Supplementary Information for

Evidence of microbial rhodopsins in Antarctic Dry Valley edaphic systems

Leandro D. Guerrero, Surendra Vikram, Thulani P. Makhalanyane and Don A. Cowan

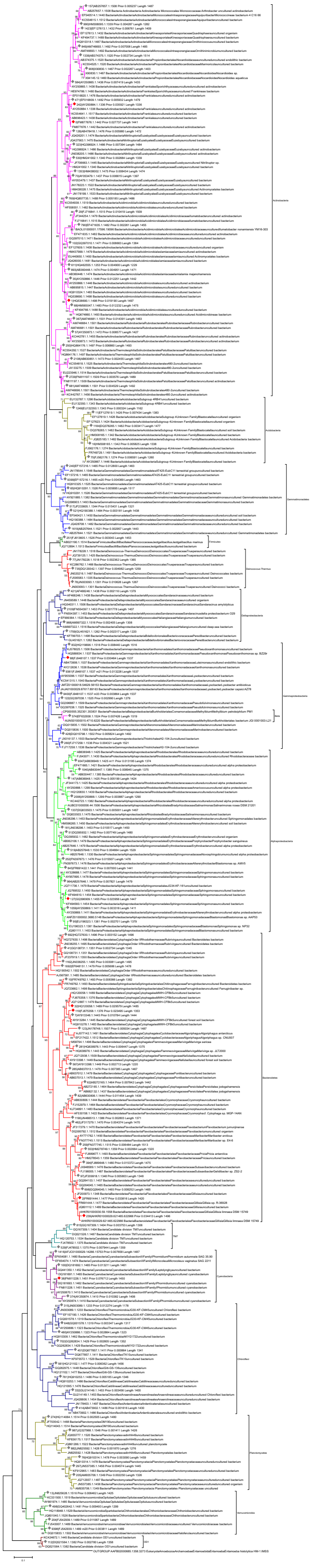
Centre of Microbial Ecology and Genomics, Department of Genetics, University of Pretoria, Pretoria, South Africa.

* Corresponding Author: Prof. Don A. Cowan,
Centre for Microbial Ecology and Genomics,
University of Pretoria,
Lynwood Road, 0028 Pretoria,
South Africa

