Dupont et al.

Detailed sequencing and assembly methods

Single cell sample collection and cell sorting: Seawater was collected at the Scripps Institution of Oceanography (SIO) Pier on April 9th, 2009 at 8:30 am. The water temperature was 13.5°C with a salinity of 33.5 PSU and a chlorophyll *a* fluorescence of 2.7 (SIO Automated Shore station equipment, www.sccoos.org). The sample was filtered (0.8 μm pore size), amended with glycerol (final concentration 15% v/v), flash frozen, and stored at −80°C. Prior to sorting, the sample was thawed and stained with SYBR Green I nucleic acid stain (at 10X, Invitrogen). Single cells were sorted using a FACS Aria II flow cytometer equipped with a 488nm laser and custom forward scatter (FSC)-PMT (BD Biosciences) using detection by the side scatter (SSC-PMT) and green fluorescence (512 nm), with the highest purity setting and the lowest flow rate to avoid sorting of coincident events. Phosphate buffered saline (PBS), sterilized by 0.2 μm filtration, was used as sheath fluid. Single cells were sorted into 384-well plates containing 4 μl of TE (10 mM Tris, 0.1 mM EDTA, pH 8.0) buffer in each well, and stored at −80°C until MDA.

Multiple Displacement Amplification (MDA): Measures taken to avoid introduction of contaminating DNAs during MDA include: preparation of lysis and neutralization reagents and TE from ultrapure TrisCl, EDTA, and H_20 (Ambion); 0.1 μm filtration of reagents; preparation of reagents and master mixes in a dedicated DNA-free UV cabinet. MDA was performed using the GenomiPhi HY kit (GE Healthcare). MDA reactions were prepared using the epMotion 5075 LH Automated Pipetting System (Eppendorf). Cells in 4 μl of TE were lysed by addition of 2 μl of alkaline lysis solution (645 mM KOH, 265 mM DTT, 2.65 mM EDTA pH 8.0) followed by incubation on ice for 10 min. The lysis reaction was neutralized with 7 μl of neutralization solution (2 μl of 1290 mM TrisCl pH 4.5 and 5 μl GenomiPhi Sample Buffer), and 12.5 ul of GenomiPhi master mix was then added (10.8 μl of GenomiPhi Reaction Buffer and 1.2 ul GenomiPhi Enzyme Mix) for a total reaction volume of 25 μl. After a brief centrifugation, the reactions were incubated at 30°C for 16 h followed by a 10 min incubation at 65°C in a GeneAmp PCR System 9700 (Applied Biosystems). PicoGreen DNA quantification of MDA yield was performed as on a diluted fraction using the Quant-iT dsDNA Assay Kit (Invitrogen) as per manufacturer's protocol with detection on a Flexstation3 fluorometer (Molecular Dynamics).

Further amplification of single cell MDAs for sequencing was performed in 8 replicates with 650-750 ng of the original MDA as template. Reactions were as above except that the template DNA was denatured by addition of 1 μl of alkaline lysis solution (1075 mM KOH, 265 mM DTT, 26.5 mM EDTA pH 8.0) followed by incubation on ice for 10 min. One μl of neutralization solution (2150 mM TrisCl pH 4.5), and 19 μl of GenomiPhi master mix was then added (11.25 μl of GenomiPhi Reaction Buffer, 6.5 μl GenomiPhi Sample Buffer, and 1.25 μl GenomiPhi Enzyme Mix) for a total reaction volume of 25 μl. MDA products for each template were pooled and yield determined by PicoGreen assay.

16S PCR analysis: Bacterial 16S rDNA PCR was performed using a 1:20 dilution of the single cell MDA products as template and universal bacterial primers 27f and 1492r (Lane, 1991). Cycling conditions were as follows: 94°C for 3 min, 35 cycles of 94°C for 30 sec, 55°C for 30 sec, 72°C for 90 sec, with a final extension at 72°C for 10 min. 16S PCR products were treated with Exo I and shrimp alkaline phosphatase (both from Fermentas) to remove excess primers and nucleotides as follows: 10 µl of PCR product, 0.5 ul Exo I, 1 µl SAP and 3.5 µl nuclease-free water were combined and incubated for 37°C for 20 min, then 80°C for 15 min. ExoI/SAP treated products were sequenced by direct cycle sequencing at the Joint Technology Center (JTC, J. Craig Venter Institute, Rockville, MD). Forward and reverse reads were trimmed using default settings (CLC bio, Muehltal, Germany), and 16S sequences were identified and classified taxonomically by BLAST against the NCBI nr database, and the Ribosomal Database Project (RDP) Classifier (Cole *et al.*, 2009; Wang *et al.*, 2007), respectively. SAR86 taxonomy was confirmed by clustering and BLAST analysis using a curated 16S database, and two MDAs with 16S sequences with >97% identity to SAR86 were selected for 454 sequencing.

Preparation of 454 sequencing libraries: 454-FLX Titanium 3 kb paired end libraries were generated and sequenced at the JTC using 10 µg of pooled reamplified MDA as template. MDA DNA was hydrosheared (HydroShear, Genomic Solutions) and fragments of approximately 3 kb were size selected using agarose gel electrophoresis and gel purification. Library construction, emPCR, enrichment and 454 sequencing were performed following the vendor's (Roche) standard operating procedures with some modifications. qPCR was used to estimate the number of molecules needed for emPCR. Automation (BioMek FX) was used to "break" the emulsions after emPCR and butanol was added to facilitate sample handling during the breaking process. The REM e (Robotic Enrichment Module) from Roche was used to automate the bead enrichment process. Two single cell MDA reactions were barcoded and sequenced on the same plate, generating 700,000+ reads of >200 bp for each genome.

Assembly of SAR86 single cell genomes: Data derived from single cells should confirm the gene content found in the metagenomic assemblies and could potentially provide information about the variable and hyper-variables segments associated with microbial genomes that cannot be easily acquired from metagenomic assemblies. Assembly of the single cells was carried out using the Celera Assembler (4) available at http://sourceforge.net/apps/mediawiki/wgs-assembler with error cutoffs set to 0.05, a word size of 14, and the bog unitiger. Assemblies will be deposited to the NCBI genome project database prior to publication; individual reads will be deposited in the short read archive.

Assembly, binning, and validation of SAR86 metagenomes: A global assembly of the GOS dataset was performed with the Celera Assembler. Metagenomic assemblies were executed aggressively by using default parameters, except error cutoffs were increased to 15% and the utgGenomeSize set to 150,000 bps as described previously (5). Some of these scaffolds were highly similar in gene organization (synteny) with previously sequenced SAR86 BAC sequences and a 60-70% nucleotide identity implied that these scaffolds were closely related to the SAR86 lineage. In an effort to isolate more of the

SAR86 genomes, we constructed proportional sample distribution of reads recruited (see fragment recruitment described below) above 90% identity for all the GOS scaffolds over 5kbp in length. This distribution was used to calculate a distance matrix (using the R statistical package) between all the GOS scaffolds. Manual inspection of scaffolds closest to the already identified SAR86B scaffolds produced a set of 36 scaffolds that had a consistent distribution with the scaffolds that were syntenic with known SAR86 BACs. The total span of these scaffolds was under 2Mbp which was consistent with the sizes of other known planktonic microbes. In an effort to find more complete assemblies and order and orient these smaller putative SAR86 scaffolds we searched for other GOS scaffolds that were syntenic to these putative SAR86 scaffolds. While many scaffolds were identified, two much larger scaffolds (806kb and 443kb) were syntenic with the previously identified putative SAR86 scaffolds and known SAR86 BACs while sharing a similar sample distribution and coverage characteristics with each other. We were unable to discover any other scaffolds that had similar coverage and sample distribution with this second set of putative SAR86 scaffolds suggesting that any missing sequence is either small or associated with variable regions that would not be identified using these approaches. The two long SAR86 scaffolds and 36 smaller contigs were used to largely order and orient each other, producing a meta-scaffold that we designated SAR86A. A separate set of SAR86-like scaffolds were identified by their unique GOS sample distribution and binned to produce the SAR86B genome. Reads recruited to the SAR86B scaffolds are derived almost from exclusively GS-100 and GS-102 and to a lesser extent from GS-148 and GS-149 (Fig. 1C).

