

1. GOS metagenomic data: These data are from samples collected from 52 GOS sites. The data set consists of 10.97 million reads, of which 8.4 million were produced using Sanger sequencing and 2.54 million were produced using 454 Titanium sequencing. All sequence data used here are available at NCBI (under project ID 13694) and have also been submitted to CAMERA¹.

The Sanger data used here include published data² from the following GOS sites: GS00b (Sargasso Stations 13 and 11), GS00c (Sargasso Station 3), GS00d (Sargasso Station 13), GS01a (Sorcerer Hydrostation S), GS01b (Sorcerer Hydrostation S), GS01c (Sorcerer Hydrostation S), GS02 (Just outside Canadian water), GS03 (Browns bank), GS04 (Outside Halifax), GS05 (Bedford Basin), GS06 (Bay of Fundee), GS07 (Transit outside of Maine), GS08 (Newport Harbor, Rhode Island), GS09 (Block Island), GS10 (Cape May), GS13 (11.3 miles off of Nags Head N.C.), GS14 (South of Charleston S.C), GS15 (Off Dry Tortugas), GS16 (Northwest of western most Cuba), GS17 (Yucatan Channel, Mexico), GS18 (Rosario bank, Honduras), GS19 (Northeast of Colon, Panama Canal), GS21 (Gulf of Panama), GS22 (250 miles from Panama City), GS23 (30 miles from Coco's Island, Ecuador), GS25 (Dirty Rock Coco's Island, Ecuador), GS26 (134 miles NE of Galapagos), GS27 (Devil's Crown Off of Floreana Island, Galapagos), GS28 (Off the Coast of Floreana, Galapagos), GS29 (North James Bay, Santiago), GS30 (Thermal Vent off of Roca Redonda, Galapagos), GS31 (Off Fernandina Island, Galapagos), GS34 (North Seamore Island, Galapagos), GS35 (Wolf Island, Galapagos Islands), GS36 (Cabo Marshall, Isabella Island Galapagos), GS37 (Equatorial Pacific TAO Buoy), and GS51 (Rangiroa Atoll, French Polynesia). For the purposes of this study we excluded the GS00a sample from the analysis because it contained an exceptionally high abundance of the non-planktonic *Burkholderia* and *Shewanella*-like genomic sequences^{2,3}.

We also used the following Sanger data, which have been submitted to NCBI's Trace Archive and CAMERA: GS48 - Inside Cook's Bay, Moorea (0.1µm, 0.8µm, and 3.0µm); GS49 - Outside Cook's Bay, Moorea (0.1µm); GS108 - Inside Lagoon, Cocos Keeling, Australia (0.1µm, 0.8µm, and 3.0µm); GS109 - Indian Ocean (0.1µm); GS110 - Indian Ocean (0.1µm, 0.8µm, and 3.0µm); GS111 - Indian Ocean (0.1µm); GS112 - Indian Ocean (0.1µm, 0.8µm, and 3.0µm); GS113 - Indian Ocean (0.1µm); GS114 - Indian Ocean, 500 miles west of Seychelles (0.1µm); GS115 - Indian Ocean (0.1µm); GS116 - Indian Ocean, Outside Seychelles (0.1µm); GS117 - Indian Ocean, St. Anne Island, Seychelles (0.1µm and 0.8µm); GS119 - Outside of Reunion Island (0.1µm); GS120 - Madagascar Waters (0.1µm); GS121 - Region between Madagascar and South Africa (0.1µm); GS122 - International Waters between Madagascar and S. Africa (0.1µm, 0.8µm, and 3.0µm); GS123 - International Waters Between Madagascar and S. Africa (0.1µm); GS148 - E. coast Zanzibar, open ocean (0.1µm); GS149 - W. coast Zanzibar, harbor region (0.1µm).

In addition, the following 0.1µm filter 454 data sets were used in this study and have been submitted to NCBI's Sequence Read Archive and CAMERA:

GS108 (library identifier GOS108XLRVAL-4F-1-400) and GS112 (library identifier GOS112XLRVAL-4F-1-400).

2. GOS 16S PCR libraries: This includes data from 35 sites - GS02, GS03, GS04, GS05, GS06, GS07, GS08, GS09, GS10, GS11, GS12, GS13, GS14, GS15, GS16, GS17, GS18, GS19, GS21, GS22, GS23, GS25, GS27, GS28, GS29, GS30, GS31, GS34, GS35, GS36, GS108, GS110, GS112, GS117, and GS122. The library data for sites GS08, GS11, GS12, GS19, and GS22 have been published⁴. The remaining data are available in NCBI's Trace Archive under the following library identifiers:

GOS_GS02-R-16S-01-01-1P5KB, GOS_GS03-R-16S-01-01-1P5KB, GOS_GS04-R-16S-01-01-1P5KB, GOS_GS05-R-16S-01-01-1P5KB, GOS_GS06-R-16S-01-01-1P5KB, GOS_GS07-R-16S-01-01-1P5KB, GOS_GS09-R-16S-01-01-1P5KB, GOS_GS10-R-16S-01-01-1P5KB, GOS_GS10-R-16S-01-02-1P5KB, GOS_GS13-R-16S-01-01-1P5KB, GOS_GS14-R-16S-01-01-1P5KB, GOS_GS15-R-16S-01-01-1P5KB, GOS_GS16-R-16S-01-01-1P5KB, GOS_GS17-R-16S-01-01-1P5KB, GOS_GS18-R-16S-01-01-1P5KB, GOS_GS21-R-16S-01-01-1P5KB, GOS_GS21-R-16S-01-03-1P5KB, GOS_GS21-R-16S-01-04-1P5KB, GOS_GS23-R-16S-01-01-1P5KB, GOS_GS25-R-16S-01-01-1P5KB, GOS_GS27-R-16S-01-01-1P5KB, GOS_GS28-R-16S-01-01-1P5KB, GOS_GS29-R-16S-01-01-1P5KB, GOS_GS30-A-A1-01-01-0P6KB, GOS_GS-30-A-A1-01-02-0P6KB, GOS_GS-30-B-B1-01-01-1P5KB, GOS_GS-30-B-B1-02-1P5KB, GOS_GS-30-PCR-01-01-1P6KB, GOS_GS31-R-16S-01-01-1P5KB, GOS_GS34-R-16S-01-01-1P5KB, GOS_GS34-R-16S-01-02-1P5KB, GOS_GS34-R-16S-01-04-1P5KB, GOS_GS34-R-16S-01-05-1P5KB, GOS_GS35-R-16S-01-01-1P5KB, GOS_GS35-R-16S-01-03-1P5KB, GOS_GS35-R-16S-01-05-1P5KB, GOS_GS36-R-16S-01-01-1P5KB, GOS_GS36-R-16S-01-05-1P5KB, GOS_GS110-R-01-1P5KB, GOS_GS117-R-01-1P5KB, GOS_GS117-R-02-1P5KB, GOS_GS117-R-03-1P5KB, GOS_GS122-R-01-1P5KB, GOS_GSIOSM108-R-01-1P5KB, GOS_GSIOSM110-R-01-1P5KB, GOS_GSIOSM110-R-02-1P5KB, GOS_GSIOSM112-R-01-1P5KB, GOS_GSIOSM117-R-01-1P5KB, GOS_GSIOSM122-R-01-1P5KB, and GOS_GSIOSM122-R-02-1P5KB.