Email: don.cowan@up.ac.za

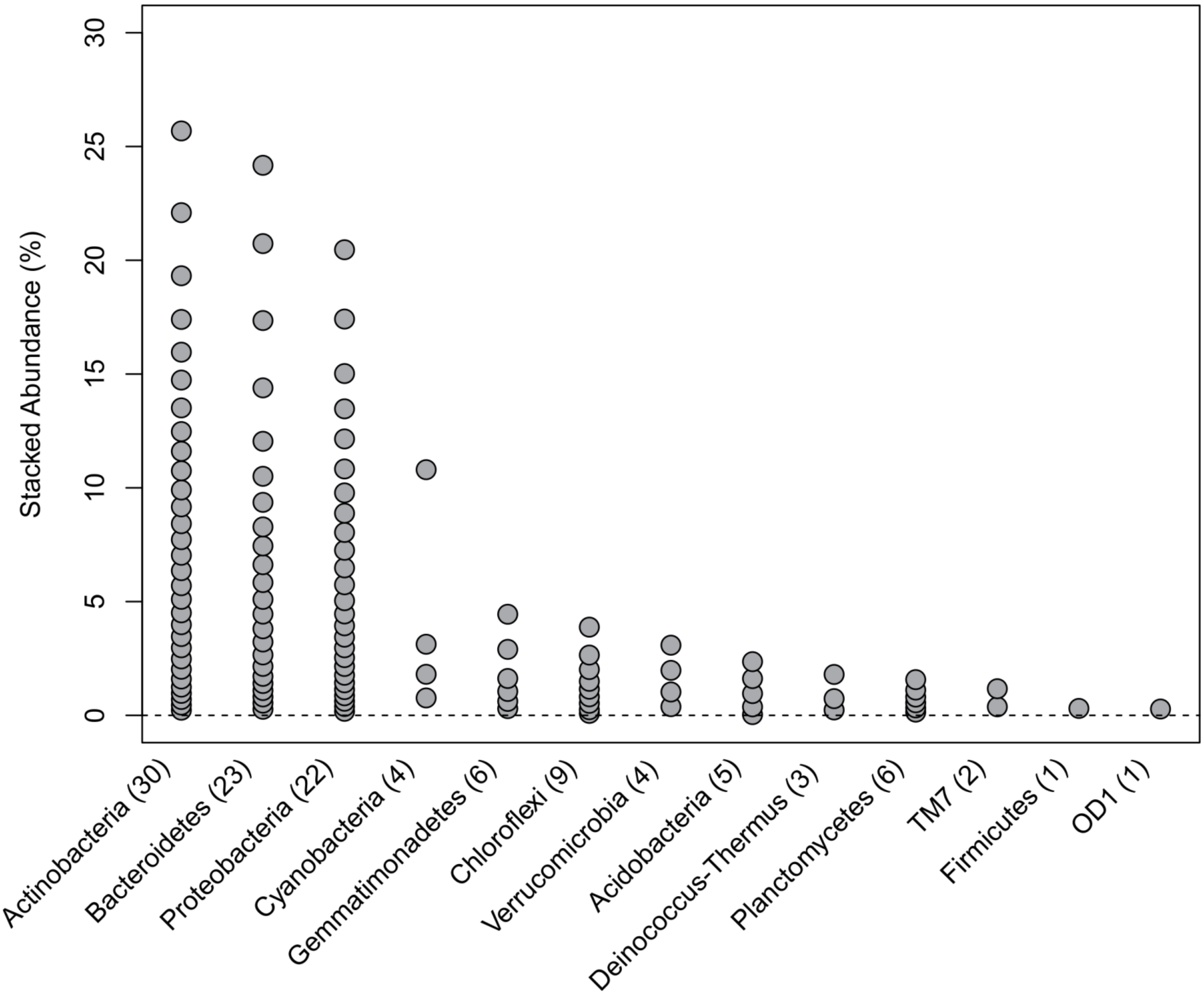
Supplementary Figures 1 to 11

Supplementary Tables 1 to 5



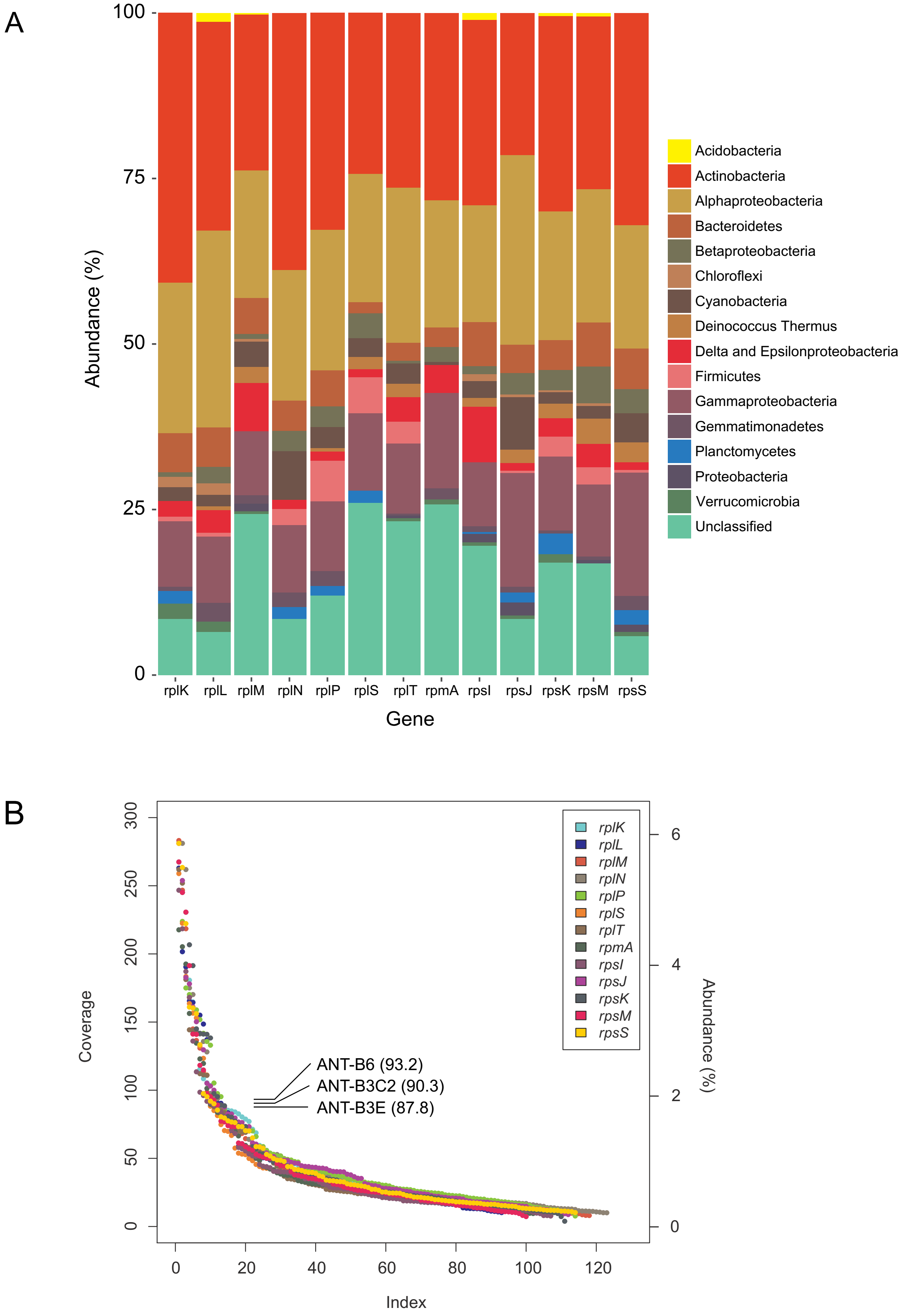
Supplementary Fig. 1

Maximum Likelihood tree of small rRNA genes reconstructed with EMIRGE from the metagenome. OTUs from the metagenome are marked with diamonds. The ten most abundant OTUs are shown in red. For each metagenome sequence the accession number of the starting candidate sequence, the relative abundance estimated by EMIRGE and the sequence length are indicated after the sequence number. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. All positions containing gaps and missing data were eliminated. There were a total of 649 positions in the final dataset. Bootstrap support values >50% for nodes are given.



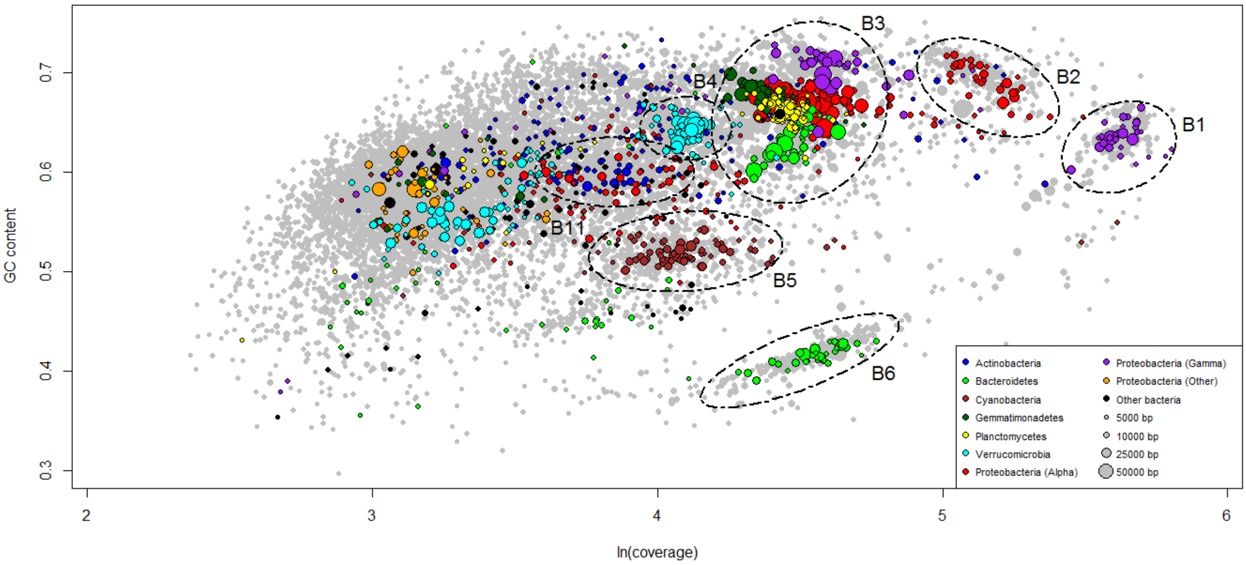
Supplementary Fig. 2

Stacked abundance of metagenomics 16S rRNA genes. OTUs (97% similarity) were classified according with phylum affiliation. The number of OTUs in each phylum is indicated in brackets. Detailed information of the abundance of each OTU is shown in Supplementary Fig. 1.