Metagenomic assemblies are not equivalent to complete or even draft genomes produced from a clonal population, there is no demonstrably "correct" assembly possible from a metagenomic dataset. Indeed, re-assembly with the Celera Assembler or using other assemblers using either the entire GOS dataset or subsets associated with specific SAR86 populations produces subtly different assemblies in which synteny is largely maintained but the scaffold lengths vary significantly (data not shown). The assemblies that were used for the publication were chosen because they possessed the longest scaffolds/contigs. While "correctness" cannot be demonstrated explicitly, a number of different factors were examined to demonstrate that the assemblies are broadly representative of a SAR86 genome population and are sufficient to draw reasonable conclusions about the genomic and physiological capabilities of this population. The evidence that suggests that the assemblies are reasonable substrates for biological inference include:

1) Gene organization, or synteny, determined using promer is conserved between both metagenomic SAR86 assemblies, the single cell assemblies, and the previously isolated BAC clones demonstrating that the assemblies are largely self-consistent and correspond with independently derived data from clonal sources (Fig. S1).

2) Recruitment to the assembled scaffolds shows a consistent sample distribution and depth of coverage over the length of the genome (excluding the highly conserved large and small ribosomal rRNA operon) consistent with recruitment patterns seen for other planktonic marine microbial genomes derived from cultivated isolates (Fig. 2).

3) Clonal analysis of the mated reads recruited to the SAR86A assembly above 90% identity show that there are large numbers of well mated reads (in proper orientation and the correct distance apart) across the length of the genome except across the break between the two scaffold (Fig. 2). The coherence of hundreds of thousands of mate pairs with the proposed indicates that these individual scaffolds are well largely consistent with SAR86 genome organization in diverse oceanic locations.

4) Analysis of single copy core marker genes indicates that they are indeed single copy in the SAR86A and B assemblies indicating that we are not binning multiple genomes into a single assembly (Table 1).

5) Scatter plots derived from the principal component analysis (PCA) of the oligonucleotide frequencies (k = 4) show that the SAR86 assemblies form tight clusters that are consistent with a clonally derived genome (see ref (4) for technical details, Fig. S2). The PCA analysis was used to identify 5 scaffolds in the SAR86B assembly that did not cluster with the bulk of the data. Outliers from this cluster were further examined by blast to determine their taxonomic affiliation and five that were not consistent with γ-proteobacteria were eliminated to produce the set of scaffolds that we designate as SAR86B (http://gos.jcvi.org/openAccess/scatterPlotViewer2.html select "SAR86.genomes.new" from pulldown). It is worth noting that these eliminated scaffolds are dominated by conserved hypothetical proteins and do not convey any potentially insightful metabolic capabilities. The assemblies and annotation will be deposited in the NCBI genome project collection prior to publication.

Supplemental Figure Captions

Figure S1: Promer alignments between the SAR86A genome assembly and SAR86B, C, and D. Also shown is the alignment between the SAR86A assembly and 3 bacterial artificial chromosomes for SAR86; from top to bottom, EBACHOT4E07 (Sabehi *et al.*, 2004), EBAC31A08 (Beja *et al.*, 2000), and EBAC20E09 (Sabehi *et al.*, 2004).

Figure S2: Three dimensional scatter plots of the 1st three principle components of tetranucleotide usage. In panel A, 10kb overlapping segments of SAR86 genome assemblies A (black), B (blue), C (red), and D (purple) are shown, along with putative SAR86B contaminant contigs that while having very similar sample abundance profiles were excluded from the final analysis because they do not cluster with the SAR86 genomes (green; see methods for additional details). In panel B, the PCA analysis was repeated with 20kb overlapping segments of SAR86A (black), *Prochlorococcus* AS9601 (red), and *Pelagibacter ubique* HTCC7211 (purple) to show the separation of the three most abundant marine microbes based on oligonucleotide frequencies. While these genomes can be separated with smaller overlapping segments, clean separation relying on oligonucleotide frequencies alone requires larger segments and indicates the value of assembly. The variance for these PCA plots are shown in Panel C and D respectively with the greatest variance found in the first three principal components.

Figure S3: Phylogeny of SAR86. A maximum likelihood phylogeny of 12 concantenated proteins found in nearly all γ-proteobacterial genomes and SAR86A and SAR86B genomes described here is shown.

Figures S4: Amino acid biosynthesis in SAR86 A) The histidine biosynthesis operons in SAR86B and D. Also noted is how much more of the corresponding metagenomic and single cell contig is present upstream and downstream of the locus displayed. The percentages are the number of orfs on these neigboring scaffolds orthologous to another SAR86 protein. B) The arginine biosynthesis operon in SAR86B. C) Abundance of Arg, His, and Met biosynthesis pathways in natural SAR86 populations. Each point is the abundance of the relevant amino acid biosynthesis genes and SAR86 core genes in a single GOS metagenome. The units on the axes are the average number of reciprocal best blast hits to either 107 core genes (x axis) or the genes involved in the synthesis of each amino acid (y-axis), followed by normalization for the number of genes in that category.

Figure S5: A nucleotide alignment of several small non-coding RNAs found in a metatranscriptomic library from the Pacific Ocean (Shi *et al.*, 2009) and a genomic region found in both SAR86A and SAR86B genome assemblies. The genomic context of this small RNA within SAR86 genomes is shown in figure 3.

Figure S6: Number of TonB-dependent outer membrane receptors (TBDR) in microbial genomes relative to total number of open reading frames. Most data is from Blanvillian et al. (Blanvillian *et al.*, 2008), while for the SAR86 genomes the gene models were manually examined to verify the presence of a tonB box in each TBDR.

Figure S7: TonB-dependent receptor containing genomic loci in SAR86. A) The genome structure of a TBDR locus possibly catalyzing sulfolipid uptake and degradation. Also shown is the chemical structure of a sulfolipid with theoretical cleavage sites that are color-coded according to the protein in the locus. B) Similarly for a TBDR locus possibly involved in glycine-betaine polar lipid degradation. While speculative, the specific color-coded cleavage reactions are consistent with the enzymatic activity of the putative proteins found adjacent to the indicated TBDR.

Supplemental Refereneces

Beja O, Aravind L, Koonin EV, Suzuki M, Hadd A, Nguyen LP *et al* (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289:** 1902-1906.

Blanvillian S, Meyer D, Boulanger A, Lautier M, Guynet C, Denance N *et al* (2008). Plant carbohydrate scavenging through Ton-B dependent receptors: A feature shared by phytopathogenic and aquatic bacteria. *PloS One* **2:** e224.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* **37:** D141-145.

Lane DJ (ed.) (1991) *16S/23S rRNA sequencing*. John Wiley: Chichester, UK.

Sabehi G, Beja O, Suzuki MT, Preston CM, DeLong EF (2004). Different SAR86 groups harbour divergent proteorhodopsins. *Environmental Microbiology* **6:** 903-910.

Shi Y, Tyson GW, DeLong EF (2009). Metatranscriptomics reveal unique microbial small RNAs in the ocean's water column. *Nature* **459:** 266-269.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73:** 5261-7.

