Supplementary Information

New Abundant Microbial Groups in Aquatic Hypersaline Environments

Rohit Ghai[1], Lejla Pašić[1,2], Ana Beatriz Fernández[3], Ana-Belen Martin-Cuadrado[1], Carolina Megumi Mizuno[1], Katherine D. McMahon[4], R. Thane Papke[5], Ramunas Stepanauskas[6], Beltran Rodriguez-Brito[7], Forest Rohwer[7], Cristina Sánchez-Porro[3], Antonio Ventosa[3] and Francisco Rodríguez-Valera[1*]

[1]*Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, San Juan de Alicante, Alicante, Spain*
[2]*Department of Biology, Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia*
[3]*Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Sevilla, Sevilla, Spain*
[4]*Departments of Civil and Environmental Engineering and Bacteriology, University of Wisconsin Madison, Madison, USA*
[5]*Department of Molecular Cell Biology, University of Connecticut, Storrs, Connecticut, USA*
[6]*Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine, USA.*
[7]*Department of Biology, San Diego State University, San Diego, California, USA*

*Correspondence and requests for materials should be addressed to F.R-V. (frvalera@umh.es.)

Supplementary Table S1. General features of the datasets

Datasets	Salinity (%)	Temperature (°C)	pH	# of Reads	Dataset Size (Mb)	Average Read Length (bp)
Santa Pola Saltern (SS37)	37	41	8.0	760740	309	417
Santa Pola Saltern (SS19)	19	34	8.0	1315302	475	361
Deep Chlorophyll Maximum (DCM3)	3.8	15.9	8.1	1204321	312	259
Punta Cormoran (PC6)	6.4	37.6		60391	64	1073

Supplementary Table S2. Most abundant organisms in PC6.

Organism Name	Number of Hits	% of Classified Hits	Taxonomy	GC %
***Synechococcus* sp. RS9917**	33710	5.66	Cyanobacteria, Chroococcales	64
Marine actinobacterium PHSC20C1	26691	4.48	Actinobacteria, unclassified Actinobacteria	59
Clavibacter michiganensis* subsp. *michiganensis	20972	3.52	Actinobacteria, Actinobacteridae	72
Leifsonia xyli* subsp. *xyli	13453	2.26	Actinobacteria, Actinobacteridae	70
Acidothermus cellulolyticus	6652	1.11	Actinobacteria, Actinobacteridae	66
Dinoroseobacter shibae	9857	1.65	Alphaproteobacteria, Rhodobacterales	66
Congregibacter litoralis	10511	1.76	Gammaproteobacteria, unclassified Gammaproteobacteria	57
Blastopirellula marina	14375	2.41	Planctomycetes, Planctomycetacia	57
Marinobacter hydrocarbonoclasticus aquaeolei	8694	1.46	Gammaproteobacteria, Alteromonadales	57
Parvibaculum lavamentivorans	7518	1.26	Alphaproteobacteria, Rhizobiales	62
Chromohalobacter salexigens	6379	1.07	Gammaproteobacteria, Oceanospirillales	63
Silicibacter pomeroyi	6727	1.13	Alphaproteobacteria, Rhodobacterales	60
***Roseovarius* sp. 217**	7379	1.24	Alphaproteobacteria, Rhodobacterales	60
Rhodobacterales bacterium HTCC2654	6065	1.01	Alpharoteobacteria, Rhodobacterales	64
Kineococcus radiotolerans	6257	1.05	Actinobacteria, Actinobacteridae	74
Hahella chejuensis	7225	1.21	Gammaproteobacteria, Oceanospirillales	53
Rhodopirellula baltica	6998	1.17	Planctomycetes, Planctomycetacia	55

The analysis was performed using the MG-RAST server. 86% of all sequences in the dataset could be classified. The '% hits classified' column represents the fraction of all classified hits that could be ascribed to the microbe. Only hits with e-value <1e-5 and minimum alignment length of 50 were considered classified. A brief taxonomy, and the GC% of the organism's genome are shown in the last two columns.

Supplementary Table S3. Most abundant organisms in SS19.

Organism Name	Number of Hits	% of Classified Hits	Taxonomy	GC%
Haloquadratum walsbyi	114437	14.708	Euryarchaeota, Halobacteria	48
Salinibacter ruber	64657	8.31	Bacteroidetes, Sphingobacteria	66
Halogeometricum borinquense	63268	8.131	Euryarchaeota, Halobacteria	61
Halomicrobium mukohataei	36975	4.752	Euryarchaeota, Halobacteria	65
Nitrococcus mobilis	34287	4.407	Gammaproteobacteria, Chromatiales	59
Halorhabdus utahensis	28199	3.624	Euryarchaeota, Halobacteria	62
Alkalilimnicola ehrlichei	25078	3.223	Gammaproteobacteria, Chromatiales	67
Gramella forsetii	13582	1.746	Bacteroidetes, Flavobacteria	36
Halorhodospira halophila	12215	1.57	Gammaproteobacteria, Chromatiales	67
Croceibacter atlanticus	11429	1.469	Bacteroidetes, Flavobacteria	33
Marine actinobacterium PHSC20C1	9147	1.176	Actinobacteria, unclassified Actinobacteria	59
Clavibacter michiganensis* subsp. *michiganensis	8022	1.031	Actinobacteria, Actinobacteridae	72
***Flavobacterium* sp. MED217**	7750	0.996	Bacteroidetes, Flavobacteria	39
Rhodothermus marinus	7134	0.917	Bacteroidetes, Sphingobacteria	64
Leifsonia xyli* subsp. *xyli	6616	0.85	Actinobacteria, Actinobacteridae	67
***Roseovarius* sp. 217**	6468	0.831	Alpharoteobacteria, Rhodobacterales	60
Chromohalobacter salexigens	6041	0.776	Gammaproteobacteria, Oceanospirillales	63
***Cellulophaga* sp. MED134**	5412	0.696	Bacteroidetes, Flavobacteria	38

The analysis was performed using the MG-RAST server. 59% of all sequences in the dataset could be classified. The '% hits classified' column represents the fraction of all classified hits that could be ascribed to the microbe. Only hits with e-value <1e-5 and minimum alignment length of 50 were considered classified. A brief taxonomy, and the GC% of the organism's genome are shown in the last two columns.

Supplementary Table S4. Most abundant organisms in SS37

Organism Name	Number of Hits	% of Classified Hits	Taxonomy	GC%
Haloquadratum walsbyi	304850	63.661	Euryarchaeota, Halobacteria	48
Halogeometricum borinquense	37327	7.795	Euryarchaeota, Halobacteria	61
Halomicrobium mukohataei	25699	5.367	Euryarchaeota, Halobacteria	65
Halorhabdus utahensis	21271	4.442	Euryarchaeota, Halobacteria	62
Salinibacter ruber	18663	3.897	Bacteroidetes, Sphingobacteria	66
***Anaeromyxobacter* sp. Fw109-5**	1014	0.212	Deltaproteobacteria, Myxococcales	73
Chitinophaga pinensis	974	0.203	Bacteroidetes, Sphingobacteria	45
Pedobacter heparinus	889	0.186	Bacteroidetes, Sphingobacteria	42
Methanocaldococcus jannaschii	870	0.182	Euryarchaeota, Methanococci	31
Methanopyrus kandleri	859	0.179	Euryarchaeota, Methanopyri	61
Spirosoma linguale	840	0.175	Bacteroidetes, Cythophagia	50
Methanococcoides burtonii	814	0.17	Euryarchaeota, Methanomicrobia	40
Rhodothermus marinus	791	0.165	Bacteroidetes, Sphingobacteria	64
Archaeoglobus fulgidus	771	0.161	Euryarchaeota, Archaeoglobi	48
Cytophaga hutchinsonii	752	0.157	Bacteroidetes, Cythophagia	38
Thermococcus kodakarensis	752	0.157	Euryarchaeota, Thermococci	51
Sorangium cellulosum	744	0.155	Deltaproteobacteria, Myxococcales	71
Methanothermobacter thermautotrophicus	719	0.15	Euryarchaeota, Methanobacteria	49

The analysis was performed using the MG-RAST server. 64.6% of all sequences in the dataset could be classified. The '% hits classified' column represents the fraction of all classified hits that could be ascribed to the microbe. Only hits with e-value <1e-5 and minimum alignment length of 50 were considered classified. A brief taxonomy, and the GC% of the organism's genome are shown in the last two columns.