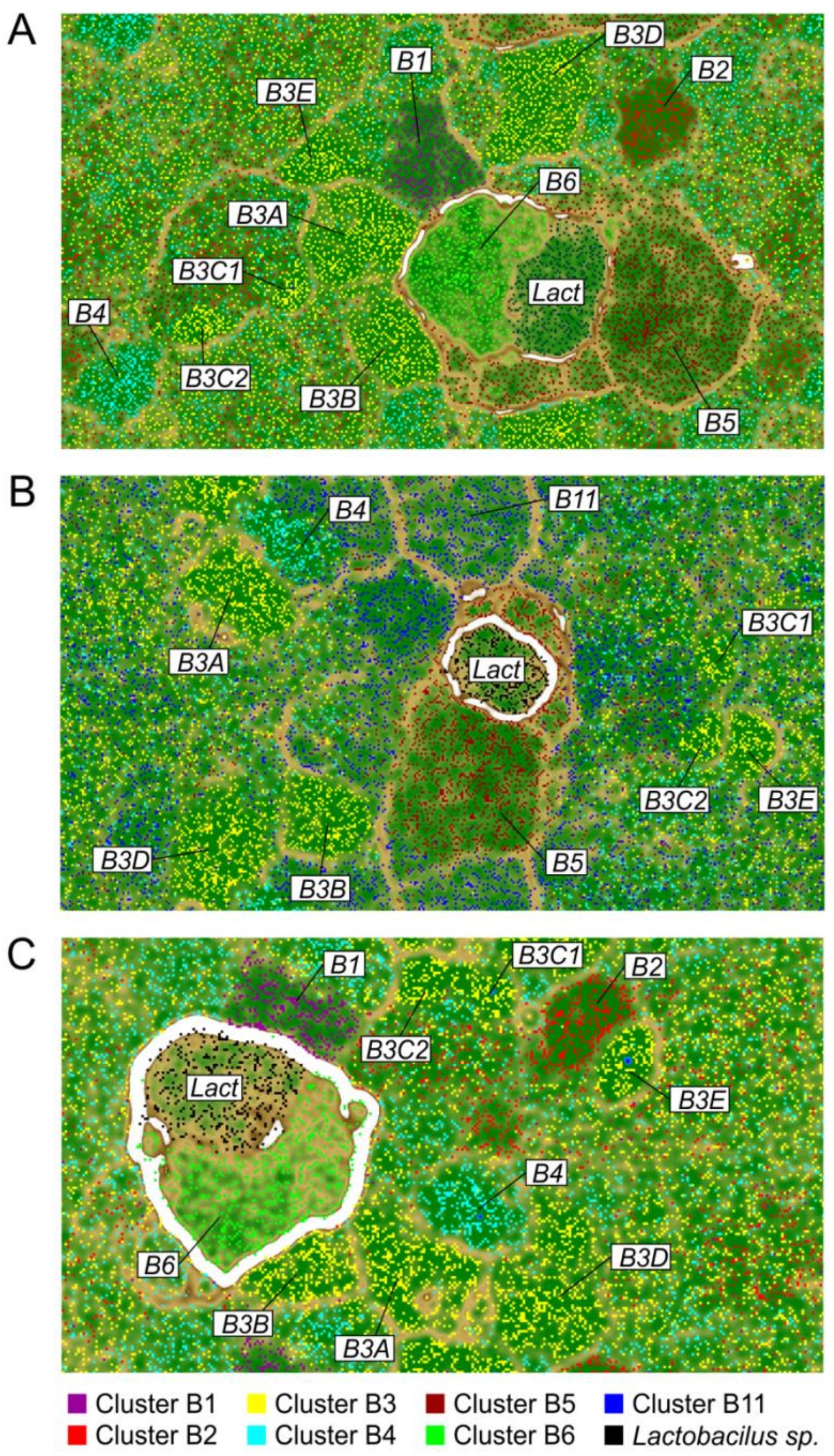
Supplementary Fig. 3

(A) Relative abundance for single copy marker genes shorter than 150 aa identified in the metagenome. (B) Coverage of marker genes. For each marker gene, observed hits were sorted according with their coverage value (x axis). Average genome coverage values of identified genomes with rhodopsin-like genes are shown by arrows.



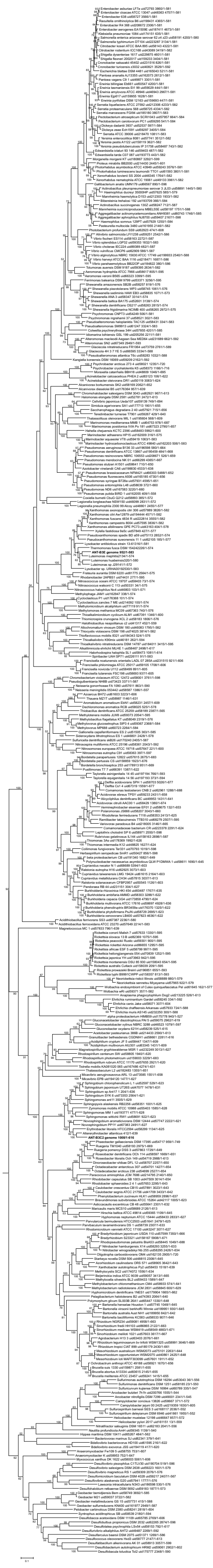
Supplementary Fig. 4

Metagenome contigs plotted according to their GC content and coverage. Coloured dots show contigs with essential genes classified at phylum level (Class level for Alpha and Gamma-proteobacteria). The dotted line ellipses show clusters with identified genomes. Groups of same colour contigs inside clusters belong to single genomes or phylogenetically close genomes with similar abundance (coverage) and GC content. ANT-B3C2 and ANT-B3E genomes were extracted from the B3 cluster after been resolved using tetranucleotide frequencies. Circles represent only contigs longer than 5000 bp and the size is proportional to the contig length.