3. Previously sequenced marine genomes used in this study: we used 60 previously sequenced marine genomes. These genomes were sequenced to closure. Their names and GenBank accessions are: *Acaryochloris marina* MBIC11017 (158303474, 158309729, 158310112, 158310521, 158310884, 158311158, 158311381, 158311570, 158311744), *Alcanivorax borkumensis* SK2 (110645972), *Candidatus Pelagibacter ubique* HTCC1062 (71061822), *Chlorobaculum parvum* NCIB 8327 (193085153), *Chlorobium phaeobacteroides* BS1 (189494441), *Cyanothece* ATCC 51142 (172034820, 172034859, 172034885, 172034917, 172054853), *Desulfovibrio vulgaris* Miyazaki F (218756063), *Dinoroseobacter shibae* DFL 12 (157910316, 157913894,

157914090, 157914228, 157914364, 157914435), *Geobacillus kaustophilus* HTA426 (56378334, 56378377), *Gramella forsetii* KT0803 (120434372), *Haloarcula marismortui* ATCC 43049 (55228832, 55228869, 55228912, 55228953, 55229005, 55229137, 55229304, 55229667, 55232799), *Hyphomonas neptunium* ATCC 15444 (114737225), *Jannaschia CCS1* (88862040, 88866253), *Magnetococcus MC-1* (117607074), *Maricaulis maris MCS10* (114339016), *Marinobacter aquaeolei* VT8 (120322793, 120326657, 120326872), *Marinomonas MWYL1* (150834967), *Microcystis aeruginosa* NIES 843 (166085114), *Nitrosococcus oceani* ATCC 19707 (76881831, 76881875), *Nitrosopumilus maritimus* SCM1 (160338813), *Petrogobus mobilis* SJ95 (160359135), *Photobacterium profundum* SS9 (46911589, 47419843, 47419844), *Pirellula* sp (32456059), *Prochlorococcus marinus* CCMP1375 (33238865), *Prochlorococcus marinus* MED4 (33772317), *Prochlorococcus marinus* MIT9313 (33772318), *Prochlorococcus marinus* MIT 9215 (157386913), *Prochlorococcus marinus* MIT 9301 (126542380), *Prochlorococcus marinus* MIT 9312 (78711884), *Prochlorococcus marinus* NATL2A (154949252), *Prosthecochloris aestuarii* DSM 271 (194310671, 194312935), *Pseudoalteromonas haloplanktis* TAC125 (76873893, 76876834), *Psychromonas ingrahamii* 37 (119862398), *Rhodopseudomonas palustris* TIE 1 (192282182), *Roseobacter denitrificans* OCh 114 (109453537, 115292851, 115292957, 115293014, 115293031), *Saccharophagus degradans* 2-40 (89949249), *Shewanella* MR-4 (113883030), *Shewanella* MR-7 (113886955, 113890962), *Shewanella baltica* OS155 (125995462, 125999774, 125999867, 125999939, 125999952), *Shewanella baltica* OS185 (151363173, 151367497), *Shewanella baltica* OS195 (160858456, 160858524, 160858598, 160858648), *Shewanella baltica* OS223 (217496351, 217500742, 217500901, 217501047), *Shewanella denitrificans* OS217 (91713371), *Shewanella frigidimarina* NCIMB 400 (114332481), *Shewanella loihica* PV-4 (126636230), *Shewanella woodyi* ATCC 51908 (169809717), *Silicibacter* TM1040 (99034845, 99035147, 99036121), *Silicibacter pomeroyi* DSS-3 (56676665, 56680476), *Sphingopyxis alaskensis* RB2256 (95105268, 98975575), *Synechococcus* CC9311 (113880062), *Synechococcus* CC9605 (78196017), *Synechococcus* CC9902 (78167878), *Synechococcus* RCC307 (147849409), *Synechococcus* WH 7803 (147846875), *Synechococcus elongatus* PCC 7942 (81167692, 81170305), *Synechococcus* sp WH8102 (33772316), *Thermosiphon africanus* TCF52B (217035986), *Thermotoga* RQ2 (170175310), *Trichodesmium erythraeum* IMS101 (110164990), *Vibrio* *harveyi* ATCC BAA-1116 (156523975, 156527546, 156529921).

4. Previously sequenced non-marine genomes used in this study: we used 740 previously sequenced non-marine genomes. These genomes were sequenced to closure. Their names and GenBank accessions are: *Acholeplasma laidlawii* PG 8A (161984995), *Acidiphilium cryptum* JF-5 (146400198, 146400370, 146400540, 146400618, 146400651, 146400690, 146400702), *Acidithiobacillus ferrooxidans* ATCC 23270 (218517237), *Acidithiobacillus ferrooxidans* ATCC 53993 (198246669), *Acidobacterium bacterium* Ellin345 (94549081), *Acidotherrmus cellulolyticus* 11B (117647674), *Acidovorax* JS42

(120604516, 120608524, 120608609), *Acidovorax avenae* citrulli AAC00-1 (120587178), *Acinetobacter baumannii* AB0057 (213054530, 213058321), *Acinetobacter baumannii* AB307 0294 (213985689), *Acinetobacter baumannii* ACICU (183207914, 183211582, 183211611), *Acinetobacter baumannii* ATCC 17978 (126373828, 126373843, 126385999), *Acinetobacter baumannii* AYE (169147024, 169147032, 169147044, 169147050, 169147133), *Acinetobacter baumannii* SDF (169150741, 169150750, 169150781, 169150821), *Acinetobacter* sp ADP1 (49529273), *Actinobacillus pleuropneumoniae* L20 (126096280), *Actinobacillus pleuropneumoniae* serovar 3 JL03 (165875558), *Actinobacillus pleuropneumoniae* serovar 7 AP76 (189914400, 189916532, 189916535, 189916540), *Actinobacillus succinogenes* 130Z (150839411), *Aeromonas hydrophila* ATCC 7966 (117558854), *Aeromonas salmonicida* A449 (142849896, 142855988, 142856267, 32186830), *Aeropyrum pernix* (116062158), *Agrobacterium radiobacter* K84 (221721649, 221725460, 221727862, 221728079, 221728482), *Agrobacterium tumefaciens* C58 Cereon (159139455, 159140696, 159141512, 159141737), *Agrobacterium vitis* S4 (221734005, 221737306, 221738391, 221738525, 221738762, 221739013, 221739573), *Akkermansia muciniphila* ATCC BAA 835 (187424568), *Aliivibrio salmonicida* LFI1238 (208007585, 208010409, 208011468, 208011498), *Alkalilimnicola ehrlichei* MLHE-1 (114225560), *Alkaliphilus metalliredigens* QYMF (149947715), *Alkaliphilus oremlandii* OhILAs (158139258), *Anabaena variabilis* ATCC 29413 (75699950, 75705040, 75705385, 75705417), *Anaerocellum thermophilum* DSM 6725 (222454904), *Anaeromyxobacter* Fw109-5 (152026452), *Anaeromyxobacter* K (196170270), *Anaeromyxobacter dehalogenans* 2CP-C (85772941), *Anaeromyxobacter dehalogenans* 2CP 1 (219952977), *Anaplasma marginale* Florida (222418877), *Anaplasma marginale* St Maries (56387602), *Anaplasma phagocytophilum* HZ (88597753), *Anoxybacillus flavithermus* WK1 (212559329), *Aquifex aeolicus* (6626248), *Archaeoglobus fulgidus* (6626247), *Arcobacter butzleri* RM4018 (157698125), *Aromatoleum aromaticum* EbN1 (56311475, 56315609, 56315884), *Arthrobacter* FB24 (116608677, 116612825, 116612996, 116613120), *Arthrobacter aureus* TC1 (119947346, 119951388, 119951681), *Arthrobacter chlorophenolicus* A6 (219857661, 219861547, 219862102), *Aster yellows witches-broom phytoplasma* AYWb (84789385), *Azoarcus* BH72 (119668705), *Azorhizobium caulinodans* ORS 571 (158328513), *Bacillus amyloliquefaciens* FZB42 (154350369), *Bacillus anthracis* Ames (30260185), *Bacillus anthracis* Ames 0581 (47552137, 50082967, 50118566), *Bacillus anthracis* str Sterne (49176966), *Bacillus cereus* AH187 (217062892, 217068328, 217068352, 217068626), *Bacillus cereus* AH820 (218534755, 218540236, 218540557), *Bacillus cereus* ATCC14579 (29899096), *Bacillus cereus* ATCC 10987 (42740913, 42741405), *Bacillus cereus* B4264 (218158707), *Bacillus cereus* G9842 (218540569, 218546058, 218546310), *Bacillus cereus* Q1 (221237819, 221243025, 221243254), *Bacillus cereus* ZK (51973633, 66970195, 66970632, 66970687, 66970696), *Bacillus cereus* cytotoxis NVH 391-98 (152022606), *Bacillus clausii* KSM-K16 (56908016), *Bacillus halodurans* (47118318), *Bacillus licheniformis* ATCC 14580 (145902672), *Bacillus licheniformis* DSM 13 (52346357), *Bacillus subtilis*

