankton (76). Since only one Bacteroidetes MAG cluster was biased toward the >3 um filter, attachment to organic matter doesn't seem to be the main reason behind the difference in filter capture between Bacteroidetes and Actinobacteria, unless the particles are mainly in the 0.8 - 3.0 um size range. Another possibility is that the bias reflects cell size distributions; each population has a specific size distribution that will influence what proportion of cells will pass through the membrane, as indicated previously (77). Interestingly, the (0.8 - 3.0 um)/(0.1 - 0.8 um) read count ratio is correlated to size of the MAGs (Spearman rho = 0.76; P = 10-5; Fig. 3F), indicating a positive correlation between cell size and genome size.

The reason for streamlining of genomes in oligotrophs is not known (62). Lowered energetic costs for replication is one possibility. Despite the energetic requirements for DNA replication being low (<2% of the total energy budget; (78)), the extremely large effective population sizes of oligotrophic pelagic bacteria could explain selection for this trait (62). Another possibility is spatial constraints. In *Pelagibacter* the genome occupies 30% of the cell volume (64), hence cell size minimisation may be constrained by the genome size. A strong correlation between cell- and genome size for oligotrophic microbes would favor such an explanation. Further analysis with more reconstructed genomes and higher resolution of filter sizes, combined with e.g. FISH-analysis, could shed more light on the mechanisms behind genome streamlining.

Seasonal succession

Pronounced seasonal changes in environmental conditions with phytoplankton spring blooms are characteristic of temperate coastal waters. As typical for the central Baltic Sea, in 2012 an early spring bloom of diatoms was followed by a dinoflagellate bloom, causing inorganic nitrogen to decrease rapidly; later in summer, diazotrophic filamentous cyanobacteria bloomed (Fig. 4 and Fig. S10).

The only reconstructed picocyanobacteria genome peaked in early summer, between the spring and summer blooms of the larger phytoplankton. A similar pattern was previously observed for an operational taxonomic unit identical to the 16S rRNA gene of the reconstructed genome (31). The seasonal dynamics of heterotrophic MAGs were highly influenced by the phytoplankton blooms, with different populations co-varying with different phytoplankton (Fig. 4). Phylum-level patterns were present, with a Bacteroidetes-dominated community in spring and early summer (7/9 Bacteroidetes genome MAG clusters), coinciding with the spring phytoplankton blooms, and Actinobacteria being more predominant in the second half of the year (8/10 Actinobacterial genome MAG clusters). This is largely in agreement with what is known for Bacteroidetes: being better adapted to feeding on complex carbohydrates abundant for the duration of phytoplankton blooms (38, 79). This was also reflected in the functional annotations where Bacteroidetes MAGs contained several enzymes for degradation of polysaccharides and were enriched for certain aminopeptidases (COG 1506 and 2234). For Actinobacteria, no such general correlation pattern to phytoplankton has been shown, but there are indications of associations with, and active uptake of photosynthates from cyanobacterial blooms (80–82). Actinobacteria MAGs, which were enriched in genes for the uptake and metabolism of monosaccharides such as galactose and xylose, became abundant as levels of dissolved organic carbon increased in the water.

Interestingly, pilus assembly genes were most abundant in MAGs belonging to Gammaproteobacteria (SAR92) and Verrucomicrobia (Opitutaceae and LD19). During the spring diatom bloom, SAR92 and Opitutaceae peaked in relative abundance, while the LD19 MAG cluster was most abundant in July and August, during the bloom of filamentous cyanobacteria. These lineages may thus have specific adaptations to attach to different types of phytoplankton, as they appear throughout the season and utilize nutrients present on the surface or leaked from these large cells.

Besides these phylum- and order-level trends, temporal patterns were also observed at finer phylogenetic scales. The peaks of Luna clades coincide with phytoplankton blooms, while acI and acIV are more abundant in autumn, after the blooms. AcI also appears to have a more stable abundance profile than acIV, in agreement with Newton et al (2011). As previously reported for acI SAGs (14), cyanophycinase (COG4242) was found in two of the three acI MAG clusters, potentially allowing degradation of the storage compound cyanophycin synthesized by cyanobacteria. The two acI MAG clusters (BACL2 and 4) encoding cyanophycinase became abundant in late July, as filamentous cyanobacteria, which typically produce cyanophycin, started to peak in abundance (Fig. 4). In contrast, all acIV and Luna MAGs lacked this gene.

Furthermore, contrasting dynamics between members of the same clade, as exemplified by one acIV population blooming in spring, highlight that, despite the general similarities in their functional repertoire, lineage specific adaptations allow different microniches to be occupied by different strains (Fig. 2, Fig. 4). As an example, the spring blooming acIV BACL6 contained several genes for nucleotide degradation that were missing in the summer blooming acIV MAG clusters, such as adenine phosphoribosyltransferase (EC:2.4.2.7), thymidine phosphorylase (EC:2.4.2.4) and pyrimidine utilization protein B (TIGR03614). In addition, BACL6 contained genes *sulP* (TIGR00815) and *phnA* (TIGR00686) for uptake of sulfate and uptake and utilization of alkylphosphonate, respectively. These genes were also found in the spring blooming BACL25 (Luna clade), but were notably absent from the summer blooming acI, acIV and Luna MAG clusters. The capacity to utilize nucleotides and phosphonates as major carbon and phosphorous sources thus probably set BACL6 and 25 apart from other closely related lineages.

The two SAR11 MAG clusters also showed contrasting seasonal patterns, with BACL20 being abundant in spring and peaking in early summer, while BACL5 appeared later and showed a stable profile from July onwards. Functional analysis showed that, of these two populations, BACL5 contained several genes related to phosphate acquisition and storage that were missing from BACL20. These included the high-affinity *pstS* transporter, the polyphosphate kinase and exopolyphosphatase as well as the phosphate starvation-inducible gene *phoH*. BACL5 therefore appears better adapted to the low concentrations of phosphate found in mid- to late summer (Fig. S10). In addition, proteorhodopsin was found in BACL5, but not in BACL20. However, since the latter consists of only one MAG, this gene may have been missed due to incomplete genome assembly.

Biogeography of the brackish microbiome

To assess how abundant the MAGs presented here are in other marine and freshwater environments around the globe, fragment recruitment was performed from a collection of samples comprising a wide range of salinity levels. At intermediate levels of sequence identity, different phylogenetic lineages recruit preferentially fresh or marine water fragments. Most markedly, SAR11 displays a clear marine profile, while acI and acIV actinobacteria have a distinct freshwater signature (Fig. 5A, Fig S11A, Fig. S12). In addition, MAGs belonging to Bacteroidetes and Gammaproteobacteria show signs of a marine rather

than a freshwater signature that fits with the presence of the Na+-transporting NADH dehydrogenase in these lineages. However, at a high similarity level of 99%, only reads from brackish environments are recruited, including estuaries in North America, to the exclusion of fresh and marine waters much closer geographically to the Baltic sea (Fig. 5B, Fig. S11B, Fig. S12). Indeed, it is remarkable that BACL8 is placed phylogenetically as a single clade together with a SAG sampled on the North American Atlantic coast (Fig. S4, Fig. S12). Despite being separated by thousands of kilometers of salt water, these cells share >92% similarity over Blast+ high-scoring pairs. Overall, our analysis showed that the reconstructed genomes recruited sequences primarily from brackish estuary environments at various levels of sequence identity (Fig. 6).

The Baltic Sea is a young system, formed by the opening of the Danish straits to the North Sea in a long process between 13,000 and 8,000 years ago (83). The initially high salinity has slowly decreased due to the influx of freshwater from the surrounding area and the narrow connection to the open ocean, forming a stable brackish system around 4,000 years ago. Even considering fast rates of evolution for bacteria, the high degree of separation observed at the whole genome level between the Baltic metagenome and global fresh and marine metagenomes cannot be explained by isolation in the Baltic alone. Based on the rates of evolution presented by (84), it would take over 100 thousand years for free-living bacteria to accumulate 1% genome divergence. These specialists must therefore have evolved before current stable bodies of brackish water, such as the Baltic, the Black Sea and the Caspian, were formed in the end of the last glacial period. Intriguingly, brackish-typical green sulfur bacteria have been observed in sediment layers of 217,000 years in the now highly saline mediterranean (85), suggesting that brackish populations might migrate between these transient environments as salinity shifts. This is in agreement with the well documented separation between freshwater and marine species, which indicates that salinity level is a main barrier isolating populations (reviewed in (86)). Strains previously adapted to brackish environments and transported through winds, currents or migratory animals can thus proliferate and saturate all available niches before fresh and marine strains can effectively adapt to the new environment (87).

The prokaryotic populations of the Baltic Sea thus appear to have adapted to its intermediate salinity levels via a different mode than its macrobial species, which are only recently adapted to brackish environments from the surrounding fresh and marine waters (88–90). Indeed, while there is low macrobial species-richness and low intra-species diversity in the Baltic (91), suggestive of a recent evolutionary bottleneck, no such observation has been made for bacteria in the region (6, 30).

Conclusion

Here we have presented 83 genomes, corresponding to 30 clusters at >99% nucleotide identity, reconstructed from metagenomic shotgun sequencing assemblies using an unsupervised binning approach. Many of these belong to lineages with no previous reference genomes, including lineages known from 16S studies to be highly abundant. We show that the seasonal dynamics of these bacteria follow phylogenetic divisions, but with fine-grained lineage specific adaptations. We confirm previous observations on the

prevalence of genome streamlining in pelagic bacteria. Finally, we propose that brackish environments exert such strong selection on tolerance for intermediate salinity that lineages adapted to it flourish throughout the globe with little influence of surrounding aquatic communities. The new genomes are now available to the wider research community to explore further questions in microbial ecology and biogeography, solidly placing the automated reconstruction of genomes from metagenomes as an invaluable tool in ocean science.

Methods

Sample collection, library preparation and sequencing

Water samples were collected on 37 occasions between march and december in 2012, 10 km off the coast of Öland (Sweden) at 2 m depth at the Linnaeus Marine Observatory (N 56°55.851, E 17°03.640) using a Ruttner sampler. All samples are referred to in the text and figures by their sampling date, in the format yymmdd. Samples were filtered successively at 3.0 µm and 0.22 µm. The 0.22 µm fraction was used for DNA extraction as described in (92). The concentration of chlorophyll a was assesses using the method of Jespersen et al (Jespersen A.M., Christoffersen K. (1987). Measurements of chlorophyll-a from phytoplankton using ethanol as extraction solvent. Arch Hydrobiol 637 109: 445–454) Concentrations of phosphate, ammonium, nitrate and silica were measured as described by Valderrama et al (Valderrama J.C. (1995). Methods of nutrient analysis. In: Hallegraeff MG, Anderson DM, Cembella AD (eds). Manual on Harmful Marine Microalgae. 748 IOC Manuals and Guides No. 33. UNESCO: Paris. pp 251-282). Microscopic counts of phytoplankton were performed as described in (32).

2-10 ng of DNA from each sample were prepared with the Rubicon ThruPlex kit (Rubicon Genomics, Ann Harbour, Michigan, USA) according to the instructions of the manufacturer. Cleaning steps were performed with MyOne™carboxylic acid-coated superparamagnetic beads (Invitrogen, Carlsbad, CA, USA). Finished libraries were sequenced in SciLifeLab/NGI (Solna, Sweden) on a HiSeq 2500 (Illumina Inc, San Diego, CA, USA). On average, 31.9 million pair-ended reads of 2x100 bp were generated.

Quality filtering and assembly

Reads were quality trimmed using sickle (93) to eliminate stretches where average quality scores fall below 30. Cutadapt (94) was used to eliminate adapter sequences from short fragments detected by FastQC (95). Finally, FastUniq (96) was used to eliminate reads which were, on both forward and reverse strands, identical prefixes of longer reads (on average, 49% of the reads from each sample).

Each sample was then assembled separately, using Ray (97) with kmer length 21, 31, 41, 51, 61, 71 and 81. Contigs from each of these assemblies were cut up to 1000 bp in sliding windows every 100 bp using Metassemble (98) and reassembled using 454 Life Science's software Newbler (Roche, Basel, Switzerland).

Binning of sequencing data and construction of MAG

The quality-filtered reads of each sample were mapped against the contigs of all other samples using Bowtie2 (99), Samtools (100), Picard (101) and BEDTools (102). Contigs from each sample were then binned based on their tetranucleotide composition and covariation across all samples using Concoct (28) and accepting contigs over 1000, 3000 or 5000 bp in length. Prodigal (103) was used to predict proteins on contigs for each bin, and these were compared to the COG database with RPS-BLAST (104). The resulting hits were compared to a small set of 36 single-copy genes (SCG) used by Concoct, only considering a protein hit if it covered more than half of the reference length. Bins were considered good if they presented at least 30 of the 36 SCG, no more than two of which in multiple copies. Another set of phylum-specific SCG was used to evaluate each selected bin more carefully. Both the general prokaryotic SCG and phylum-specific SCG were selected such that they were present in at least 97% of sequenced representatives within that taxon and had an average gene count of less than 1.03. For the phylum-specific SCG, proteobacteria was divided down to class level for increased sensitivity.

For each sample, only one Concoct run was chosen for downstream analysis. For most samples, the 1000 bp cutoff provided the maximum number of good bins, but samples 120705, 120828 and 121004 had best results with 3000 bp. This resulted in 83 good bins in total. As the same genome could have been independently found in more than one sample, MUMmer (105) was used to compare all good bins against each other. The distance between two bins was set as one minus average nucleotide identity, given a minimum of 50% bin coverage of the larger bin in each pair. This procedure yielded 30 clearly distinct clusters (BACL) independently of the clustering method used (average-, full- or single-linkage).

Abundance estimation and comparison of MAG

The relative abundance of each MAG was estimated using the fraction of reads in each sample mapping to respective MAG. Normalized on the size of that bin this yielded the measure *fraction of reads per nucleotide in bin*. This measure was chosen since it is comparable across samples with varying sequencing output and different bin sizes.

Using the CONCOCT input table, multiplying the average coverage per nucleotide with the length of the contig in question and summing over all contigs within a bin and within a sample gave the number of reads per bin within a sample. The fraction of reads in each sample mapping to each bin was then calculated by dividing this value with the total number of reads from each sample, after having removed duplicated reads.

Functional analysis

Contigs in each genome cluster were annotated using PROKKA (v. 1.7, (103, 106–112), modified so that partial genes covering edges of contigs were included, to suit metagenomic datasets, and extended with additional annotations so that Pfam (113), TIGRFAMs (114), COG (115) and Enzyme Commission numbers were given for all sequences where applicable. The extended annotation was performed using homology search with RPS-Blast (Camacho 2009). Metabolic pathways were predicted in MAGs using MinPath (v. 1.2, (116)

with the Metacyc database (v. 18.1, (117)) as a reference. Non-metric multidimensional scaling (NMDS) analysis was applied to the genome clusters based on their annotations as well as a subset of transporter genes (Supplementary file 5), and the metabolic pathways. Abundances of functional features were explored, and statistical analyses of functional differences between groups of MAGs performed using STAMP (v. 2.0.9, (118)) with multiple test correction using the Benjamini-Hochberg FDR method.

Taxonomic and phylogenetic annotation

Initial taxonomy assignment for each MAG was done with Phylosift (119). To improve the resolution of annotations, classification of 16S rRNA genes was also used. Complete or partial 16S genes were identified on contigs using WebMGA (120). Further, since rRNA is difficult both to assemble and to bin, a complementary approach was used where partial 16S rRNA genes were assembled for each MAG using reads classified as SSU rRNA by SortMeRNA (121), but whose paired-end read was assembled in another contig belonging to the same MAG . The identified and reconstructed 16S fragments were classified with stand-alone SINA 1.2.13 (122) and by Blasting against the data by (59).

Using the information provided by Phylosift and 16S analysis , relevant isolate genomes and single-amplified genomes (SAG) were selected. These were combined with all complete prokaryotic genomes in NCBI. Prodigal was used for protein prediction in each genome. These proteomes, together with the proteomes of our MAGs, were used for phylogenetic tree reconstruction using Phylophlan (123). Phylophlan's reference database was not used as we noticed that, in instances where genomes that were already present in the reference were processed by us and added, they tended to branch closer to the MAG than otherwise, thus indicating a role of protein prediction method in the phylogeny. The tree visualizations displayed here were generated with iTOL v2 (124). For the sake of clarity, not all species included in the tree are maintained in the overview or clade-specific insets. Since the distance between MAGs and their nearest neighbours in NCBI were as a rule too large for ANI calculation, we adopted Genome BLAST Distance for this comparison (125).

Genome streamlining analysis

The dataset of marine microbial isolates from (126) was downloaded from CAMERA (127). These were functionally annotated in the same way as the MAGs. For streamlining analysis, the GC content, genome length, and average fraction of non-coding nucleotides were calculated. To avoid bias of shorter contigs, the average fraction of non-coding nucleotides was only based on sequences longer than 5000 nucleotides. For clarity, only genomes belonging to the same phyla as our reconstructed MAGs were included in the analysis.

Fragment recruitment

Fragment recruitment (5) was used to estimate the presence of the reconstructed MAGs in various locations around the globe. We selected a total of 86 metagenomic samples obtained from a wide range of salinity levels and geographic locations (Table 2). Missing salinity value

for Delaware Bay (GS011) was set to 15 PSU after consulting the Delaware Bay Operational Forecast System (DBOFS) (http://tidesandcurrents.noaa.gov/ofs/dbofs/dbofs.html).

All samples were sub-sampled to 10.000 sequences each 350 bp in length and all reads were queried against a database of the reconstructed genome bins using Blast+ (v. 2.2.30, Altschul 1997). Non-coding intergenic sequences were excluded by using only the nucleotide sequences of predicted open reading frames. Only samples comprising the 0.1-0.8 µm filter fraction were used and only hits with e-value cutoff < 1e-5 and alignment length > 200 bp were considered. For visualizations, the number of hits for genome bins in each sample was normalized against the total size (in bp) of the bin. These normalized counts were then averaged over the bins of each reconstructed genome cluster.

Table 2. Metagenomic projects used as queries

Project	Project ID	Samples	Salinity range (PSU)	Reference
Global Ocean Sampling Expedition	CAM_PROJ_GOS	56	0.1 - 37, 63 (hypersaline)	(5)
Global Ocean Sampling Baltic Sea	CAM_P_0001109	19	0 - 34	(6)
Freshwater metagenomes	PRJEB4844	9	0	(128)
Lake Lanier metagenome by 454	SRR063691	1	0	(129))
Metagenomics of the Amazon	SRR091234	1	0	(130)

Acknowledgements

We thank Anders Månsson and Kristofer Bergström for their knowledgeable and persistent sampling effort, and Sabina Arnautovic and Emmelie Nilsson for their careful processing of samples. This work was funded by the Swedish Research Councils VR (grant 2011-5689) and EC BONUS project BLUEPRINT, partially funded by FORMAS through grants to A.F.A, as well as Formas project ECOCHANGE (Strategic Grant for Marine Research) through a grant to C.L. and J.P. All computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) and the PDC Center for High Performance Computing at KTH. Sequencing was conducted at the Swedish National Genomics Infrastructure (NGI) at SciLifeLab in Stockholm.

References

1. Connon SA, Giovannoni SJ (2002) High-Throughput Methods for Culturing Microorganisms in Very-Low-Nutrient Media Yield Diverse New Marine Isolates. *Appl*

Environ Microbiol 68:3878–3885.

2. Lauro FM et al. (2009) The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci U S A* 106:15527–15533.

3. Swan BK et al. (2013) Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci U S A* 110:11463–11468.

4. Venter JC et al. (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* 304:66–74.

5. Rusch DB et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol*:DOI: 10.1371/journal.pbio.0050077.

6. Dupont CL et al. (2014) Functional Tradeoffs Underpin Salinity-Driven Divergence in Microbial Community Composition. *PLoS One*:DOI: 10.1371/journal.pone.0089549.

7. Goldberg SM et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc Natl Acad Sci U S A* 103:11240–11245.

8. Gilbert JA et al. (2010) Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel. *Stand Genomic Sci* 3:183–193.

9. Luo H, Moran MA (2013) Assembly-free metagenomic analysis reveals new metabolic capabilities in surface ocean bacterioplankton. *Environ Microbiol Rep* 5:686–696.

10. Tyson GW et al. (2004) ommunity structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.

11. Marcy Y et al. (2007) Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A* 104:11889–11894.

12. Woyke T et al. (2009) Assembling the marine metagenome, one cell at a time. *PLoS One* 4:e5299.

13. Martinez-Garcia M et al. (2011) High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISMEJ* 6:113–123.

14. Ghylin TW et al. (2014) Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *ISMEJ* 8:2503–2516.

15. McLean JS et al. (2013) Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc Natl Acad Sci U S A* 110:E2390–9.

16. Dodsworth JA et al. (2013) Single-cell and metagenomic analyses indicate a fermentative

and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun* 4:1854.

17. Podar M et al. (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl Environ Microbiol* 73:3205–3214.

18. Wallner G, Fuchs B, Spring S, Beisker W, Amann R (1997) Flow sorting of microorganisms for molecular analysis. *Appl Environ Microbiol* 63:4223–4231.

19. Sekar R, Fuchs BM, Amann R, Pernthaler J (2004) Flow sorting of marine bacterioplankton after fluorescence in situ hybridization. *Appl Environ Microbiol* 70:6210–6219.

20. Wilson MC et al. (2014) An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* 506:58–62.

21. Yilmaz S, Singh AK (2012) Single cell genome sequencing. *Curr Opin Biotechnol* 23:437–443.

22. Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res* 12:281–290.

23. Dick GJ et al. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10:R85.

24. Sharon I et al. (2012) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* 23:111–120.

25. Albertsen M et al. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 31:533–538.

26. Imelfort M et al. (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*:e603.

27. Nielsen HB et al. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 32:822–828.

28. Alneberg J et al. (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* 11:1144–1146.

29. Ojaveer H et al. (2010) Status of Biodiversity in the Baltic Sea. *PLoS One*:DOI: 10.1371/journal.pone.0012467.

30. Herlemann DPR et al. (2011) Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J* 5:1571–1579.

31. Andersson AF, Rieman L, Bertilsson S (2010) Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *ISMEJ*

4:171–181.

32. Lindh MV et al. (2015) Disentangling seasonal bacterioplankton population dynamics by high frequency sampling. *Environ Microbiol*:doi: 10.1111/1462–2920.12720.

33. Forth M, Liljebladh B, Stigebrandt A, Hall PO, Treusch AH (2015) Effects of ecological engineered oxygenation on the bacterial community structure in an anoxic fjord in western Sweden. *ISMEJ*:doi: 10.1038/ismej.2014.172.

34. Sinkko H et al. (2013) Bacteria contribute to sediment nutrient release and reflect progressed eutrophication-driven hypoxia in an organic-rich continental sea. *PLoS One* 8:e67061.

35. Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* 361:1929–1940.

36. Cardman Z et al. (2014) Verrucomicrobia are candidates for polysaccharide-degrading bacterioplankton in an arctic fjord of Svalbard. *Appl Environ Microbiol* 80:3749–3756.

37. Herlemann DPR et al. (2013) Metagenomic De Novo Assembly of an Aquatic Representative of the Verrucomicrobial Class Spartobacteria. *MBio* 4:e00569–12.

38. Teeling H et al. (2012) Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* 336:608–611.

39. Abell GCJ, Bowman JP Colonization and community dynamics of class Flavobacteria on diatom detritus in experimental mesocosms based on Southern Ocean seawaterâ. *FEMS Microbiol Ecol* 53:379–391.

40. Tang K, Jiao N, Liu K, Zhang Y, Li S (2012) Distribution and functions of TonB-dependent transporters in marine bacteria and environments: implications for dissolved organic matter utilization. *PLoS One* 7:e41204.

41. Alonso-Sáez L et al. (2012) Role for urea in nitrification by polar marine Archaea. *Proc Natl Acad Sci U S A* 109:17989–17994.

42. Giovannoni SJ et al. (2008) The small genome of an abundant coastal ocean methylotroph. *Environ Microbiol* 10:1771–1782.

43. Walsh DA, Lafontaine J, Grossart H-P (2013) in *Lateral Gene Transfer in Evolution* (Springer New York), pp 55–77.

44. Noinaj N, Guillier M, Barnard TJ, Buchanan SK (2010) TonB-dependent transporters: regulation, structure, and function. *Annu Rev Microbiol* 64:43–60.

45. Iverson V et al. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335:587–590.

46. Di Rienzi SC et al. (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife* 2:e01102.

47. Sekiguchi Y et al. (2015) First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking. *PeerJ* 3:e740.

48. Cho J-C, Giovannoni SJ (2004) Cultivation and growth characteristics of a diverse group of oligotrophic marine Gammaproteobacteria. *Appl Environ Microbiol* 70:432–440.

49. Meon B, Kirchman DL (2001) Dynamics and molecular composition of dissolved organic material during experimental phytoplankton blooms. *Mar Chem* 75:185–199.

50. Oh H-M, Kang I, Ferriera S, Giovannoni SJ, Cho J-C (2010) Genome sequence of the oligotrophic marine Gammaproteobacterium HTCC2143, isolated from the Oregon Coast. *J Bacteriol* 192:4530–4531.

51. Jang Y, Oh H-M, Kim H, Kang I, Cho J-C (2011) Genome sequence of strain IMCC1989, a novel member of the marine gammaproteobacteria. *J Bacteriol* 193:3672–3673.

52. Zwart G, Crump BC, Kamst-van Agterveld MP, Hagen F, Han SK (2002) Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquat Microb Ecol* 28. Available at: http://depot.knaw.nl/11265.

53. Zwart G et al. (1998) Divergent members of the bacterial division Verrucomicrobiales in a temperate freshwater lake. *FEMS Microbiol Ecol* 25:159–169.

54. Serkebaeva YM, Kim Y, Liesack W, Dedysh SN (2013) Pyrosequencing-based assessment of the bacteria diversity in surface and subsurface peat layers of a northern wetland, with focus on poorly studied phyla and candidate divisions. *PLoS One* 8:e63994.