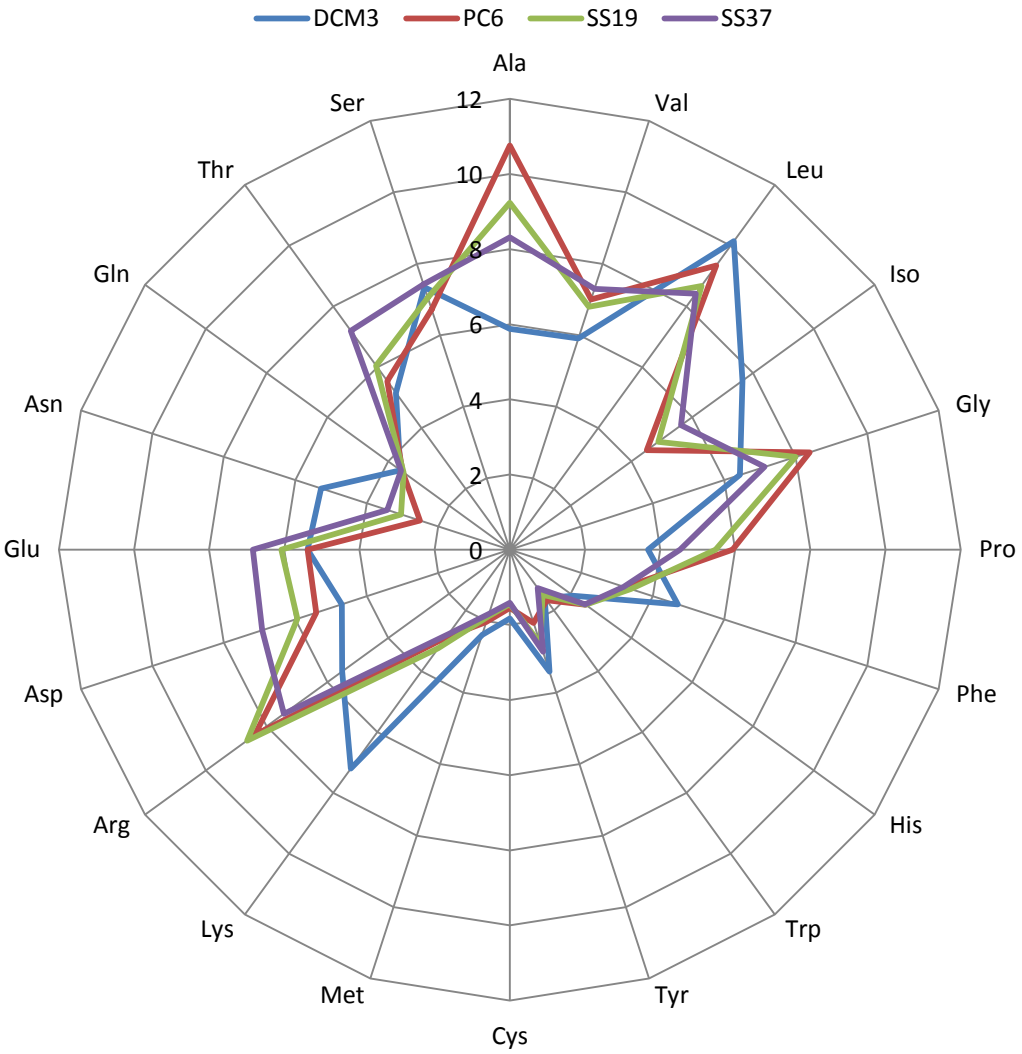
Supplementary Table S5. Percentage of acidic proteins encoded by the genera detected in SS19 dataset.

		% of proteins with pI< 4.5	Number of proteins with pI< 4.5	Total number of proteins
Halophiles	*Natronomonas*	72	1933	2659
	Halorubrum	71	2281	3184
	Halomicrobium	70	2234	3173
	Haloarcula	69	2389	3415
	Haloferax	66	1956	2945
	Halobacterium	64	1332	2075
	Haloquadratum	62	1628	2610
	Salinibacter	34	958	2801
Non-halophiles	*Dinoroseobacter*	16	660	4187
	Silicibacter	15	580	3864
	Roseobacter	15	605	4129
	Flavobacterium MED217	13	502	3735
	Gramella	12	426	3584
	Renibacterium	12	416	3507
	Alkalilimnicola	10	296	2865
	*Prochlorococcus**	6	107	1717
	Nitrococcus	6	194	3503
	*Polynucleobacter**	5	71	1508

*Genomes of typically marine bacteria (*Prochlorococcus*) and typically freshwater bacteria (*Polynucleobacter*) are also included for reference.