(38680335), *Bacillus thuringiensis* Al Hakam (118415003, 118419740), *Bacillus thuringiensis* konkukian (49328240, 56800306), *Bacillus weihenstephanensis* KBAB4 (163860210, 163865366, 163865656, 163865796), *Bacteroides fragilis* NCTC 9434 (60491031, 60495221), *Bacteroides fragilis* YCH46 (52214156, 52218735), *Bacteroides thetaiotaomicron* VPI-5482 (29335914, 29342101), *Bacteroides vulgatus* ATCC 8482 (149931032), *Bartonella bacilliformis* KC583 (120613812), *Bartonella henselae* Houston-1 (49237636), *Bartonella quintana* Toulouse (49239191), *Bartonella tribocorum* CIP 105476 (161016928, 161016947), *Baumannia cicadellinicola* Homalodisca coagulata (94219610), *Bdellovibrio bacteriovorus* (42494925), *Beijerinckia indica* ATCC 9039 (182632885, 182636455, 182636632), *Bifidobacterium adolescentis* ATCC 15703 (118764602), *Bifidobacterium animalis lactis* AD011 (219620149), *Bifidobacterium longum* (23395284, 58012118), *Bifidobacterium longum* DJO10A (189427298, 23194406, 23194417), *Bifidobacterium longum infantis* ATCC 15697 (213522389), *Bordetella avium* 197N (115421100), *Bordetella bronchiseptica* (33591071), *Bordetella parapertussis* (33591070), *Bordetella pertussis* (33591069), *Bordetella petrii* (163258032), *Borrelia afzelii* PKo (110890097, 110890956, 110891092, 110891129, 110891229, 110891264), *Borrelia burgdorferi* (2690008, 2690079, 2690123, 2690175, 2690224, 3253098, 6382081, 6382257, 6382302, 6382346, 6382393, 6626249), *Borrelia burgdorferi* ZS7 (218164230, 218164353, 218165262, 218165308, 218165344), *Borrelia duttonii* Ly (201083369, 201084190, 201084199, 201084213, 201084348, 201084532, 201084551, 201084580, 201084601), *Borrelia garinii* PBI (51572834, 51593839, 51593840), *Borrelia hermsii* DAH (119860731), *Borrelia turicatae* 91E135 (119861577), *Bradyrhizobium* ORS278 (146189981), *Bradyrhizobium japonicum* (47118316), *Brucella abortus* 9-941 (62195123, 62197154), *Brucella abortus* S19 (189018840, 189020808), *Brucella canis* ATCC 23365 (161334802, 161336905), *Brucella melitensis* (17986242, 17986243), *Brucella melitensis* biovar Abortus (82615033, 82939231), *Brucella ovis* (148369114, 148370077), *Brucella suis* ATCC 23445 (163673000, 163674922), *Buchnera aphidicola* (27904416), *Buchnera aphidicola* 5A *Acyrtosiphon pisum* (219624232), *Buchnera aphidicola* Cc *Cinara cedri* (116256804), *Buchnera aphidicola* Sg (21672292), *Buchnera aphidicola* Tuc7 *Acyrtosiphon pisum* (219621678), *Buchnera* sp (47118299, 7707813, 7707817), *Burkholderia* 383 (77964193, 77965403, 77968738), *Burkholderia ambifaria* MC40 6 (171991584, 171994659, 171997042, 171998010), *Burkholderia cenocepacia* AU 1054 (105891751, 105894717, 105897395), *Burkholderia cenocepacia* HI2424 (116646113, 116649273, 116651960, 116652879), *Burkholderia cenocepacia* J2315 (190714214, 190714216, 190714218, 190714220), *Burkholderia cenocepacia* MC0 3 (169814598, 169817759, 169820555), *Burkholderia cepacia* AMMD (115280044, 115283258, 115285645, 115286659), *Burkholderia mallei* ATCC 23344 (52426793, 66967517), *Burkholderia mallei* NCTC 10229 (124289161, 124291339), *Burkholderia mallei* NCTC 10247 (126238441, 126240836), *Burkholderia mallei* SAVP1 (121224242, 121226989), *Burkholderia multivorans* ATCC 17616 JGI (160340609, 160343756, 160345909, 160346733), *Burkholderia multivorans* ATCC 17616 Tohoku (189332915, 189336000,