Figure S1A. SAR86A (x) vs SAR86B (y)

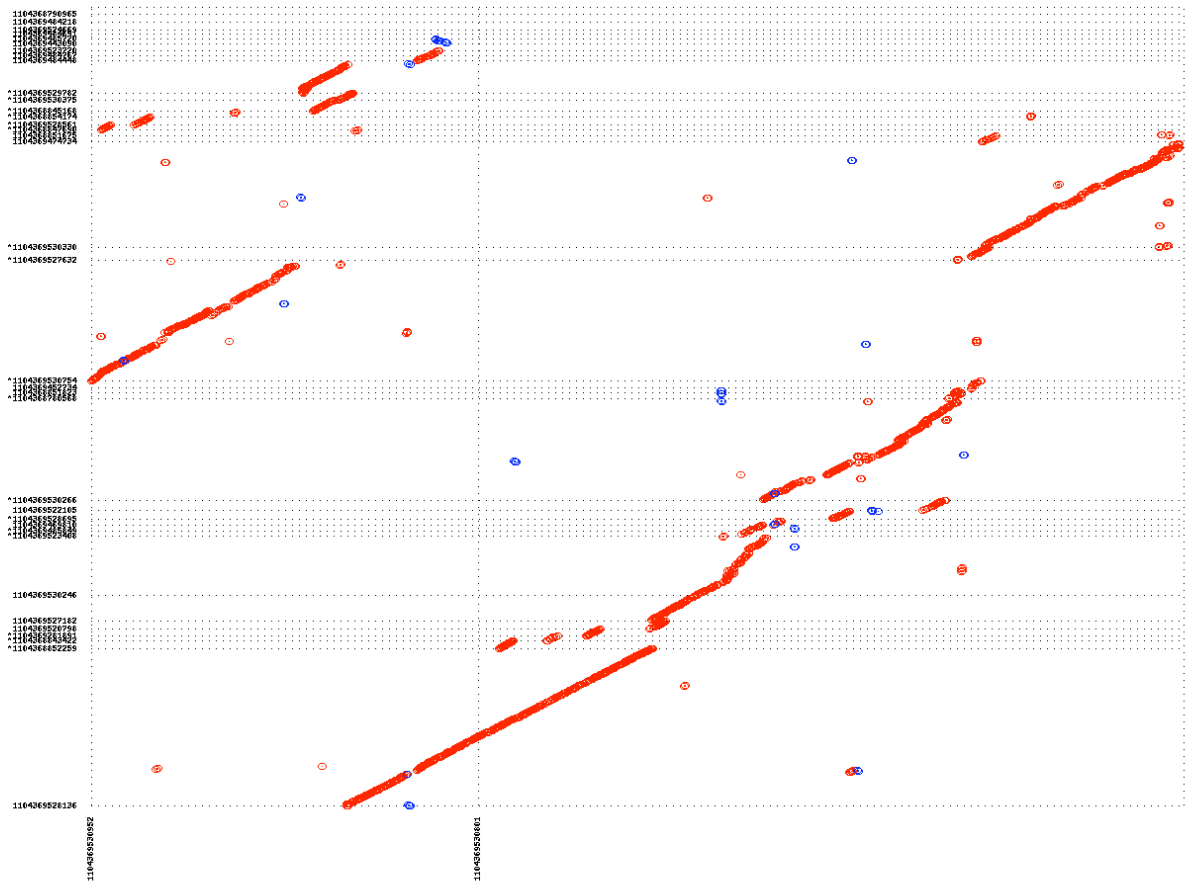


Figure 1B: SAR86A (x) vs SAR86 BACs (y)

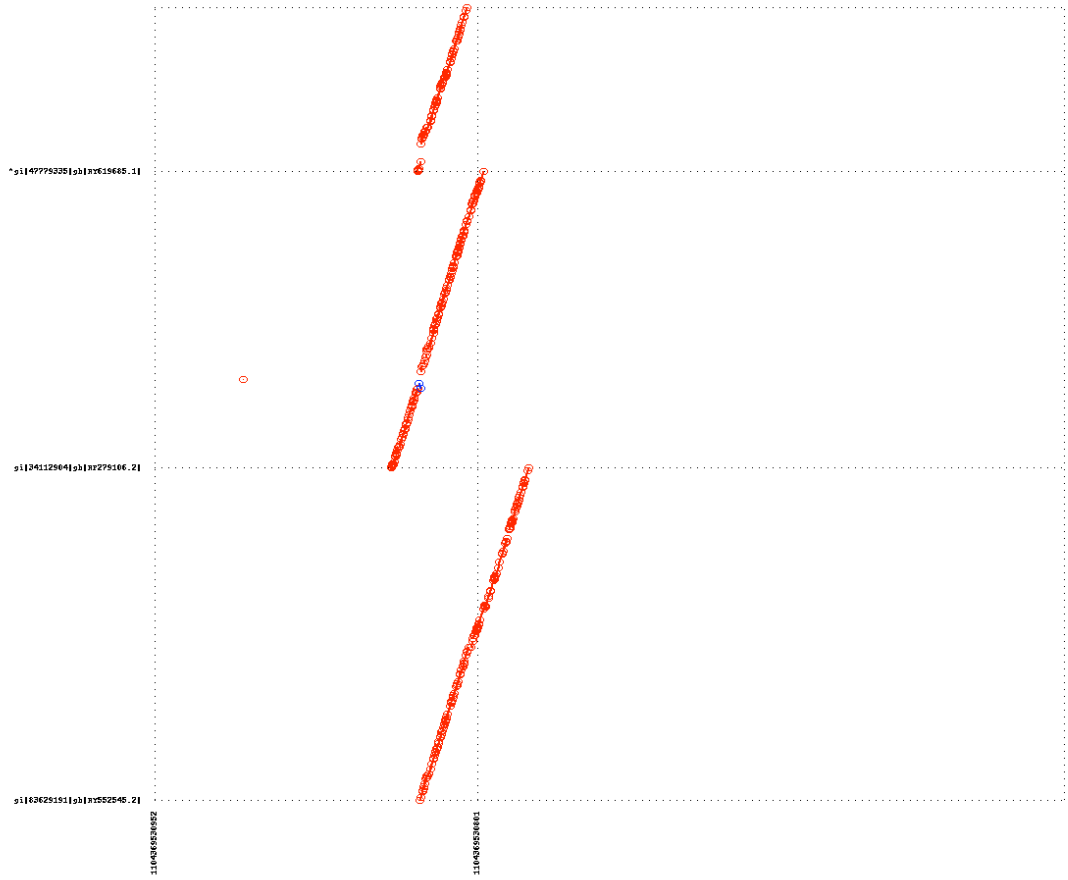


Figure 1C: SAR86A (x) vs SAR86D (y)