55. Dunfield PF et al. (2007) Methane oxidation by an extremely acidophilic bacterium of the phylum Verrucomicrobia. *Nature* 450:879–882.

56. Zhu Y et al. (2007) Staphylococcus aureus biofilm metabolism and the influence of arginine on polysaccharide intercellular adhesin synthesis, biofilm formation, and pathogenesis. *Infect Immun* 75:4219–4226.

57. Welander PV et al. (2009) Hopanoids play a role in membrane integrity and pH homeostasis in Rhodopseudomonas palustris TIE-1. *J Bacteriol* 191:6145–6156.

58. Doughty DM et al. (2011) The RND-family transporter, HpnN, is required for hopanoid localization to the outer membrane of Rhodopseudomonas palustris TIE-1. *Proc Natl Acad Sci U S A* 108:E1045–51.

59. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S (2011) A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* 75:14–49.

60. Larsson J et al. (2014) Picocyanobacteria containing a novel pigment gene cluster dominate the brackish water Baltic Sea. *ISMEJ* 8:1892–1903.

61. Palenik B et al. (2006) Genome sequence of Synechococcus CC9311: Insights into adaptation to a coastal environment. *Proc Natl Acad Sci U S A* 103:13555–13559.

62. Giovannoni SJ, Thrash JC, Temperton B (2014) Implications of streamlining theory for

microbial ecology. *ISMEJ* 8:1553–1565.

63. Dufresne A et al. (2003) Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* 100:10020–10025.

64. Giovannoni SJ et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309:1242–1245.

65. Button DK (1998) Nutrient uptake by microorganisms according to kinetic parameters from theory as related to cytoarchitecture. *Microbiol Mol Biol Rev* 62:636–645.

66. Nedashkovskaya OI et al. (2005) Polaribacter butkevichii sp. nov., a novel marine mesophilic bacterium of the family Flavobacteriaceae. *Curr Microbiol* 51:408–412.

67. Yoon J-H, Kang S-J, Oh T-K (2006) Polaribacter dokdonensis sp. nov., isolated from seawater. *Int J Syst Evol Microbiol* 56:1251–1255.

68. Lee YS, Lee D-H, Kahng H-Y, Sohn SH, Jung JS (2011) Polaribacter gangjinensis sp. nov., isolated from seawater. *Int J Syst Evol Microbiol* 61:1425–1429.

69. Kim B-C et al. (2013) Polaribacter sejongensis sp. nov., isolated from Antarctic soil, and emended descriptions of the genus Polaribacter, Polaribacter butkevichii and Polaribacter irgensii. *Int J Syst Evol Microbiol* 63:4000–4005.

70. Hyun D-W et al. (2014) Polaribacter atrinae sp. nov., isolated from the intestine of a comb pen shell, Atrina pectinata. *Int J Syst Evol Microbiol* 64:1654–1661.

71. Nedashkovskaya OI, Kukhlevskiy AD, Zhukova NV (2013) Polaribacter reichenbachii sp. nov.: a new marine bacterium associated with the green alga Ulva fenestrata. *Curr Microbiol* 66:16–21.

72. Li H et al. (2014) Polaribacter huanghezhanensis sp. nov., isolated from Arctic fjord sediment, and emended description of the genus Polaribacter. *Int J Syst Evol Microbiol* 64:973–978.

73. Park S et al. (2014) Polaribacter marinivivus sp. nov., a member of the family Flavobacteriaceae isolated from seawater. *Antonie Van Leeuwenhoek* 106:1139–1146.

74. Azam F, Malfatti F (2007) Microbial structuring of marine ecosystems. *Nat Rev Microbiol* 5:782–791.

75. Gómez-Pereira PR et al. (2012) Genomic content of uncultured Bacteroidetes from contrasting oceanic provinces in the North Atlantic Ocean. *Environ Microbiol* 14:52–66.

76. Wong E et al. (2012) The Vibrio cholerae colonization factor GbpA possesses a modular structure that governs binding to different host surfaces. *PLoS Pathog* 8:e1002373.

77. Zeigler Allen L et al. (2012) Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. *ISME J* 6:1403–1414.

78. Sabath N, Ferrada E, Barve A, Wagner A (2013) Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol Evol* 5:966–977.

79. Buchan A, LeCleir GR, Gulvik CA, González JM (2014) Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat Rev Microbiol* 12:686–698.

80. Wu X, Xi W, Ye W, Yang H (2007) Bacterial community composition of a shallow hypertrophic freshwater lake in China, revealed by 16S rRNA gene sequences. *FEMS Microbiol Ecol* 61:85–96.

81. Parveen B et al. (2011) Diversity and dynamics of free-living and particle-associated Betaproteobacteria and Actinobacteria in relation to phytoplankton and zooplankton communities. *FEMS Microbiol Ecol* 77:461–476.

82. Penn K, Wang J, Fernando SC, Thompson JR (2014) Secondary metabolite gene expression and interplay of bacterial functions in a tropical freshwater cyanobacterial bloom. *ISME J* 8:1866–1878.

83. Björck S (1995) A review of the history of the Baltic Sea, 13.0-8.0 ka BP. *Quat Int* 27:19–40.

84. Kuo C, Ochman H (2009) Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biol Direct* 4:doi:10.1186/1745–6150.

85. Coolen M, Overmann J (2007) 217 000-year-old DNA sequences of green sulfur bacteria in Mediterranean sapropels and their implications for the reconstruction of the paleoenvironment. *Environ Microbiol* 9:238–249.

86. Logares R et al. (2009) Infrequent marine–freshwater transitions in the microbial world. *Trends Microbiol* 17:414–422.

87. De Meester L, Gómez A, Okamura B, Schwenk K (2002) The Monopolization Hypothesis and the dispersal–gene flow paradox in aquatic organisms. *Acta Oecol* 23:121–135.

88. Väinölä R, Hvilsom MM (1991) Genetic divergence and a hybrid zone between Baltic and North Sea Mytilus populations (Mytilidae: Mollusca). *Biol J Linn Soc Lond* 43:127–148.

89. Nilsson J et al. (2001) Matrilinear phylogeography of Atlantic salmon (Salmo salar L.) in Europe and postglacial colonization of the Baltic Sea area. *Mol Ecol* 10:89–102.

90. Luttikhuizen PC, Drent J, Baker AJ (2003) Disjunct distribution of highly diverged mitochondrial lineage clade and population subdivision in a marine bivalve with pelagic larval dispersal. *Mol Ecol* 12:2215–2229.

91. Johannesson K, André C (2006) Life on the margin: genetic isolation and diversity loss in a peripheral marine ecosystem, the Baltic Sea. *Mol Ecol* 15:2013–2029.

92. Riemann L, Steward GF, Azam F (2000) Dynamics of bacterial community composition

and activity during a mesocosm diatom bloom. *Appl Environ Microbiol* 66:578–587.

93. Joshi NA, Fass JN (2011) *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)*.

94. Martin M (2012) Cutadapt removes adapter sequences from high-throughput sequencing reads. *Bioinformatics in Action* 17:10–12.

95. Andrews S (2009) *FastQC. A quality control tool for high throughput sequence data*.

96. Xu H et al. (2012) FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS One* 7:e52249.

97. Boisvert S, Laviolette F, Corbeil J (2010) Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. *J Comput Biol* 17:1519–1533.

98. De Bruijn I (2012) *Metassemble: Scripts to run several metagenomics assembly programs*.

99. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359.

100. Li H et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:1078–1079.

101. Wysoker A, Tibbetts K, McCowan M, Homer N, Fennell T (2010) *Picard Tools*.

102. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

103. Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC (2012) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 28:2223–2230.

104. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 32:W327–31.

105. Kurtz S et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.

106. Stajich JE et al. (2002) The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Res* 12:1611–1618.

107. Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* 32:11–16.

108. Camacho C et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

109. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*:doi: 10.1093/nar/gkr367.

110. Tange O (2011) GNU Parallel: The Command-Line Power Tool. *;login: The USENIX*

Magazine:42–47.

111. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinforamtics* 29:2933–2935.

112. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069.

113. Finn RD et al. (2013) Pfam: the protein families database. *Nucleic Acids Res* 42:D222–30.

114. Mitchell A et al. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* 43:D213–21.

115. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28:33–36.

116. Ye Y, Doak TG (2009) A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. *PLoS Comput Biol* 5:e1000465.

117. Caspi R et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleci Acids Research* 40:D742–53.

118. Parks DH, Tyson GW, Hugenholtz P, Beiko RG (2014) STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30:3123–3124.

119. Darling AE et al. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243.

120. Wu S, Zhu Z, Fu L, Niu B, Li W (2011) WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 12:444.

121. Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28:3211–3217.

122. Pruesse E, Peplies J, Glöckner FO (2012) SINA: accurate high troughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*:doi: 10.1093/bioinformatics/bts252.

123. Segata N, Börnigen D, Morgan XC, Huttenhower C (2013) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 4:2304.

124. Letunic I, Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39:W475–8.

125. Auch AF, Klenk H-P, Göker M (2010) Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci*

2:142–148.

126. Yooseph S et al. (2010) Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* 468:60–66.

127. Sun S et al. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* 39:D546–51.

128. Eiler A et al. (2013) Unveiling distribution patterns of freshwater phytoplankton by a next generation sequencing based approach. *PLoS One* 8:e53516.

129. Oh S et al. (2011) Metagenomic Insights into the Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Appl Environ Microbiol* 77:6000–6011.

130. Ghai R et al. (2011) Metagenomics of the Water Column in the Pristine Upper Course of the Amazon River. *PLoS One* 6:e23785.

131. Kang I et al. (2012) Genome sequence of “Candidatus Aquiluna” sp. strain IMCC13023, a marine member of the Actinobacteria isolated from an arctic fjord. *J Bacteriol* 194:3550–3551.

132. Hahn MW, Schmidt J, Taipale SJ, Doolittle WF, Koll U (2014) Rhodoluna lacicola gen. nov., sp. nov., a planktonic freshwater bacterium with stream-lined genome. *Int J Syst Evol Microbiol* 64:3254–3263.

133. Garcia SL et al. (2013) Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *ISME J* 7:137–147.

134. Glöckner FO et al. (2000) Comparative 16S rRNA Analysis of Lake Bacterioplankton Reveals Globally Distributed Phylogenetic Clusters Including an Abundant Group of Actinobacteria. *Appl Environ Microbiol* 66:5053–5065.

135. Newton RJ, Kent AD, Triplett EW, McMahon KD (2006) Microbial community dynamics in a humic lake: differential persistence of common freshwater phylotypes. *Environ Microbiol* 8:956–970.

136. Humbert J-F et al. (2009) Comparison of the structure and composition of bacterial communities from temperate and tropical freshwater ecosystems. *Environ Microbiol* 11:2339–2350.

137. Bowman J (2014) in *The Prokaryotes*, 135., eds Rosenberg E, DeLong E, Lory S, Stackebrandt E, Thompson F (Springer Berlin Heidelberg), pp 539–550.

138. Huggett MJ, Hayakawa DH, Rappé MS (2012) Genome sequence of strain HIMB624, a cultured representative from the OM43 clade of marine Betaproteobacteria. *Stand Genomic Sci* 6:11–20.

139. Crosbie ND, Pöckl M, Weisse T (2003) Dispersal and phylogenetic diversity of nonmarine picocyanobacteria, inferred from 16S rRNA gene and cpcBA-intergenic spacer

sequence analyses. *Appl Environ Microbiol* 69:5716–5721.

140. Hugenholtz P, Pitulle C, Hershberger KL, Pace NR (1998) Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol* 180:366–376.

141. Op den Camp HJ et al. (2009) Environmental, genomic and taxonomic perspectives on methanotrophic Verrucomicrobia. *Environ Microbiol Rep* 1.

142. Yoon J et al. (2007) Coraliomargarita akajimensis gen. nov., sp. nov., a novel member of the phylum “Verrucomicrobia” isolated from seawater in Japan. *IJSEM* 57:959–963.

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

144. Wuchter C et al. (2006) Archaeal nitrification in the ocean. *Proc Natl Acad Sci U S A* 103:12317–12322.

145. Walker CB et al. (2010) Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci U S A* 107:8818–8823.

146. Park S et al. (2014) Genomes of Two New Ammonia-Oxidizing Archaea Enriched from Deep Marine Sediments. *PLoS One*:DOI: 10.1371/journal.pone.0096449.

147. Labrenz M et al. (2010) Relevance of a crenarchaeotal subcluster related to Candidatus Nitrosopumilus maritimus to ammonia oxidation in the suboxic zone of the central Baltic Sea. *ISMEJ* 4:1496–1508.

148. Pruitt KD et al. (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 42:D756–63.

Figure legends

Figure 1. Phylogenetic tree highlighting MAG. Phyla and proteobacterial classes for which MAG were generated are highlighted with coloured branches: Thaumarchaeota (dark blue), Cyanobacteria (blue-green), Actinobacteria (lime green), Alphaproteobacteria (yellow), Bacteroidetes (orange), Verrucomicrobia (purple), Gammaproteobacteria (red) and Betaproteobacteria (dark green).

Figure 2. Clustering of genome clusters based on COG abundance. Non-metric multidimensional scaling was applied to a pairwise distance matrix of the genomes and the first two dimensions are shown. Genome clusters are displayed with abbreviated lineage names and cluster numbers in parentheses and further colored according to Phyla/Class.

Figure 3. Genome properties and filter size fraction distributions of MAGs. Principal Components Analysis (PCA) on our 30 MAG clusters and 135 genomes from marine isolates (Yooseph, 2010) based on log-transformed percentages of non-coding DNA, GC-content, COG category Transcription [K], Signal transduction [T], Defense mechanism [V], Secondary metabolites biosynthesis [Q] and Lipid transport and metabolism [I]. Only isolates belonging to the phyla and proteobacterial classes as represented by MAGs were included. (A) The first- and second principal components (PC1 and PC2) explain 80% and 9% of the variance, respectively. (B) Genome size vs. percentage of non-coding DNA plotted for the same set of genomes (C) A zoom-in on the smaller and denser genomes (C). Number of sequence reads matching to our reconstructed genomes from different filter fractions from Dupont et al (6) (D-E). 10,000 random reads were sampled from each size fraction from 21 samples and aligned to the MAGs by BLAST, using 95% identity and alignment length of 100 bp as cutoff. Counts were averaged over MAGs in the same genome cluster and summed over all samples for each size fraction. (F) The ratio of matches between the 0.8 - 3.0 and the 0.1 - 08 um fraction were plotted against genome size, both in log scale.

Figure 4. Seasonal dynamics of MAG clusters and phytoplankton. The heatmap plot shows the relative abundance of each MAG cluster in the time-series, based on calculated coverage from read-mappings. In addition, the relative abundance of eukaryotic groups, assessed by microscopy, is shown for the same samples. MAG clusters and eukaryotic groups were clustered using Spearman correlations and is shown in the hierarchical clustering on the left. Coloured rows at the top indicate month and season of each sample.

Figure 5. Biogegraphical abundance profiles of MAGs. Heatmap plots showing the abundance of recruited reads from various samples and sample groups to each of the 30 MAG clusters at the 85% (A) and 99% (B) identity cutoff levels. Shown values represent number of recruited reads/kb of genome per 10.000 queried reads. For clarity, several sample groups have been collapsed with recruitment values averaged over samples in the group. Sample groups are indicated by the lower color strip above the plot and samples are ordered by salinity (shown in the upper color strip). See Figure S9 for full visualizations of samples. ‘BalticAsm’ represents a metagenomic co-assembly of all the 37 samples in the time-series.

Figure 6. Recruitment profiles at different nucleotide identity. Fragment recruitment values were calculated at various %nt cutoffs for different sample groups. For each sample in each group, the average recruitment over all 30 MAG clusters was calculated and boxes show the variation of these averages from all samples in each group. The number of samples used in each group are indicated in the legend.

Figure S1. An overview of the generation of MAGs and clusters. First, reads from each sample are assembled separately. Secondly, reads from each of the samples are mapped onto the contigs of all samples, generating a coverage profile for each contig. Together with tetranucleotide composition, these profiles are used to bin reads in each sample. Since bins

corresponding to the same genome may be generated independently in multiple samples, average nucleotide identity is used to compare bins amongst themselves, generating clusters.

Figure S2. The number of approved bins generated from each sample is directly proportional to the number of sequencing reads passing quality control, the number of contigs assembled and the number of contigs of at least 1000 bp (LargeContigs).

Figure S3. Average-linkage clustering of bins. Distance is defined as 1 minus the average nucleotide identity between bins, or set to 1 when the coverage of the larger bin over the smaller is less than 50%.

Figure S4. Phylogenetic placement of the MAGs generated in different phyla/classes
a) Actinobacteria
Reconstructed actinobacterial genomes belong to the lineages acI, acIV and Luna. Ac-I and Luna belong to the order Actinomycetales, while Ac-IV are Acidimicrobiales. Isolates of the Luna clade have had their genome sequenced (131, 132). Isolation of acI and acIV has been unsuccessful, but SAGs have recently been published for acI-A and acI-B (14, 133). Reconstructed 16S fragments placed MAGs in lineages acI-A, acI-B, acI-C, acIV-A, acIV-C and Luna (59, 80, 134–136). Consistent with previous estimates based on a few contigs identified as acIV (130) these MAGs are relatively low GC (average 53%).
b) Bacteroidetes
The reconstructed Bacteroidetes genomes belong to the families Flavobacteriaceae, Cryomorphaceae and the order Sphingobacteriales. Cryomorphaceae is a poorly studied family with just three isolate genomes and one SAG available (137), not closely related to the clusters that form three major branches presented here.
c) Alphaproteobacteria
Typically marine SAR11 and Rhodobacteraceae clusters were constructed. The two Pelagibacteriaceae clusters are clearly placed in the marine Ia clade, contrary to a previous study that suggested Baltic Pelagibacteriaceae belonged to the brackish clade IIIa or the freshwater clade LD12 (6).
d) Betaproteobacteria
One OM43 MAG was constructed. Up to now, only two reference genomes were available for this methylotrophic clade, both from the Pacific Ocean (42, 138)
e) Gammaproteobacteria
Known marine clades of Gammaproteobacteria were reconstructed, including SAR86, SAR92, and OM182, for which no reference genome had been previously available.
f) Cyanobacteria
The cyanobacteria genome reconstructed here belongs to a picocyanobacteria with 100% 16S identity to various freshwater *Synechococcus/Cyanobium* (139). An operational taxonomic unit of identical 16S has also been observed to be abundant across the Baltic with a strong spatial correlation with the previously described Verrucomicrobia MAG “Candidatus Spartobacteria baltica” (30).
g) Verrucomicrobia
The Verrucomicrobia phylum has been divided into five monophyletic subdivisions (140), but shortly afterwards a novel freshwater subdivision (LD19; (53)) and a subdivision

consisting of acidophilic methanotrophs (141) were found. Clusters 24 is phylogenetically placed within the family Opitutaceae (Subdivision 4), closest to the freshwater isolate *Coraliomargarita akajimensis (142)*; Fig. S4). Cluster 9 forms a sister branch to *Methylacidiphilum infernorum*, but their 16S rRNA genes are only 84% identical. Instead, 16S positions Cluster 9 as a member of LD19.

h) Thaumarchaeota

Thaumarchaeota are very abundant in the global ocean (143), where they play important roles in the nitrogen and carbon cycles by driving ammonia oxidation (144). A single archaeal genome cluster was reconstructed, whose 16S rRNA gene is 99% identical to marine lineages of *Nitrosopumilus maritimus (145, 146)*). *N. maritimus* are known to be abundant in Baltic Sea suboxic waters and this MAG has 98% identity to the previously described Baltic *N. maritimus* GD2 (98%; (147)). However, a whole genome comparison does not support the placement of these MAGs in the same species as *N. maritimus*, but rather in the same genus or family.

Figure S5. Non-metric multidimensional scaling plots of functional annotations. Non-metric multidimensional scaling was applied to a pairwise distance matrix of the genomes and the first two dimensions are shown. Genome clusters are displayed with abbreviated lineage names and cluster numbers in parentheses and further colored according to Phyla/Class. Based on enzyme and metabolic pathway predictions (enzyme and Minpath, respectively) and hidden Markov model profile designations (PFAM and TIGRFAM).

Figure S6. Abundance of glycoside hydrolase genes in genome clusters. Counts of were averaged over MAGs in each genome cluster.

Figure S7. Abundance of transporters in genome clusters. Counts of transporter genes were averaged over MAGs in each genome cluster and over transport types if applicable (see Supplementary file 6).

Figure S8. Core genome and pigment phylogeny of picocyanobacteria and Cluster 30. Phylogenetic trees are shown for core-genome proteins in picocyanobacteria (as in Larsson et al 2014) as well as concatenated alignments for phycocyanin (*cpcBA*) and phycoerythrin (*cpeBA*) gene products. In the core phylogeny, clade colors indicate subcluster designations for picocyanobacteria. In the pigment phylogenies, clade colors show pigment type designations. Scale bars indicate expected substitutions per site.

Figure S9. Pairwise scatter-plots showing correlations of genomes features among MAGs and isolate genomes. PC1 is the first principal component of the PCA of Figure 4A (that included all parameters except genome-size). Genomes are color-coded and shaped as in Figure 4.

Figure S10. Variation in temperature (in ℃) and concentrations of nutrients (in µg/L for Chlorophyll a and in µM for other nutrients).

Figure S11. Biogegraphical abundance profiles of MAGs. Heatmap plots showing the abundance of recruited reads from various samples and sample groups to each of the 30 MAG clusters at the 85% (A) and 99% (B) identity cutoff levels. Shown values represent number of recruited reads/kb of genome per 10.000 queried reads. As in Figure 3 but with all samples in each sample group shown.

Figure S12. Fragment recruitment over genes in a select number of clusters. For each genome cluster, the largest bin is shown with the x-axis representing predicted open reading frames and the y-axis representing the nucleotide identity in %. Consequently, each recruited read is shown at a specific gene (x coordinate) and % identity (y coordinate) and further colored by sample group shown in the legend.

Supplementary table 1: ANOSIM analysis of clustering of MAGs based on functional potential.

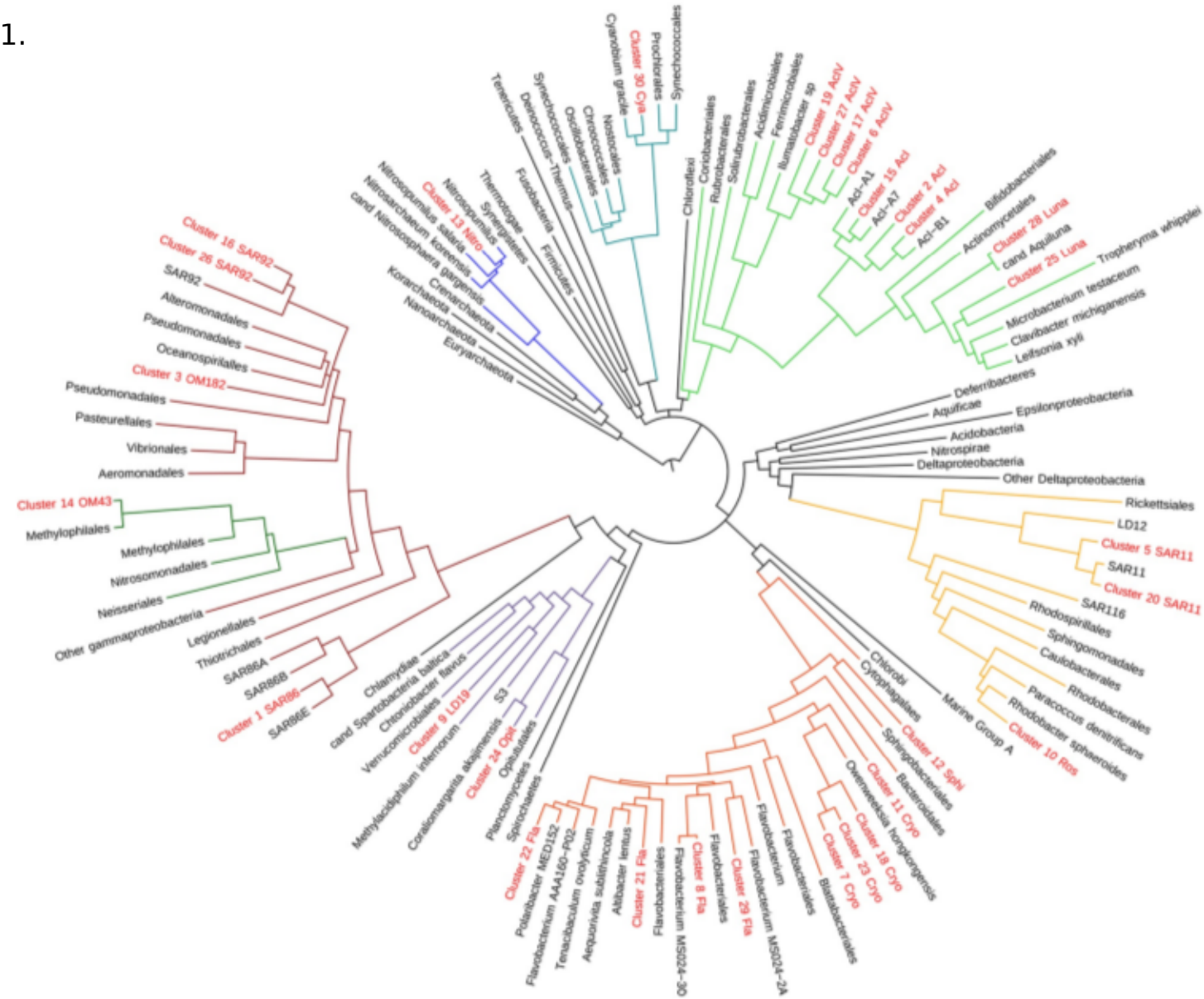
File S1: For each approved reconstructed bin, the number of bases, number of contigs, taxonomy and copy number of core COGs is reported.