189338131, 189338899), *Burkholderia phymatum* STM815 (184191232, 184194305, 184196655, 184198282), *Burkholderia pseudomallei* 1106a (126225085, 126229231), *Burkholderia pseudomallei* 1710b (76577973, 76581710), *Burkholderia pseudomallei* K96243 (52208053, 52211453), *Burkholderia vietnamiensis* G4 (134131840, 134132180, 134132444, 134132552, 134132802, 134134073, 134135188, 134137285), *Burkholderia xenovorans* LB400 (91685338, 91689770, 91692731), *Caldivirga maquilingensis* IC-167 (157919427), *Campylobacter concisus* 13826 (157101370), *Campylobacter curvus* 525 92 (153792987), *Campylobacter fetus* 82-40 (118413283), *Campylobacter jejuni* (30407139), *Campylobacter jejuni* 81-176 (121504137, 121505102), *Campylobacter jejuni* 81116 (157385286), *Campylobacter jejuni* RM1221 (57165696), *Campylobacter jejuni* doylei 269 97 (152938384), *Campylobacter lari* RM2100 (222538267, 222539772), *Candidatus Amoebophilus asiaticus* 5a2 (189496911), *Candidatus Azobacteroides pseudotrichonymphae genomovar* CFP2 (212548595, 212549421), *Candidatus Blochmannia floridanus* (33518905), *Candidatus Blochmannia pennsylvanicus* BPEN (71795899), *Candidatus Carsonella ruddii* PV (116235183), *Candidatus Desulfococcus oleovorans* Hxd3 (158508843), *Candidatus Desulfuridis audaxviator* MP104C (169637063), *Candidatus Korarchaeum cryptofilum* OPF8 (170173707), *Candidatus Methanoregula boonei* 6A8 (153998101), *Candidatus Methanosphaerula palustris* E1 9c (219544946), *Candidatus Phytoplasma australiense* (171853496), *Candidatus Phytoplasma mali* (193806760), *Candidatus Ruthia magnifica* Cm *Calyptogena magnifica* (118566999), *Candidatus Sulcia muelleri* GWSS (158633136), *Carboxydotherrmus hydrogenoformans* Z-2901 (77994731), *Caulobacter* K31 (167346403, 167351469, 167351685), *Caulobacter crescentus* (19172958), *Caulobacter crescentus* NA1000 (220962111), *Cellvibrio japonicus* Ueda107 (190684944), *Chlamydia muridarum* (29251571, 7190951), *Chlamydia trachomatis* (12057206), *Chlamydia trachomatis* 434 Bu (165929961), *Chlamydia trachomatis* A HAR-13 (76167241), *Chlamydia trachomatis* L2b UCH 1 proctitis (165930836), *Chlamydophila abortus* S26 3 (62147714), *Chlamydophila caviae* (29835118, 29835126), *Chlamydophila felis* Fe C-56 (89331179, 89332185), *Chlamydophila pneumoniae* AR39 (12057210), *Chlamydophila pneumoniae* CWL029 (6626250), *Chlamydophila pneumoniae* J138 (47118320), *Chlamydophila pneumoniae* TW 183 (33236960), *Chlorobium chlorochromatii* CaD3 (78170183), *Chlorobium limicola* DSM 245 (189339705), *Chlorobium phaeobacteroides* DSM 266 (119353206), *Chlorobium tepidum* TLS (21672293), *Chloroflexus* Y 400 fl (222447181), *Chloroflexus aggregans* DSM 9485 (219541215), *Chloroflexus aurantiacus* J 10 fl (163666892), *Chloroherpeton thalassium* ATCC 35110 (193087197), *Chromobacterium violaceum* (34105712), *Chromohalobacter salexigens* DSM 3043 (91795226), *Citrobacter koseri* ATCC BAA-895 (157081501, 157086505, 157086521), *Clavibacter michiganensis* NCPPB 382 (147829011, 147829040, 147829108), *Clavibacter michiganensis* sepedonicus (167728869, 169155030, 169157974), *Clostridium acetobutylicum* (14994317, 25168256), *Clostridium beijerinckii* NCIMB 8052 (149901357), *Clostridium botulinum* A (148287495, 148291070), *Clostridium botulinum* A3 Loch Maree

(169405180, 169408836), *Clostridium botulinum* A ATCC 19397 (152926829), *Clostridium botulinum* A Hall (152930382), *Clostridium botulinum* B1 Okra (169119423, 169123081), *Clostridium botulinum* B Eklund 17B (187720473, 187723947), *Clostridium botulinum* E3 Alaska E43 (188497817), *Clostridium botulinum* F Langeland (152933790, 152937426), *Clostridium cellulolyticum* H10 (219997787), *Clostridium difficile* 630 (115249003), *Clostridium kluyveri* DSM 555 (146336987, 152206095), *Clostridium kluyveri* NBRC 12016 (219567067, 219570536), *Clostridium novyi* NT (118133308), *Clostridium perfringens* ATCC 13124 (110673209), *Clostridium perfringens* SM101 (110681940, 110684521), *Clostridium phytofermentans* ISDg (160426828), *Clostridium tetani* E88 (28204652, 28208715), *Clostridium thermocellum* ATCC 27405 (125712750), *Colwellia psychrerythraea* 34H (71143482), *Coprothermobacter proteolyticus* DSM 5265 (206737695), *Corynebacterium diphtheriae* (38231477), *Corynebacterium efficiens* YS-314 (47118314), *Corynebacterium glutamicum* ATCC 13032 Bielefeld (41349502), *Corynebacterium glutamicum* ATCC 13032 Kitasato (42602314), *Corynebacterium jeikeium* K411 (68262661), *Corynebacterium urealyticum* DSM 7109 (171850984), *Coxiella burnetii* (29542648, 71066702), *Coxiella burnetii* CbuG Q212 (212010078), *Coxiella burnetii* Dugway 7E9-12 (154355007, 154357066), *Coxiella burnetii* RSA 331 (161761524, 161761570), *Cupriavidus taiwanensis* (170937689, 170938221, 170938722), *Cyanobacteria bacterium* Yellowstone A-Prime (86553275), *Cyanobacteria bacterium* Yellowstone B-Prime (86556045), *Cyanothece* PCC 7424 (218169741, 218174969, 218175206, 218175382, 218175446), *Cyanothece* PCC 7425 (219862254, 219867221, 219867382, 219867555), *Cyanothece* PCC 8801 (218165370, 218169631, 218169684), *Cytophaga hutchinsonii* ATCC 33406 (110279108), *Dechloromonas aromatica* RCB (71845263), *Dehalococcoides* CBDB1 (73659672), *Dehalococcoides ethenogenes* 195 (57223978), *Deinococcus geothermalis* DSM 11300 (110746139, 158342661, 94554390), *Deinococcus radiodurans* (11612676, 11612677, 6460827, 6460959), *Delftia acidovorans* SPH-1 (160361034), *Desulfatibacillum alkenivorans* AK 01 (218759244), *Desulfotobacterium hafniense* DCB 2 (219536331), *Desulfotalea psychrophila* LSv54 (50874889, 50878008, 50878110), *Desulfotomaculum reducens* MI-1 (134050581), *Desulfovibrio desulfuricans* ATCC 27774 (219867585), *Desulfovibrio desulfuricans* G20 (78217452), *Desulfovibrio vulgaris* DP4 (120561280, 120564222), *Desulfovibrio vulgaris* Hildenborough (46447666, 46451220), *Desulfurococcus kamchatkensis* 1221n (218764930), *Diaphorobacter* TPSY (221728669), *Dichelobacter nodosus* VCS1703A (146232099), *Dictyoglomus thermophilum* H 6 12 (206739197), *Dictyoglomus turgidum* DSM 6724 (217335548), *Ehrlichia canis* Jake (72393774), *Ehrlichia chaffeensis* Arkansas (88599018), *Ehrlichia ruminantium* Gardel (58416339), *Ehrlichia ruminantium* Welgevonden UPISA (57160810), *Ehrlichia ruminantium* str. Welgevonden CIRAD (58417290), *Elusimicrobium minutum* Pei191 (186970583), *Enterobacter* 638 (145316543, 145320659), *Enterobacter sakazakii* ATCC BAA-894 (156530483, 156534761, 156534800), *Erwinia carotovora atroseptica* SCRI1043 (49609491), *Erwinia tasmaniensis* (188027007, 188027015, 188027120, 188027167, 188027207), *Escherichia coli*