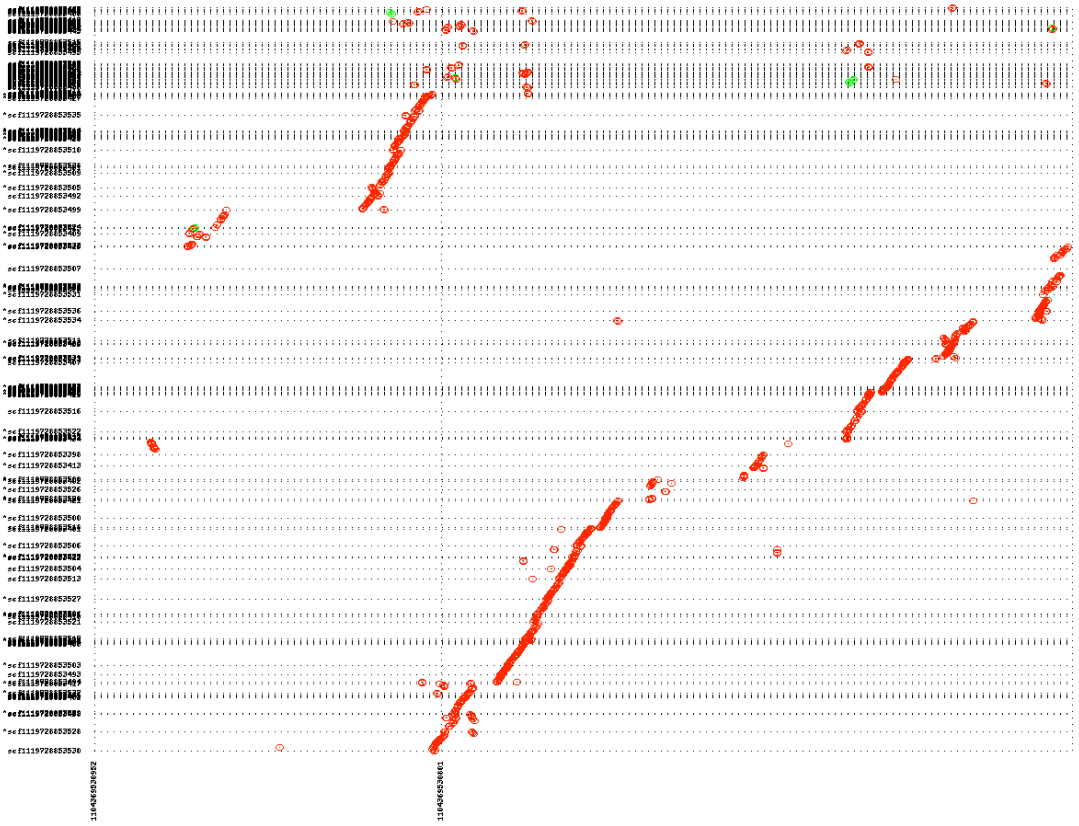


Figure S1D: SAR86A (x) vs SAR86C (y)

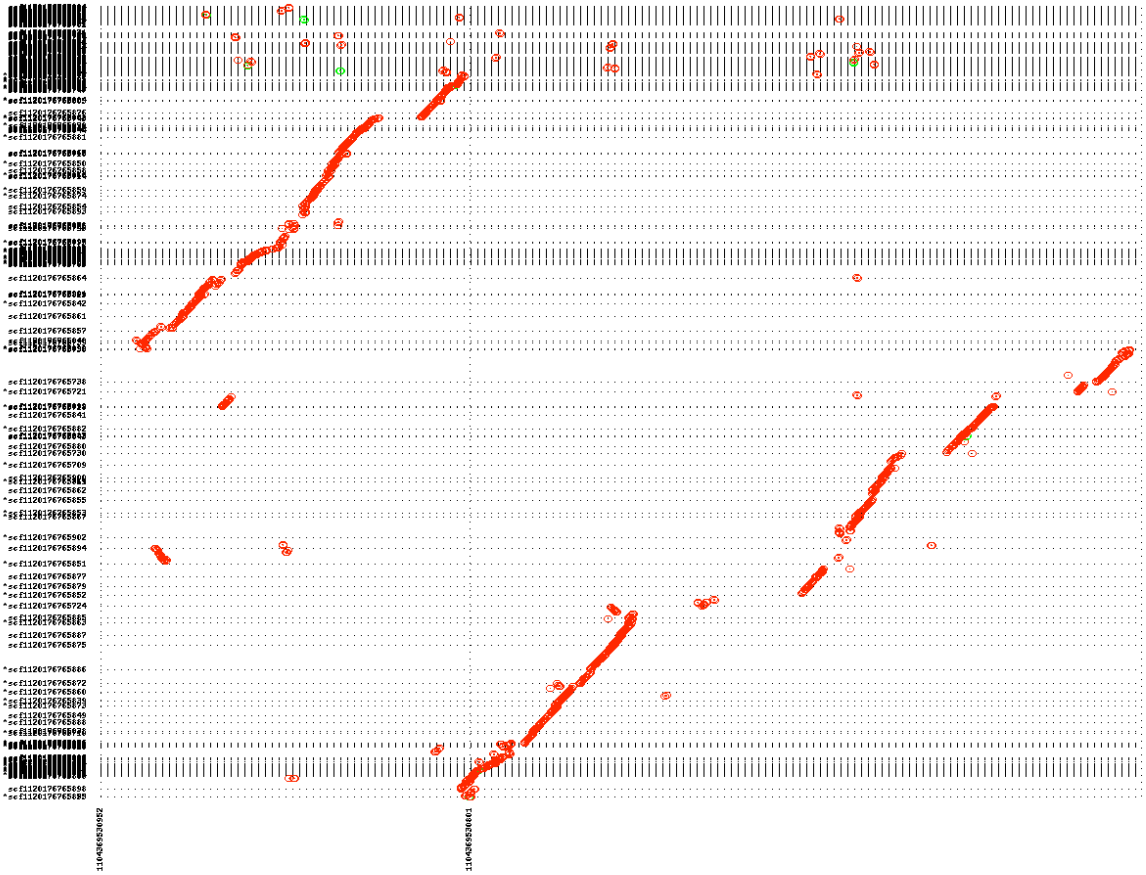


Figure S2

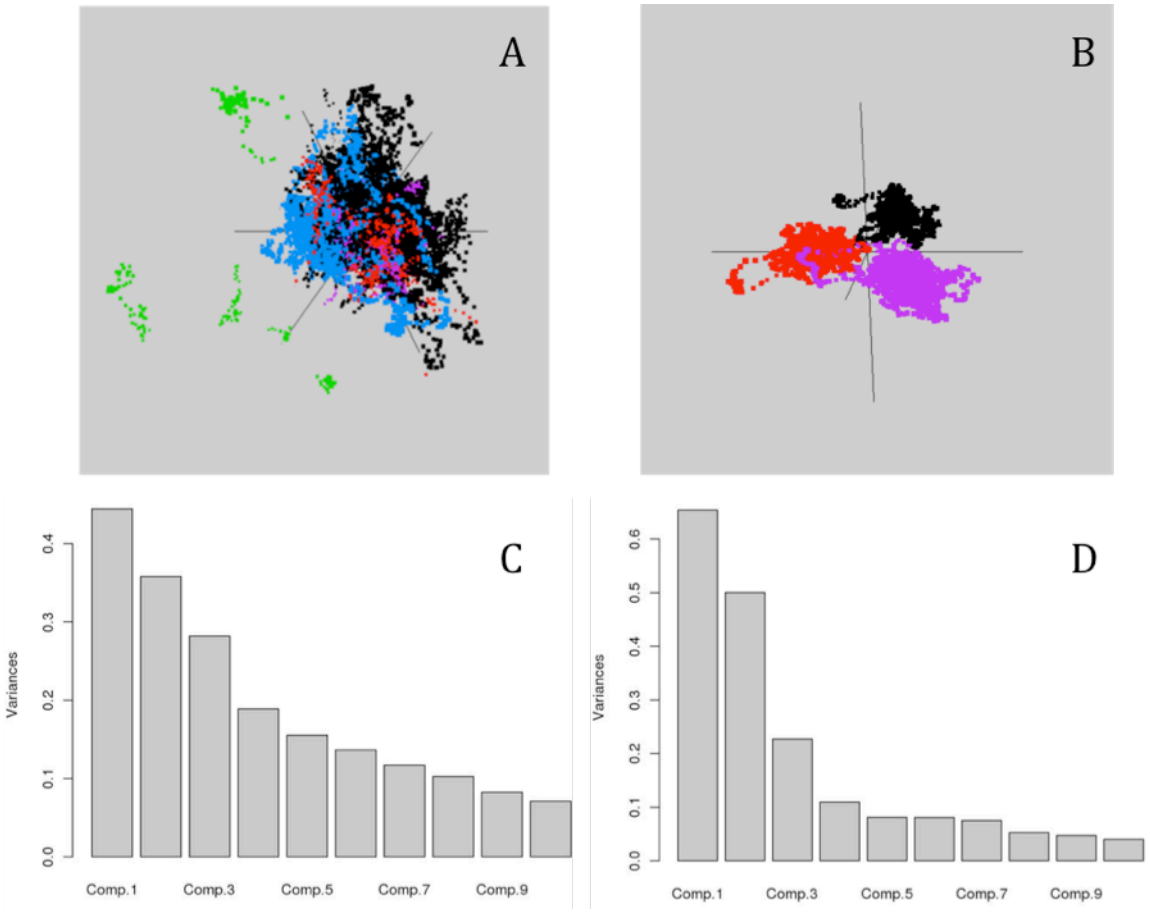

A
B
C
D
Variances
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Comp.1
Comp.3
Comp.5
Comp.7
Comp.9