File S2: Phylogenetic tree in Newick format including all genomes in NCBI RefSeq (148), all approved reconstructed bins and selected single amplified genomes and isolate genomes

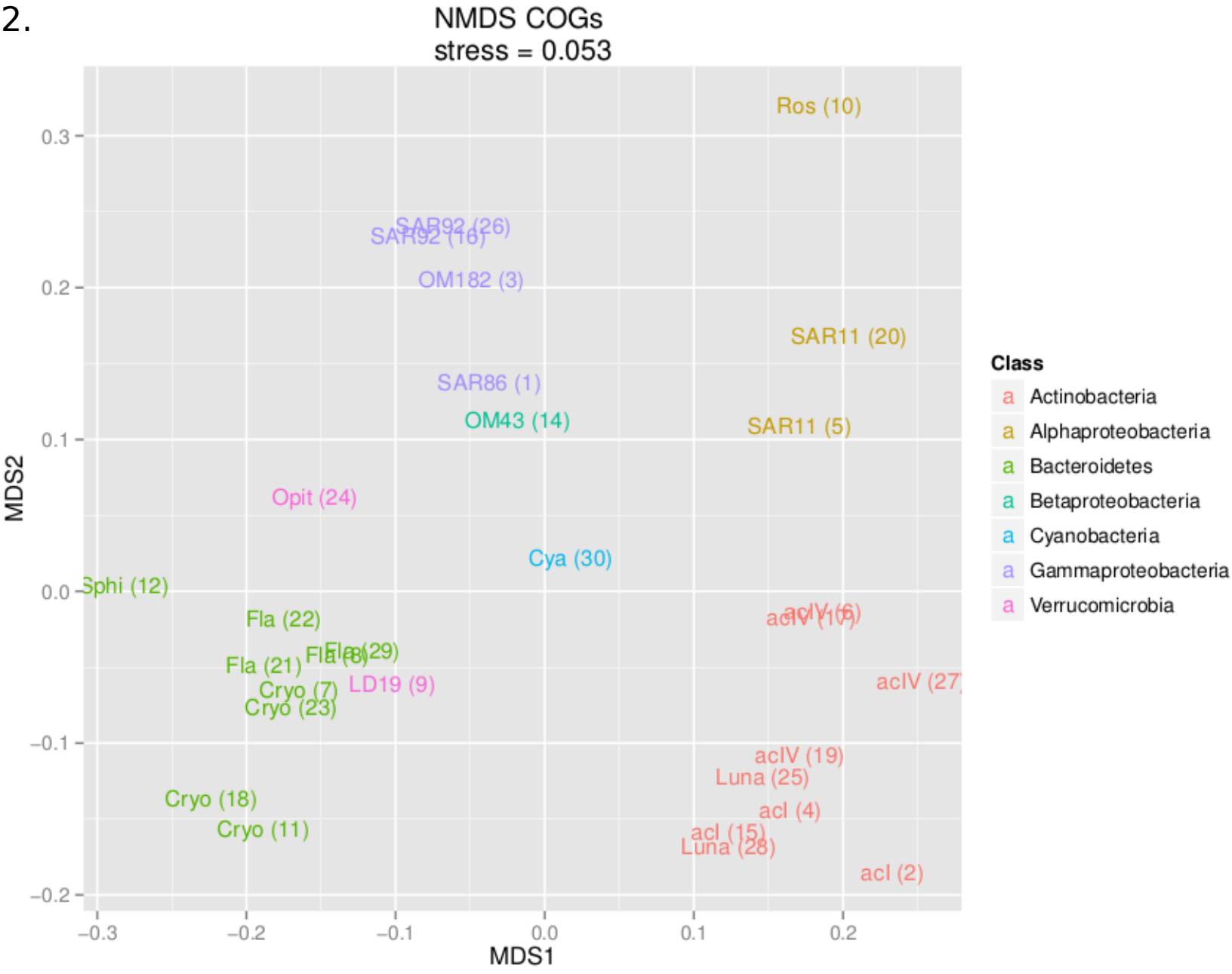
File S3: Presence or absence of every possible COG/PFAM/TIGRFAM and enzyme in each approved bin.

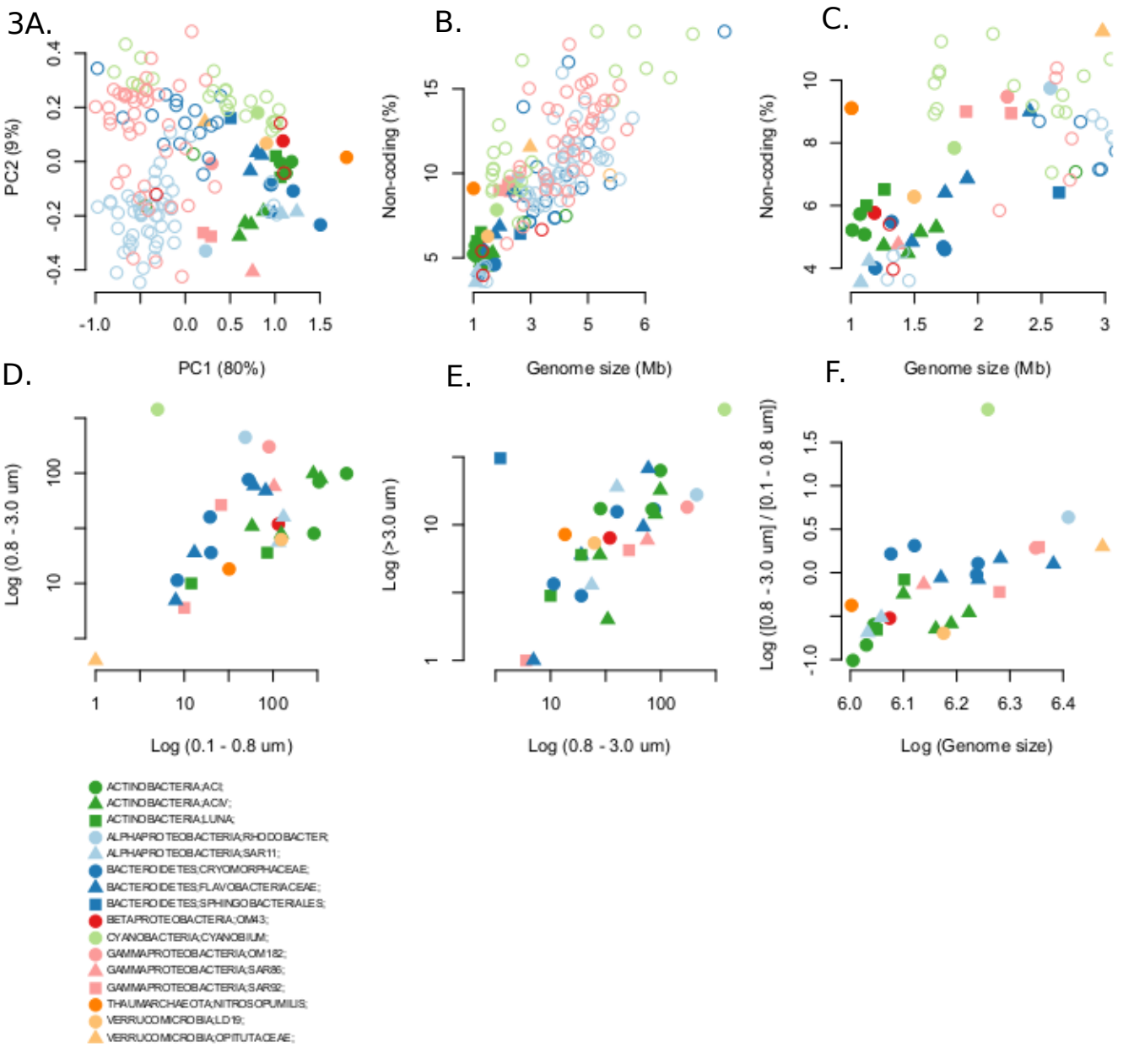
File S4: Transporter genes included in each gene cluster

1.

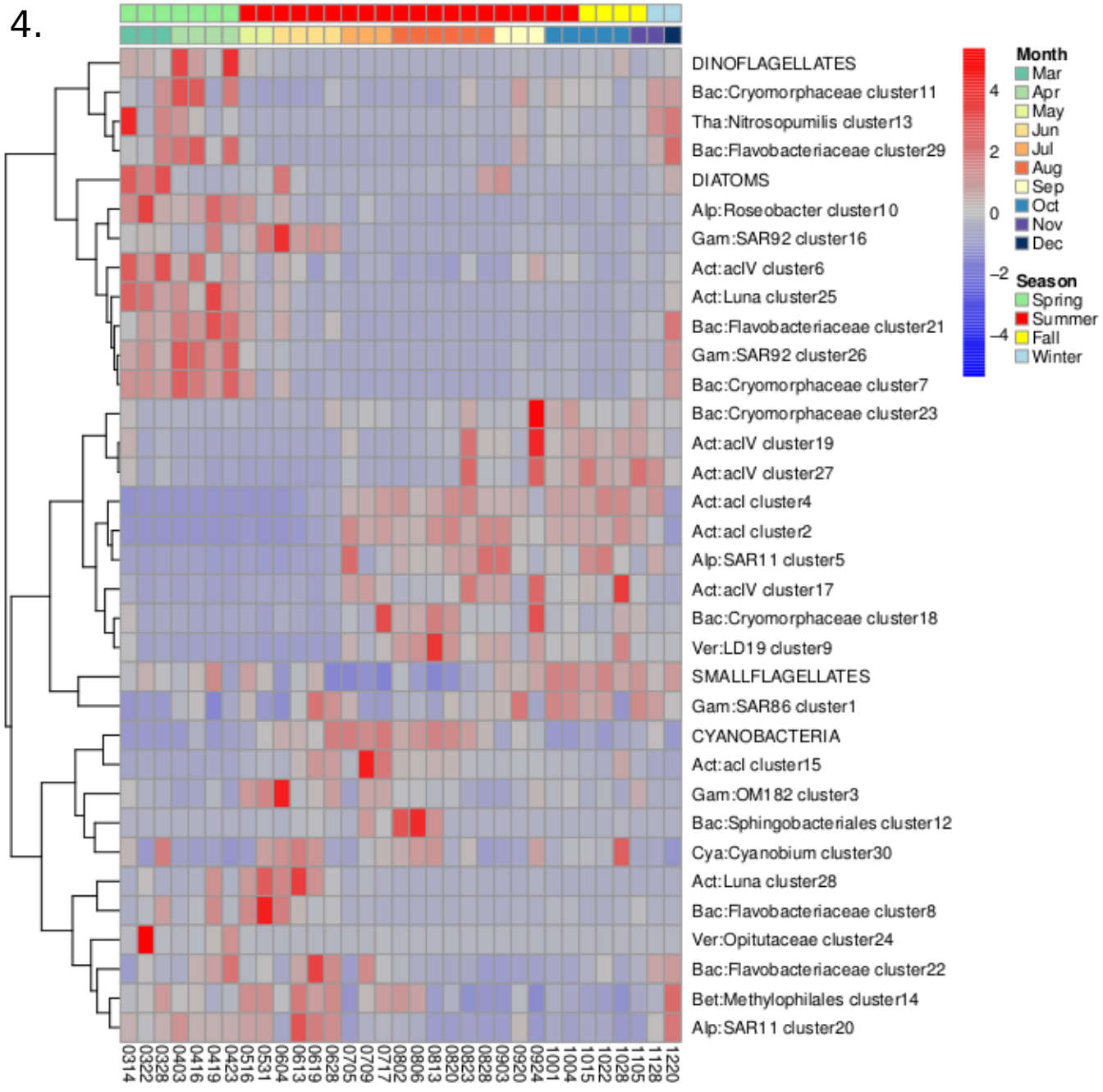

Cyanobium gracile
Cluster 30 Cya
Prochlorales
Synechococcales
Chlorobi
Coriobacteriales
Rubrobacterales
Solirubrobacterales
Acidimicrobiales
Ferrimicrobiales
Ilumatobacter sp
Cluster 19 AcIV
Cluster 27 AcIV
Cluster 17 AcIV
Cluster 6 AcIV
Acl-A1
Acl-A7
Acl-B1
Cluster 15 Acl
Cluster 2 Acl
Cluster 4 Acl
Bifidobacteriales
Actinomycetales
Cluster 28 Luna
cand Aquiluna
Cluster 25 Luna
Tropheryma whipplei
Microbacterium testaceum
Clavibacter michiganensis
Leifsonia xyli
Deferribacteres
Aquificae
Epsilonproteobacteria
Acidobacteria
Nitrospirae
Deltaproteobacteria
Other Deltaproteobacteria
Rickettsiales
LD12
Cluster 5 SAR11
SAR11
Cluster 20 SAR11
SAR116
Rhodospirillales
Sphingomonadales
Caulobacterales
Rhodobacterales
Paracoccus denitrificans
Rhodobacter sphaeroides
Cluster 10 Ros
Marine Group A
Chlorobi
Cytophagales
Cluster 12 Sphi
Sphingobacteriales
Bacteroidales
Cluster 11 Cryo
Owenweeksia hongkongensis
Cluster 18 Cryo
Cluster 23 Cryo
Cluster 7 Cryo
Bacteroidales
Flavobacteriales
Flavobacterium
Flavobacterium MS024-2A
Cluster 29 Fla
Flavobacteriales
Cluster 8 Fla
Flavobacterium MS024-3C
Flavobacteriales
Cluster 21 Fla
Aequorivita sublithincola
Altibacter lentus
Tenacibaculum ovolyticum
Flavobacterium AAA160-P02
Polaribacter MED152
Cluster 22 Fla
Spirochaetes
Planctomycetes
Opitutales
Cluster 24 Opit
Coralissmargarita akajimensis
S3
Methylacidiphilum infernorum
Cluster 9 LD19
Verrucomicrobiales
Chlorobiales
cand Spartobacteria baltica
Chlamydiae
Euryarchaeota
Nanoarchaeota
Korarchaeota
Crenarchaeota
cand Nitrosocaldaceae gargensis
Nitrosoarchaeum koreensis
Nitrosopumilus salaria
Cluster 13 Nitro
Nitrosopumilus
Thermoplasmatales
Synergistales
Thermotogae
Fusobacteria
Firmicutes
Tenericutes
Deinococcus-Thermus
Synechococcales
Oscillatoriales
Chroococcales
Nostocales
Cluster 16 SAR92
Cluster 26 SAR92
SAR92
Alteromonadales
Pseudomonadales
Oceanospirillales
Cluster 3 OM182
Pseudomonadales
Pasteurellales
Vibrionales
Aeromonadales
Cluster 14 OM43
Methylophilales
Methylophilales
Nitrosomonadales
Neisseriales
Other gammaproteobacteria
Legionellales
Thiotrichales
SAR86A
SAR86B
Cluster 1 SAR86
SAR86E

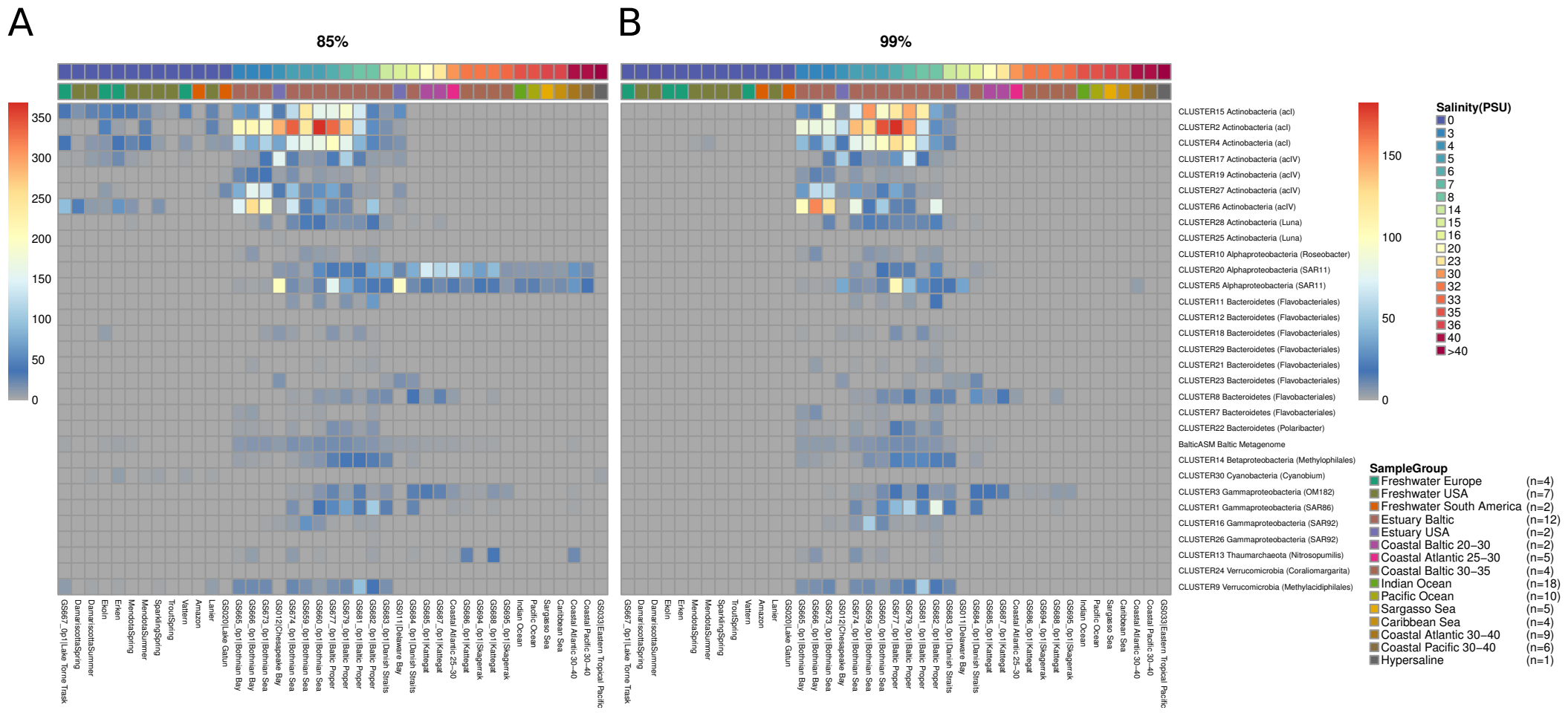
2.


NMDS COGs
stress = 0.053
MDS2
MDS1
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
Ros (10)
SAR92 (26)
SAR92 (16)
OM182 (3)
SAR11 (20)
SAR86 (1)
OM43 (14)
SAR11 (5)
Opit (24)
Cya (30)
Sphi (12)
Fla (22)
Fla (21)
Fla (8)
Fla (29)
Cryo (7)
Cryo (23)
LD19 (9)
acIV (6)
acIV (10)
acIV (27)
acIV (19)
Luna (25)
Cryo (18)
Cryo (11)
acI (4)
acI (15)
Luna (28)
acI (2)
Class
a Actinobacteria
a Alphaproteobacteria
a Bacteroidetes
a Betaproteobacteria
a Cyanobacteria
a Gammaproteobacteria
a Verrucomicrobia

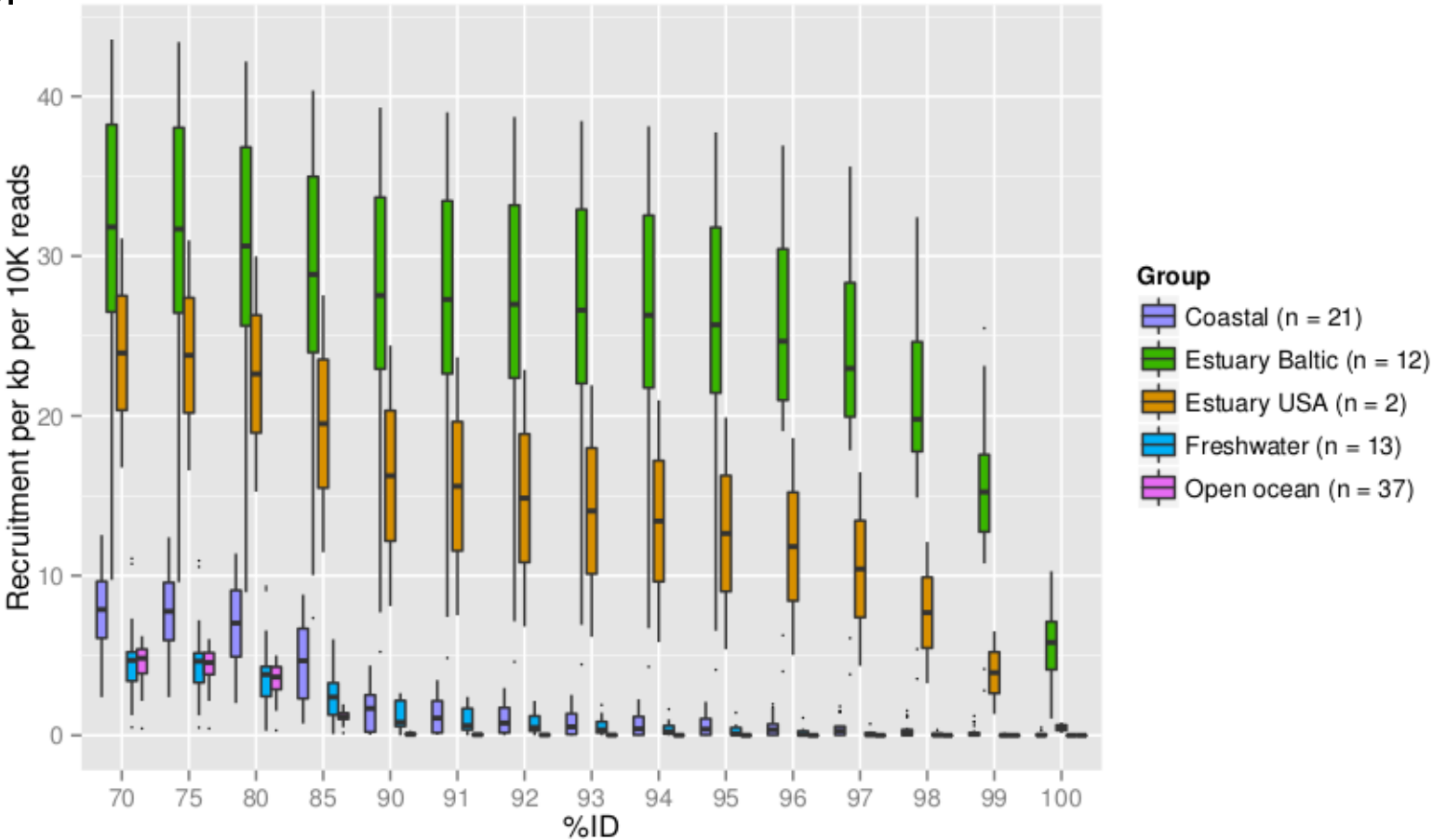

3A.
PC2 (9%)
PC1 (80%)
B.
Non-coding (%)
Genome size (Mb)
C.
Non-coding (%)
Genome size (Mb)
D.
Log (0.8 - 3.0 um)
Log (0.1 - 0.8 um)
E.
Log (>3.0 um)
Log (0.8 - 3.0 um)
F.
Log ([0.8 - 3.0 um] / [0.1 - 0.8 um])
Log (Genome size)
ACTINOBACTERIA;ACI;
ACTINOBACTERIA;ACIV;
ACTINOBACTERIA;LUNA;
ALPHAPROTEOBACTERIA;RHODOBACTER;
ALPHAPROTEOBACTERIA;SAR11;
BACTEROIDETES;CRYOMORPHACEAE;
BACTEROIDETES;FLAVOBACTERIACEAE;
BACTEROIDETES;SPHINGOBACTERIALES;
BETAPROTEOBACTERIA;OM43;
CYANOBACTERIA;CYANOBIUM;
GAMMAPROTEOBACTERIA;OM182;
GAMMAPROTEOBACTERIA;SAR86;
GAMMAPROTEOBACTERIA;SAR92;
THAUMARCHAEOTA;NITROSOPUMILUS;
VERRUCOMICROBIA;LD19;
VERRUCOMICROBIA;OPITUTACEAE;

4.