0127 H6 E2348 69 (215263233, 215267788, 215267879), *Escherichia coli* 536 (110341805), *Escherichia coli* 55989 (218350208), *Escherichia coli* APEC O1 (115511419, 88770133, 99867038), *Escherichia coli* CFT073 (26111730), *Escherichia coli* C ATCC 8739 (169752989), *Escherichia coli* E24377A (157076497, 157076566, 157076596, 157076665, 157076671, 157076741, 157081497), *Escherichia coli* ED1a (218425442), *Escherichia coli* HS (157065147), *Escherichia coli* IAI1 (218359353), *Escherichia coli* IAI39 (218368405), *Escherichia coli* K 12 substr DH10B (169887498), *Escherichia coli* K 12 substr W3110 (85674274), *Escherichia coli* O157H7 EDL933 (3822114, 56384585), *Escherichia coli* O157 H7 EC4115 (209156984, 209157093, 209162409), *Escherichia coli* S88 (218363708), *Escherichia coli* SE11 (209910450, 209915130, 209915255, 209915368, 209915436, 209915447, 209915455), *Escherichia coli* SMS 3 5 (170517292, 170522036, 170522190, 170522201, 170522206), *Escherichia coli* UMN026 (218349957, 218430358), *Escherichia coli* UTI89 (91070629, 91075696), *Escherichia fergusonii* ATCC 35469 (218354981), *Exiguobacterium sibiricum* 255 15 (171988566), *Fervidobacterium nodosum* Rt17-B1 (154152641), *Finnegoldia magna* ATCC 29328 (167830502, 167832134), *Flavobacterium johnsoniae* UW101 (146152184), *Flavobacterium psychrophilum* JIP02 86 (149770655), *Francisella tularensis* FSC 198 (110319990), *Francisella tularensis* WY96-3418 (134048946), *Francisella tularensis* holarctica (89143280), *Francisella tularensis* holarctica OSU18 (115128880), *Francisella tularensis* mediasiatica FSC147 (187711822), *Francisella tularensis* novicida U112 (118422521), *Francisella tularensis* tularensis (56603679), *Frankia EAN1pec* (158107272), *Frankia alni* ACN14a (111147037), *Fusobacterium nucleatum* (20095250), *Geobacillus thermodenitrificans* NG80-2 (134265192, 134268585), *Geobacter* FRC 32 (221562406), *Geobacter lovleyi* SZ (189419341, 189422948), *Geobacter metallireducens* GS-15 (78192483, 78196003), *Geobacter sulfurreducens* (39985517), *Geobacter uraniumreducens* Rf4 (146395585), *Gloeobacter violaceus* (37508091), *Gluconacetobacter diazotrophicus* PAI5 (161784420, 161788202), *Gluconacetobacter diazotrophicus* PAI 5 (209529865, 209533338), *Gluconobacter oxydans* 621H (58000668, 58000832, 58000862, 58000881, 58000905), *Granulobacter bethesdensis* CGDNIH1 (114314838), *Haemophilus ducreyi* 35000HP (33149228), *Haemophilus influenzae* (6626252), *Haemophilus influenzae* 86 028NP (156617157), *Haemophilus influenzae* PittEE (148715293), *Haemophilus influenzae* PittGG (148717999), *Haemophilus parasuis* SH0165 (219690483), *Haemophilus somnus* 129PT (112822192), *Hahella chejuensis* KCTC 2396 (83630956), *Halobacterium salinarum* R1 (167726115, 167728226, 167728379, 167728580, 167728829), *Halobacterium* sp (12057215, 12057216, 2822278), *Haloquadratum walsbyi* (109624723, 109627334), *Halorhodospira halophila* SL1 (121588215), *Halorubrum lacusprofundi* ATCC 49239 (222451341, 222454053, 222454527), *Halothermothrix orenii* H 168 (219992160), *Helicobacter acinonychis* Sheeba (109713861, 109715474), *Helicobacter hepaticus* (32263428), *Helicobacter pylori* G27 (208431905, 208433399), *Helicobacter pylori* HPAG1 (107836197, 107837734), *Helicobacter pylori* J99 (12057207), *Helicobacter pylori* P12 (210132169, 210133737), *Helicobacter*

pylori Shi470 (189491895), *Heliobacterium modesticaldum* Ice1 (171696369), *Herminiimonas arsenicoxydans* (133737197), *Herpetosiphon aurantiacus* ATCC 23779 (159889572, 159894549, 159894781), *Hydrogenobaculum* Y04AAS1 (195931992), *Hyperthermus butylicus* (123977600), *Idiomarina loihiensis* L2TR (56178122), *Ignicoccus hospitalis* KIN4 I (156565779), *Janthinobacterium* Marseille (151279845), *Kineococcus radiotolerans* SRS30216 (151362970, 151363155), *Klebsiella pneumoniae* 342 (206564770, 206570325, 206570572), *Klebsiella pneumoniae* MGH 78578 (150953431, 150958208, 150958389, 150958513, 150958612, 150958618), *Kocuria rhizophila* DC2201 (183579876), *Lactobacillus acidophilus* NCFM (158967071), *Lactobacillus brevis* ATCC 367 (116098028, 116100214, 116100226), *Lactobacillus casei* (190711126), *Lactobacillus delbrueckii bulgaricus* (103422338), *Lactobacillus delbrueckii bulgaricus* ATCC BAA-365 (116092543), *Lactobacillus fermentum* IFO 3956 (183225820), *Lactobacillus gasseri* ATCC 33323 (116094265), *Lactobacillus johnsonii* NCC 533 (41584196), *Lactobacillus plantarum* (30407129, 45723550), *Lactobacillus reuteri* F275 Kitasato (183223999), *Lactobacillus sakei* 23K (78609255), *Lactobacillus salivarius* UCC118 (33321051, 33321103, 90820184, 90821902), *Lactococcus lactis* (13400022), *Lactococcus lactis cremoris* MG1363 (124491690), *Lawsonia intracellularis* PHE MN1-00 (94730694, 94731880, 94731910, 94731935), *Legionella pneumophila* Corby (148279912), *Legionella pneumophila* Lens (53752796, 53755815), *Legionella pneumophila* Paris (53749768, 53755675), *Leifsonia xyli xyli* CTCB0 (50950407), *Leptospira biflexa* serovar Patoc Patoc 1 Ames (167774246, 167777524, 167777791), *Leptospira biflexa* serovar Patoc Patoc 1 Paris (167777849, 167781241, 167781518), *Leptospira borgpetersenii* serovar Hardjo-bovis JB197 (116123488, 116127312), *Leptospira borgpetersenii* serovar Hardjo-bovis L550 (116119596, 116122449), *Leptospira interrogans* serovar Copenhageni (45602555, 45602556), *Leptospira interrogans* serovar Lai (24204616, 24204676), *Leptothrix cholodnii* SP 6 (170774137), *Leuconostoc citreum* KM20 (169803215, 169804918, 169804955, 169804976, 169804990), *Leuconostoc mesenteroides* ATCC 8293 (116096021, 116097992), *Listeria innocua* (16415787, 30407126), *Listeria monocytogenes* (30407125), *Listeria monocytogenes* 4b F2365 (83316099), *Listeria monocytogenes* HCC23 (217332573), *Listeria welshimeri* serovar 6b SLCC5334 (116740300), *Lysinibacillus sphaericus* C3 41 (168990106, 168994691), *Macrococcus caseolyticus* JCSC5402 (222119372, 222121345, 222121415, 222121420, 222121436), *Magnetospirillum magneticum* AMB-1 (82943940), *Mannheimia succiniciproducens* MBEL55E (52306107), *Mesoplasma florum* L1 (50363371), *Mesorhizobium* BNC1 (110283028, 110283346, 110287411, 110287525), *Mesorhizobium loti* (11995001, 47118328, 47118329), *Metallosphaera sedula* DSM 5348 (145701035), *Methanobacterium thermoautotrophicum* (6626257), *Methanobrevibacter smithii* ATCC 35061 (148551077), *Methanococcoides burtonii* DSM 6242 (91711097), *Methanococcus aeolicus* Nankai-3 (150013142), *Methanococcus jannaschii* (1500644, 6626255), *Methanococcus maripaludis* C5 (132662671), *Methanococcus maripaludis* C6 (159885889), *Methanococcus maripaludis* C7 (150032960), *Methanococcus maripaludis* S2 (45050763), *Methanococcus*