Figure S3
A. 12 protein phylogeny

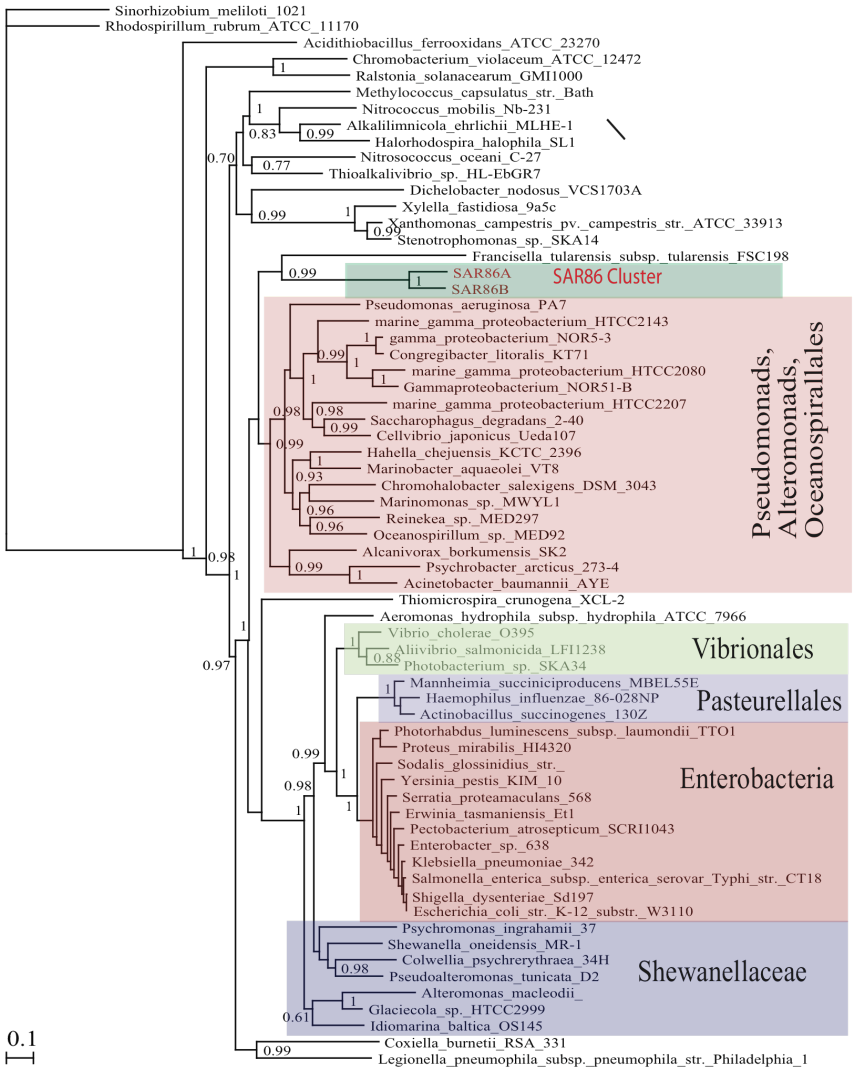
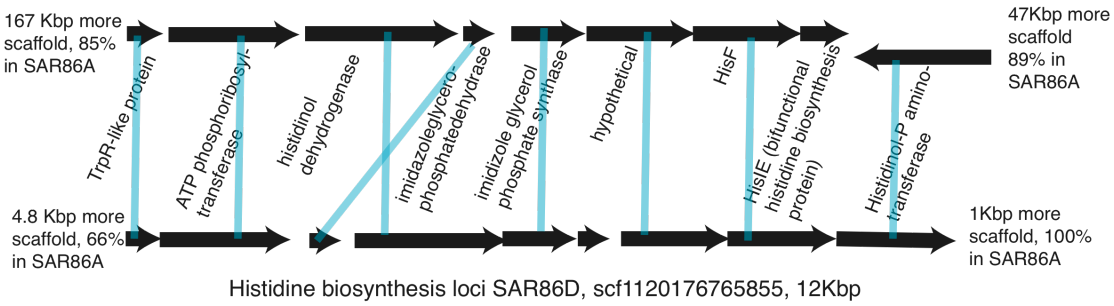