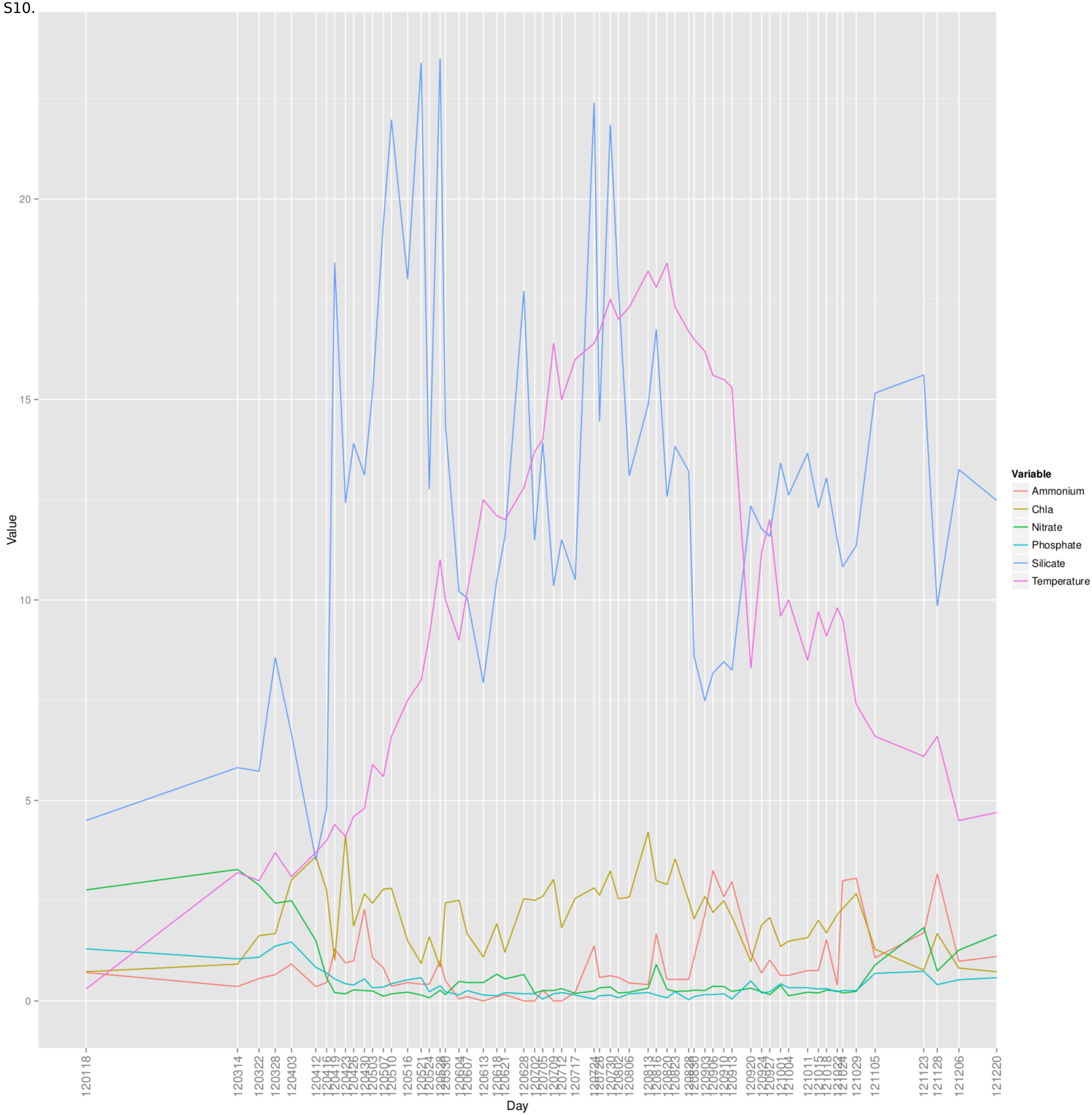

DINOFLAGELLATES
Bac:Cryomorphaceae cluster11
Tha:Nitrosopumilis cluster13
Bac:Flavobacteriaceae cluster29
DIATOMS
Alp:Roseobacter cluster10
Gam:SAR92 cluster16
Act:aclV cluster6
Act:Luna cluster25
Bac:Flavobacteriaceae cluster21
Gam:SAR92 cluster26
Bac:Cryomorphaceae cluster7
Bac:Cryomorphaceae cluster23
Act:aclV cluster19
Act:aclV cluster27
Act:acl cluster4
Act:acl cluster2
Alp:SAR11 cluster5
Act:aclV cluster17
Bac:Cryomorphaceae cluster18
Ver:LD19 cluster9
SMALLFLAGELLATES
Gam:SAR86 cluster1
CYANOBACTERIA
Act:acl cluster15
Gam:OM182 cluster3
Bac:Sphingobacteriales cluster12
Cya:Cyanobium cluster30
Act:Luna cluster28
Bac:Flavobacteriaceae cluster8
Ver:Opitutaceae cluster24
Bac:Flavobacteriaceae cluster22
Bet:Methylophilales cluster14
Alp:SAR11 cluster20
0314 0322 0328 0403 0416 0419 0423 0516 0531 0604 0613 0619 0628 0705 0709 0717 0802 0806 0813 0820 0823 0828 0903 0920 0924 1001 1004 1015 1022 1028 1105 1128 1220
4
2
0
−2
−4
Month
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Season
Spring
Summer
Fall
Winter


A
85%
B
99%
350
300
250
200
150
100
50
0
CLUSTER15 Actinobacteria (acI)
CLUSTER2 Actinobacteria (acI)
CLUSTER4 Actinobacteria (acI)
CLUSTER17 Actinobacteria (acIV)
CLUSTER19 Actinobacteria (acIV)
CLUSTER27 Actinobacteria (acIV)
CLUSTER6 Actinobacteria (acIV)
CLUSTER28 Actinobacteria (Luna)
CLUSTER25 Actinobacteria (Luna)
CLUSTER10 Alphaproteobacteria (Roseobacter)
CLUSTER20 Alphaproteobacteria (SAR11)
CLUSTER5 Alphaproteobacteria (SAR11)
CLUSTER11 Bacteroidetes (Flavobacteriales)
CLUSTER12 Bacteroidetes (Flavobacteriales)
CLUSTER18 Bacteroidetes (Flavobacteriales)
CLUSTER29 Bacteroidetes (Flavobacteriales)
CLUSTER21 Bacteroidetes (Flavobacteriales)
CLUSTER23 Bacteroidetes (Flavobacteriales)
CLUSTER8 Bacteroidetes (Flavobacteriales)
CLUSTER7 Bacteroidetes (Flavobacteriales)
CLUSTER22 Bacteroidetes (Polaribacter)
BalticASM Baltic Metagenome
CLUSTER14 Betaproteobacteria (Methylophilales)
CLUSTER30 Cyanobacteria (Cyanobium)
CLUSTER3 Gammaproteobacteria (OM182)
CLUSTER1 Gammaproteobacteria (SAR86)
CLUSTER16 Gammaproteobacteria (SAR92)
CLUSTER26 Gammaproteobacteria (SAR92)
CLUSTER13 Thaumarchaeota (Nitrosopumilis)
CLUSTER24 Verrucomicrobia (Coraliomargarita)
CLUSTER9 Verrucomicrobia (Methylacidiphilales)
GS067_0p1|Lake Torne Trask
DamariscottaSpring
DamariscottaSummer
Ekoln
Erken
MendotaSpring
MendotaSummer
SparklingSpring
TroutSpring
Vattern
Amazon
Lanier
GS020|Lake Gatun
GS665_0p1|Bothnian Bay
GS666_0p1|Bothnian Bay
GS673_0p1|Bothnian Sea
GS012|Chesapeake Bay
GS674_0p1|Bothnian Sea
GS659_0p1|Bothnian Sea
GS660_0p1|Bothnian Sea
GS677_0p1|Baltic Proper
GS679_0p1|Baltic Proper
GS681_0p1|Baltic Proper
GS682_0p1|Baltic Proper
GS683_0p1|Danish Straits
GS011|Delaware Bay
GS684_0p1|Danish Straits
GS685_0p1|Kattegat
GS687_0p1|Kattegat
Coastal Atlantic 25–30
GS686_0p1|Kattegat
GS694_0p1|Skagerrak
GS688_0p1|Kattegat
GS695_0p1|Skagerrak
Indian Ocean
Pacific Ocean
Sargasso Sea
Caribbean Sea
Coastal Atlantic 30–40
Coastal Pacific 30–40
GS033|Eastern Tropical Pacific
150
100
50
0
Salinity(PSU)
0
3
4
5
6
7
8
14
15
16
20
23
30
32
33
35
36
40
>40
SampleGroup
Freshwater Europe (n=4)
Freshwater USA (n=7)
Freshwater South America (n=2)
Estuary Baltic (n=12)
Estuary USA (n=2)
Coastal Baltic 20–30 (n=2)
Coastal Atlantic 25–30 (n=5)
Coastal Baltic 30–35 (n=4)
Indian Ocean (n=18)
Pacific Ocean (n=10)
Sargasso Sea (n=5)
Caribbean Sea (n=4)
Coastal Atlantic 30–40 (n=9)
Coastal Pacific 30–40 (n=6)
Hypersaline (n=1)

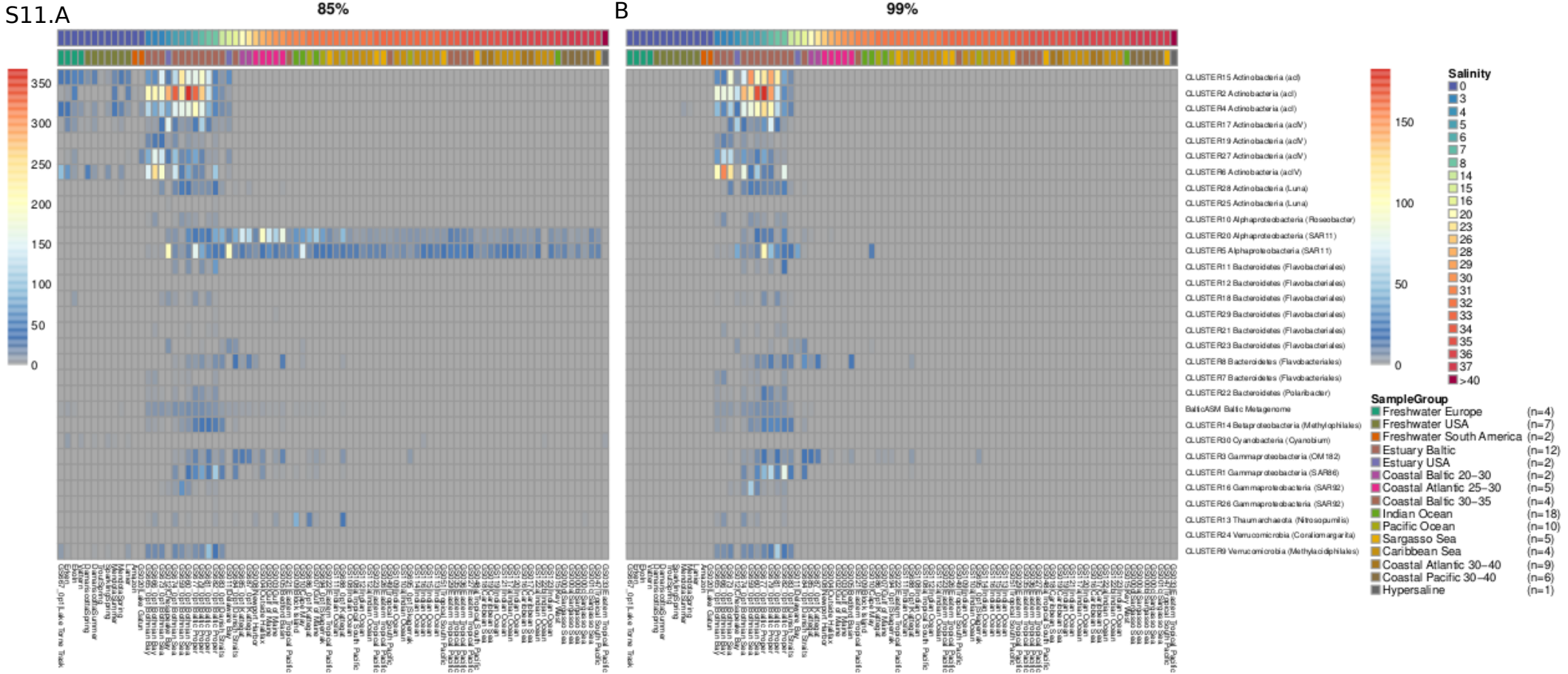
6.

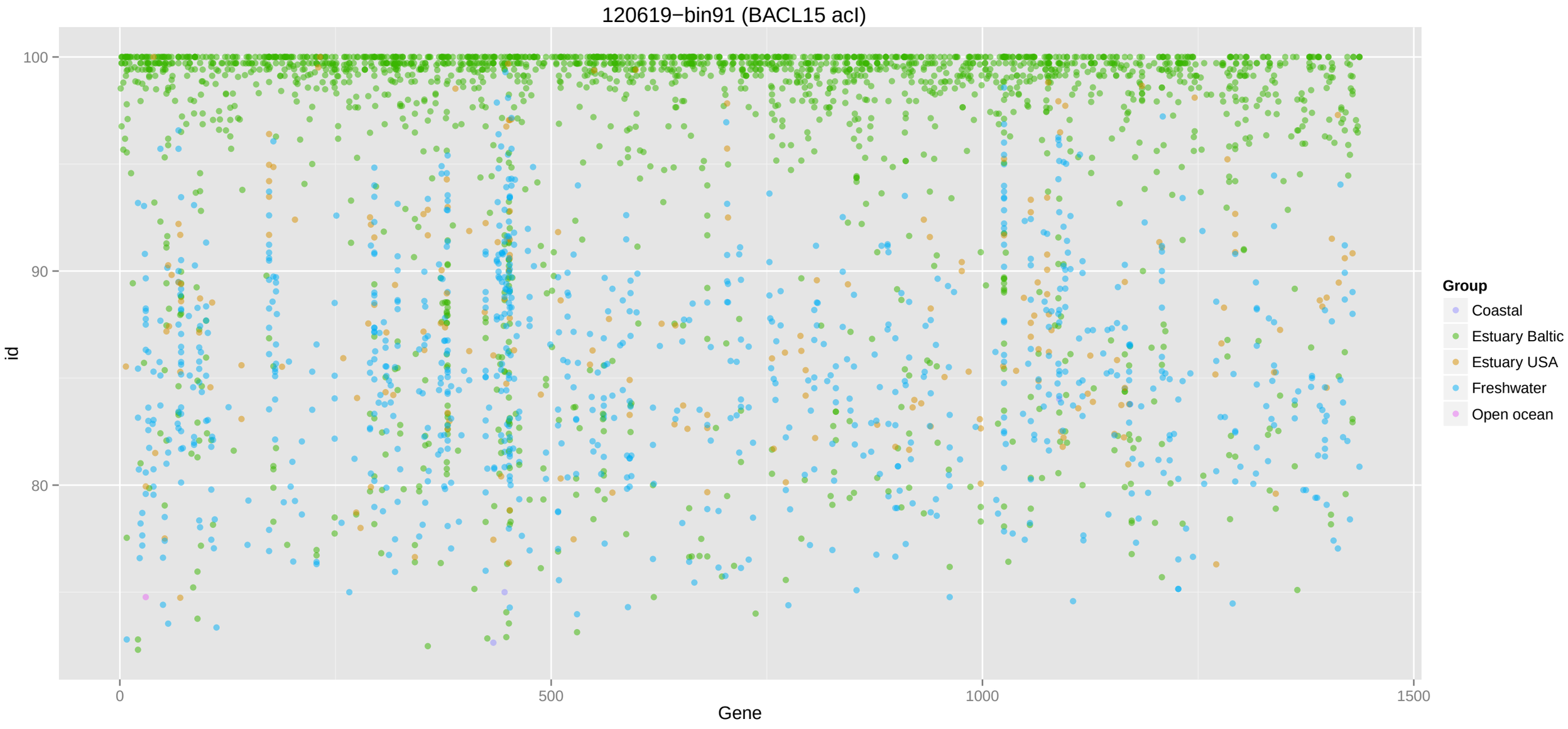

Recruitment per kb per 10K reads
0
10
20
30
40
70
75
80
85
90
91
92
93
94
95
96
97
98
99
100
%ID
Group
Coastal (n = 21)
Estuary Baltic (n = 12)
Estuary USA (n = 2)
Freshwater (n = 13)
Open ocean (n = 37)

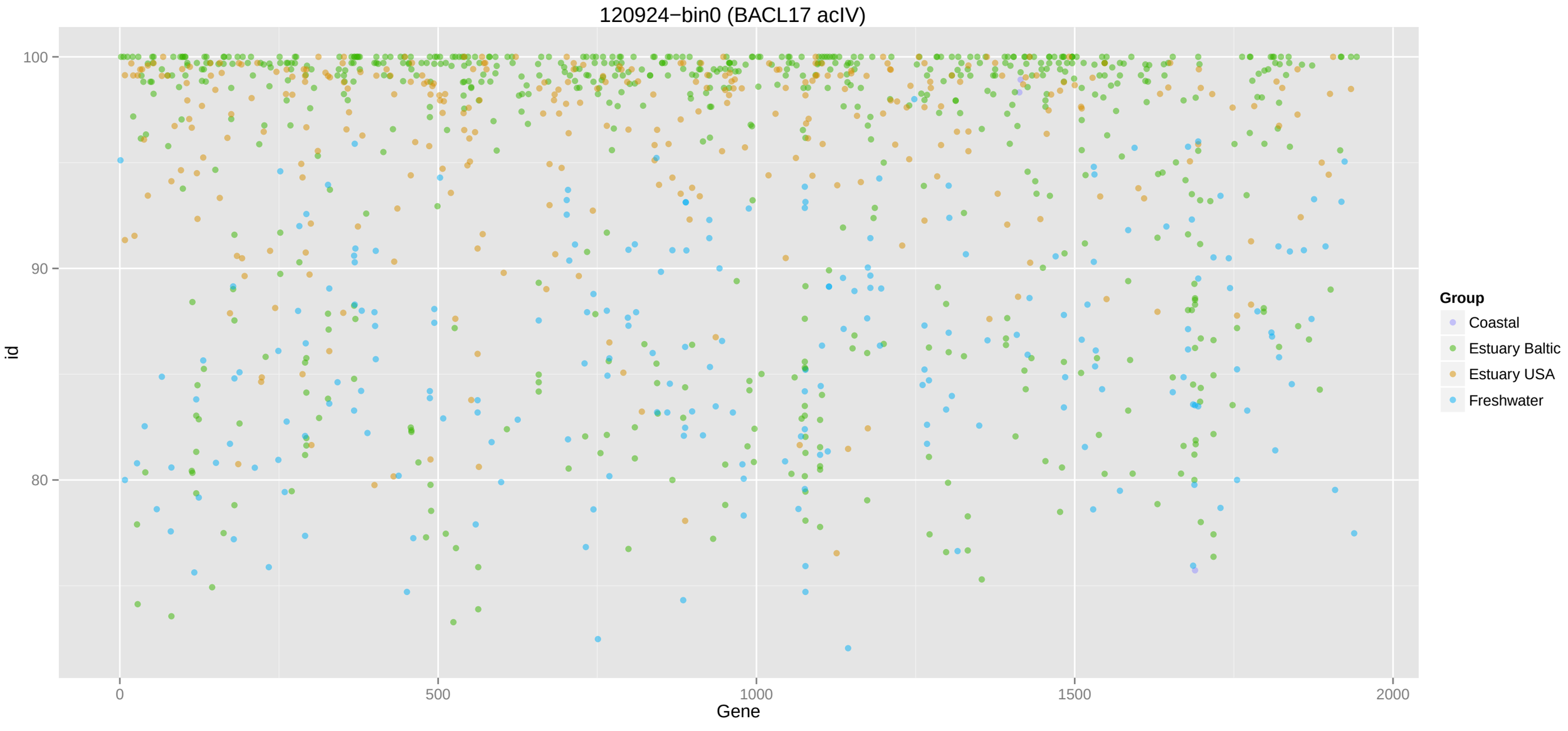
S10.

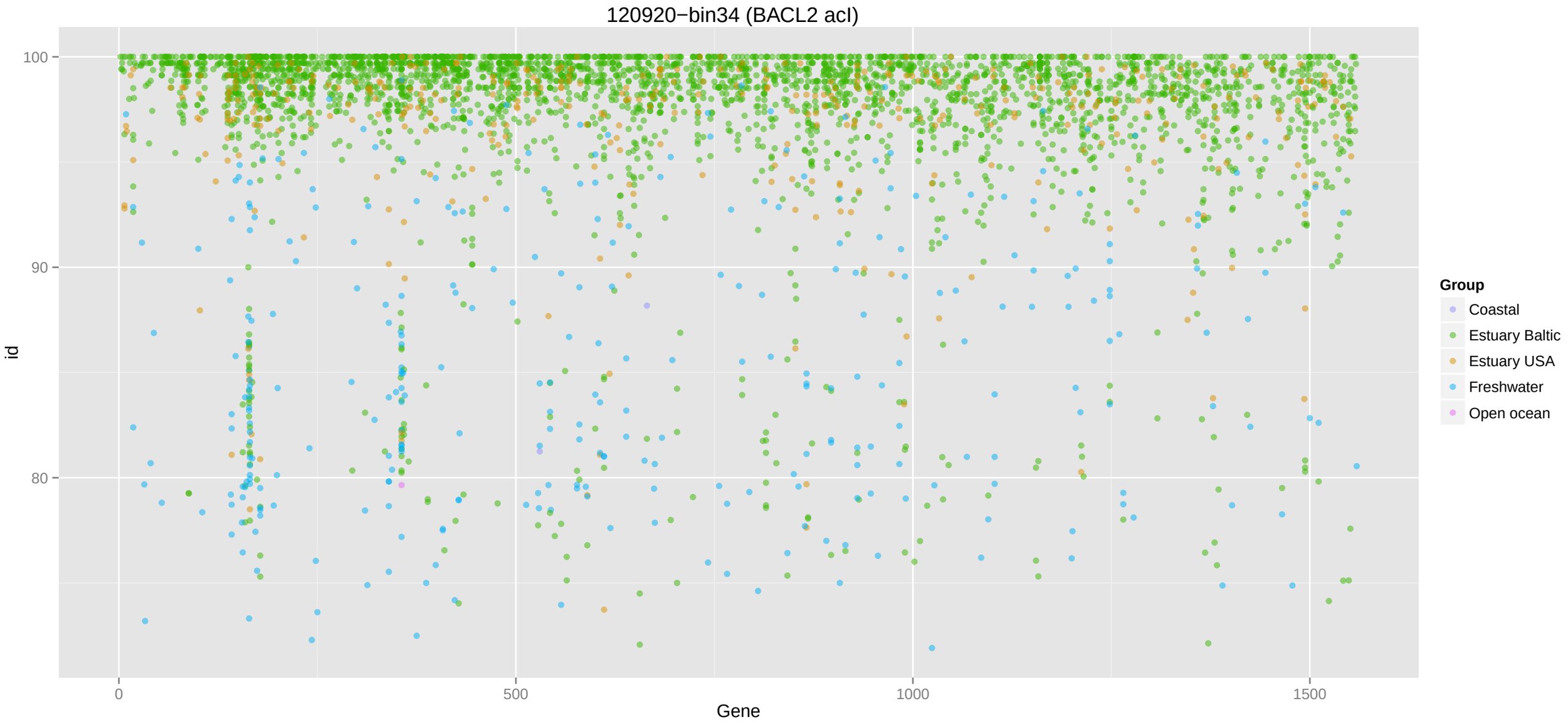

Value
20
15
10
5
0
Day
Variable
Ammonium
Chla
Nitrate
Phosphate
Silicate
Temperature