Supplementary Fig. 5

Tetranucleotide frequencies ESOM maps. Each graph (A, B and C) represents a different ESOM run. Dots represent contigs coloured according with the clusters identification in Fig. 2. Isolated areas (bins) in a single colour may represent a possible individual genome. *Lactococcus* sp. was used as a control (*Lact*).

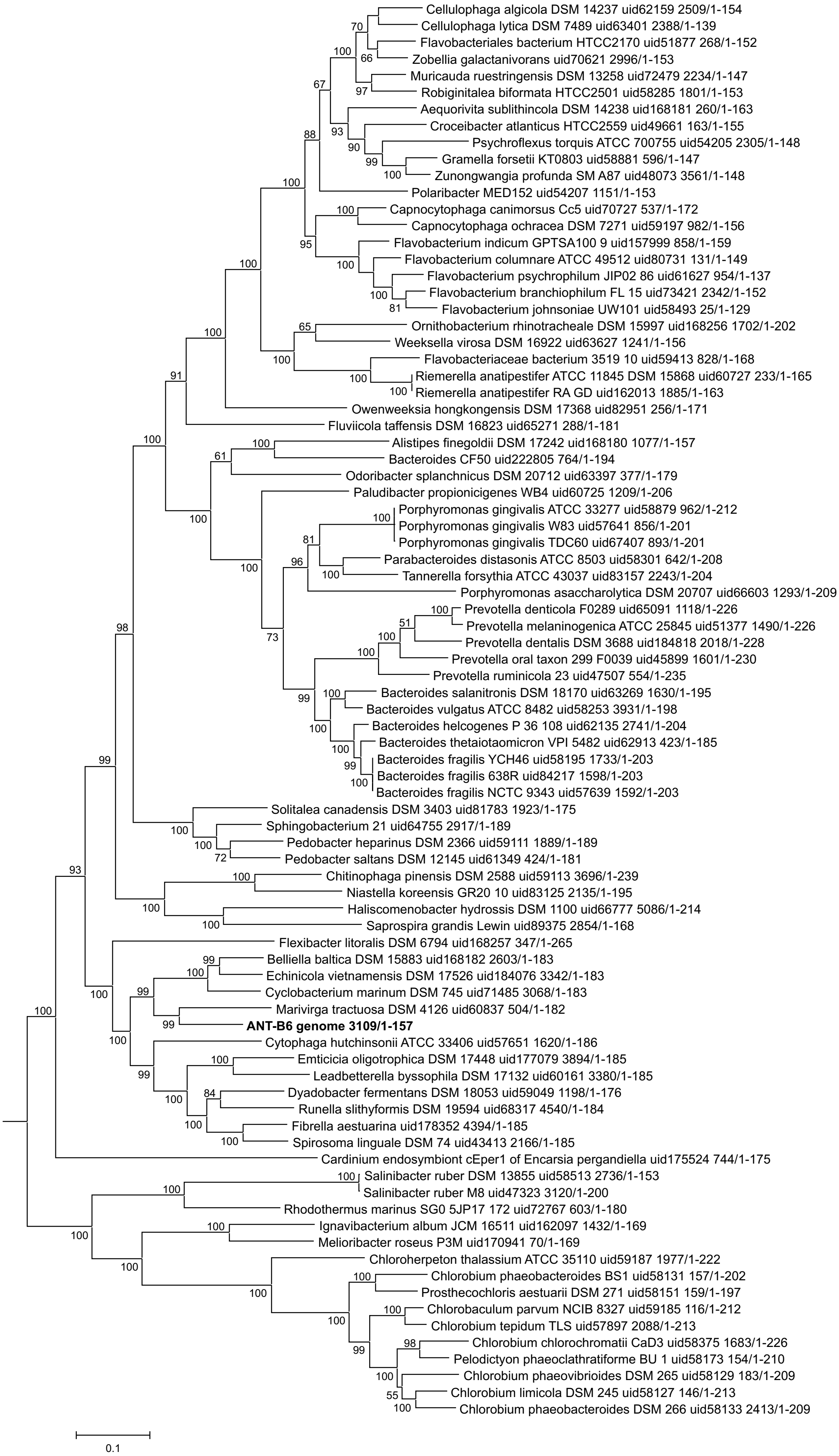


Supplementary Fig. 6

Proteobacteria maximum likelihood tree based on 12 concatenated marker proteins. Alignment of 370 Proteobacteria genomes, including ANT-B3C2 and ANT-B3E genomes. Values higher than 50%, based on 500 bootstrap replicates are shown.