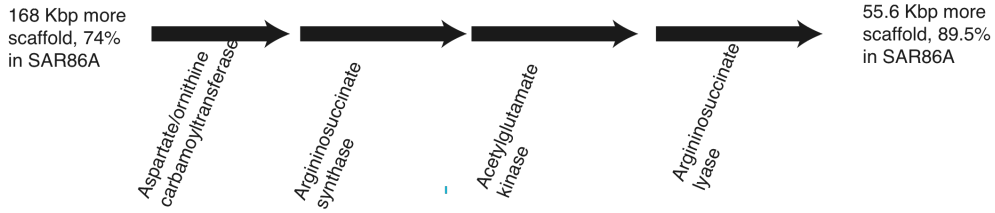
Figure S4

A. Histidine biosynthesis loci SAR86B, Scaf. 1104369530266, 220 Kbp



Histidine biosynthesis loci SAR86D, scf1120176765855, 12Kbp

B. Arginine Biosynthesis loci SAR86B, Scaf. 1104369530330, 230 Kbp



C. Abundance of the SAR86 Met, His, and Arg synthesis loci relative to SAR86 core genes

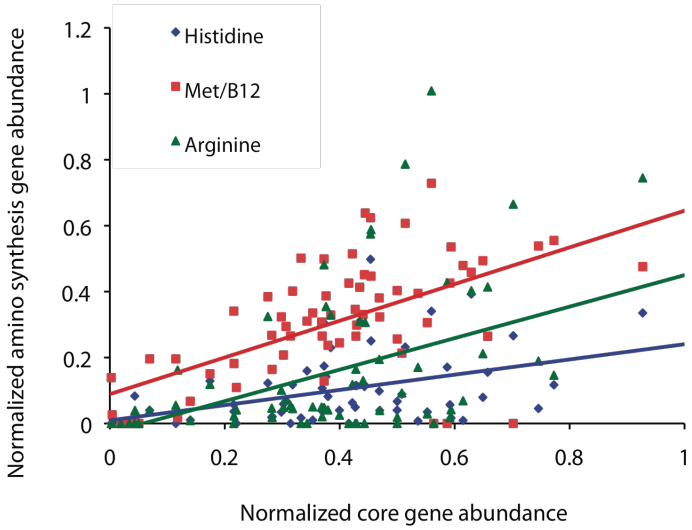


Figure S5

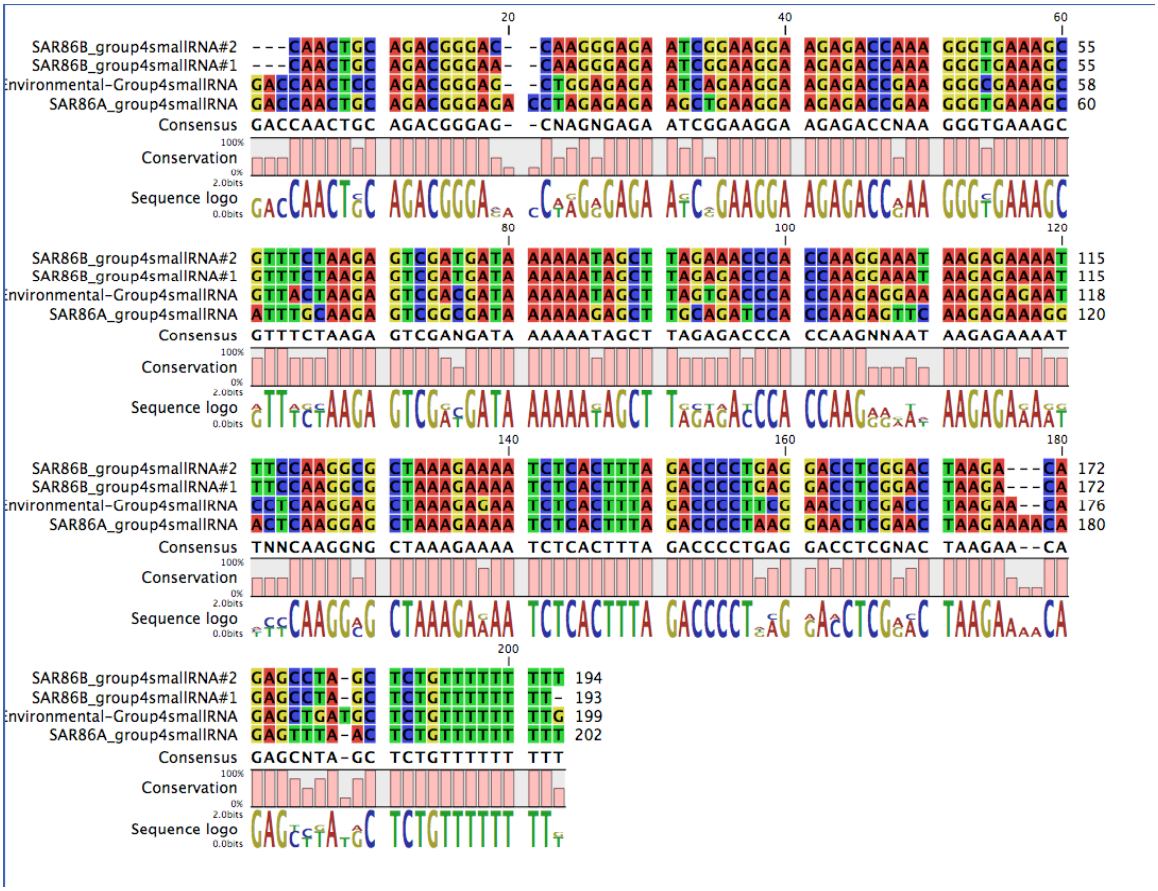

SAR86B_group4smallRNA#2
SAR86B_group4smallRNA#1
Environmental-Group4smallRNA
SAR86A_group4smallRNA
Consensus
Conservation
Sequence logo

Figure S6

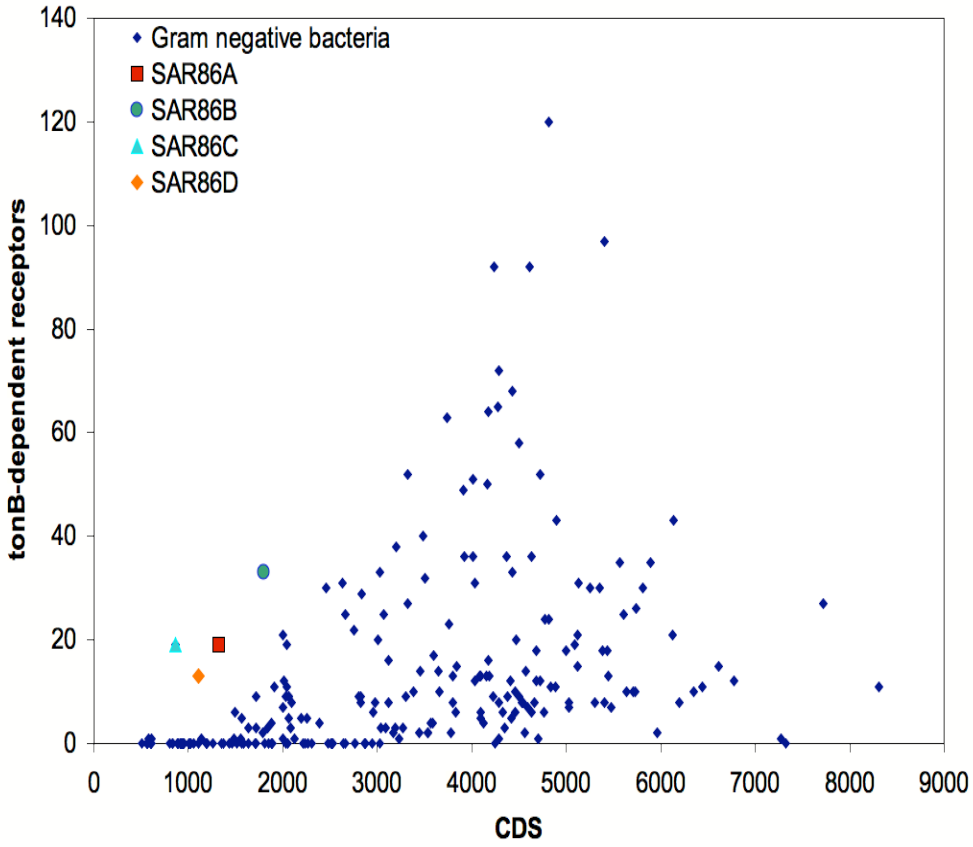

Gram negative bacteria
SAR86A
SAR86B
SAR86C
SAR86D
tonB-dependent receptors
CDS
0
20
40
60
80
100
120
140
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

Figure S7

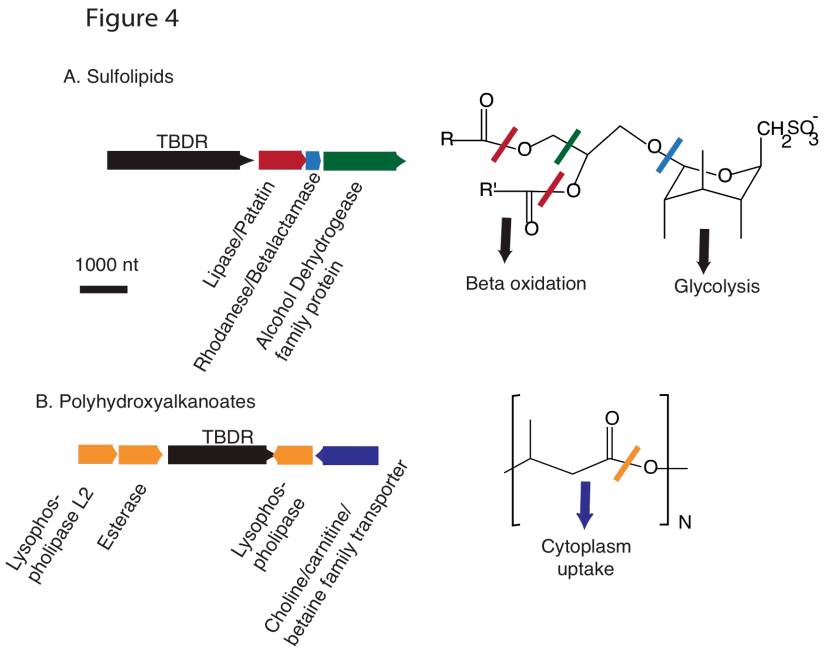


Figure 4
A. Sulfolipids
TBDR
1000 nt
Lipase/Patatin
Rhodanese/Betalactamase
Alcohol Dehydrogease family protein
R
R'
O
CH2SO3-
Beta oxidation
Glycolysis
B. Polyhydroxyalkanoates
TBDR
Lysophos-pholipase L2
Esterase
Lysophos-pholipase
Choline/carnitine/ betaine family transporter
N
Cytoplasm uptake

Table S1: The single copy genes used to determine genome completeness.