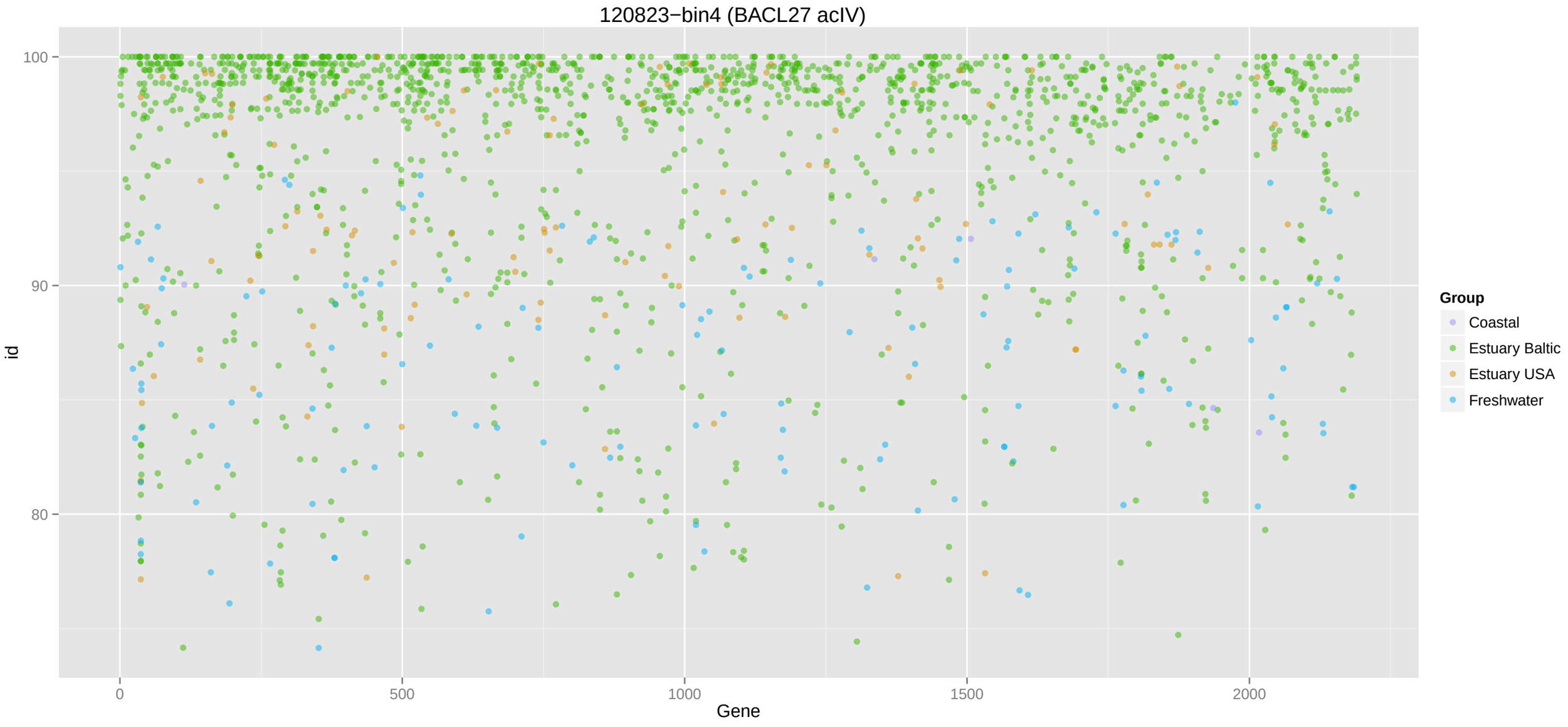
S11.A

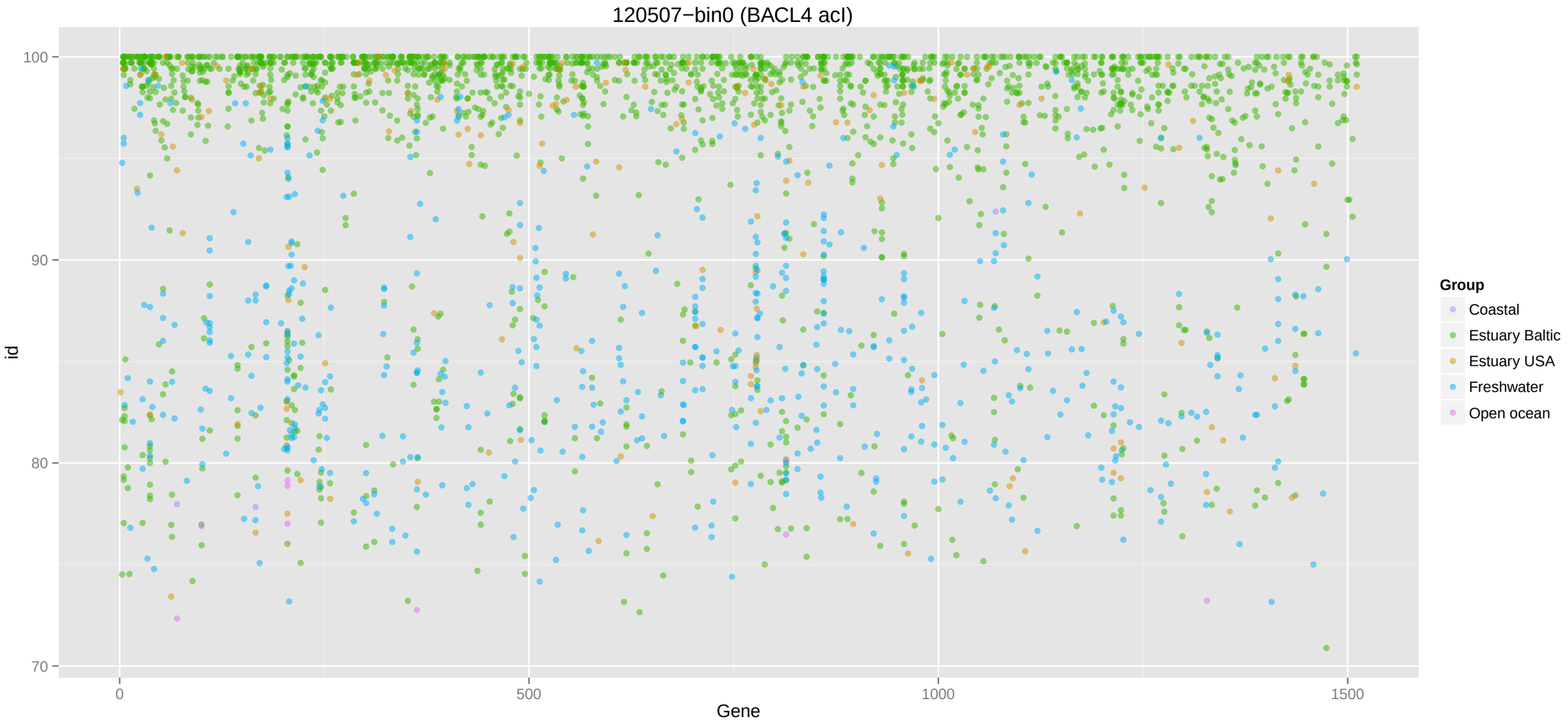

85%
99%
B
350
300
250
200
150
100
50
0
CLUSTER15 Actinobacteria (acI)
CLUSTER2 Actinobacteria (acI)
CLUSTER4 Actinobacteria (acI)
CLUSTER17 Actinobacteria (acIV)
CLUSTER19 Actinobacteria (acIV)
CLUSTER27 Actinobacteria (acIV)
CLUSTER6 Actinobacteria (acIV)
CLUSTER28 Actinobacteria (Luna)
CLUSTER25 Actinobacteria (Luna)
CLUSTER10 Alphaproteobacteria (Roseobacter)
CLUSTER20 Alphaproteobacteria (SAR11)
CLUSTER5 Alphaproteobacteria (SAR11)
CLUSTER11 Bacteroidetes (Flavobacteriales)
CLUSTER12 Bacteroidetes (Flavobacteriales)
CLUSTER18 Bacteroidetes (Flavobacteriales)
CLUSTER29 Bacteroidetes (Flavobacteriales)
CLUSTER21 Bacteroidetes (Flavobacteriales)
CLUSTER23 Bacteroidetes (Flavobacteriales)
CLUSTER8 Bacteroidetes (Flavobacteriales)
CLUSTER7 Bacteroidetes (Flavobacteriales)
CLUSTER22 Bacteroidetes (Polaribacter)
BalticASM Baltic Metagenome
CLUSTER14 Betaproteobacteria (Methylophilales)
CLUSTER30 Cyanobacteria (Cyanobium)
CLUSTER3 Gammaproteobacteria (OM182)
CLUSTER1 Gammaproteobacteria (SAR86)
CLUSTER16 Gammaproteobacteria (SAR92)
CLUSTER26 Gammaproteobacteria (SAR92)
CLUSTER13 Thaumarchaeota (Nitrosopumilis)
CLUSTER24 Verrucomicrobia (Coraliomargarita)
CLUSTER9 Verrucomicrobia (Methylacidiphilales)
150
100
50
0
Salinity
0
3
4
5
6
7
8
14
15
16
20
23
26
28
29
30
31
32
33
34
35
36
37
>40
SampleGroup
Freshwater Europe (n=4)
Freshwater USA (n=7)
Freshwater South America (n=2)
Estuary Baltic (n=12)
Estuary USA (n=2)
Coastal Baltic 20–30 (n=2)
Coastal Atlantic 25–30 (n=5)
Coastal Baltic 30–35 (n=4)
Indian Ocean (n=18)
Pacific Ocean (n=10)
Sargasso Sea (n=5)
Caribbean Sea (n=4)
Coastal Atlantic 30–40 (n=9)
Coastal Pacific 30–40 (n=6)
Hypersaline (n=1)

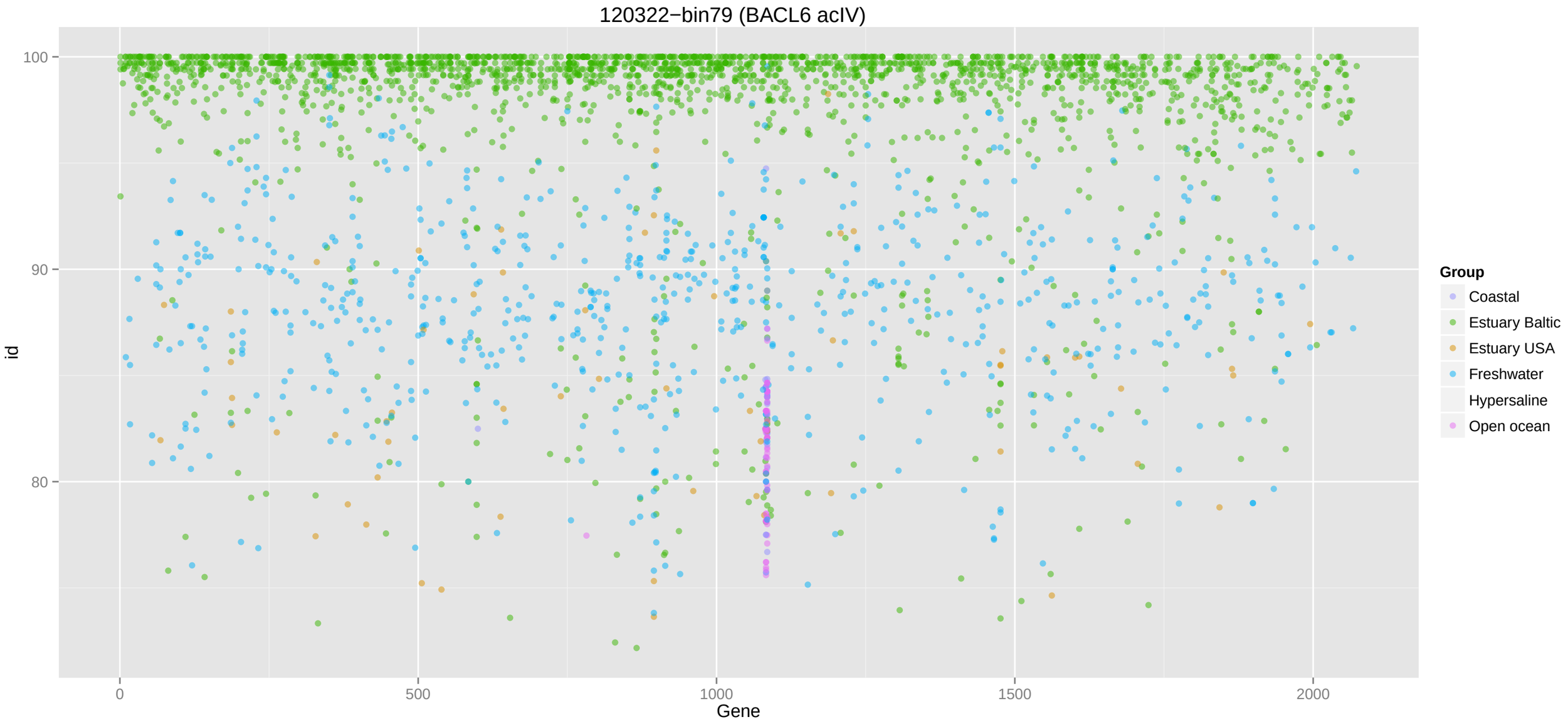

120619−bin91 (BACL15 acI)
100
90
80
id
0
500
1000
1500
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Open ocean

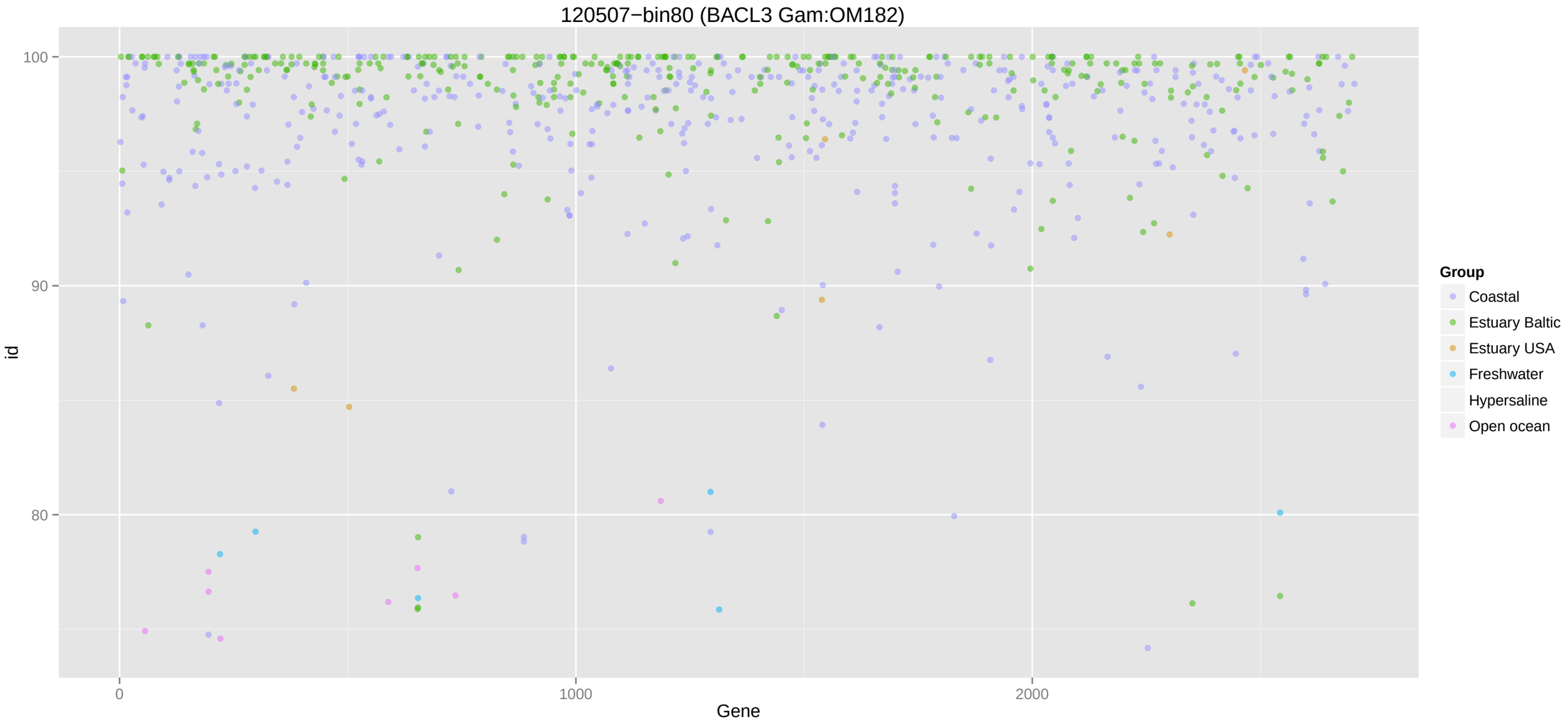

120924−bin0 (BACL17 acIV)
id
100
90
80
0
500
1000
1500
2000
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater


120920−bin34 (BACL2 acl)
100
90
80
id
0
500
1000
1500
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Open ocean

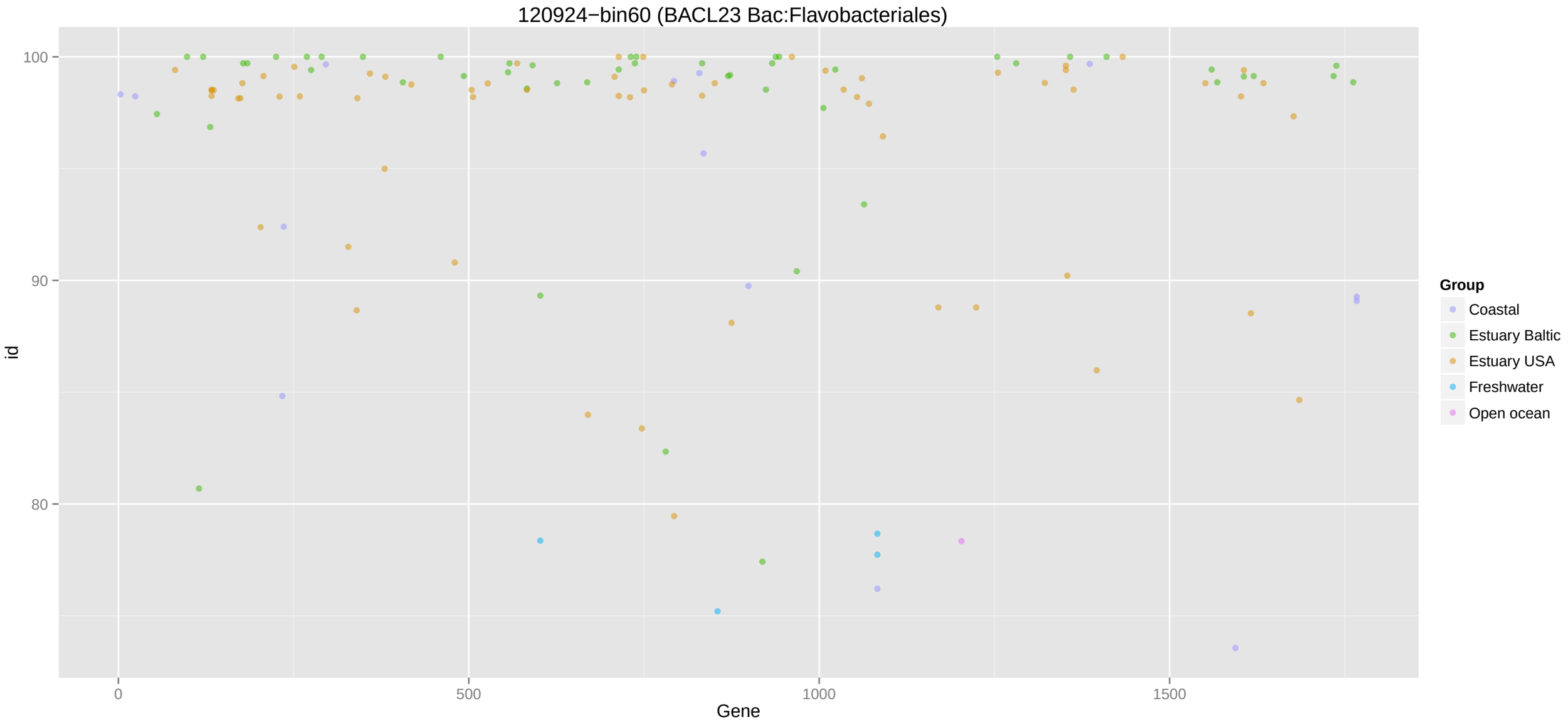

120823−bin4 (BACL27 acIV)
id
100
90
80
0
500
1000
1500
2000
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater

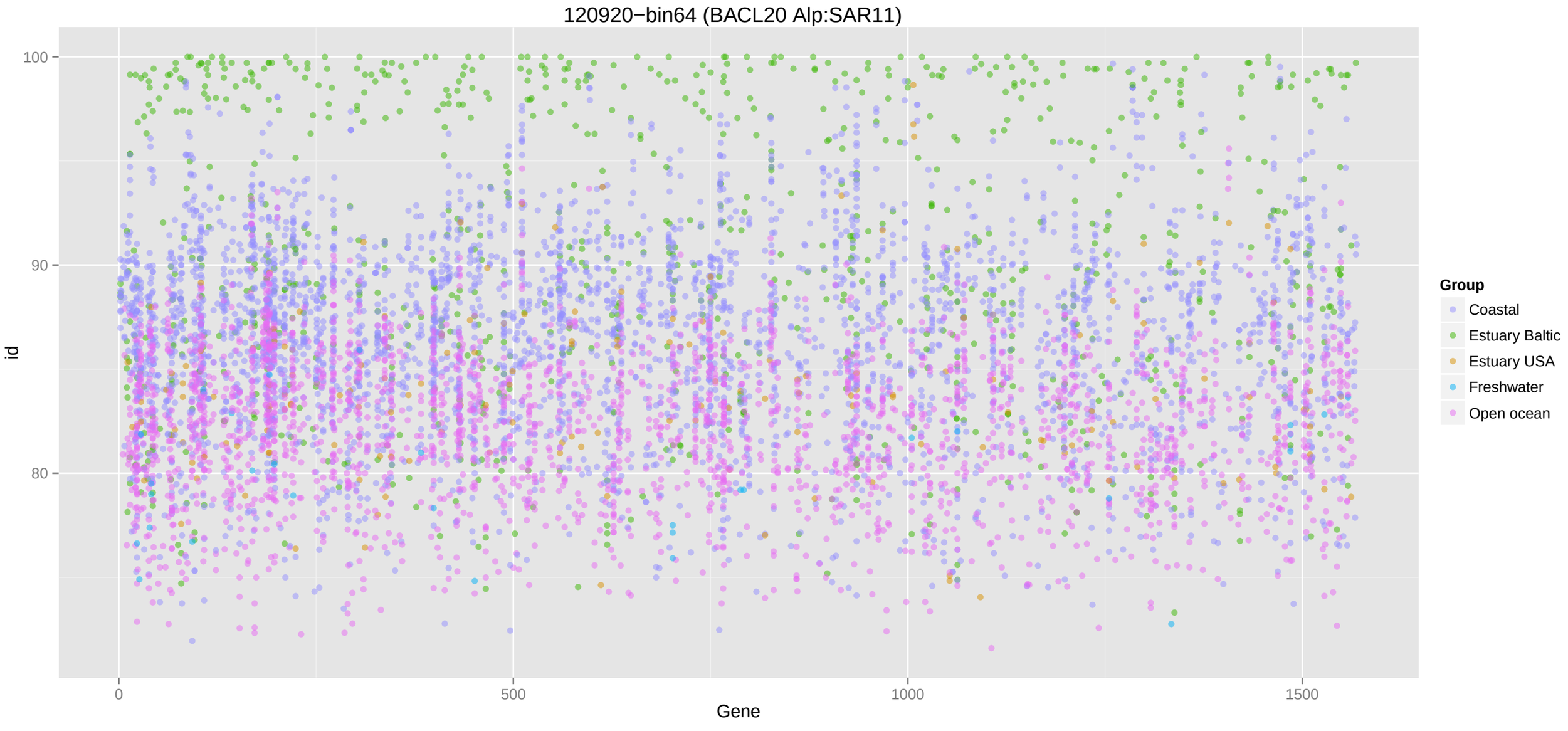

120507−bin0 (BACL4 acl)
id
100
90
80
70
0
500
1000
1500
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Open ocean

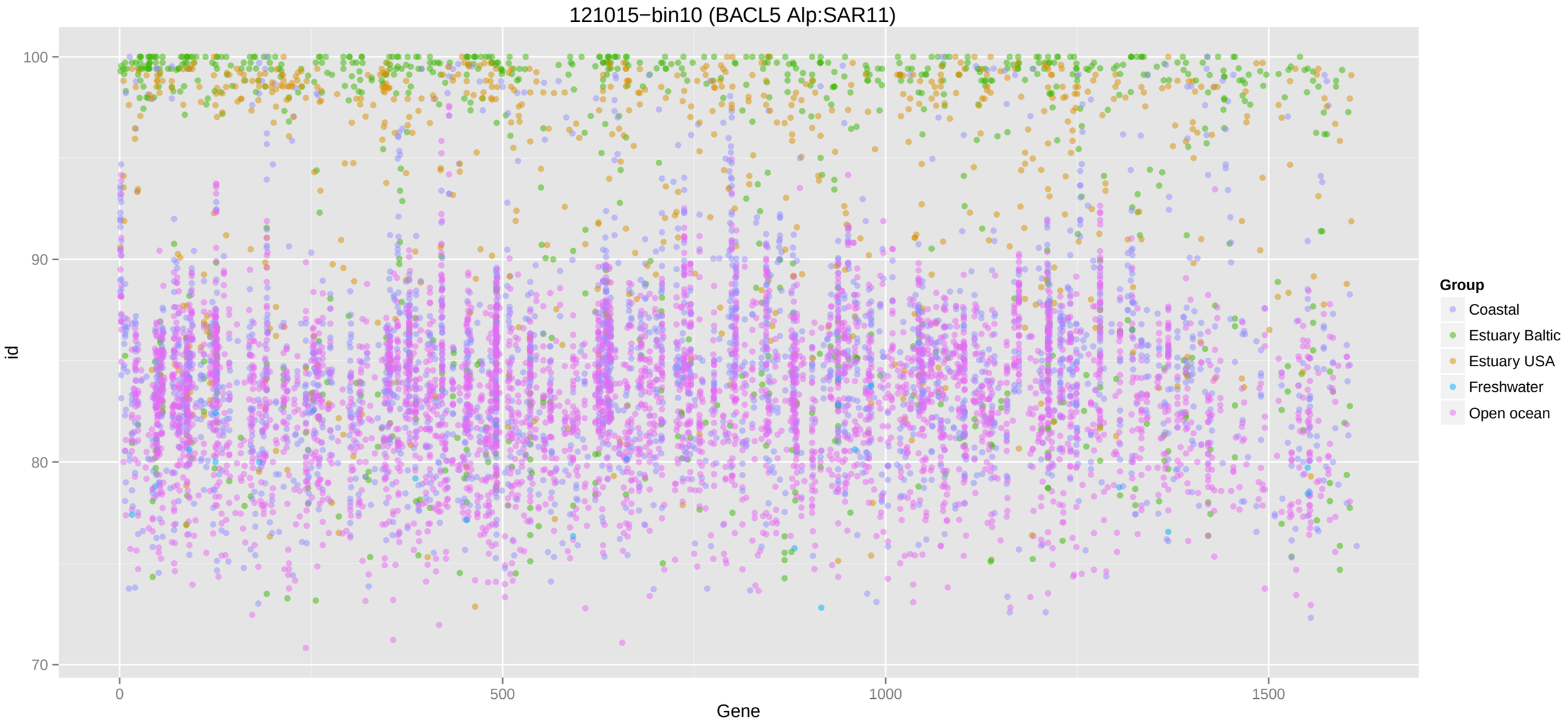

120322−bin79 (BACL6 acIV)
id
100
90
80
0
500
1000
1500
2000
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Hypersaline
Open ocean