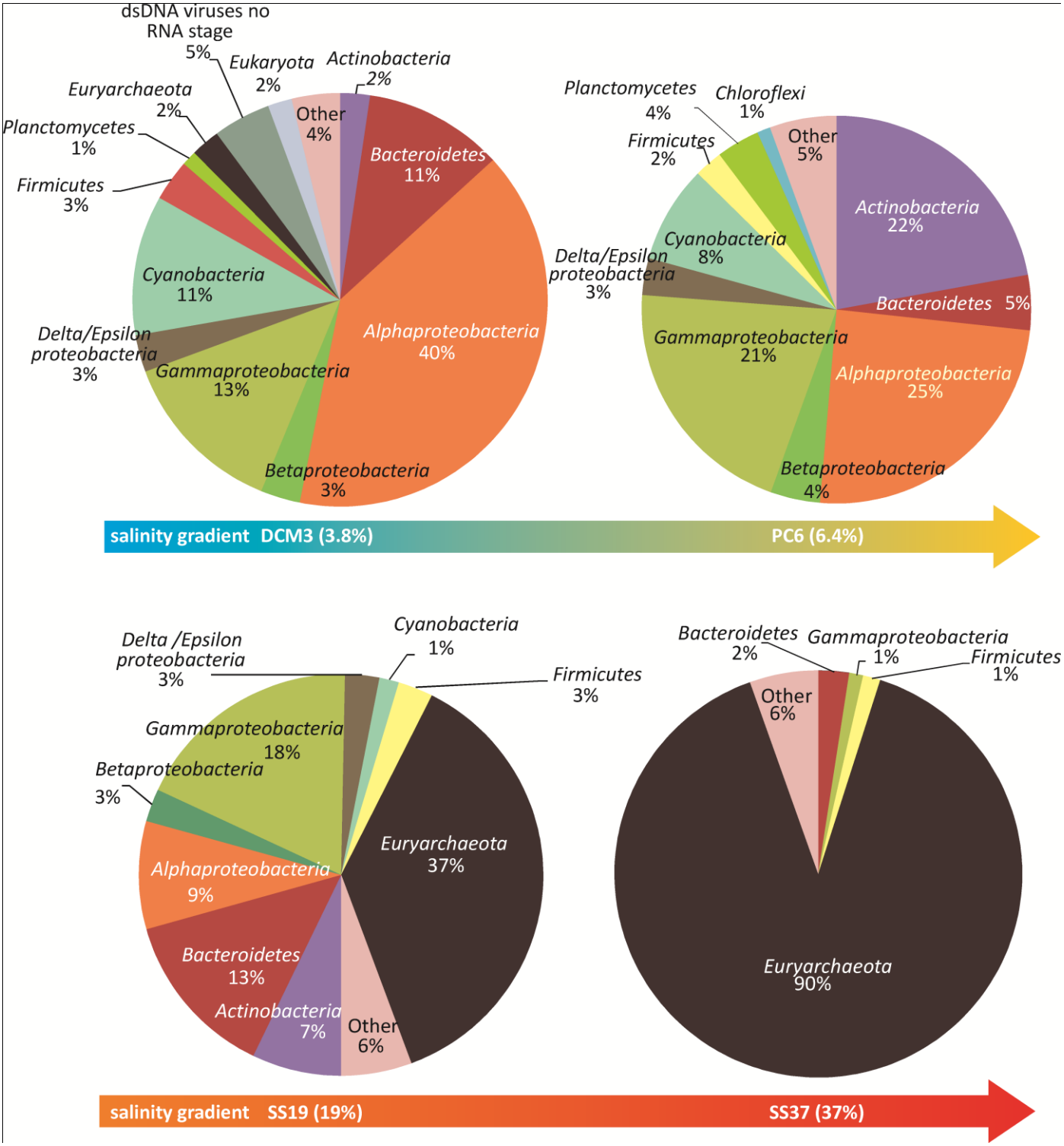
Supplementary Figure S1.

Amino acid composition of proteins from four metagenomic datasets



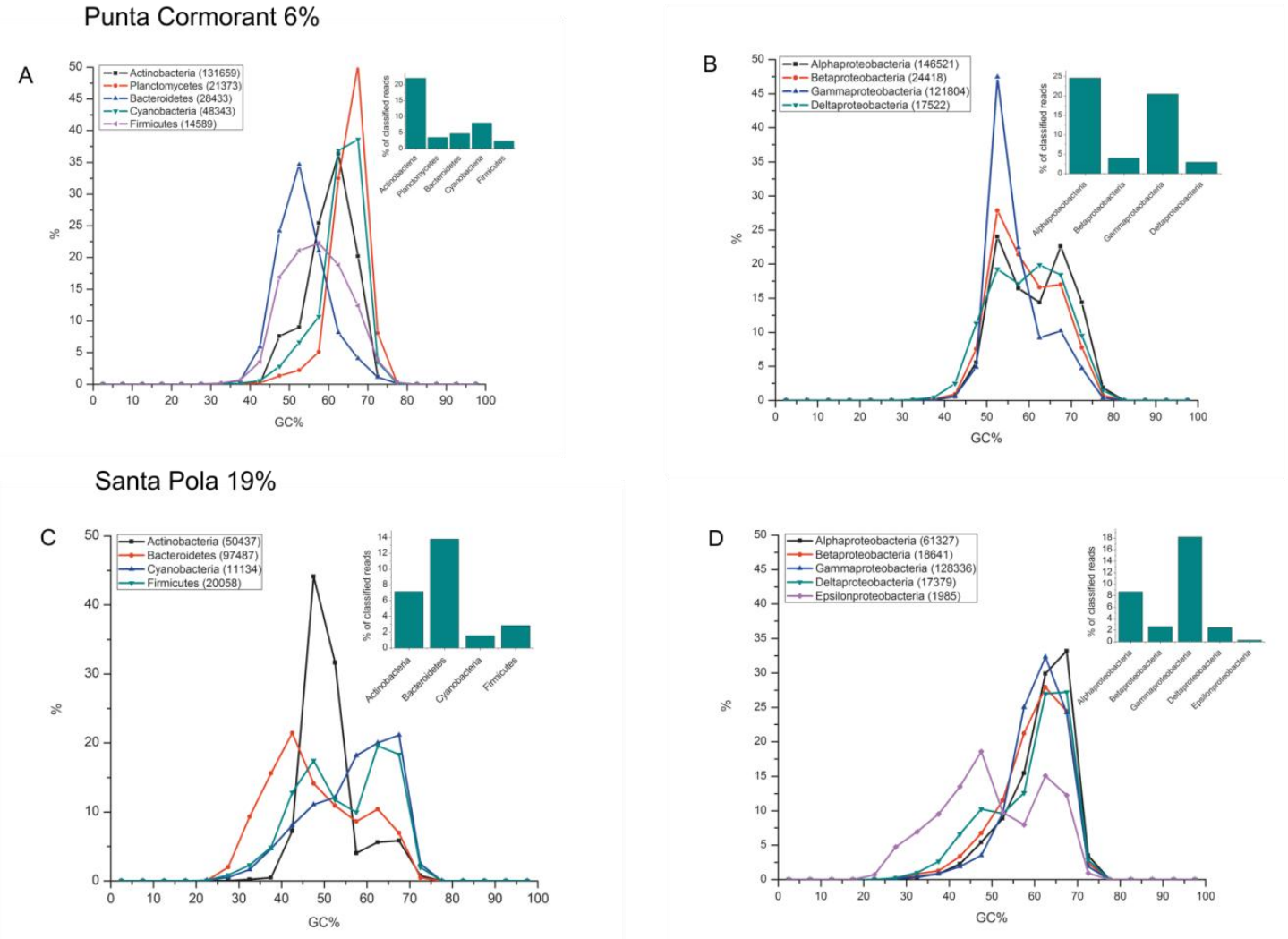
Supplementary Figure S2.

Community structure from all reads for the four metagenomic datasets (using the MG-RAST server)