vanniellii SB (150011463), Methanocorpusculum labreanum Z (124362370), Methanoculleus marisnigri JR1 (125860746), Methanopyrus kandleri (19913449), Methanosaeta thermophila PT (116664776), Methanosarcina acetivorans (19918815), Methanosarcina barkeri fusaro (72394702, 72394721), Methanosarcina mazei (20986624), Methanosphaera stadtmanae (84372150), Methanospirillum hungatei JF-1 (88186784), Methylacidiphilum infernorum V4 (189184875), Methylibium petroleiphilum PM1 (124257968, 124261792), Methylobacillus flagellatus KT (91708343), Methylobacterium 4 46 (168192641, 168199251, 168199314), Methylobacterium chloromethanicum CM4 (218520385, 218525559), Methylobacterium extorquens PA1 (163661062), Methylobacterium nodulans ORS 2060 (219944660, 219952017, 219952420, 219952830, 219952884, 219952933), Methylobacterium populi BJ001 (179342784, 179348463, 179348605), Methylobacterium radiotolerans JCM 2831 (170652972, 170658659, 170659167, 170659220, 170659264, 170659306, 170659338, 170659384), Methylocella silvestris BL2 (217501576), Methylococcus capsulatus Bath (66270661), Moorella thermoacetica ATCC 39073 (83571788), Mycobacterium JLS (126232413), Mycobacterium KMS (119692146, 119697607, 119697891), Mycobacterium MCS (108767400, 108772792), Mycobacterium abscessus ATCC 19977T (169239053, 169239075), Mycobacterium avium 104 (118163506), Mycobacterium avium paratuberculosis (41400296), Mycobacterium bovis (31742509), Mycobacterium bovis BCG Pasteur 1173P2 (121491530), Mycobacterium gilvum PYR-GCK (145213092, 145218334, 145218633, 145218658), Mycobacterium leprae (30407142), Mycobacterium leprae Br4923 (219932450), Mycobacterium marinum M (183173361, 183178785), Mycobacterium smegmatis MC2 155 (118168627), Mycobacterium tuberculosis CDC1551 (50952454), Mycobacterium tuberculosis F11 (148719718), Mycobacterium tuberculosis H37Ra (148503909), Mycobacterium tuberculosis H37Rv (41353971), Mycobacterium ulcerans Agy99 (118568029), Mycobacterium vanbaalenii PYR-1 (119953846), Mycoplasma agalactiae PG2 (148291314), Mycoplasma arthritidis 158L3 1 (193001753), Mycoplasma capricolum ATCC 27343 (83283139), Mycoplasma gallisepticum (31541778), Mycoplasma genitalium (84626123), Mycoplasma hyopneumoniae 232 (53987142), Mycoplasma hyopneumoniae 7448 (71913466), Mycoplasma hyopneumoniae J (71851486), Mycoplasma mobile 163K (47458799), Mycoplasma mycoides (126252003), Mycoplasma penetrans (47118307), Mycoplasma pneumoniae (26117688), Mycoplasma pulmonis (30407144), Mycoplasma synoviae 53 (71850813), Myxococcus xanthus DK 1622 (108460647), Nanoarchaeum equitans (40068520), Natranaerobius thermophilus JW NM WN LF (179349329, 179352484), Natronomonas pharaonis (76556520, 76559182, 76559308), Neisseria gonorrhoeae FA 1090 (59717368), Neisseria gonorrhoeae NCCP11945 (193932879), Neisseria meningitidis 053442 (161594571), Neisseria meningitidis FAM18 (120865607), Neisseria meningitidis MC58 (66731897), Neisseria meningitidis Z2491 (30407145), Neorickettsia sennetsu Miyayama (88600124), Nitratiruptor SB155-2 (151421614), Nitrobacter hamburgensis X14 (91798527, 91802332, 91802572, 91802745), Nitrobacter winogradskyi Nb-255 (74419069),

Nitrosomonas europaea (30407130), Nitrosomonas eutropha C71 (114307050, 114309494, 114309550), Nitrospira multiformis ATCC 25196 (82409200, 82411958, 82411976, 82411993), Nocardia farcinica IFM10152 (54013472, 54019156, 54019317), Nocardioides JS614 (119525916, 119534933), Nostoc punctiforme PCC 73102 (186463002, 186469090, 186469351, 186469549, 186469623), Nostoc sp (17134586, 17134773, 17134864, 17134931, 47118302, 47118303), Novosphingobium aromaticivorans DSM 12444 (145322134, 145322317, 87133707), Oceanobacillus iheyensis (42632302), Ochrobactrum anthropi ATCC 49188 (151559234, 151561966, 151563660, 151563822, 151563914, 151564001), Oenococcus oeni PSU-1 (116090851), Oligotropha carboxidovorans OM5 (209871356), Onion yellows phytoplasma (162809319), Opitutus terrae PB90 1 (177839040), Orientia tsutsugamushi Boryong (146739436), Orientia tsutsugamushi Ikeda (189179678), Parabacteroides distasonis ATCC 8503 (149935098), Parachlamydia sp UWE25 (46399275), Paracoccus denitrificans PD1222 (119372524, 119376152, 119377815), Parvibaculum lavamentivorans DS-1 (154154406), Pasteurella multocida (13400023), Pediococcus pentosaceus ATCC 25745 (116101968), Pelobacter carbinolicus (90823168), Pelobacter propionicus DSM 2379 (118501159, 118504736, 118504932), Pelodictyon luteolum DSM 273 (78165794), Pelodictyon phaeoclathratiforme BU 1 (194307659), Pelotomaculum thermopropionicum SI (146272432), Phenylobacterium zucineum HLK1 (196476886, 196480416), Photorhabdus luminescens (37510968), Picrophilus torridus DSM 9790 (48429720), Polaromonas JS666 (91695138, 91700611, 91700939), Polaromonas naphthalenivorans CJ2 (120591888, 120595973, 120596279, 120596441, 120596597, 120596734, 120596787, 120596808), Polynucleobacter necessarius STIR1 (171192370), Polynucleobacter necessarius asymbioticus QLW P1DMWA 1 (145046595), Porphyromonas gingivalis ATCC 33277 (188593544), Porphyromonas gingivalis W83 (34398108), Propionibacterium acnes KPA171202 (50839098), Proteus mirabilis (172046403, 172046405), Pseudoalteromonas atlantica T6c (109698613), Pseudomonas aeruginosa (110227054), Pseudomonas aeruginosa LESB58 (218768969), Pseudomonas aeruginosa PA7 (150958624), Pseudomonas aeruginosa UCBPP-PA14 (115583796), Pseudomonas entomophila L48 (95101722), Pseudomonas fluorescens Pf-5 (68342549), Pseudomonas fluorescens Pf0 1 (77380231), Pseudomonas mendocina ymp (145573243), Pseudomonas putida F1 (148509317), Pseudomonas putida GB 1 (166857509), Pseudomonas putida KT2440 (24987239), Pseudomonas putida W619 (169757190), Pseudomonas stutzeri A1501 (145568602), Pseudomonas syringae phaseolicola 1448A (71553748, 71558731, 71558859), Pseudomonas syringae pv B728a (63253978), Pseudomonas syringae tomato DC3000 (28855971, 28856039, 28856110), Psychrobacter PRwf-1 (148570901, 148573276, 148573295), Psychrobacter arcticum 273-4 (71037566), Psychrobacter cryohalolentis K5 (92392509, 92395348), Pyrobaculum aerophilum (18308975), Pyrobaculum arsenaticum DSM 13514 (145282035), Pyrobaculum caldifontis JCM 11548 (126248349), Pyrobaculum islandicum DSM 4184 (119672928), Pyrococcus abyssi (30407140), Pyrococcus furiosus