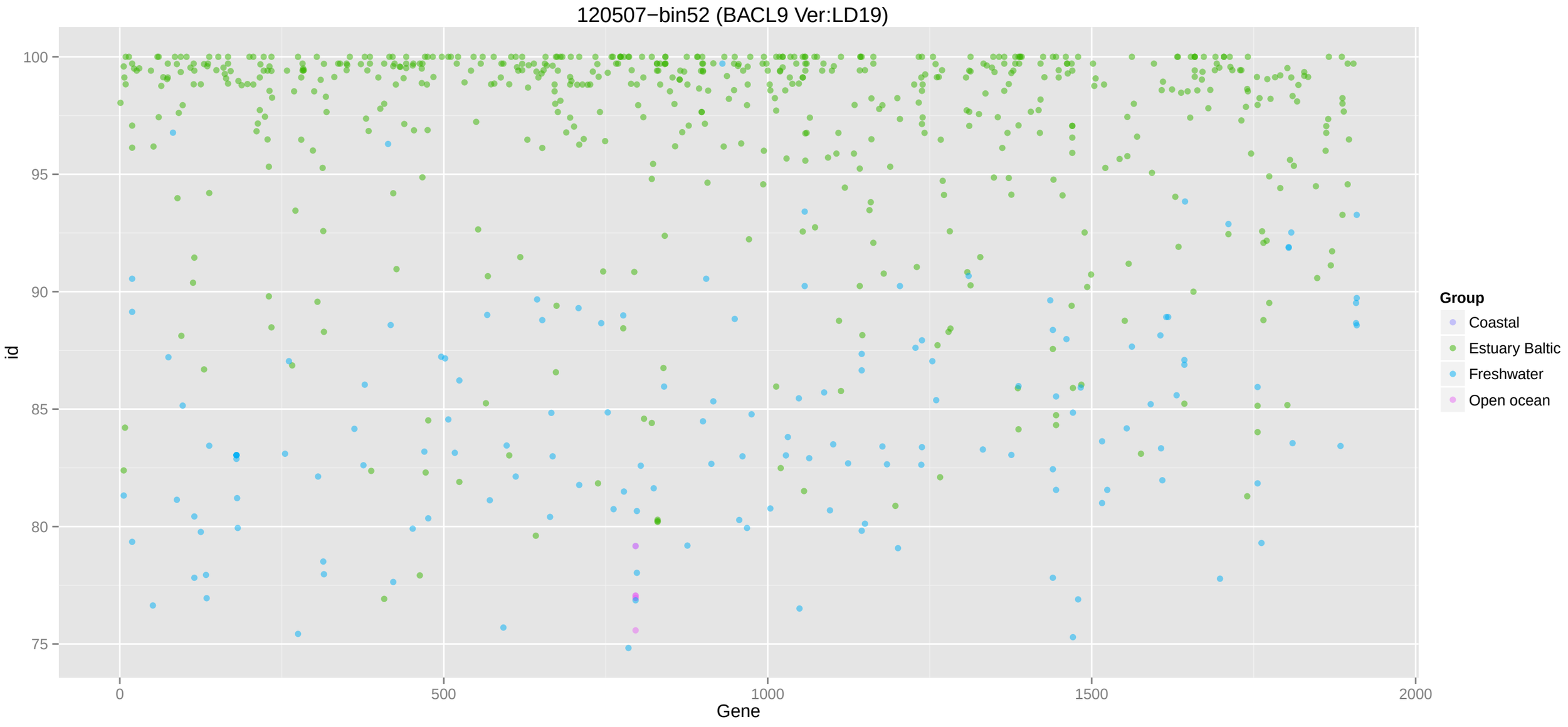

120507−bin80 (BACL3 Gam:OM182)
100
90
80
id
0
1000
2000
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Hypersaline
Open ocean

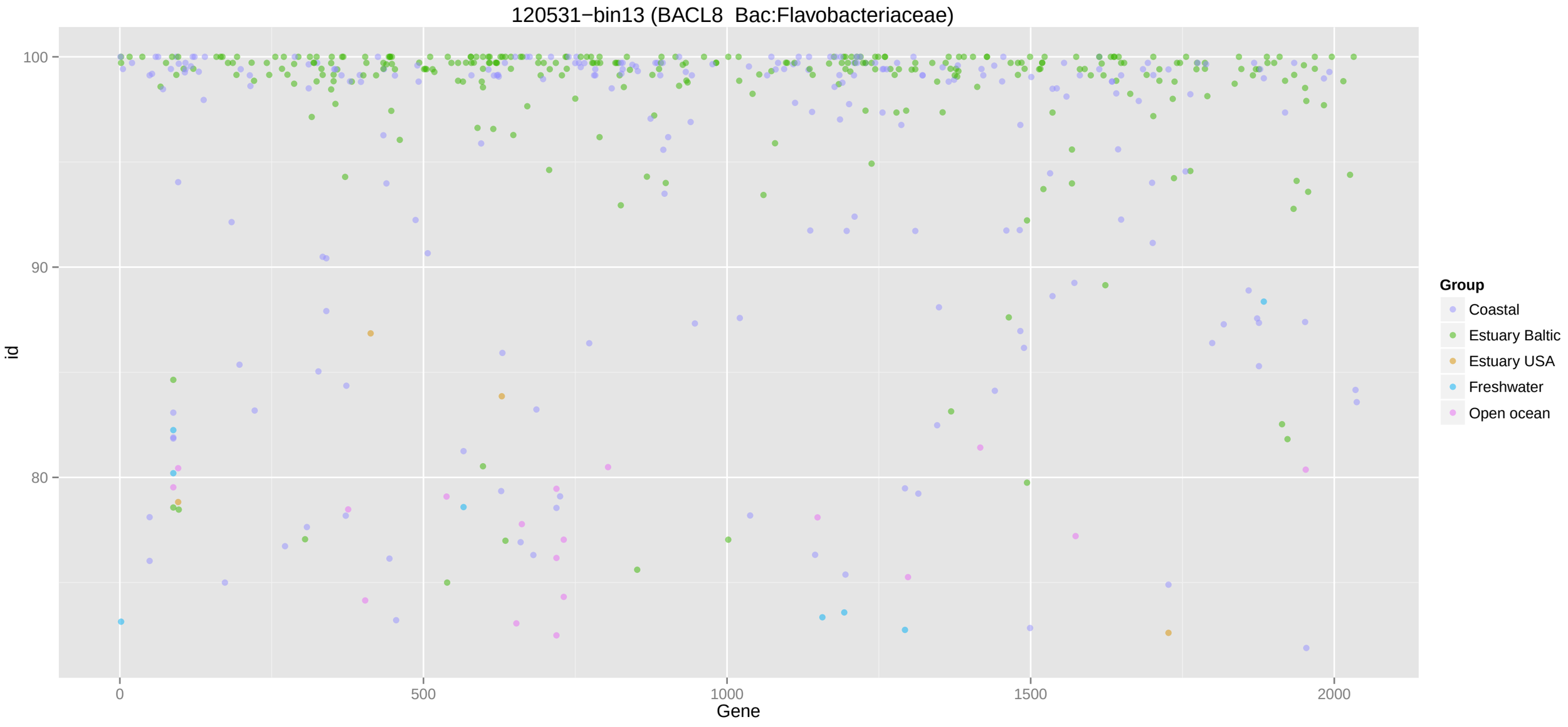
120619−bin26 (BACL1 Gam:SAR86)
id
100
90
80
70
0
500
1000
1500
Gene
Group
Coastal
Estuary Baltic
Open ocean

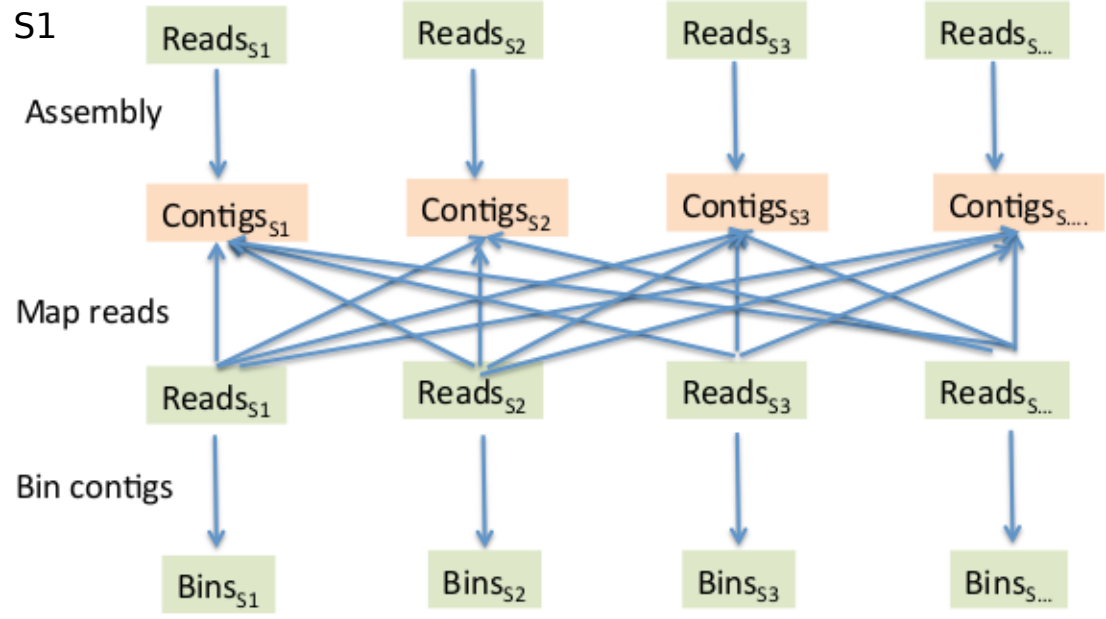

120924−bin60 (BACL23 Bac:Flavobacteriales)
id
100
90
80
0
500
1000
1500
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Open ocean


120920−bin64 (BACL20 Alp:SAR11)
100
90
80
id
0
500
1000
1500
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Open ocean

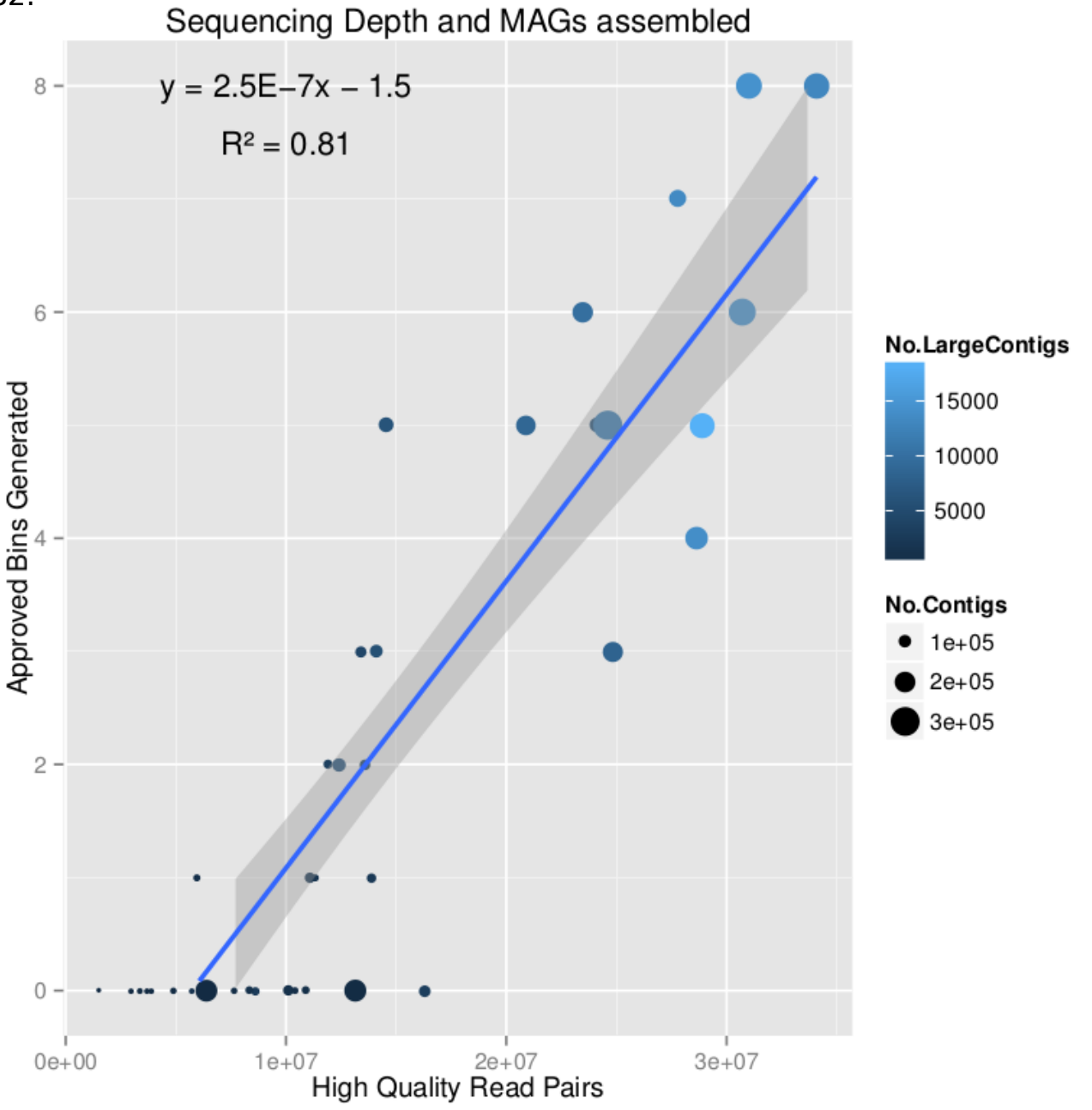

121015−bin10 (BACL5 Alp:SAR11)
id
100
90
80
70
0
500
1000
1500
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Open ocean


120507−bin52 (BACL9 Ver:LD19)
id
100
95
90
85
80
75
0
500
1000
1500
2000
Gene
Group
Coastal
Estuary Baltic
Freshwater
Open ocean

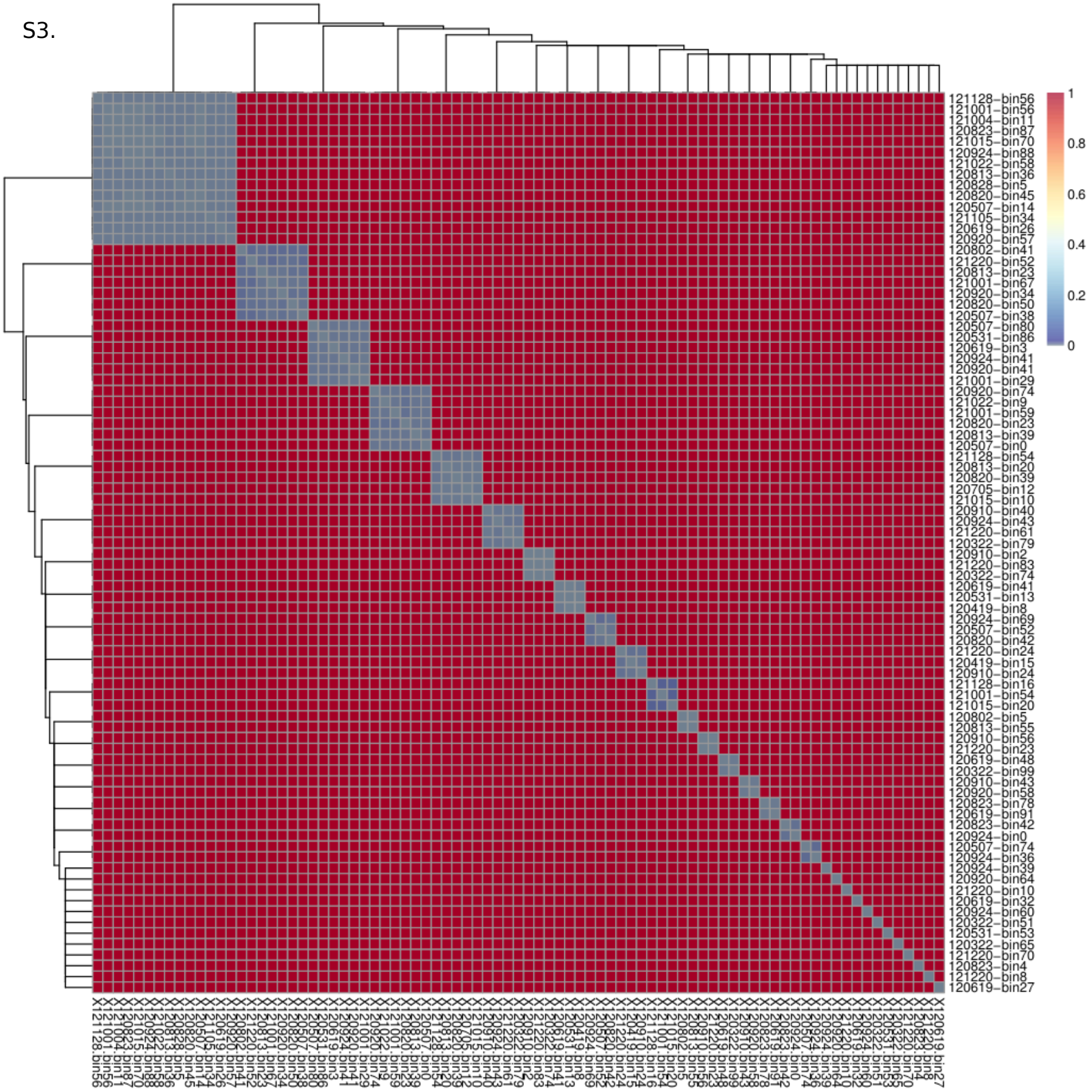

120531−bin13 (BACL8 Bac:Flavobacteriaceae)
id
100
90
80
0
500
1000
1500
2000
Gene
Group
Coastal
Estuary Baltic
Estuary USA
Freshwater
Open ocean


S1
Reads$_{S1}$
Reads$_{S2}$
Reads$_{S3}$
Reads$_{S...}$
Assembly
Contigs$_{S1}$
Contigs$_{S2}$
Contigs$_{S3}$
Contigs$_{S....}$
Map reads
Reads$_{S1}$
Reads$_{S2}$
Reads$_{S3}$
Reads$_{S...}$
Bin contigs
Bins$_{S1}$
Bins$_{S2}$
Bins$_{S3}$
Bins$_{S...}$

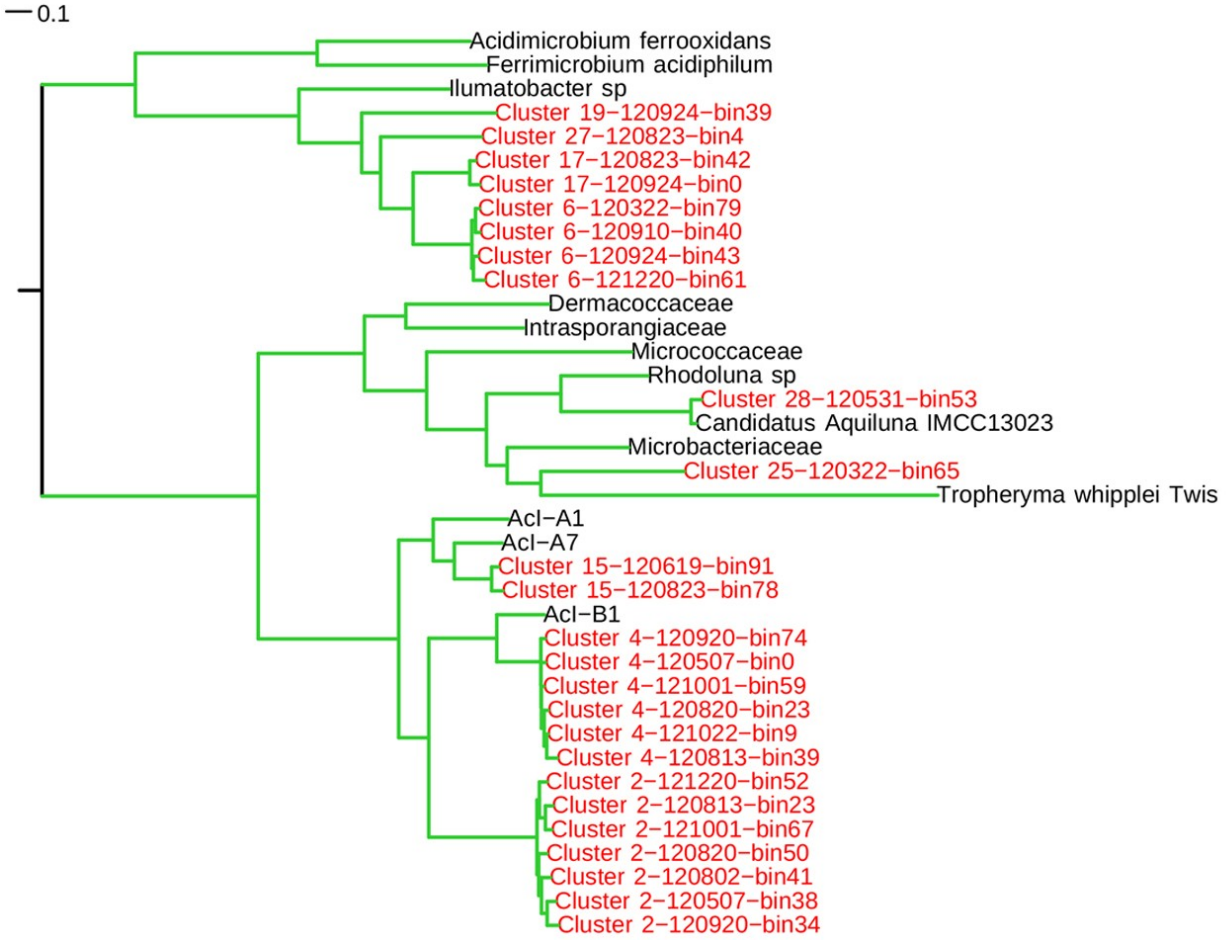
S2.


Sequencing Depth and MAGs assembled
y = 2.5E−7x − 1.5
R² = 0.81
Approved Bins Generated
High Quality Read Pairs
0
2
4
6
8
0e+00
1e+07
2e+07
3e+07
No.LargeContigs
15000
10000
5000
No.Contigs
1e+05
2e+05
3e+05

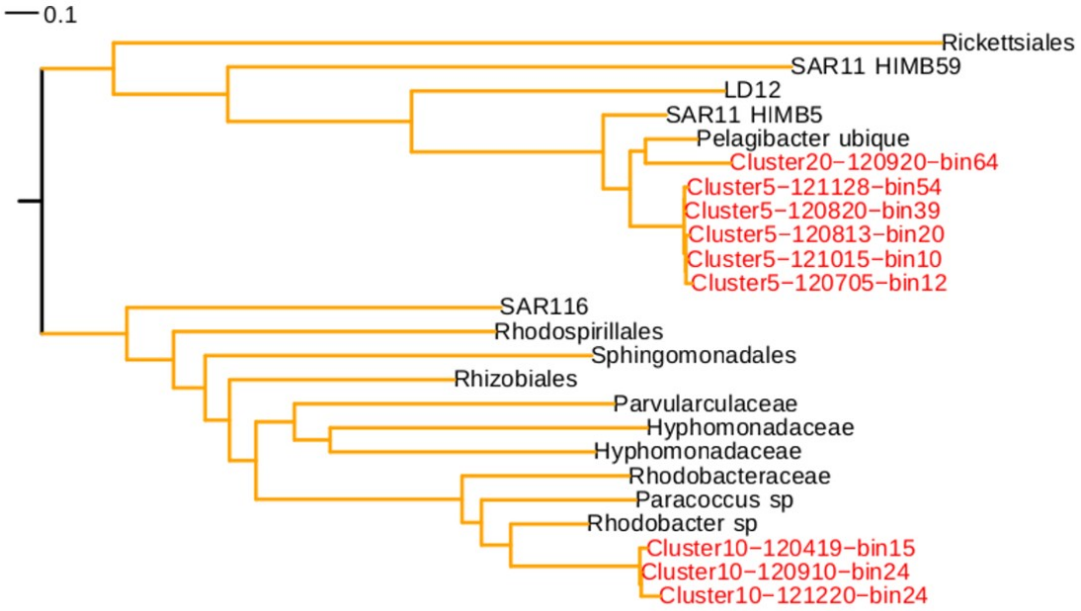
S3.