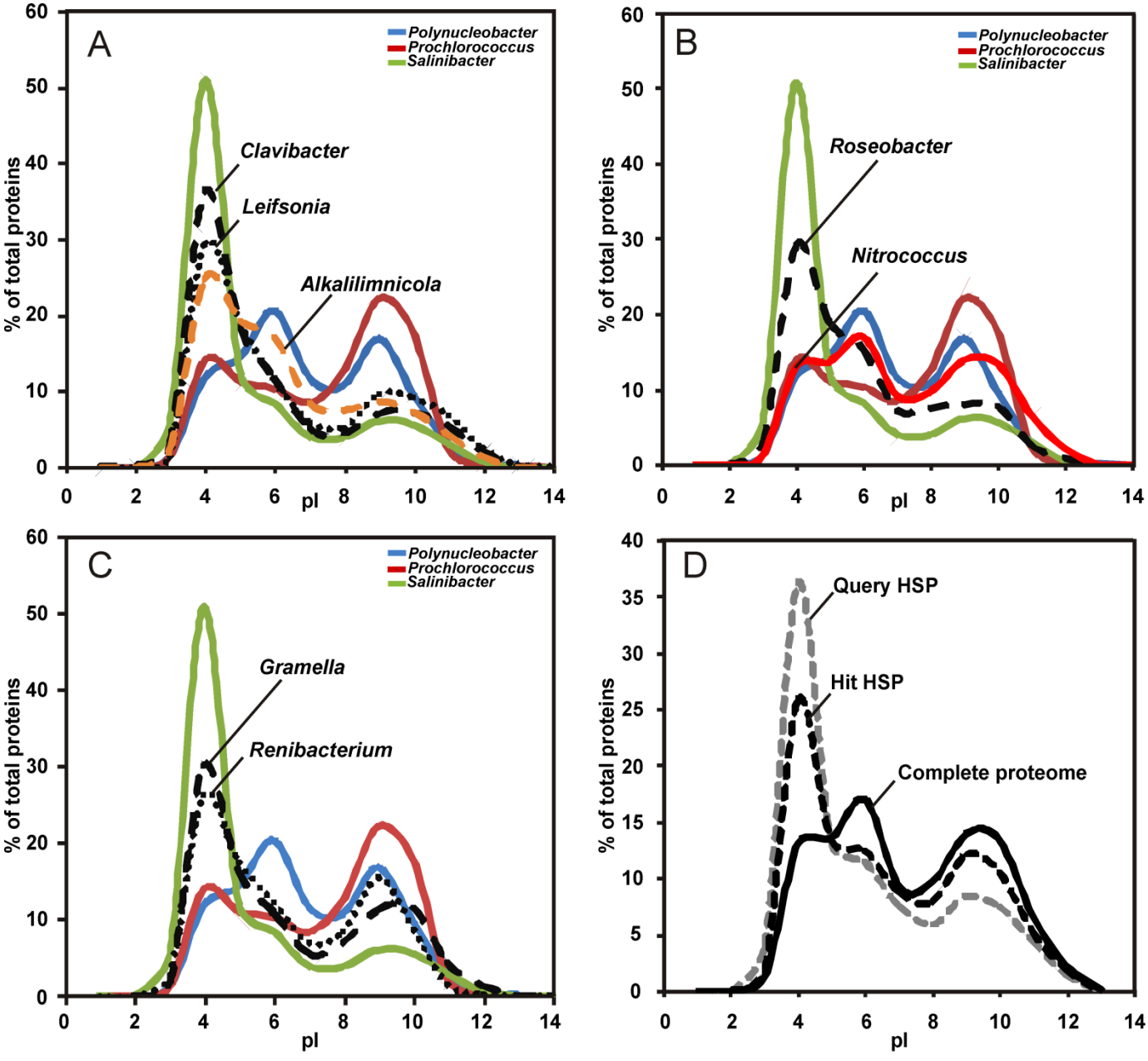
Supplementary Figure S3.

GC% of reads assigned to the dominant taxa in Punta Cormoran and Santa Pola 19% datasets. Actinobacteria, Planctomycetes, Bacteroidetes, Cyanobacteria are shown in (A) for both metagenomes, while all Proteobacteria (Alpha, Beta, Gamma, Delta) are shown in (B). The % of reads in the entire Dataset assigned to each taxa are indicated in the bar charts in the inset for each figure. The numbers in brackets indicate number of reads that were assigned to each of these taxa.



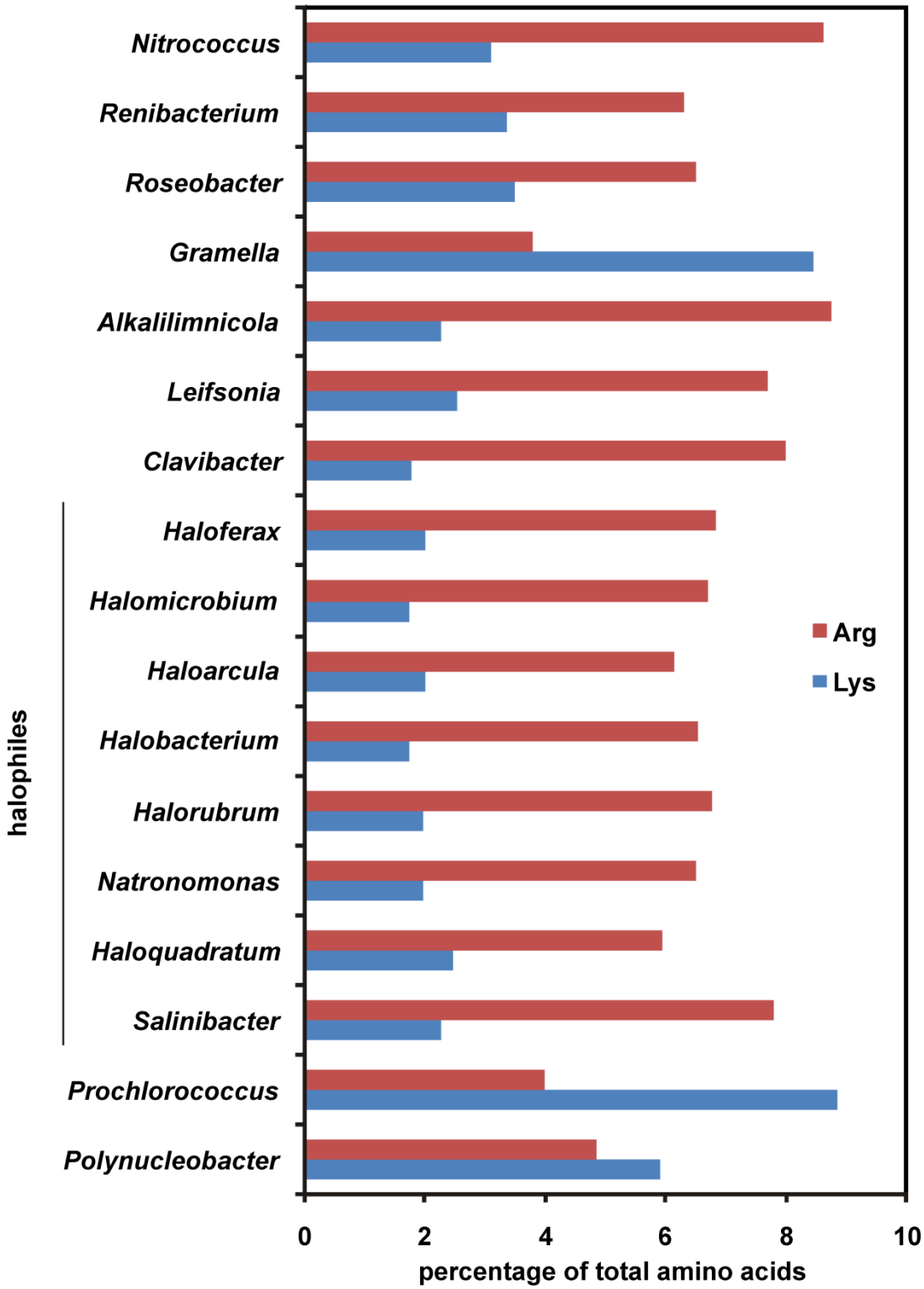
Supplementary Figure S4.