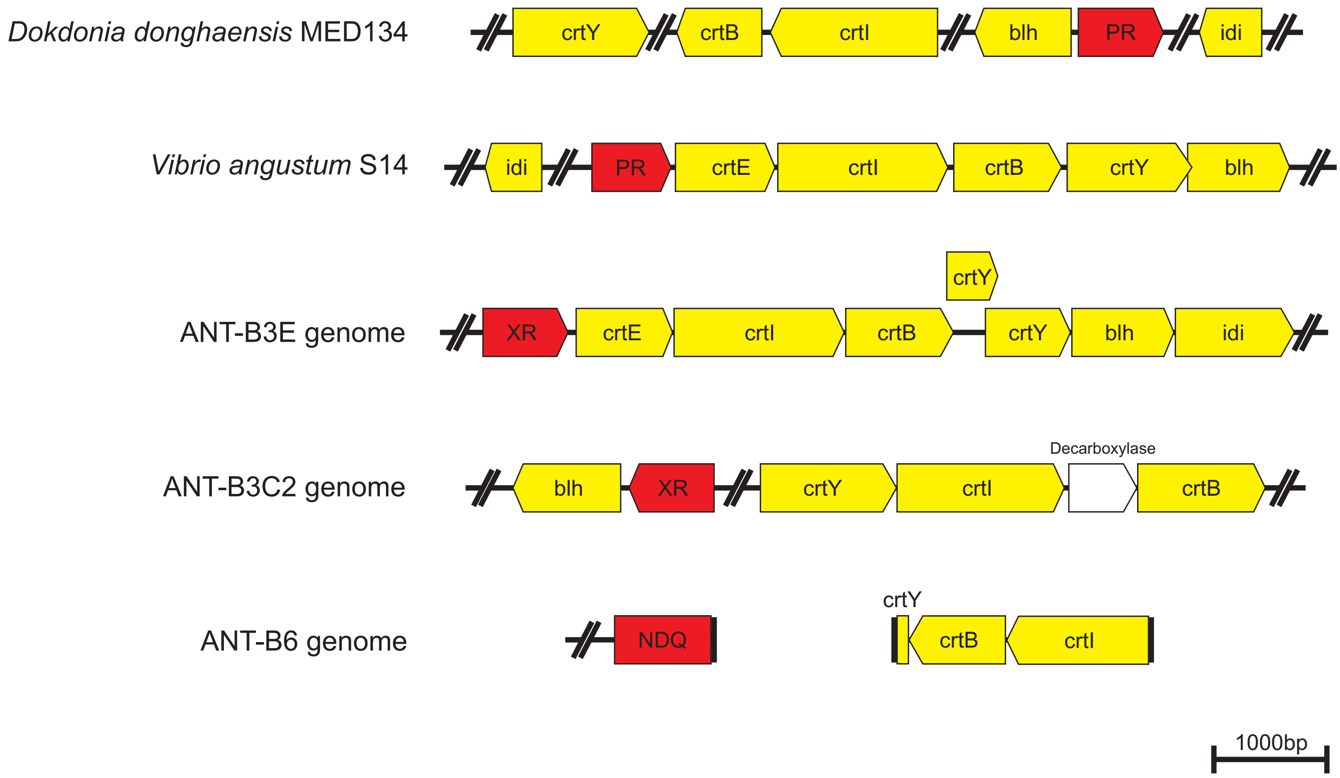
Supplementary Fig. 7

Maximum likelihood tree of assembled genomes based on the 16S rRNA gene alignment. (A) Sphingomonadales group tree including the ANT-B3C2 genome. (B) Lysobacteriales group tree including the ANT-B3E genome. Values higher than 50%, based on 500 bootstrap replicates are shown.



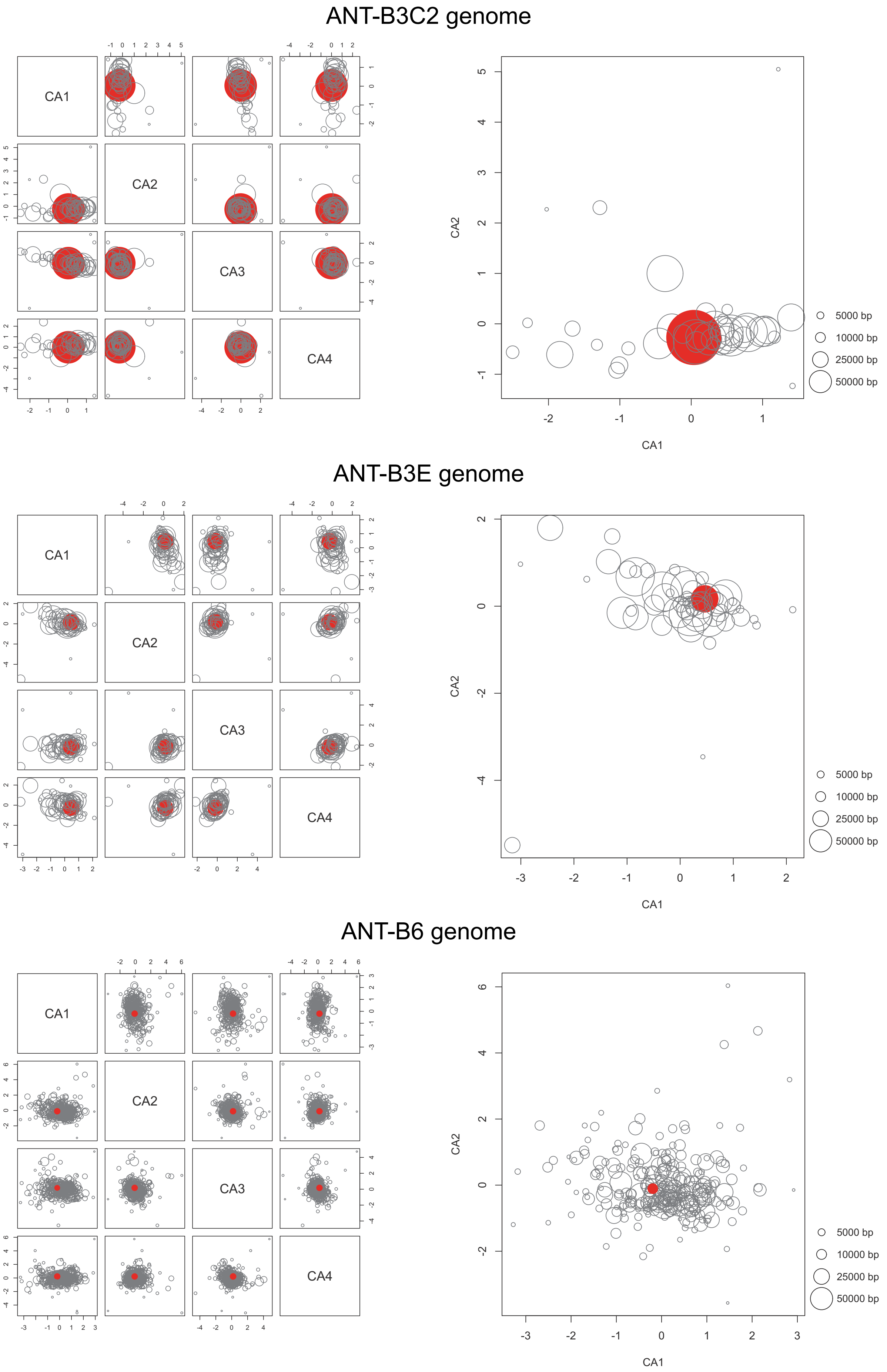
Supplementary Fig. 8

Bacteroidetes maximum likelihood tree based on 27 concatenated marker proteins. Alignment of 85 Bacteroidetes genomes, including the ANT-B6 genome. Values higher than 50%, based on 500 bootstrap replicates are shown.