121128–bin56
121001–bin56
121004–bin11
120823–bin87
121015–bin70
120924–bin88
121022–bin58
120813–bin36
120828–bin5
120820–bin45
120507–bin14
121105–bin34
120619–bin26
120920–bin57
120802–bin41
121220–bin52
120813–bin23
121001–bin67
120920–bin34
120820–bin50
120507–bin38
120507–bin80
120531–bin86
120619–bin3
120924–bin41
120920–bin41
121001–bin29
120920–bin74
121022–bin9
121001–bin59
120820–bin23
120813–bin39
120507–bin0
121128–bin54
120813–bin20
120820–bin39
120705–bin12
121015–bin10
120910–bin40
120924–bin43
121220–bin61
120322–bin79
120910–bin2
121220–bin83
120322–bin74
120619–bin41
120531–bin13
120419–bin8
120924–bin69
120507–bin52
120820–bin42
121220–bin24
120419–bin15
120910–bin24
121128–bin16
121001–bin54
121015–bin20
120802–bin5
120813–bin55
120910–bin56
121220–bin23
120619–bin48
120322–bin99
120910–bin43
120920–bin58
120823–bin78
120619–bin91
120823–bin42
120924–bin0
120507–bin74
120924–bin36
120924–bin39
120920–bin64
121220–bin10
120619–bin32
120924–bin60
120322–bin51
120531–bin53
120322–bin65
121220–bin70
120823–bin4
121220–bin8
120619–bin27
1
0.8
0.6
0.4
0.2
0

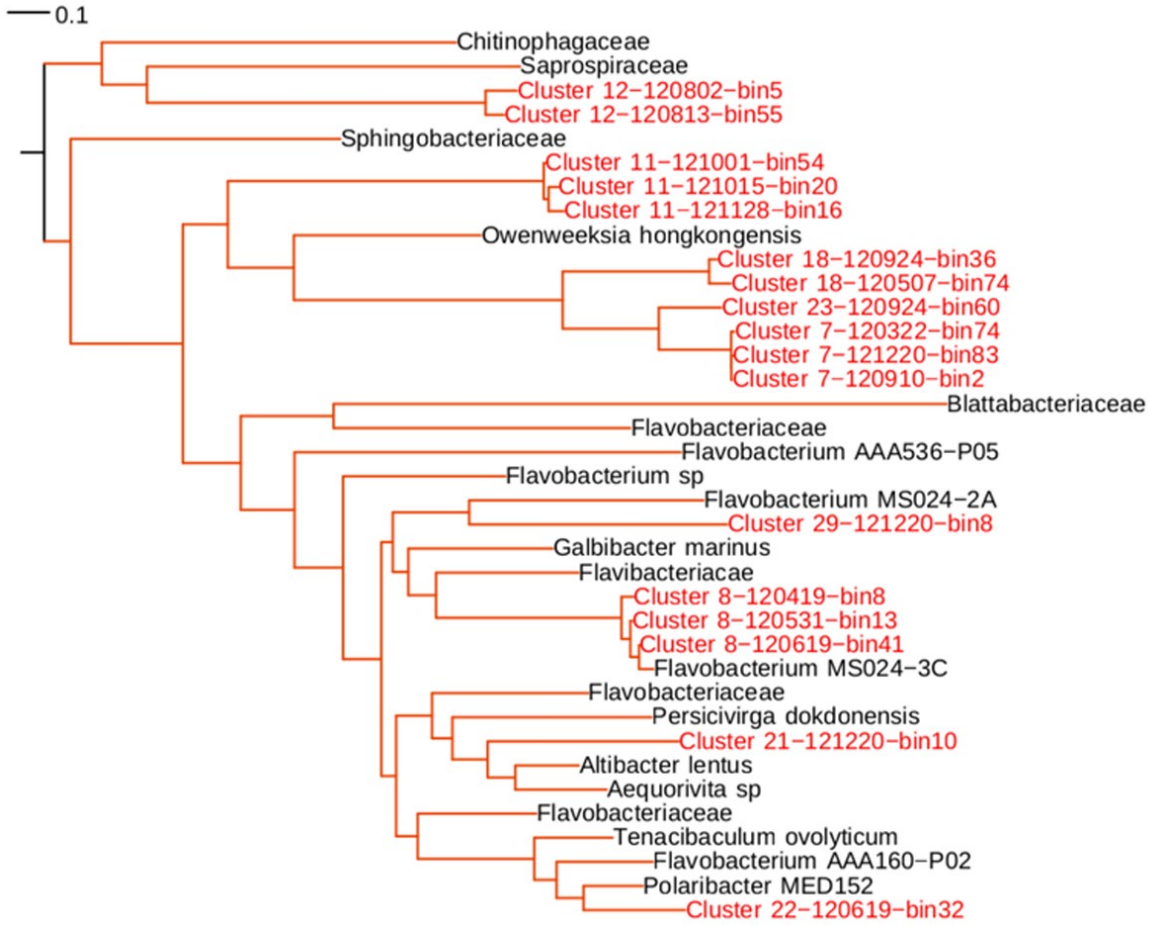
a)


0.1
Acidimicrobium ferrooxidans
Ferrimicrobium acidiphilum
Ilumatobacter sp
Cluster 19−120924−bin39
Cluster 27−120823−bin4
Cluster 17−120823−bin42
Cluster 17−120924−bin0
Cluster 6−120322−bin79
Cluster 6−120910−bin40
Cluster 6−120924−bin43
Cluster 6−121220−bin61
Dermacoccaceae
Intrasporangiaceae
Micrococcaceae
Rhodoluna sp
Cluster 28−120531−bin53
Candidatus Aquiluna IMCC13023
Microbacteriaceae
Cluster 25−120322−bin65
Tropheryma whipplei Twis
AcI−A1
AcI−A7
Cluster 15−120619−bin91
Cluster 15−120823−bin78
AcI−B1
Cluster 4−120920−bin74
Cluster 4−120507−bin0
Cluster 4−121001−bin59
Cluster 4−120820−bin23
Cluster 4−121022−bin9
Cluster 4−120813−bin39
Cluster 2−121220−bin52
Cluster 2−120813−bin23
Cluster 2−121001−bin67
Cluster 2−120820−bin50
Cluster 2−120802−bin41
Cluster 2−120507−bin38
Cluster 2−120920−bin34

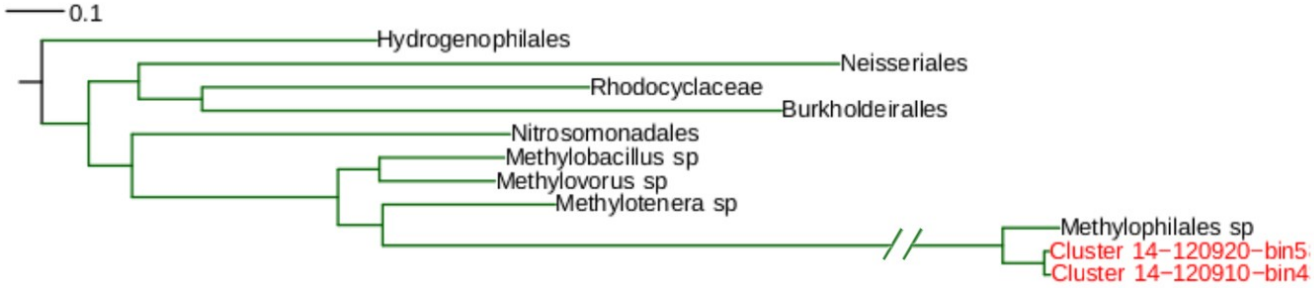
b)


0.1
Rickettsiales
SAR11 HIMB59
LD12
SAR11 HIMB5
Pelagibacter ubique
Cluster20−120920−bin64
Cluster5−121128−bin54
Cluster5−120820−bin39
Cluster5−120813−bin20
Cluster5−121015−bin10
Cluster5−120705−bin12
SAR116
Rhodospirillales
Sphingomonadales
Rhizobiales
Parvularculaceae
Hyphomonadaceae
Hyphomonadaceae
Rhodobacteraceae
Paracoccus sp
Rhodobacter sp
Cluster10−120419−bin15
Cluster10−120910−bin24
Cluster10−121220−bin24

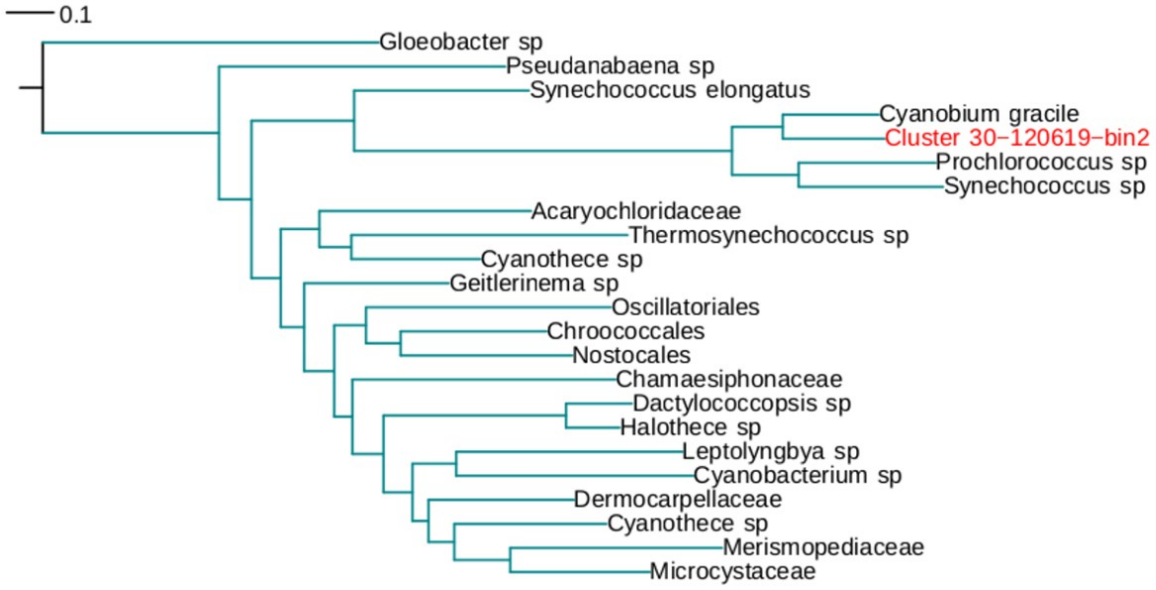
c)


0.1
Chitinophagaceae
Saprospiraceae
Cluster 12−120802−bin5
Cluster 12−120813−bin55
Sphingobacteriaceae
Cluster 11−121001−bin54
Cluster 11−121015−bin20
Cluster 11−121128−bin16
Owenweeksia hongkongensis
Cluster 18−120924−bin36
Cluster 18−120507−bin74
Cluster 23−120924−bin60
Cluster 7−120322−bin74
Cluster 7−121220−bin83
Cluster 7−120910−bin2
Blattabacteriaceae
Flavobacteriaceae
Flavobacterium AAA536−P05
Flavobacterium sp
Flavobacterium MS024−2A
Cluster 29−121220−bin8
Galbibacter marinus
Flavibacteriacae
Cluster 8−120419−bin8
Cluster 8−120531−bin13
Cluster 8−120619−bin41
Flavobacterium MS024−3C
Flavobacteriaceae
Persicivirga dokdonensis
Cluster 21−121220−bin10
Altibacter lentus
Aequorivita sp
Flavobacteriaceae
Tenacibaculum ovolyticum
Flavobacterium AAA160−P02
Polaribacter MED152
Cluster 22−120619−bin32

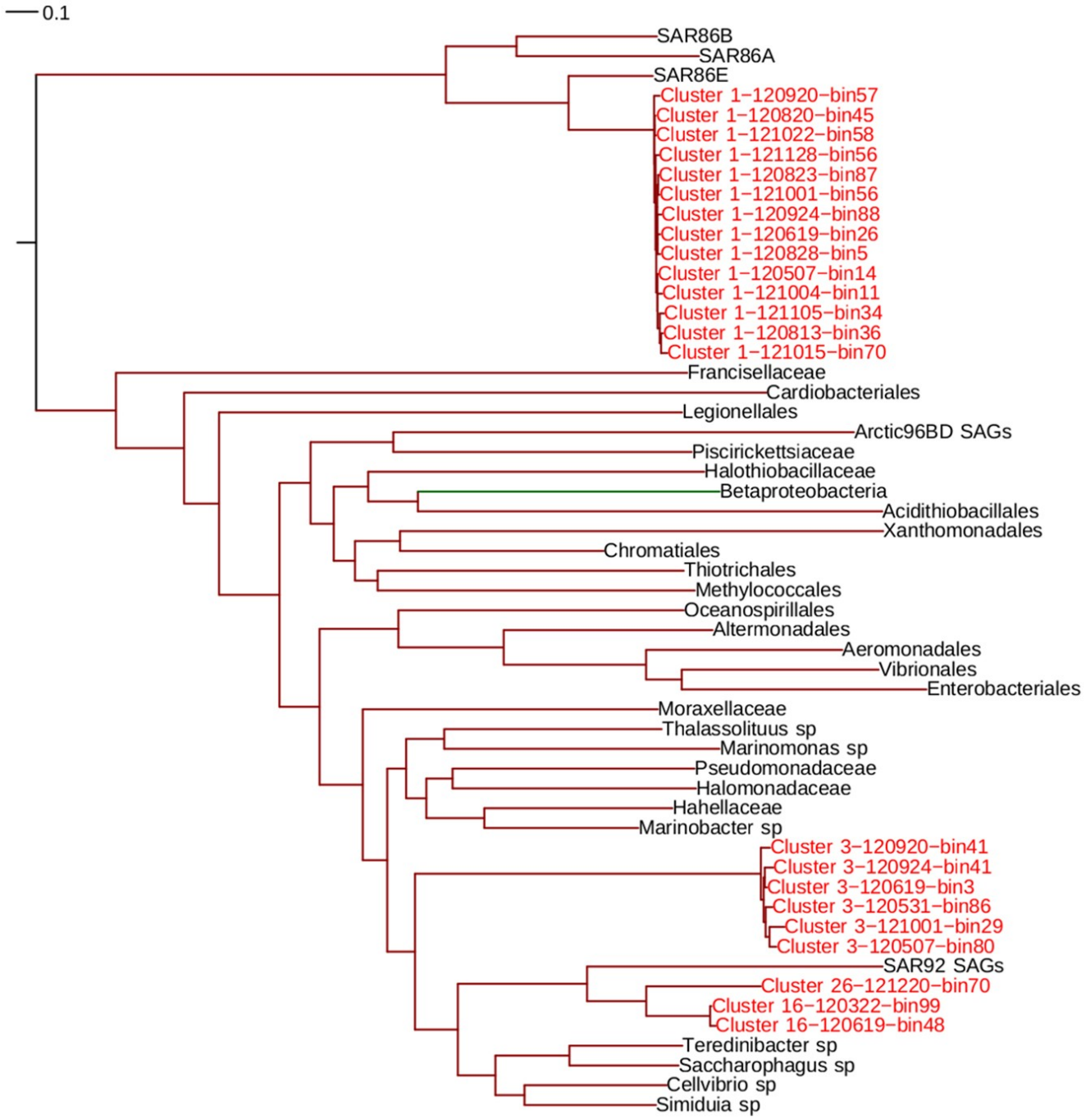
d)


0.1
Hydrogenophilales
Neisseriales
Rhodocyclaceae
Burkholderiales
Nitrosomonadales
Methylobacillus sp
Methylovorus sp
Methylotenera sp
Methylophilales sp
Cluster 14−120920−bin5
Cluster 14−120910−bin4

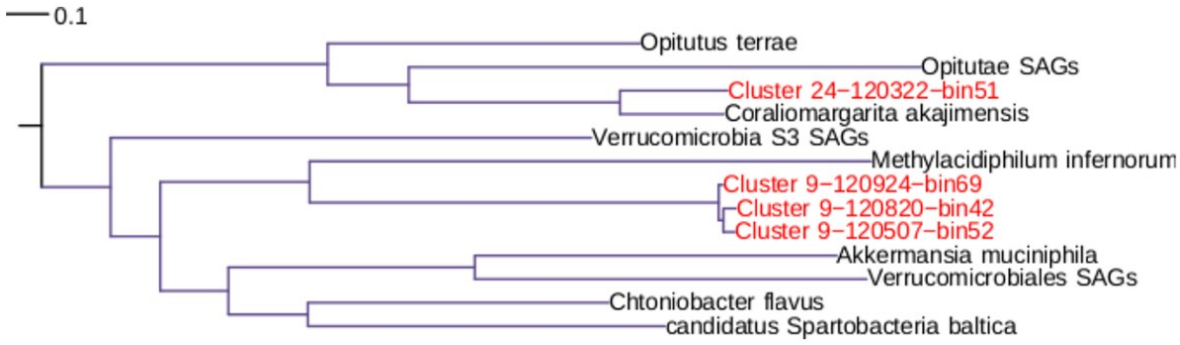
e)


0.1
Gloeobacter sp
Pseudanabaena sp
Synechococcus elongatus
Cyanobium gracile
Cluster 30−120619−bin2
Prochlorococcus sp
Synechococcus sp
Acaryochloridaceae
Thermosynechococcus sp
Cyanothece sp
Geitlerinema sp
Oscillatoriales
Chroococcales
Nostocales
Chamaesiphonaceae
Dactylococcopsis sp
Halothece sp
Leptolyngbya sp
Cyanobacterium sp
Dermocarpellaceae
Cyanothece sp
Merismopediaceae
Microcystaceae

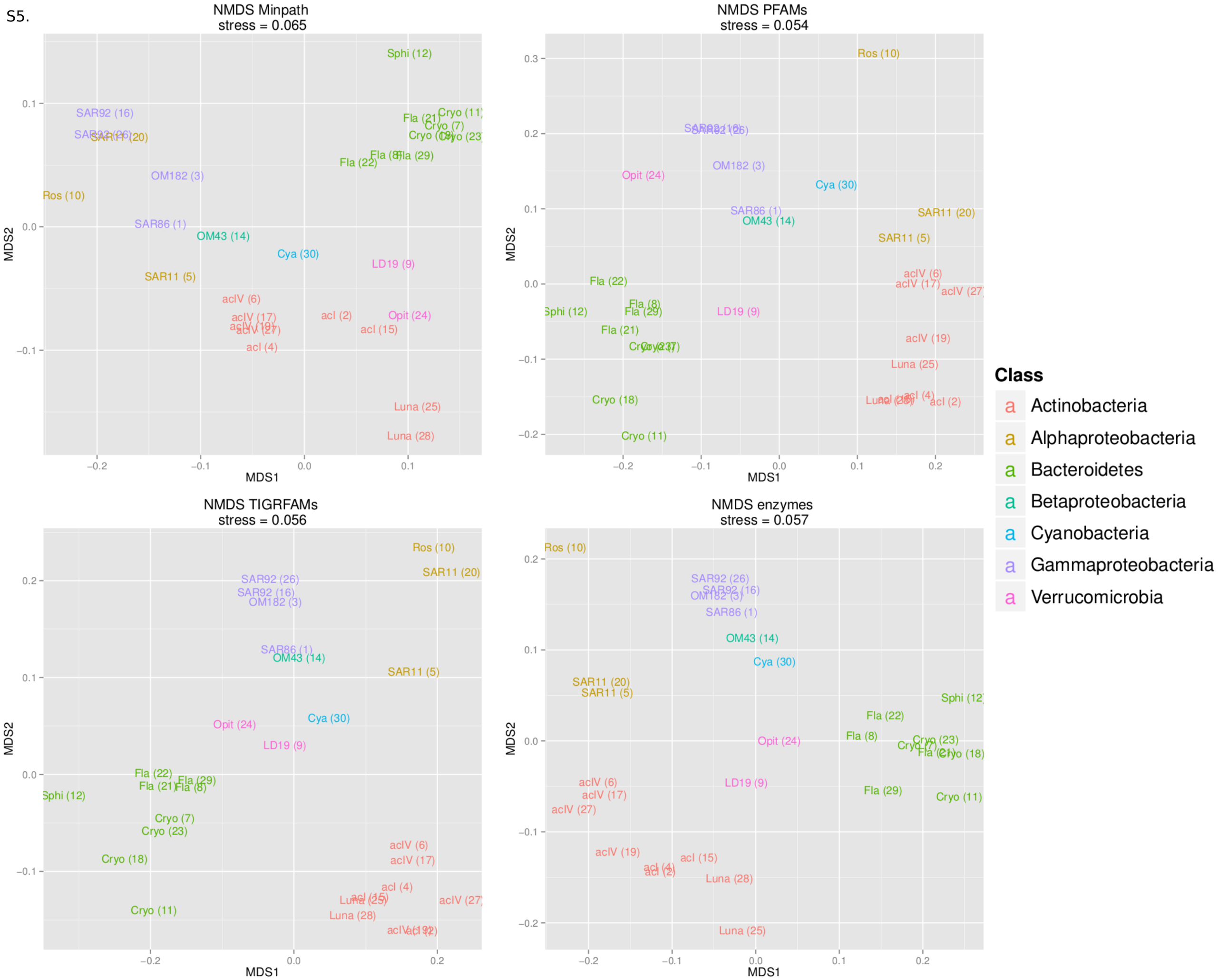
f)


0.1
SAR86B
SAR86A
SAR86E
Cluster 1−120920−bin57
Cluster 1−120820−bin45
Cluster 1−121022−bin58
Cluster 1−121128−bin56
Cluster 1−120823−bin87
Cluster 1−121001−bin56
Cluster 1−120924−bin88
Cluster 1−120619−bin26
Cluster 1−120828−bin5
Cluster 1−120507−bin14
Cluster 1−121004−bin11
Cluster 1−121105−bin34
Cluster 1−120813−bin36
Cluster 1−121015−bin70
Francisellaceae
Cardiobacteriales
Legionellales
Arctic96BD SAGs
Piscirickettsiaceae
Halothiobacillaceae
Betaproteobacteria
Acidithiobacillales
Xanthomonadales
Chromatiales
Thiotrichales
Methylococcales
Oceanospirillales
Altermonadales
Aeromonadales
Vibrionales
Enterobacteriales
Moraxellaceae
Thalassolituus sp
Marinomonas sp
Pseudomonadaceae
Halomonadaceae
Hahellaceae
Marinobacter sp
Cluster 3−120920−bin41
Cluster 3−120924−bin41
Cluster 3−120619−bin3
Cluster 3−120531−bin86
Cluster 3−121001−bin29
Cluster 3−120507−bin80
SAR92 SAGs
Cluster 26−121220−bin70
Cluster 16−120322−bin99
Cluster 16−120619−bin48
Teredinibacter sp
Saccharophagus sp
Cellvibrio sp
Simiduia sp

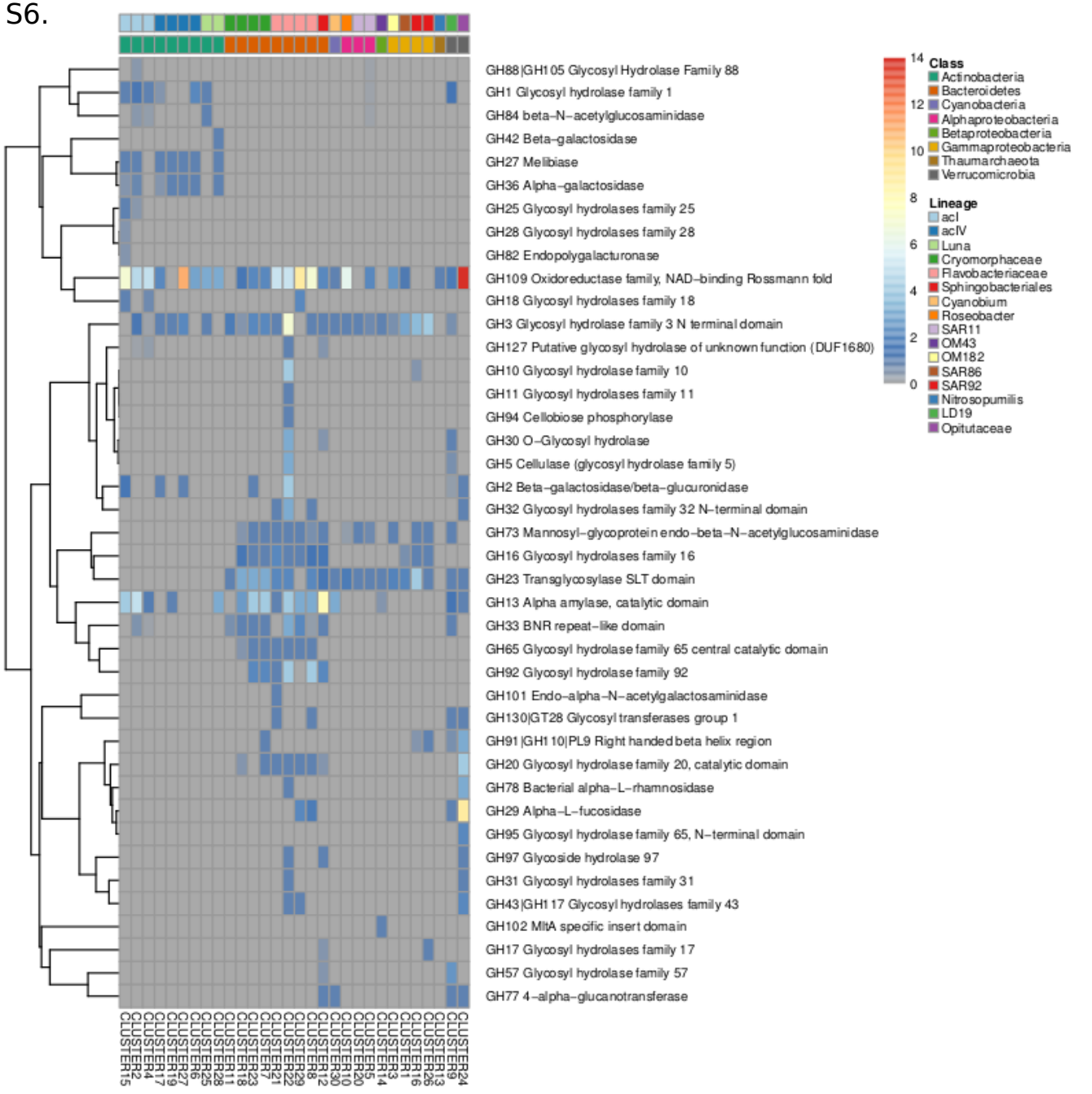
g)


0.1
Opitutus terrae
Opitutae SAGs
Cluster 24−120322−bin51
Coraliomargarita akajimensis
Verrucomicrobia S3 SAGs
Methylacidiphilum infernorum
Cluster 9−120924−bin69
Cluster 9−120820−bin42
Cluster 9−120507−bin52
Akkermansia muciniphila
Verrucomicrobiales SAGs
Chtoniobacter flavus
candidatus Spartobacteria baltica

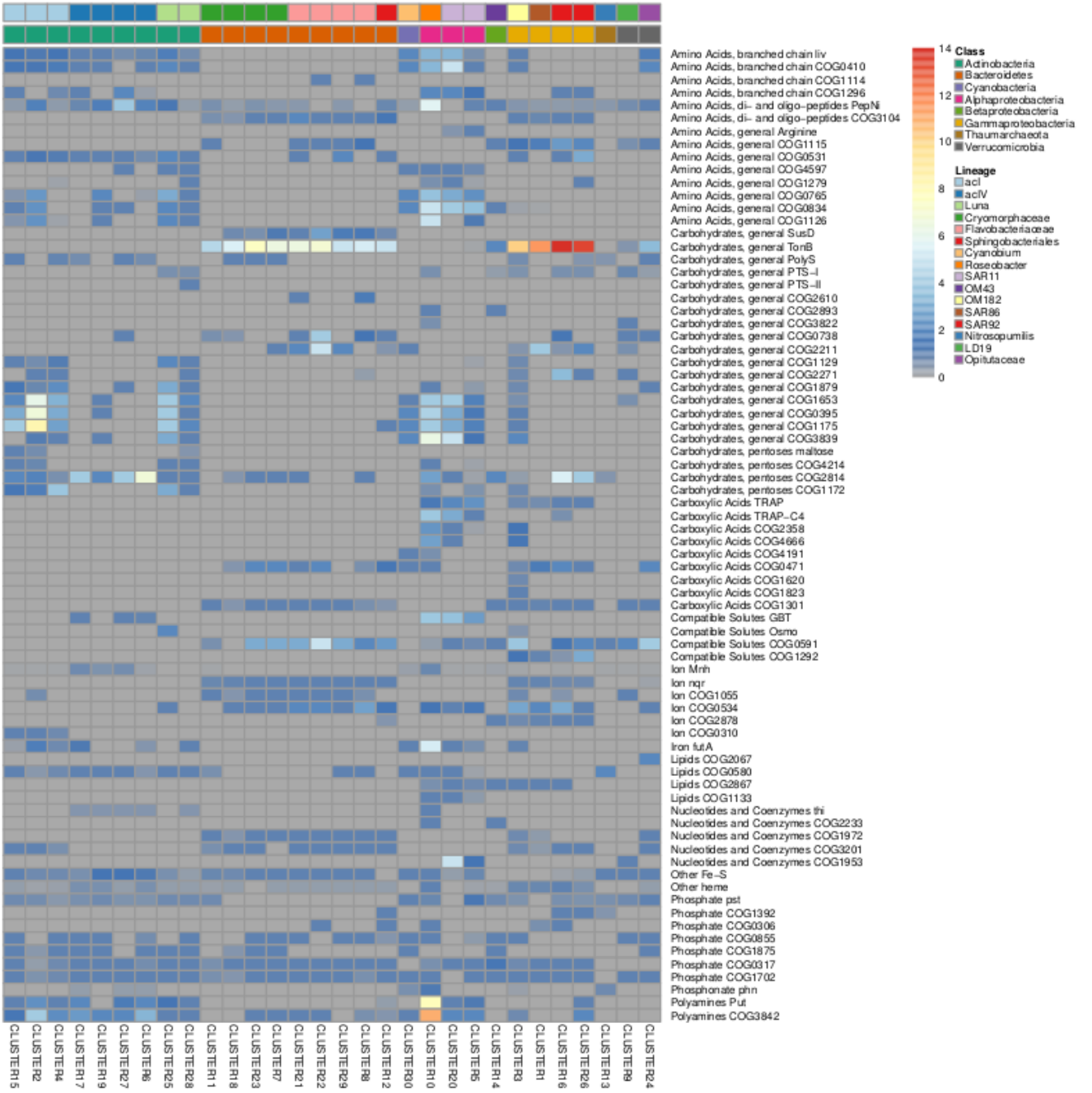
S5.