Comparison of the isoelectric profile of (A) *Clavibacter* and *Leifsonia* (B) *Roseobacter* and *Nitrococcus* (C) *Gramella* and *Renibacterium* with a typical halophilic, marine and a freshwater bacteria. (D) Comparison of the isoelectric profile of translated protein queries to their hits in the *Nitrococcus mobilis* proteome. Query HSP: The pI profile of the protein translation of the 454 reads. Hit HSP: The pI profile of the blast hits in the *Nitrococcus* genome. The profile of the complete proteome of *Nitrococcus mobilis* is also shown for comparison



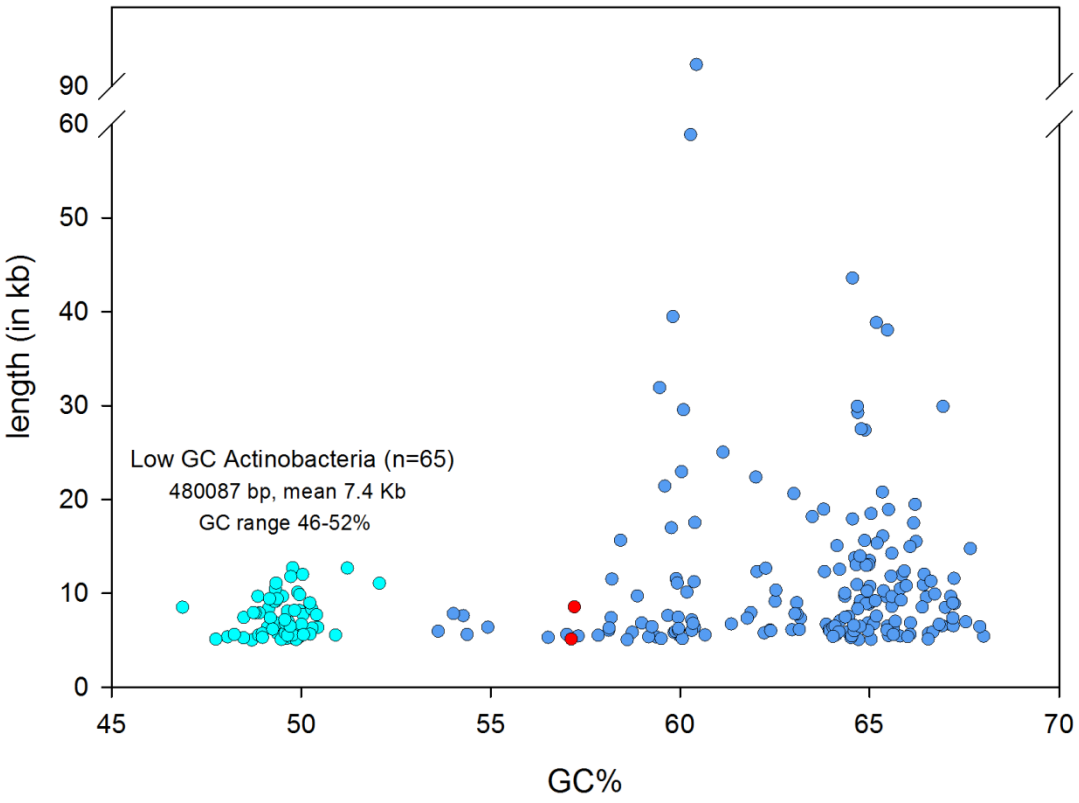
Supplementary Figure S5.

Preference for arginine in place of lysine in proteomes of typical halophilic microbes, and in several other genera detected in SS19. A typical freshwater bacteria (*Polynucleobacter*) and a typical marine bacteria (*Prochlorococcus*) are included for reference



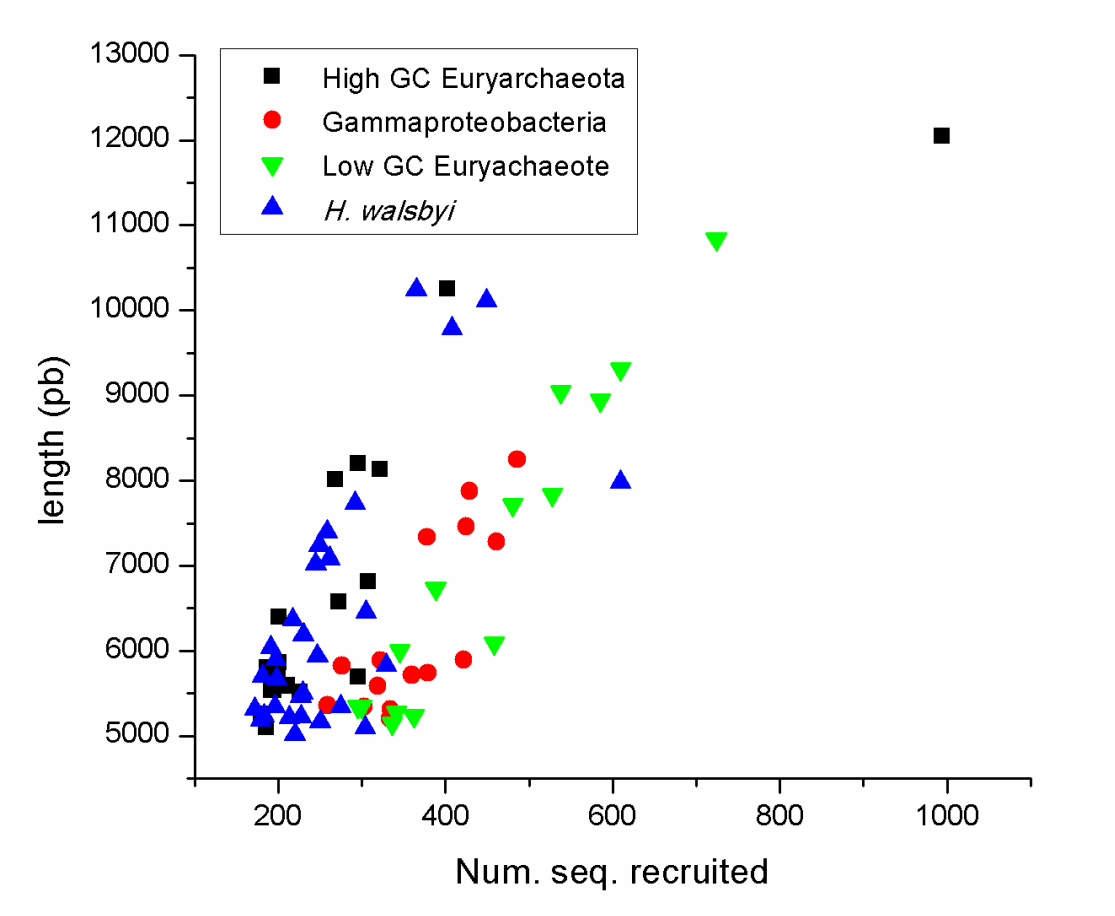
Supplementary Figure S6.

GC% versus length of Actinobacterial scaffolds from Punta Cormoran. Two types of clusters are indicated, light blue: Low GC Actinobacterial Contigs, and blue: High GC actinobacterial contigs. Two contigs that are shown in red are those containing a 16S rRNA sequence.