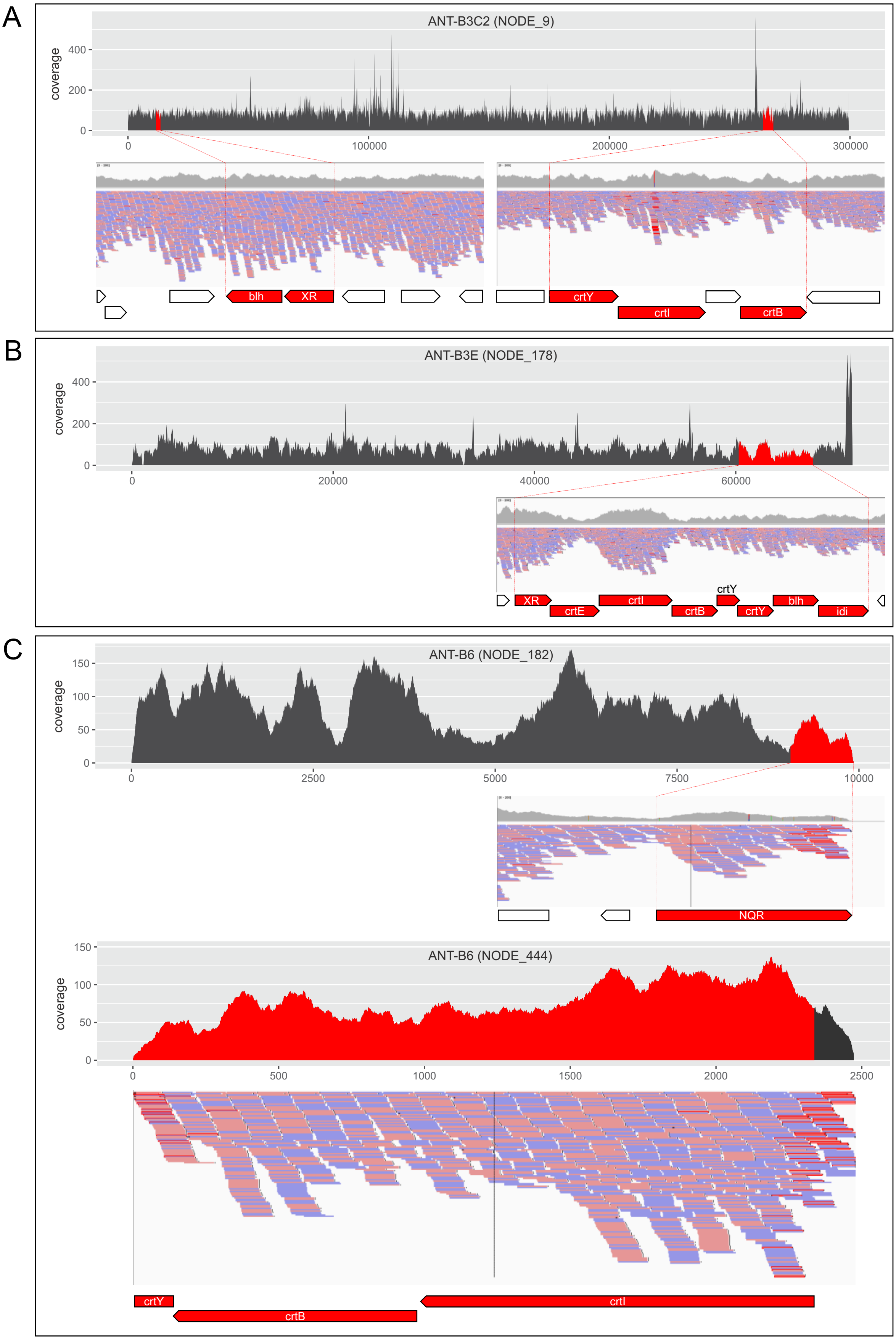
Supplementary Fig. 9

Schematic representation of the rhodopsin and retinal biosynthesis genes. Genes involves in the retinal biosynthesis are coloured in yellow. All the genes were identified in a single contig for the ANT-B3E and ANT-B3C2 genomes, whereas in ANT-B6 only four genes were identified in two different contigs. Single vertical bars indicate contig ends.



Supplementary Fig. 10

Correspondence analysis based on the contigs nucleotide composition of assembled genomes. Contigs carrying the rhodopsin coding sequence and most of the retinal synthesis genes are coloured red.



Supplementary Fig. 11

Schematic representation of the rhodopsin and retinal synthesis genes harbouring contigs. A) ANT-B3C2 genome. B) ANT-B3E genome. C) ANT-B6 genome. Whole contig coverage is represented by histograms. Coding sequences of rhodopsin and retinal synthesis genes are coloured red and zoomed to show paired-end reads assemble detail and genes disposition. Unrelated genes in the neighbourhood are shown as empty boxes.