NMDS Minpath
stress = 0.065
NMDS PFAMs
stress = 0.054
NMDS TIGRFAMs
stress = 0.056
NMDS enzymes
stress = 0.057
MDS1
MDS2
Class
Actinobacteria
Alphaproteobacteria
Bacteroidetes
Betaproteobacteria
Cyanobacteria
Gammaproteobacteria
Verrucomicrobia

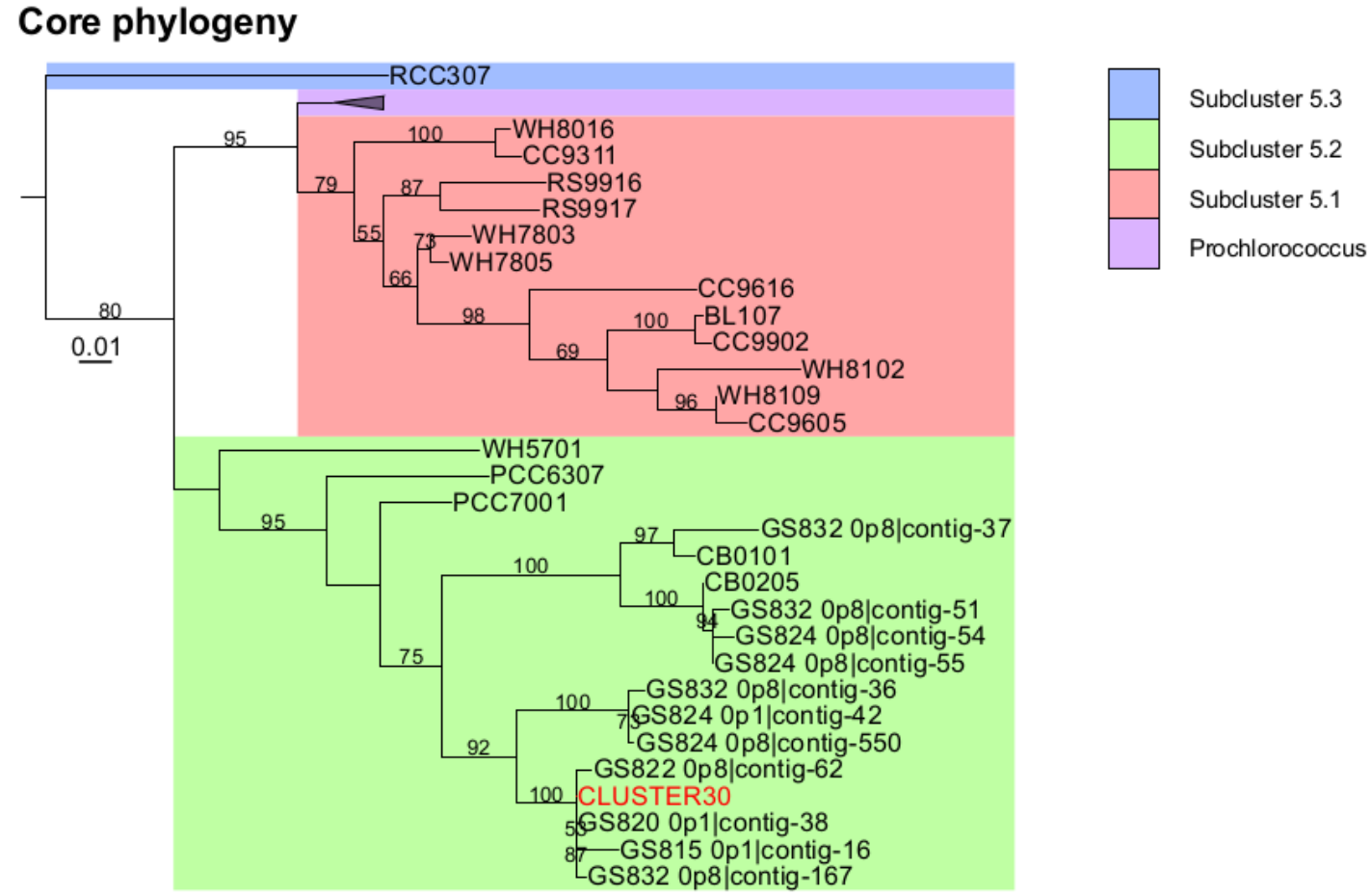
S6.

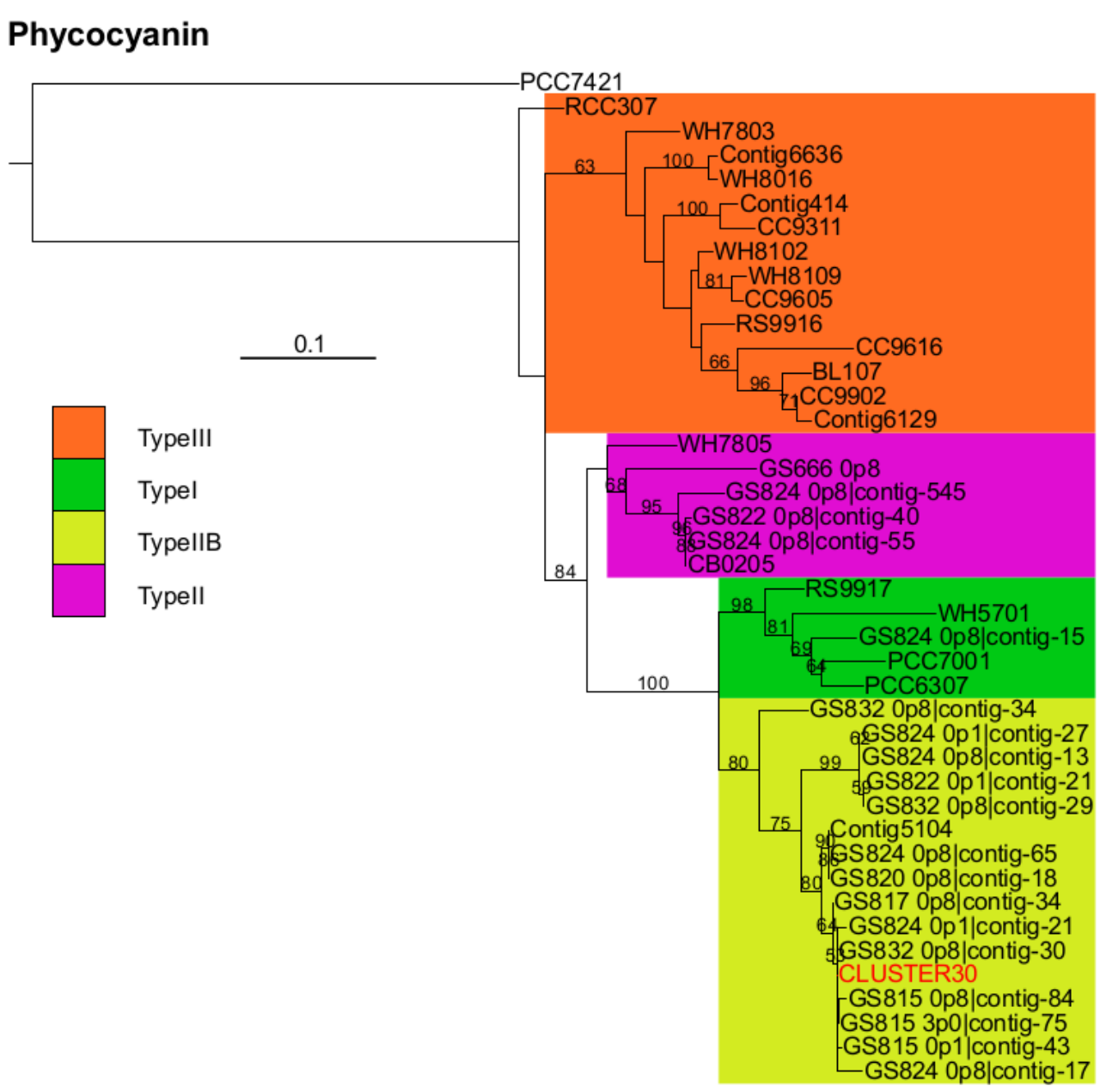

GH88|GH105 Glycosyl Hydrolase Family 88
GH1 Glycosyl hydrolase family 1
GH84 beta-N-acetylglucosaminidase
GH42 Beta-galactosidase
GH27 Melibiase
GH36 Alpha-galactosidase
GH25 Glycosyl hydrolases family 25
GH28 Glycosyl hydrolases family 28
GH82 Endopolygalacturonase
GH109 Oxidoreductase family, NAD-binding Rossmann fold
GH18 Glycosyl hydrolases family 18
GH3 Glycosyl hydrolase family 3 N terminal domain
GH127 Putative glycosyl hydrolase of unknown function (DUF1680)
GH10 Glycosyl hydrolase family 10
GH11 Glycosyl hydrolases family 11
GH94 Cellobiose phosphorylase
GH30 O-Glycosyl hydrolase
GH5 Cellulase (glycosyl hydrolase family 5)
GH2 Beta-galactosidase/beta-glucuronidase
GH32 Glycosyl hydrolases family 32 N-terminal domain
GH73 Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase
GH16 Glycosyl hydrolases family 16
GH23 Transglycosylase SLT domain
GH13 Alpha amylase, catalytic domain
GH33 BNR repeat-like domain
GH65 Glycosyl hydrolase family 65 central catalytic domain
GH92 Glycosyl hydrolase family 92
GH101 Endo-alpha-N-acetylgalactosaminidase
GH130|GT28 Glycosyl transferases group 1
GH91|GH110|PL9 Right handed beta helix region
GH20 Glycosyl hydrolase family 20, catalytic domain
GH78 Bacterial alpha-L-rhamnosidase
GH29 Alpha-L-fucosidase
GH95 Glycosyl hydrolase family 65, N-terminal domain
GH97 Glycoside hydrolase 97
GH31 Glycosyl hydrolases family 31
GH43|GH117 Glycosyl hydrolases family 43
GH102 MltA specific insert domain
GH17 Glycosyl hydrolases family 17
GH57 Glycosyl hydrolase family 57
GH77 4-alpha-glucanotransferase
CLUSTER15
CLUSTER2
CLUSTER4
CLUSTER17
CLUSTER19
CLUSTER27
CLUSTER6
CLUSTER25
CLUSTER28
CLUSTER11
CLUSTER18
CLUSTER23
CLUSTER7
CLUSTER21
CLUSTER22
CLUSTER29
CLUSTER8
CLUSTER12
CLUSTER30
CLUSTER10
CLUSTER20
CLUSTER5
CLUSTER14
CLUSTER3
CLUSTER1
CLUSTER16
CLUSTER26
CLUSTER13
CLUSTER9
CLUSTER24
14
12
10
8
6
4
2
0
Class
Actinobacteria
Bacteroidetes
Cyanobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Thaumarchaeota
Verrucomicrobia
Lineage
acI
acIV
Luna
Cryomorphaceae
Flavobacteriaceae
Sphingobacteriales
Cyanobium
Roseobacter
SAR11
OM43
OM182
SAR86
SAR92
Nitrosopumilis
LD19
Opitutaceae

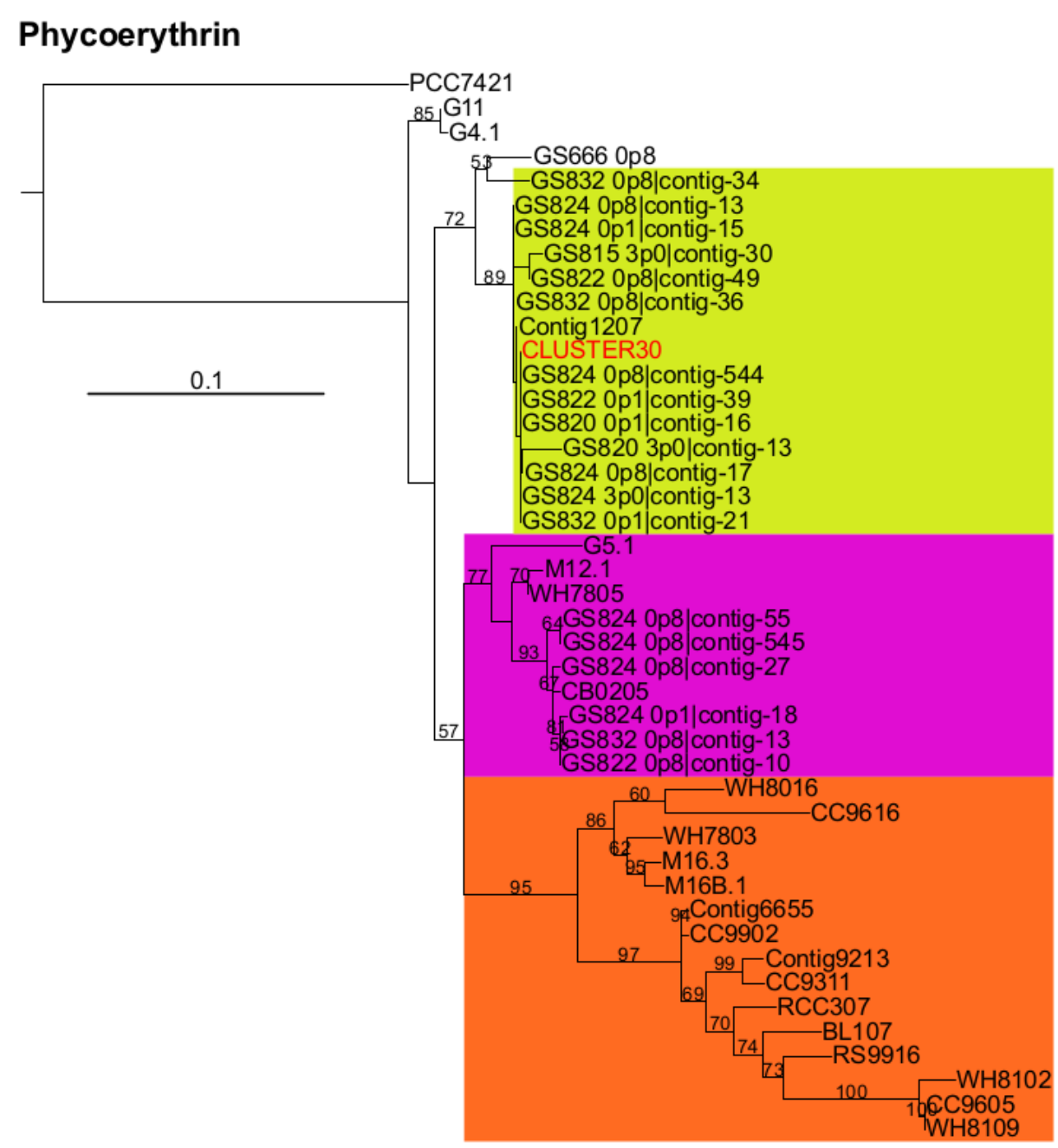
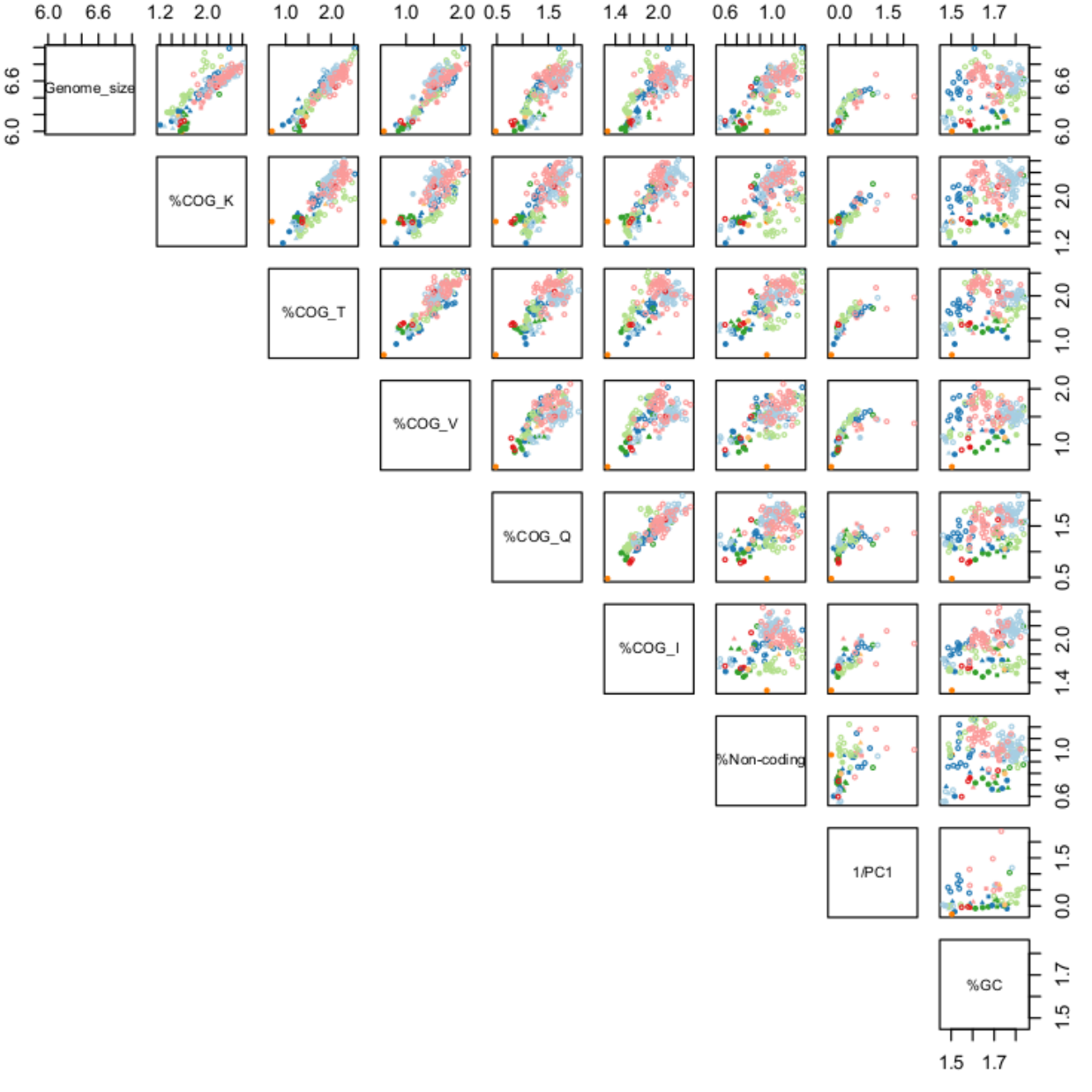
S7.


Amino Acids, branched chain liv
Amino Acids, branched chain COG0410
Amino Acids, branched chain COG1114
Amino Acids, branched chain COG1296
Amino Acids, di- and oligo-peptides PepNi
Amino Acids, di- and oligo-peptides COG3104
Amino Acids, general Arginine
Amino Acids, general COG1115
Amino Acids, general COG0531
Amino Acids, general COG4597
Amino Acids, general COG1279
Amino Acids, general COG0765
Amino Acids, general COG0834
Amino Acids, general COG1126
Carbohydrates, general SusD
Carbohydrates, general TonB
Carbohydrates, general PolyS
Carbohydrates, general PTS-I
Carbohydrates, general PTS-II
Carbohydrates, general COG2610
Carbohydrates, general COG2893
Carbohydrates, general COG3822
Carbohydrates, general COG0738
Carbohydrates, general COG2211
Carbohydrates, general COG1129
Carbohydrates, general COG2271
Carbohydrates, general COG1879
Carbohydrates, general COG1653
Carbohydrates, general COG0395
Carbohydrates, general COG1175
Carbohydrates, general COG3839
Carbohydrates, pentoses maltose
Carbohydrates, pentoses COG4214
Carbohydrates, pentoses COG2814
Carbohydrates, pentoses COG1172
Carboxylic Acids TRAP
Carboxylic Acids TRAP-C4
Carboxylic Acids COG2358
Carboxylic Acids COG4666
Carboxylic Acids COG4191
Carboxylic Acids COG0471
Carboxylic Acids COG1620
Carboxylic Acids COG1823
Carboxylic Acids COG1301
Compatible Solutes GBT
Compatible Solutes Osmo
Compatible Solutes COG0591
Compatible Solutes COG1292
Ion Mnh
Ion nqr
Ion COG1055
Ion COG0534
Ion COG2878
Ion COG0310
Iron futA
Lipids COG2067
Lipids COG0580
Lipids COG2867
Lipids COG1133
Nucleotides and Coenzymes thi
Nucleotides and Coenzymes COG2233
Nucleotides and Coenzymes COG1972
Nucleotides and Coenzymes COG3201
Nucleotides and Coenzymes COG1953
Other Fe-S
Other heme
Phosphate pst
Phosphate COG1392
Phosphate COG0306
Phosphate COG0855
Phosphate COG1875
Phosphate COG0317
Phosphate COG1702
Phosphonate phn
Polyamines Put
Polyamines COG3842
CLUSTER15
CLUSTER2
CLUSTER4
CLUSTER17
CLUSTER19
CLUSTER27
CLUSTER6
CLUSTER25
CLUSTER28
CLUSTER11
CLUSTER18
CLUSTER23
CLUSTER7
CLUSTER21
CLUSTER22
CLUSTER29
CLUSTER8
CLUSTER12
CLUSTER30
CLUSTER10
CLUSTER20
CLUSTER5
CLUSTER14
CLUSTER3
CLUSTER1
CLUSTER16
CLUSTER26
CLUSTER13
CLUSTER9
CLUSTER24
14
12
10
8
6
4
2
0
Class
Actinobacteria
Bacteroidetes
Cyanobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Thaumarchaeota
Verrucomicrobia
Lineage
acI
acIV
Luna
Cryomorphaceae
Flavobacteriaceae
Sphingobacteriales
Cyanobium
Roseobacter
SAR11
OM43
OM182
SAR86
SAR92
Nitrosopumilis
LD19
Opitutaceae

S8.


Core phylogeny
RCC307
WH8016
CC9311
RS9916
RS9917
WH7803
WH7805
CC9616
BL107
CC9902
WH8102
WH8109
CC9605
WH5701
PCC6307
PCC7001
GS832 0p8|contig-37
CB0101
CB0205
GS832 0p8|contig-51
GS824 0p8|contig-54
GS824 0p8|contig-55
GS832 0p8|contig-36
GS824 0p1|contig-42
GS824 0p8|contig-550
GS822 0p8|contig-62
CLUSTER30
GS820 0p1|contig-38
GS815 0p1|contig-16
GS832 0p8|contig-167
0.01
Subcluster 5.3
Subcluster 5.2
Subcluster 5.1
Prochlorococcus


Phycocyanin
PCC7421
RCC307
WH7803
Contig6636
WH8016
Contig414
CC9311
WH8102
WH8109
CC9605
RS9916
CC9616
BL107
CC9902
Contig6129
WH7805
GS666 0p8
GS824 0p8|contig-545
GS822 0p8|contig-40
GS824 0p8|contig-55
CB0205
RS9917
WH5701
GS824 0p8|contig-15
PCC7001
PCC6307
GS832 0p8|contig-34
GS824 0p1|contig-27
GS824 0p8|contig-13
GS822 0p1|contig-21
GS832 0p8|contig-29
Contig5104
GS824 0p8|contig-65
GS820 0p8|contig-18
GS817 0p8|contig-34
GS824 0p1|contig-21
GS832 0p8|contig-30
CLUSTER30
GS815 0p8|contig-84
GS815 3p0|contig-75
GS815 0p1|contig-43
GS824 0p8|contig-17
0.1
TypeIII
TypeI
TypeIIB
TypeII


Phycoerythrin
PCC7421
G11
G4.1
GS666 0p8
GS832 0p8|contig-34
GS824 0p8|contig-13
GS824 0p1|contig-15
GS815 3p0|contig-30
GS822 0p8|contig-49
GS832 0p8|contig-36
Contig1207
CLUSTER30
GS824 0p8|contig-544
GS822 0p1|contig-39
GS820 0p1|contig-16
GS820 3p0|contig-13
GS824 0p8|contig-17
GS824 3p0|contig-13
GS832 0p1|contig-21
G5.1
M12.1
WH7805
GS824 0p8|contig-55
GS824 0p8|contig-545
GS824 0p8|contig-27
CB0205
GS824 0p1|contig-18
GS832 0p8|contig-13
GS822 0p8|contig-10
WH8016
CC9616
WH7803
M16.3
M16B.1
Contig6655
CC9902
Contig9213
CC9311
RCC307
BL107
RS9916
WH8102
CC9605
WH8109
0.1

S9.