Gene symbol	HMM	Cutoff of HMMER 2	Cutoff for HMMER 3	Description
alaS	TIGR00344	250	250	Alanyl-tRNA synthetase
argS	PF00750	-22.8	22.3	Arginyl-tRNA synthetase
aspS	TIGR00459	265	600	Aspartyl-tRNA synthetase
cgtA	TIGR02729	300	300	Obg family GTPase CgtA
coaE	TIGR00152	37	37	dephospho-CoA kinase
cysS	TIGR00435	250	250	cysteinyl-tRNA synthetase
dnaA	TIGR00362	320	320	chromosomal replication initiator protein DnaA
dnaG	TIGR01391	275	275	DNA primase
dnaK	TIGR02350	1110	1110	chaperone protein DnaK
dnaN	TIGR00663	30	30	DNA polymerase III, beta subunit
dnaX	TIGR02397	200	150	DNA polymerase III, subunits gamma and tau
engA	TIGR03594	250	250	ribosome-associated GTPase EngA
era	TIGR00436	100	100	GTP-binding protein Era
ffh	TIGR00959	540	540	signal recognition particle protein
fmt	TIGR00460	160	160	methionyl-tRNA formyltransferase
frr	TIGR00496	150	150	ribosome recycling factor
ftsY	TIGR00064	290	290	signal recognition particle-docking protein FtsY
glyS	TIGR00388	350	350	glycyl-tRNA synthetase
glyS	TIGR00389	300	300	glycyl-tRNA synthetase
gmk	TIGR03263	175	175	guanylate kinase
grpE	PF01025	-6.4	26.6	co-chaperone GrpE
gyrA	TIGR01063	945	945	DNA gyrase, A subunit
gyrB	TIGR01059	1170	1170	DNA gyrase, B subunit
hisS	TIGR00442	270	270	histidyl-tRNA synthetase
ileS	TIGR00392	600	600	isoleucyl-tRNA synthetase
infB	TIGR00487	425	425	translation initiation factor IF-2
infC	TIGR00168	40	40	translation initiation factor IF-3
ksgA	TIGR00755	210	210	dimethyladenosine transferase
lepA	TIGR01393	550	550	GTP-binding protein LepA
leuS	TIGR00396	700	700	leucyl-tRNA synthetase
ligA	TIGR00575	400	400	DNA ligase, NAD-dependent
mnmA	TIGR00420	250	250	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransl
mraW	PF01795	-102.2	19.8	MraW methylase family
nusA	TIGR01953	200	200	transcription termination factor NusA
nusG	TIGR00922	140	140	transcription termination/antitermination factor NusG
pgk	PF00162	-15.7	21.9	phosphoglycerate kinase
pheS	TIGR00468	280	280	phenylalanyl-tRNA synthetase, alpha subunit
pheT	TIGR00471	220	220	phenylalanyl-tRNA synthetase, beta subunit
pheT	TIGR00472	200	200	phenylalanyl-tRNA synthetase, beta subunit
prfA	TIGR00019	475	475	peptide chain release factor 1
proS	TIGR00408	480	480	prolyl-tRNA synthetase
proS	TIGR00409	300	300	prolyl-tRNA synthetase
pyrG	TIGR00337	500	500	CTP synthase
recA	TIGR02012	235	235	recA protein
rfbA	TIGR00082	25	25	ribosome-binding factor A
rnc	TIGR02191	165	165	ribonuclease III
rplA	TIGR01169	200	200	ribosomal protein L1
rplB	TIGR01171	65	65	ribosomal protein L2
rplC	PF00297	-79.6	26.8	ribosomal protein L3
rplD	PF00573	-42.9	20.8	ribosomal protein L4
rplE	PF00281	6.2	20.9	ribosomal protein L5
rplF	PF00347	3.5	22.5	ribosomal protein L6
rplI	TIGR00158	50	50	ribosomal protein L9
rplJ	PF00466	-12.8	21	ribosomal protein L10
rplK	TIGR01632	150	150	ribosomal protein L11
rplL	TIGR00855	65	65	ribosomal protein L7/L12
rplM	TIGR01066	30	30	ribosomal protein L13
rplN	TIGR01067	132	132	ribosomal protein L14
rplO	TIGR01071	60	60	ribosomal protein L15
rplP	TIGR01164	40	40	ribosomal protein L16
rplQ	TIGR00059	75	75	ribosomal protein L17
rplR	TIGR00060	80	80	ribosomal protein L18
rplS	TIGR01024	70	70	ribosomal protein L19
rplT	TIGR01032	19	19	ribosomal protein L20
rplU	TIGR00061	17	17	ribosomal protein L21
rplV	TIGR01044	60	60	ribosomal protein L22
rplW	PF00276	19.9	21.6	ribosomal protein L23
rplX	TIGR01079	16	16	ribosomal protein L24
rpmA	TIGR00062	110	110	ribosomal protein L27
rpmB	TIGR00009	45	45	ribosomal protein L28
rpmC	TIGR00012	8	8	ribosomal protein L29
rpmF	TIGR01031	13	13	ribosomal protein L32
rpmH	TIGR01030	30	30	ribosomal protein L34
rpmI	TIGR00001	40	40	ribosomal protein L35
rpoA	TIGR02027	180	180	DNA-directed RNA polymerase, alpha subunit
rpoB	TIGR02013	1200	1200	DNA-directed RNA polymerase, beta subunit
rpoC	TIGR02386	1250	1250	DNA-directed RNA polymerase, beta' or beta'' subunit
rpoC	TIGR02387	1150	1150	DNA-directed RNA polymerase, beta' or beta'' subunit
rpsB	TIGR01011	-20	-20	ribosomal protein S2
rpsC	TIGR01009	15	15	ribosomal protein S3
rpsD	TIGR01017	180	180	ribosomal protein S4
rpsE	TIGR01021	190	190	ribosomal protein S5
rpsF	TIGR00166	40	40	ribosomal protein S6
rpsG	TIGR01029	120	120	ribosomal protein S7
rpsH	PF00410	-24.4	25.5	ribosomal protein S8
rpsI	PF00380	-23.4	21.4	ribosomal protein S9
rpsJ	TIGR01049	70	70	ribosomal protein S10
rpsK	PF00411	-29.8	21.8	ribosomal protein S11
rpsL	TIGR00981	80	80	ribosomal protein S12
rpsM	PF00416	-5.6	21.3	ribosomal protein S13
rpsO	TIGR00952	30	30	ribosomal protein S15
rpsP	TIGR00002	35	35	ribosomal protein S16
rpsQ	PF00366	12.4	20.9	ribosomal protein S17
rpsR	TIGR00165	35	35	ribosomal protein S18
rpsS	TIGR01050	100	100	ribosomal protein S19
rpsT	TIGR00029	40	40	ribosomal protein S20
secA	TIGR00963	1000	1000	preprotein translocase, SecA subunit
secE	TIGR00964	13	13	preprotein translocase, SecE subunit
secG	TIGR00810	14	14	preprotein translocase, SecG subunit
secY	TIGR00967	200	200	preprotein translocase, SecY subunit
serS	TIGR00414	225	225	seryl-tRNA synthetase
smpB	TIGR00086	110	100	SmpB protein
thrS	TIGR00418	250	250	threonyl-tRNA synthetase
tig	TIGR00115	75	75	trigger factor
tilS	TIGR02432	80	80	tRNA(Ile)-lysidine synthetase
tsf	TIGR00116	8	8	translation elongation factor Ts
tyrS	TIGR00234	60	60	tyrosyl-tRNA synthetase
uvrB	TIGR00631	1000	1000	excinuclease ABC, B subunit
valS	TIGR00422	525	525	valyl-tRNA synthetase
ybeY	TIGR00043	10	10	conserved hypothetical protein YbeY
ychF	TIGR00092	225	225	GTP-binding protein YchF