BR position	55	56	57	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	212	213	214	215	216
GPR position	74	75	76	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	227	228	229	230	231
BR *Halobacterium*	T	**M**	Y	**R**	Y	A	**D**	W	L	F	T	T	P	L	L	L	L	**D**	**D**	V	S	A	**K**
PR *P. ubique*	I	**H**	Y	**R**	Y	I	**D**	W	L	I	T	V	P	L	L	M	L	**E**	**D**	F	L	N	**K**
NODE_65683 (1108)	M	**H**	Y	**R**	Y	I	**D**	W	T	L	T	V	P	L	M	C	V	**E**	**D**	A	V	N	**K**
NDQ *N. dokdonensis*	L	**L**	L	**R**	Y	L	**N**	W	L	I	D	V	P	M	L	L	F	**Q**	**D**	V	C	S	**K**
NODE_103019 (2518)	F	**L**	L	**R**	Y	L	**N**	W	L	I	D	V	P	L	L	L	V	**Q**	**D**	V	V	S	**K**
NODE_186857 (623)	I	**L**	L	**R**	Y	L	**N**	W	L	I	D	V	P	L	L	L	V	**Q**	**D**	V	T	A	**K**
NODE_265058 (1358)	F	**L**	L	**R**	Y	L	**N**	W	L	I	D	V	P	L	L	L	V	**Q**	**D**	V	L	S	**K**
NODE_315601 (2788)	L	**E**	L	**R**	Y	V	**N**	W	S	I	D	V	P	V	L	L	L	**Q**	**D**	I	T	S	**K**
NODE_575713 (1101)	L	**L**	L	**R**	Y	L	**N**	W	L	I	D	V	P	M	L	L	F	**Q**	**D**	I	L	S	**K**
NODE_749161 (1365)	I	**L**	L	**R**	Y	L	**N**	W	L	I	D	V	P	L	L	L	V	**Q**	**D**	V	T	A	**K**
NODE_766858 (1161)	L	**L**	L	**R**	Y	L	**N**	W	L	I	D	V	P	M	L	L	F	**Q**	**D**	V	C	S	**K**
NODE_500788 (3258)	-	-	-	**R**	Y	V	**N**	W	S	I	D	V	P	M	L	L	I	**Q**	**D**	I	V	S	**K**
NTQ C. *bathyomarinum*	L	**T**	L	**R**	Y	G	**N**	W	T	I	T	V	P	I	L	L	A	**Q**	**D**	I	A	S	**K**
NODE_166979 (1907)	L	**L**	F	**R**	Y	V	**N**	W	L	V	T	V	P	V	L	L	V	**Q**	**D**	I	G	S	**K**
NODE_205846 (1661)	L	**S**	L	**R**	Y	G	**N**	W	T	I	T	V	P	L	L	L	V	**Q**	**D**	I	S	T	**K**
NODE_307690 (1720)	M	**M**	L	**R**	Y	V	**N**	W	L	I	T	I	P	L	L	L	V	**Q**	**D**	V	T	S	**K**
NODE_378809 (1393)	I	**L**	L	**R**	Y	L	**N**	W	I	V	T	V	P	V	L	L	S	**Q**	**D**	L	F	S	**K**
XR *S. ruber*	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	T	V	**E**	**D**	V	L	A	**K**
NODE_309213 (2093)	Y	**H**	Y	**R**	Y	V	**D**	W	F	L	T	V	P	L	L	L	V	**E**	**D**	I	L	A	**K**
NODE_351663 (1266)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	L	I	**E**	**D**	L	A	A	**K**
NODE_393394 (5845)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	L	L	**E**	**D**	I	L	A	**K**
NODE_407503 (2693)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	V	I	**E**	**D**	V	L	A	**K**
NODE_519481 (2448)	-	-	-	**R**	Y	M	**D**	W	L	I	T	V	P	L	L	V	V	**E**	**D**	L	T	A	**K**
NODE_164387 (1242)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	L	I	**E**	**D**	L	A	A	**K**
NODE_536313 (6326)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	V	I	**E**	**D**	V	L	A	**K**
NODE_197819 (1115)	Y	**H**	Y	**R**	Y	V	**D**	W	F	L	T	V	P	L	L	L	V	**E**	-	-	-	-	-
NODE_241799 (438)	Y	**H**	Y	**R**	Y	V	**D**	W	F	L	T	V	P	L	L	L	A	**E**	-	-	-	-	-
NODE_696600 (2135)	Y	**H**	Y	**R**	Y	I	**D**	W	L	L	T	V	P	L	L	L	L	**E**	**D**	I	L	A	**K**
NODE_245604 (929)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	L	I	**E**	**D**	L	I	A	**K**
NODE_769115 (1399)	Y	**H**	Y	**R**	Y	V	**D**	W	F	L	T	V	P	L	L	L	V	**E**	-	-	-	-	-
NODE_83277 (3277)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	M	V	**E**	**D**	I	L	A	**K**
Ant17 (XR)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	L	I	**E**	**D**	L	I	A	**K**
ANT-B3C2 (XR)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	L	I	**E**	**D**	I	V	A	**K**
ANT-B3E (XR)	Y	**H**	Y	**R**	Y	V	**D**	W	L	L	T	V	P	L	L	M	V	**E**	**D**	L	V	A	**K**
ANT-B6 (NDQ)	L	**L**	L	**R**	Y	L	**N**	W	L	I	D	V	P	M	L	L	F	**Q**	-	-	-	-	-

Supplementary Table 1

Rhodopsin relevant functional residues. Conserved histidine in XR and PR (56). Residues involved in proton transfer and in complex counter-ion formation (82, 85, 96 and 212). Lysine residue forming the pSB with the retinal (216). Leucine at position 93 corresponds to green absorbing rhodopsin. (Bamann *et al.*, 2014). Contig lengths are shown between brackets. Reference sequences: BR: *Halobacterium sp.* NRC-1, GPR: Green PR from the Gammaproteobacteria SAR86, PR: *Pelagibacter ubique* HTCC1062, NDQ: *Nonlabens dokdonensis*, NTQ: *Citromicrobium bathyomarinum*, XR: *Salinibacter ruber* DSM-13855 and Ant17: *Sphingobium sp.* Ant17 isolated from Antarctic soil (Adriaenssens *et al.*, 2014).