(18980902), *Pyrococcus horikoshii* (47118297), *Ralstonia eutropha* H16 (113524807, 113528459, 32527006), *Ralstonia eutropha* JMP134 (72117119, 72120635, 72123074, 72123597), *Ralstonia metallidurans* CH34 (93352797, 93356399, 93358713, 93358955), *Ralstonia pickettii* 12J (187724002, 187727712, 187728866), *Ralstonia solanacearum* (30407127, 30407128), *Renibacterium salmoninarum* ATCC 33209 (162952245), *Rhizobium etli* CFN 42 (86279771, 86283807, 86283983, 86284147, 86284380, 86284836, 89213252), *Rhizobium etli* CIAT 652 (190694918, 190699262, 190699644, 190700001), *Rhizobium leguminosarum* bv trifolii WSM2304 (209533368, 209537694, 209538856, 209539307, 209539531), *Rhizobium leguminosarum* bv viciae 3841 (115253326, 115254108, 115254414, 115259115, 115259558, 115259715, 115259848), *Rhodobacter sphaeroides* 2 4 1 (77386265, 77386353, 77386383, 77389406, 77390242, 77390343, 77390426), *Rhodobacter sphaeroides* ATCC 17025 (145554299, 145557411, 145558214, 145558483, 145558593, 145558625), *Rhodobacter sphaeroides* ATCC 17029 (126102442, 126105563, 126106620), *Rhodobacter sphaeroides* KD131 (221158880, 221161990, 221163219, 221163365), *Rhodococcus jostii* RHA1 (110816552, 110823764, 110824911, 110825366), *Rhodoferrax ferrireducens* T118 (89343559, 89347730), *Rhodopseudomonas palustris* BisA53 (115515977), *Rhodopseudomonas palustris* BisB18 (90103542), *Rhodopseudomonas palustris* BisB5 (91680938), *Rhodopseudomonas palustris* CGA009 (39748133), *Rhodopseudomonas palustris* HaA2 (86570155), *Rhodospirillum centenum* SW (209956825), *Rhodospirillum rubrum* ATCC 11170 (83574254, 83578046), *Rickettsia akari* Hartford (157799083), *Rickettsia bellii* OSU 85-389 (157801689), *Rickettsia bellii* RML369-C (91068359), *Rickettsia canadensis* McKiel (157784452), *Rickettsia conorii* (25307999), *Rickettsia felis* URRWXC12 (67003925, 67005326, 67005395), *Rickettsia massiliae* MTU5 (157843848), *Rickettsia prowazekii* (30407152), *Rickettsia rickettsii* Iowa (165907636), *Rickettsia rickettsii* Sheila Smith (157800343), *Rickettsia typhi* wilmington (51459527), *Roseiflexus* RS-1 (148566298), *Roseiflexus castenholzii* DSM 13941 (156231356), *Rubrobacter xylanophilus* DSM 9941 (108764099), *Saccharopolyspora erythraea* NRRL 2338 (133909243), *Salinibacter ruber* DSM 13855 (83755449, 83758251), *Salinispora arenicola* CNS-205 (157914509), *Salinispora tropica* CNB-440 (145301903), *Salmonella enterica* Choleraesuis (45359286, 45758072, 62126203), *Salmonella enterica* Paratyphi ATCC 9150 (56126533), *Salmonella enterica* arizonae serovar 62 z4 z23 (160863331), *Salmonella enterica* serovar Agona SL483 (197211002, 197211055), *Salmonella enterica* serovar Dublin CT 02021853 (197936152, 197936256), *Salmonella enterica* serovar Enteritidis P125109 (206707319), *Salmonella enterica* serovar Gallinarum 287 91 (205271127), *Salmonella enterica* serovar Heidelberg SL476 (194405479, 194405605, 194405610), *Salmonella enterica* serovar Newport SL254 (133905002, 194400862, 194400866), *Salmonella enterica* serovar Paratyphi A AKU 12601 (197092687), *Salmonella enterica* serovar Paratyphi B SPB7 (161361677), *Salmonella enterica* serovar Schwarzengrund CVM19633 (194709275, 194709398, 194709404), *Salmonella enterica* serovar Typhi Ty2 (29140506), *Salmonella typhi* (16505740, 16505981, 30407157), *Salmonella typhimurium* LT2