Table S2: Abundance of SAR86 genomes across individual GOS samples and environmental data

	Abundance of each genome							
Library name	SAR86A	SAR86B	SAR86C	SAR86D	Latitude&Longitude	Depth (m)	Temperature (°C)	Salinity (PSU)
Pilot_0.1	4.26	0	0	0	32°9'30"N, 64°0'36"W	5	20	35
Pilot_0.8	0.02	0	0	0	32°9'30"N, 64°0'36"W	5	20	35
Pilot_3.0	0.17	0	0	0	32°9'30"N, 64°0'36"W	5	20	35
GS2	0	0	0.6	0.74	42°30'10"N, 67°14'24"W	1	18.2	29.24
GS3	0	0	0.57	1.06	42°51'9"N, 66°13'18"W	1	11.7	29.9
GS4	0	0	0.29	1.76	44°8'14"N, 63°38'39"W	1	17.3	28.3
GS5	0	0	0.22	0	44°41'25"N, 63°38'14"W	1	15	30.2
GS6	0	0	0.99	0.13	45°6'42"N, 64°56'47"W	1	11.2	n/a
GS7	0	0	0.42	0.23	43°37'55"N, 66°50'49"W	1	17.9	31.7
GS8	0	0	0.19	0.03	41°29'9"N, 71°21'4"W	1	9.4	26.5
GS9	0	0	0.37	0.2	41°5'28"N, 71°36'7"W	1	11	31
GS10	0	0.05	0	0.1	38°56'23"N, 74°41'6"W	1	12	31
GS13	0	0	0.11	0	36°0'14"N, 75°23'41"W	1	9.3	n/a
GS14	0.18	0.07	0	0	32°30'25"N, 79°15'49"W	1	18.6	n/a
GS15	0.36	0	0	0	24°29'18"N, 83°4'12"W	2	25.3	36
GS16	0.39	0	0	0	24°10'28"N, 84°20'39"W	2	26.4	35.8
GS17	0.74	0	0	0	20°31'21"N, 85°24'48"W	2	27	35.8
GS18	0.43	0	0	0	18°2'12"N, 83°47'53"W	2	27.4	35.4
GS19	0.42	0.05	0	0	10°42'58"N, 80°15'15"W	2	27.7	35.4
GS21	0	0.63	0	0	8°7'45"N, 79°41'28"W	2	27.6	30.7
GS22	0	0.09	0	0	6°29'33"N, 82°54'13"W	2	29.3	32.26
GS23	0.41	0	0	0	5°38'23"N, 86°33'54"W	2	28.7	32.58
GS25	0.13	0	0	0	5°30'0"N, 87°4'0"W	1	28.3	31.4
GS26	0.31	0	0	0	1°15'51"N, 89°17'41"W	2	27.8	32.58
GS27	0	0.02	0	0	1°12'57"S, 90°25'22"W	2	25.5	34.89
GS30	0.09	0.02	0	0	0°16'20"N, 91°38'23"W	19	26.9	n/a
GS31	0.02	0	0	0	0°18'34"S, 91°39'6"W	12	18.6	n/a
GS34	0.08	0	0	0	0°22'59"S, 90°16'46"W	2	27.5	n/a
GS37	0.12	0	0	0	1°58'26"S, 95°0'53"W	2	28.8	n/a
GS38	0.02	0	0	0	2°34'55"S, 97°51'46"W	2	28.4	n/a
GS41	0.08	0	0	0	5°55'48"S, 108°41'13"W	2	28.03	34.99
GS42	0.04	0	0	0	7°6'26"S, 116°7'8"W	2	27.57	38.07
GS43	0.69	0	0	0	7°39'39"S, 120°24'8"W	2	27.56	35.85
GS44	0.31	0	0	0	8°24'54"S, 124°14'23"W	2	27.7	39.36
GS45	0.59	0	0	0	9°1'33"S, 127°46'16"W	2	28.26	36.95
GS46	0.15	0	0	0	9°34'16"S, 131°29'29"W	2	28.67	35.6
GS47	0.45	0	0	0	10°7'53"S, 135°26'57"W	27	28.6	37.3
GS48	0.21	0	0	0	17°28'33"S, 149°48'43"W	1	27.3	35.06
GS48B	0.08	0	0	0	17°28'33"S, 149°48'43"W	1	27.3	35.06
GS49	0.51	0	0	0	17°27'10"S, 149°47'56"W	1	28.9	32.62
GS108	0.3	0	0	0	12°5'33"S, 96°52'54"E	2	25.8	32.44
GS108B	0.16	0	0	0	12°5'33"S, 96°52'54"E	2	25.8	32.44
GS108_454	0.14	0	0	0	12°5'33"S, 96°52'54"E	2	25.8	32.44
GS109	0.43	0	0	0	10°56'37"S, 92°3'32"E	1.5	27.18	32.57
GS110	0.73	0	0	0	10°26'46"S, 88°18'10"E	1.5	27.01	32.71
GS110B	0.4	0	0	0	10°26'46"S, 88°18'10"E	1.5	27	32.71
GS110_454	0.24	0	0	0	10°26'46"S, 88°18'10"E	1.5	27	32.71
GS111	0.86	0	0	0	9°35'49"S, 84°11'51"E	1.8	26.44	32.31
GS112	0.62	0	0	0	8°30'18"S, 80°22'32"E	1.8	26.56	32.47
GS112B	0.68	0	0	0	8°30'18"S, 80°22'32"E	1.8	26.56	32.47
GS112C	0.15	0	0	0	8°30'18"S, 80°22'32"E	1.8	26.56	32.47
GS113	0.47	0	0	0	7°0'27"S, 76°19'53"E	1.8	27.48	33.31
GS114	0.36	0	0	0	4°59'25"S, 64°58'36"E	1.5	28.15	33.13
GS114B	0.17	0	0	0	4°59'25"S, 64°58'36"E	1.5	28.15	33.13
GS114C	0.41	0	0	0	4°59'25"S, 64°58'36"E	1.5	28.15	33.13
GS114D	0.8	0	0	0	4°59'25"S, 64°58'36"E	1.5	28.15	33.13
GS115	0.03	0	0	0	4°39'45"S, 60°31'23"E	1.5	27.88	33.18
GS116	0.19	0	0	0	4°38'6"S, 56°50'10"E	1.5	26.15	33.06
GS117	0.37	0	0	0	4°36'49"S, 55°30'31"E	1.8	26.42	35.5
GS117B	0.13	0.01	0	0	4°36'49"S, 55°30'31"E	1.8	26.42	35.5
GS117C	0.11	0	0	0	4°36'49"S, 55°30'31"E	1.8	26.42	35.5
GS119	0.39	0.06	0	0	23°12'58"S, 52°18'22"E	1.9	23.79	35.38
GS120	0.21	0	0	0	26°2'6"S, 50°7'23"E	2.8	22.52	35.58
GS121	0.23	0.03	0	0	29°20'56"S, 43°12'56"E	1.5	23.1	35.41
GS122	0.03	0	0	0	30°53'54"S, 40°25'13"E	1.9	20.22	35.79
GS122B	0.07	0	0	0	30°53'54"S, 40°25'13"E	1.9	20.22	35.79
GS123	0.01	0	0	0	32°23'57"S, 36°35'31"E	2.2	20.36	35.79
GS128	0.08	0.84	0	0	30°28'56"S, 9°2'18"E	1.7	27.00	36.29
GS148B	0	1.41	0	0	6°19'S, 39°33'E	1	27.00	n/a
GS148	0.14	0.68	0	0	6°19'S, 39°33'E	1	27.00	n/a
GS149B	0	0.51	0	0	6° 7'S, 39°07'E	1	27.00	n/a