Gene	**Contigs (Number of hits)**	**Total coverage**
rplK	117	5169
rplL	93	4423
rplM	118	4845
rplN	123	5512
rplP	114	5379
rplS	107	4386
rplT	107	4421
rpmA	102	4246
rpsI	107	4222
rpsJ	112	5229
rpsK	111	5079
rpsM	100	4482
rpsS	114	4952
Average (marker genes)	109.6	4796
Rhodopsins	29	961

Supplementary Table 2

Single copy marker genes shorter than 150 aa identified in the metagenome. The contigs column represents the number of contigs coding marker genes identified. The average number of hits and average coverage was estimated for each marker gene. The number and total coverage of rhodopsins was also included in the table.

bin	Contigs	ORF	Bin size (bp)	ES	CheckM	Taxonomic affiliation
B1	151	2314	2,537,318	81	77.76 (0.00)	Proteobacteria; Gammaproteobacteria; Xanthomonadales; Xanthomonadaceae
B2	198	2311	2,446,588	80	79.90 (0.15)	Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae
B3A	329	3082	3,611,287	99	90.06 (0.00)	Planctomycetes; Planctomycetia; Planctomycetales; Planctomycetaceae
B3B	198	2932	3,416,621	105	96.05 (0.30)	Bacteroidetes; Bacteroidetes Order II. Incertae sedis; Rhodothermaceae
B3C1	101	2790	2,912,805	128	96.55 (11.00)	Proteobacteria; Alphaproteobacteria; Sphingomonadales; Erythrobacteraceae
B3C2	96	2664	2,857,999	118	89.75 (1.72)	Proteobacteria; Alphaproteobacteria; Sphingomonadales; Sphingomonadaceae
B3D	383	3351	3,754,181	77	80.62 (3.30)	Gemmatimonadetes; Gemmatimonadales; Gemmatimonadaceae
B3E	174	2600	2,878,469	72	78.45 (5.17)	Proteobacteria; Gammaproteobacteria; Xanthomonadales; Xanthomonadaceae
B4	340	3637	4,005,278	104	93.02 (4.28)	Verrucomicrobia; Verrucomicrobiae; Verrucomicrobiales; Verrucomicrobiaceae
B5*	1359	7105	7,029,346	195	94.83 (57.19)	Cyanobacteria; Oscillatoriophycideae
B6	469	3466	3,747,645	122	85.86 (15.44)	Bacteroidetes; Bacteroidetes; Cytophagia; Cytophagales; Cytophagaceae
B11	400	2441	2,357,107	78	75.13 (0.44)	Actinobacteria; Rubrobacteridae; Rubrobacterales; Rubrobacterineae; Rubrobacteraceae

Supplementary Table 3

Preliminary genomes identified in the binning. Taxonomic affiliation is based on the closest neighbour species identified using RAST-SEED (Aziz *et al.*, 2008). ES: Number of essential genes identified after binning. CheckM: Percentage of completeness and contamination calculated using the CheckM software.*Two possible different close related genomes were identified according with the number of essential genes and CheckM result.

Gene*	Proteobacteria			Bacteroidetes	
	ANT-B3E	ANT-B3C2	Alignment 370 genomes†	ANT-B6	Alignment 85 genomes†
dnaG	1	1	yes	1	no
frr	1	1	yes	nd	no
infC	1	1	yes	1	yes
nusA	1	1	no	1	yes
pgk	nd	1	no	1	yes
pyrG	1	1	yes	1	yes
rplA	nd	1	no	1	yes
rplB	nd	1	no	1	yes
rplC	nd	1	no	1	yes
rplD	nd	1	no	1	yes
rplE	nd	1	no	1	yes
rplF	nd	1	no	1	yes
rplK	nd	1	no	1	yes
rplL	nd	1	no	1	yes
rplM	1	1	yes	1	yes
rplN	nd	1	no	1	yes
rplP	nd	1	no	1	yes
rplS	1	1	yes	nd	no
rplT	1	1	yes	1	yes
rpmA	1	1	yes	nd	no
rpoB	nd	1	no	1	yes
rpsB	1	1	yes	1	yes
rpsC	nd	1	no	1	yes
rpsE	nd	1	no	1	yes
rpsI	1	1	yes	1	yes
rpsJ	nd	1	no	1	yes
rpsK	nd	1	no	1	yes
rpsM	nd	1	no	1	yes
rpsS	nd	1	no	1	yes
smpB	1	1	yes	1	yes
tsf	1	1	yes	1	yes

Supplementary Table 4

Marker gene identified in the assembled genomes (ANT-B3E, ANT-B3C2 and ANT-B6) and used to construct the concatenated alignment to compare with reference genomes from NCBI GeneBank. *Genes used by Amphora2 software. †Genes that were not detected or appeared in more than one copy in the reference genomes were not used for the alignments. nd: Not detected.

Sample	ng/µl	260/280
MVHyp1	11.2	1.77
MVHyp2	12.3	1.75
MVHyp3	18.0	1.68
MVHyp4	13.3	1.83
MVHyp5	9.9	1.72
MVHyp6	8.2	1.72
MVHyp7	6.7	1.65
MVHyp8	8.3	1.59
MVHyp9	14.2	1.43
MVHyp10	7.7	2.0
MVHyp11	1.6	3.79
MVHyp12	5.8	1.83
MVHyp13	6.6	1.74
MVHyp14	17.4	1.82
MVHyp15	10.8	2.00
MVHyp16	16.7	1.80
MVHyp17	2.1	1.16
MVHyp18	4.6	1.61
MVHyp19	9.9	1.87
MVHyp20	25.0	1.86
MVHyp21	5.4	1.53
MVHyp22	9.1	1.78
MVHyp23	3.9	2.10
MVHyp24	4.4	1.62
MVHyp25	4.0	1.70
MVHyp26	4.2	1.57
MVHyp27	4.6	1.67
MVHyp28	3.4	1.81
MVHyp29	0.9	1.07
MVHyp30	4.6	1.68
MVHyp31	12.0	1.73
MVHyp32	2.2	1.23
MVHyp33	3.4	1.95
MVHyp34	8.2	1.77
MVHyp35	6.0	2.07
MVHyp36	4.1	1.46
MVHyp37	1.5	1.61
MVHyp38	1.2	1.77
MVHyp39	5.8	1.80
MVHyp40	3.5	1.61
MVHyp41	2.1	1.7
MVHyp42	4.8	1.91
MVHyp43	1.6	1.43
MVHyp44	0.7	2.87
MVHyp45	4.4	1.33
MVHyp46	3.8	1.94
MVHyp47	6.0	1.54
MVHyp48	9.4	1.61
MVHyp49	1.0	1.60
MVHyp50	0.9	1.14

Supplementary Table 5

DNA concentration and absorption relation (260/280nm) of hypolith samples quantified using the NanoDrop.