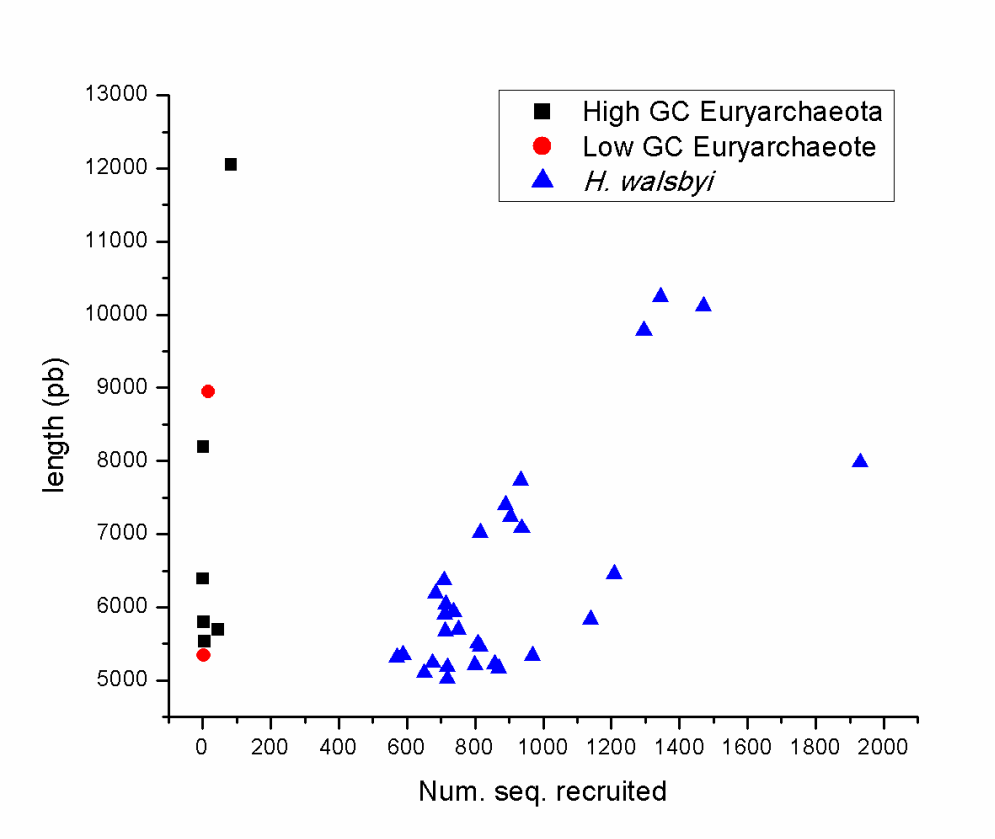
Supplementary Figure S7.

Comparative abundance of the assembled contigs from SS19 in the SS19 dataset. X- axis shows the number of reads recruited by each contig and the Y-axis shows the length of the contig. Four types of contigs are shown, those belonging to *H. walsbyi*, a Gammaproteobacteria, low GC Euryarchaeota and High GC Euryarchaeota.



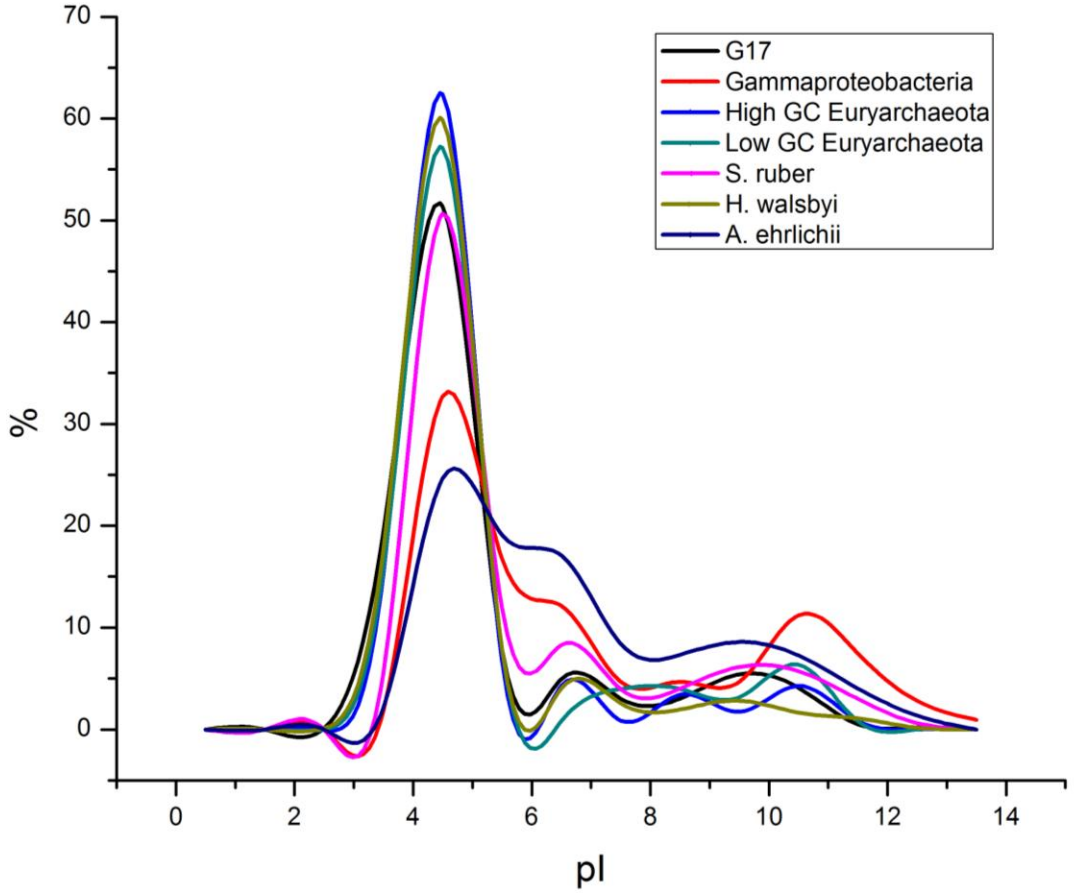
Supplementary Figure S8.

Comparative abundance of the assembled contigs from SS19 in the SS37 dataset. X- axis shows the number of reads recruited by each contig and the Y-axis shows the length of the contig. Three types of contigs are shown, those belonging to *H. walsbyi*, low GC Euryarchaeota and High GC Euryarchaeota.



Supplementary Figure S9.

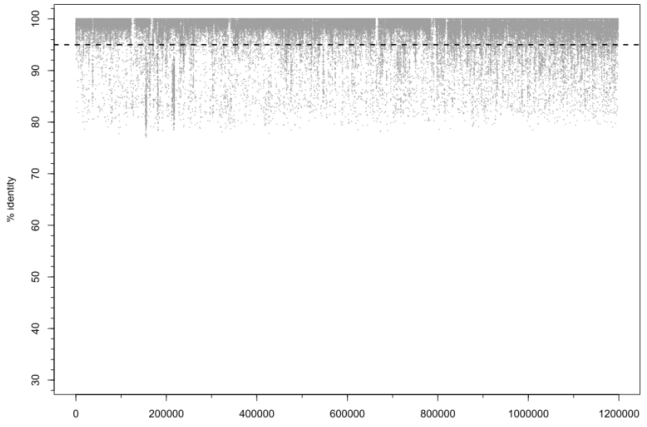
Isoelectric point profiles proteins from assembled contigs of SS19 dataset (Gammaproteobacteria, High GC Euryarchaeota, Low GC Euryarchaeota), single cell amplified genome of G17 (Candidatus *Haloredivivus*) , and reference genomes of *Haloquadratum walsbyi*, *Salinibacter ruber* and *Alkalilimnicola ehrlichii* .



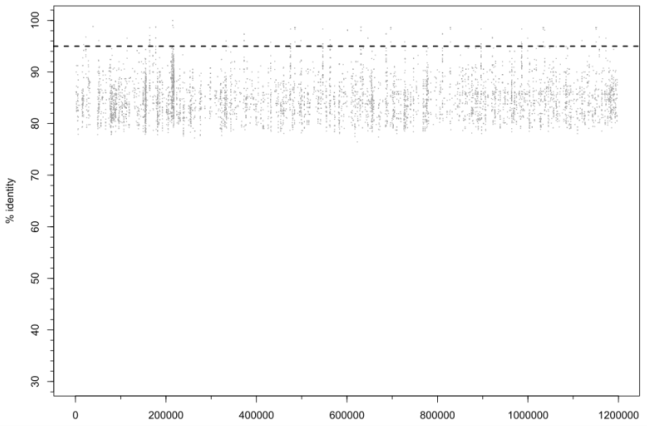
Supplementary Figure S10.

Recruitment of the metagenomic reads of SS19 and SS37 datasets by the genome of the low GC archaeon *Candidatus* Haloredivivus (also referred to as the G17 SAG). BLASTN was used to make the comparison. A horizontal line is drawn at the 95% identity levels. Only alignments that are >50bp long are shown.

Candidatus *Haloredivivus* versus SS19

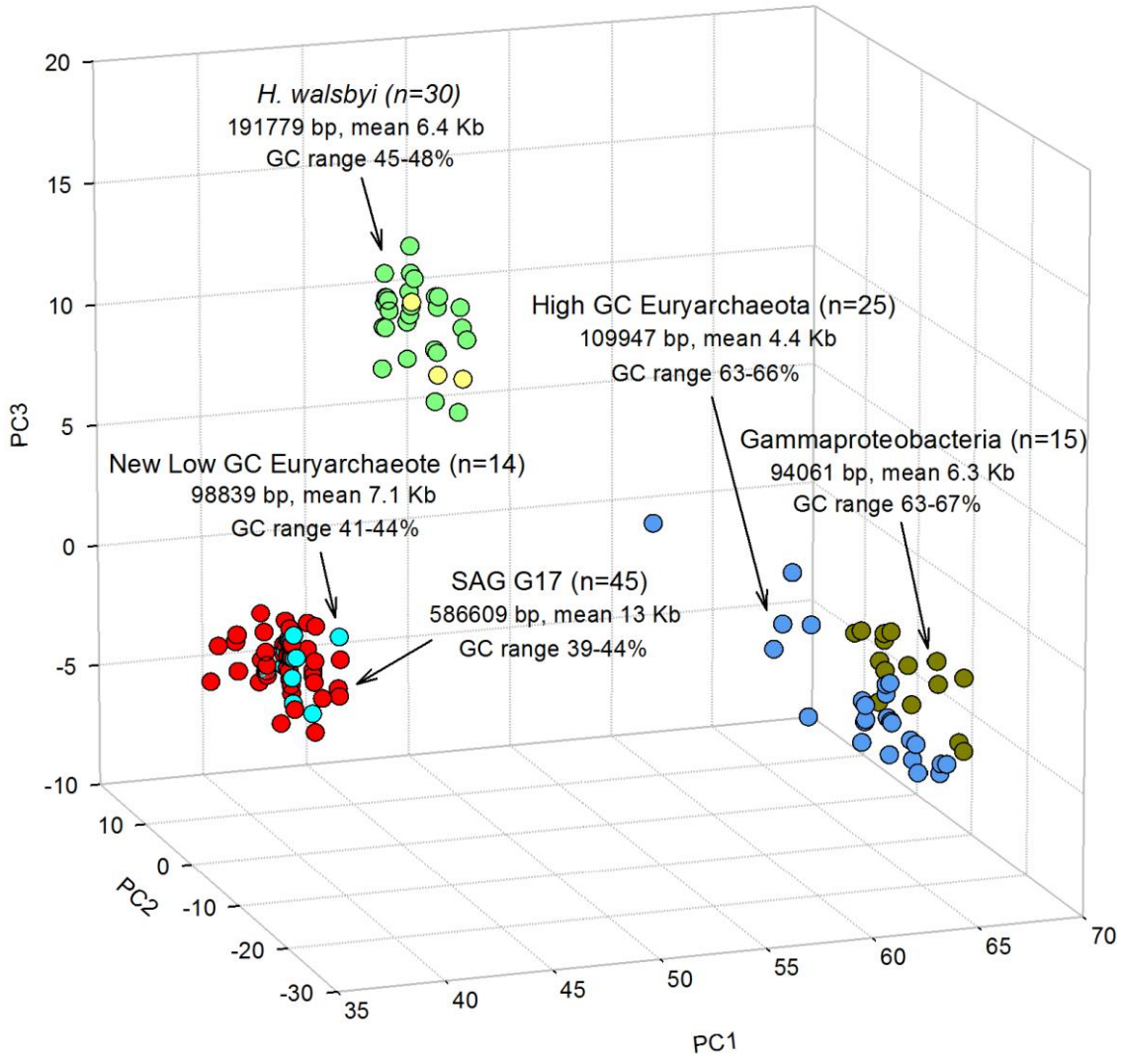


Candidatus *Haloredivivus* versus SS37



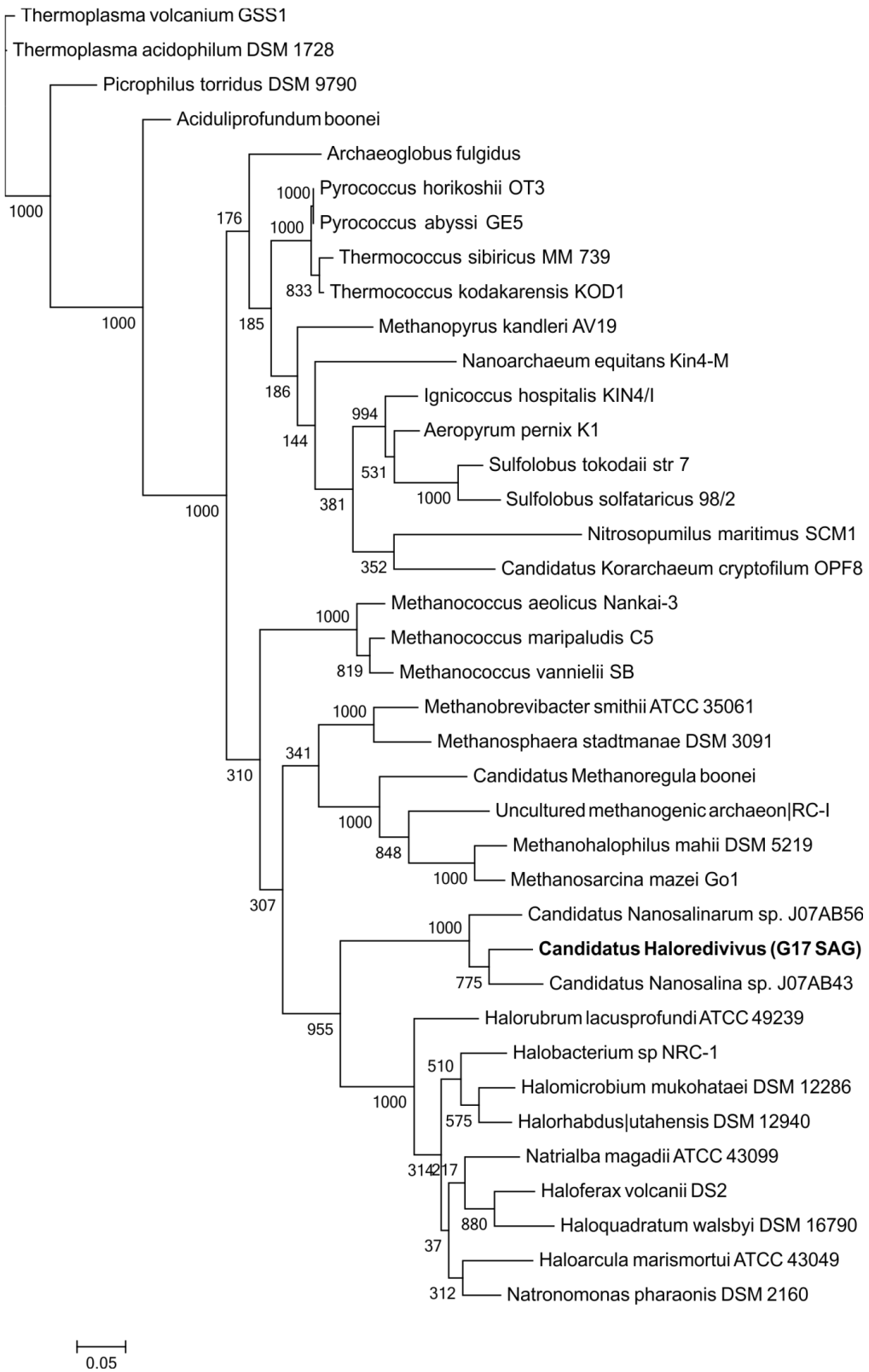
Supplementary Figure S11.

Principal component analysis of tetranucleotide frequencies of the assembled contigs of SS19 metagenome and the contigs of the assembled G17 SAG genome (*Candidatus* Haloredivivus). Red: Low GC euryarchaeote contigs, Light Blue: G17 SAG contigs, Light Green: *H. walsbyi* assembled contigs, Yellow: *H. walsbyi* assembled contigs (with a single gene not giving a hit to *H. walsbyi*, Blue: High GC Euryarchaeota assembled contigs, Dark Yellow: Gammaproteobacterial contigs.



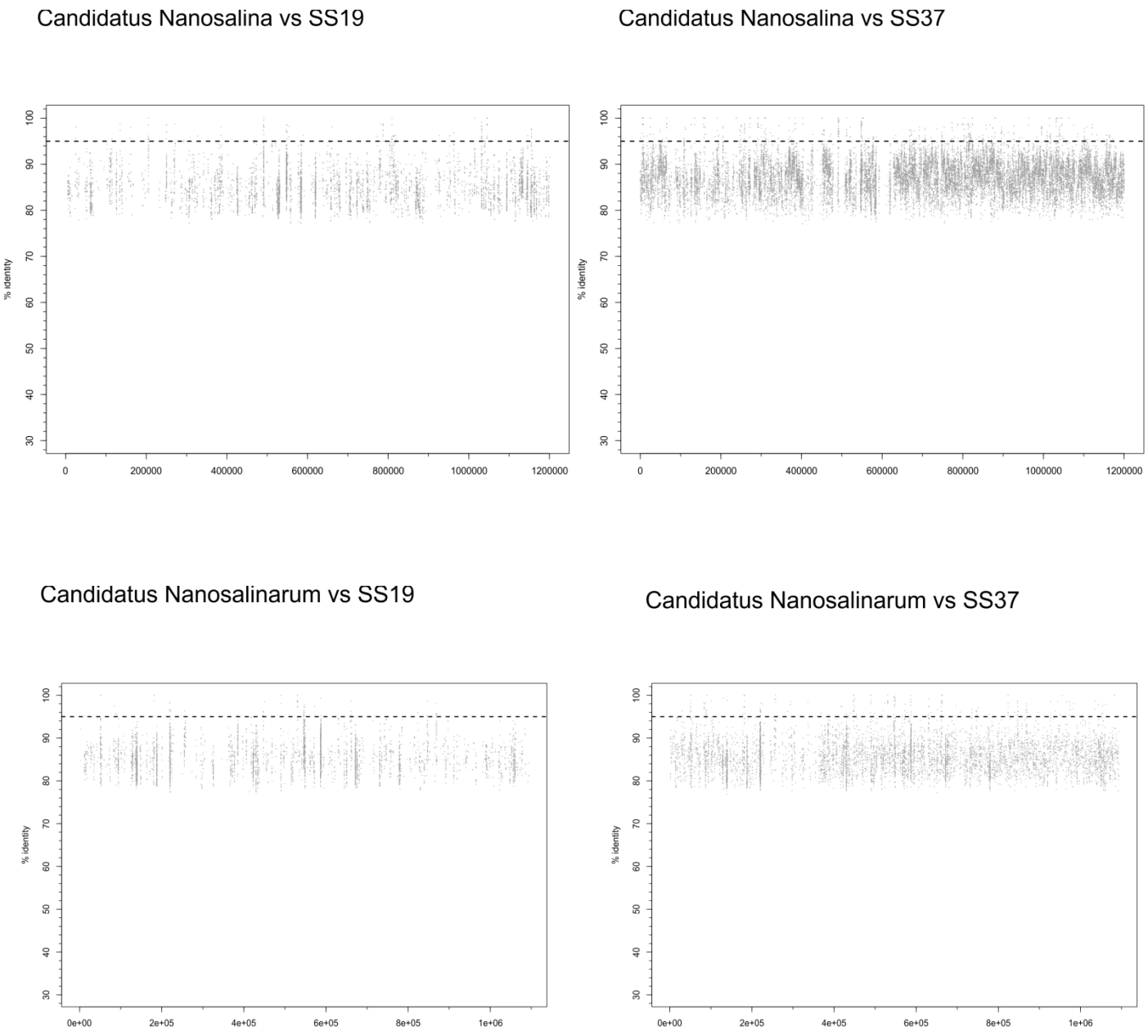
Supplementary Figure S12.

Maximum likelihood phylogenetic tree of 16S rRNA sequences from diverse archaea, including Nanohaloarchaea (*Candidatus* Haloredivivus, *Candidatus* Nanosalina sp. and *Candidatus* Nanosalinarum sp.). Bootstrap values are shown as well.



Supplementary Figure S13.

Recruitment of *Candidatus* Nanosalina and *Candidatus* Nanosalinarum against SS19 and SS37 datasets. The comparison is done using BLASTN (minimium alignment length 50). %Identity is shown on the X-axis while the Y-axis represents the genome (concatenated contigs in this case). A dashed horizontal line is shown to indicate 95% sequence identity.



Supplementary Figure S14.

Direct nucleotide comparison of *H. walsbyi* contigs assembled from the SS19 metagenome to the genome of *H. walsbyi*. The alignment length is shown on the X-axis and the % of query coverage. On the Y-axis. "Green" color indicates contigs in which all genes had a best hit to *H. walsbyi* and yellow color indicates contigs that had all genes but one that belonged to *H. walsbyi* (but still had a best hit to Euryarchaeota)