(16445224, 16445344), *Serratia proteamaculans* 568 (157320013, 157324905), *Shewanella* ANA-3 (117610791, 117614903), *Shewanella* W3-18-1 (120556926), *Shewanella amazonensis* SB2B (119765642), *Shewanella halifaxensis* HAW EB4 (167351963), *Shewanella oneidensis* (24371479, 24371484), *Shewanella pealeana* ATCC 700345 (157844830), *Shewanella piezotolerans* WP3 (212554395), *Shewanella putrefaciens* CN-32 (145562801), *Shewanella sediminis* HAW-EB3 (157315515), *Shigella boydii* CDC 3083 94 (187426711, 187426764, 187426769, 187427012, 187763155), *Shigella boydii* Sb227 (81244029, 81248166), *Shigella dysenteriae* (141327048, 81239530, 81243805), *Shigella flexneri* 2a (18462515, 24080789), *Shigella flexneri* 2a 2457T (30043918), *Shigella flexneri* 5 8401 (110613622), *Shigella sonnei* Ss046 (141327069, 141327082, 73854091, 73858315), *Sinorhizobium medicae* WSM419 (150026743, 150030273, 150031715, 150032810), *Sinorhizobium meliloti* (25168258, 30407151, 30407155), *Sodalis glossinidius morsitans* (84778498, 84780931, 84780986), *Solibacter usitatus* Ellin6076 (116222307), *Sorangium cellulosum* So ce 56 (161158851), *Sphingomonas wittichii* RW1 (148498119, 148502970, 148503256), *Staphylococcus aureus* COL (57284222, 66775708), *Staphylococcus aureus* JH1 (149944932, 149947680), *Staphylococcus aureus* JH9 (147739516, 147742268), *Staphylococcus aureus* MW2 (47118312), *Staphylococcus aureus* Mu3 (156720466), *Staphylococcus aureus* Mu50 (14020884, 47208328), *Staphylococcus aureus* N315 (13774139, 47118324), *Staphylococcus aureus* NCTC 8325 (87201381), *Staphylococcus aureus* Newman (150373012), *Staphylococcus aureus* RF122 (82655308), *Staphylococcus aureus* USA300 (87125811, 87125817, 87125821, 87125858), *Staphylococcus aureus* USA300 TCH1516 (160367075, 160369733), *Staphylococcus aureus* MRSA252 (49240382), *Staphylococcus aureus* MSSA476 (49168429, 49243355), *Staphylococcus epidermidis* ATCC 12228 (27316888, 27316889, 27316909, 27316926, 27316949), *Staphylococcus epidermidis* RP62A (57635977, 57636010), *Staphylococcus haemolyticus* (68445725), *Staphylococcus saprophyticus* (72493824, 72496326, 72496379), *Staphylothermus marinus* F1 (126013736), *Stenotrophomonas maltophilia* K279a (190010013), *Stenotrophomonas maltophilia* R551 3 (194346582), *Streptococcus agalactiae* 2603 (22535226), *Streptococcus agalactiae* A909 (76561771), *Streptococcus agalactiae* NEM316 (30407156), *Streptococcus equi* zooepidemicus MGCS10565 (195973861), *Streptococcus gordonii* Challis substr CH1 (157074445), *Streptococcus mutans* (24378526), *Streptococcus pneumoniae* ATCC 700669 (220673408), *Streptococcus pneumoniae* CGSP14 (182628304), *Streptococcus pneumoniae* D39 (116075884), *Streptococcus pneumoniae* G54 (194356312), *Streptococcus pneumoniae* Hungary19A 6 (168994879), *Streptococcus pneumoniae* R6 (25307955), *Streptococcus pneumoniae* TIGR4 (193804931), *Streptococcus pyogenes* M1 GAS (14286347), *Streptococcus pyogenes* MGAS10270 (94543017), *Streptococcus pyogenes* MGAS10394 (50902420), *Streptococcus pyogenes* MGAS10750 (94546904), *Streptococcus pyogenes* MGAS2096 (94545005), *Streptococcus pyogenes* MGAS315 (21905618), *Streptococcus pyogenes* MGAS5005 (71852596), *Streptococcus pyogenes* MGAS6180 (71801762), *Streptococcus pyogenes*

MGAS8232 (19913450), *Streptococcus pyogenes* MGAS9429 (94541139), *Streptococcus pyogenes* Manfredo (134271134), *Streptococcus pyogenes* NZ131 (209539788), *Streptococcus pyogenes* SSI-1 (47118313), *Streptococcus sanguinis* SK36 (125496804), *Streptococcus suis* 05ZYH33 (145688469), *Streptococcus suis* 98HAH33 (145690656), *Streptococcus thermophilus* CNRZ1066 (55737978), *Streptococcus thermophilus* LMD-9 (116100249), *Streptococcus thermophilus* LMG 18311 (55736088), *Streptococcus uberis* 0140J (222113012), *Streptomyces avermitilis* (148878541, 29611241), *Streptomyces coelicolor* (17026158, 30407153, 30407154), *Streptomyces griseus* NBRC 13350 (178462309), *Sulfolobus acidocaldarius* DSM 639 (68566501), *Sulfolobus solfataricus* (25307963), *Sulfolobus tokodaii* (47118305), *Sulfurihydrogenibium* YO3AOP1 (188931022), *Sulfurovum* NBC37-1 (151423458), *Symbiobacterium thermophilum* IAM14863 (51854827), *Synechococcus* PCC 7002 (169884305, 169887133, 169887151, 169887182, 169887222, 169887332), *Synechococcus elongatus* PCC 6301 (56684969), *Synechocystis* PCC6803 (38423559, 38423692, 38423799, 38423849, 47118304), *Syntrophobacter fumaroxidans* MPOB (116696516), *Syntrophomonas wolfei* Goettingen (114336511), *Syntrophus aciditrophicus* SB (85720936), *Thermoanaerobacter* X514 (166852915), *Thermoanaerobacter pseudethanolicus* ATCC 33223 (166855265), *Thermoanaerobacter tengcongensis* (20671658), *Thermobifida fusca* YX (71914138), *Thermococcus kodakaraensis* KOD1 (57158259), *Thermococcus onnurineus* NA1 (212008101), *Thermodesulfovibrio yellowstonii* DSM 11347 (206741110), *Thermofilum pendens* Hrk 5 (119524038), *Thermomicrobium roseum* DSM 5159 (221155340, 221157271), *Thermoplasma acidophilum* (30407158), *Thermoplasma volcanium* (47118327), *Thermoproteus neutrophilus* V24Sta (170933697), *Thermosiphon melanesiensis* BI429 (149792434), *Thermosynechococcus elongatus* (47118315), *Thermotoga lettingae* TMO (157313474), *Thermotoga maritima* (12057205), *Thermotoga neapolitana* DSM 4359 (221571364), *Thermotoga petrophila* RKU-1 (147734689), *Thermus thermophilus* HB27 (46197919, 46197920), *Thermus thermophilus* HB8 (55771382, 55773356, 55773608), *Thioalkalivibrio* HL EbGR7 (219994503), *Thiobacillus denitrificans* ATCC 25259 (74055513), *Thiomicrospira crunigena* XCL-2 (110744159), *Thiomicrospira denitrificans* ATCC 33889 (78496741), *Treponema denticola* ATCC 35405 (41821838), *Treponema pallidum* (6626258), *Treponema pallidum* SS14 (189017811), *Tropheryma whipplei* TW08 27 (30407160), *Tropheryma whipplei* Twist (28476818), *Ureaplasma parvum* serovar 3 ATCC 27815 (168827424), *Ureaplasma urealyticum* (32949295), *Ureaplasma urealyticum* serovar 10 ATCC 33699 (209541489), *Verminephrobacter eiseniae* EF01-2 (121551644, 121556613), *Vibrio cholerae* (12057212, 12057213), *Vibrio cholerae* O395 (146313784, 146314918), *Vibrio fischeri* ES114 (171902228, 171902372, 59482633), *Vibrio parahaemolyticus* (47118310, 47118311), *Vibrio splendidus* LGP32 (218321107, 218324054), *Vibrio vulnificus* CMCP6 (27362712, 91983532), *Vibrio vulnificus* YJ016 (37509034, 37509038), *Wigglesworthia brevipalpis* (19071887, 47118325), *Wolbachia* endosymbiont of *Brugia malayi* TRS (58418577), *Wolbachia* endosymbiont of *Culex quinquefasciatus* Pel

(190356750), *Wolbachia* endosymbiont of *Drosophila melanogaster* (42410857), *Wolinella succinogenes* (34495246), *Xanthobacter autotrophicus* Py2 (154158043, 154162790), *Xanthomonas campestris* 8004 (66571684), *Xanthomonas campestris* ATCC 33913 (21166373), *Xanthomonas campestris* B100 (167731156), *Xanthomonas campestris* vesicatoria 85-10 (78033746, 78033769, 78033813, 78033986), *Xanthomonas citri* (21110821, 21110864, 21240769), *Xanthomonas oryzae* KACC10331 (58424217), *Xanthomonas oryzae* MAFF 311018 (84365597), *Xanthomonas oryzae* PXO99A (188518722), *Xylella fastidiosa* (12057211, 48243573, 9112238), *Xylella fastidiosa* M12 (167964044), *Xylella fastidiosa* M23 (182630682, 182632844), *Xylella fastidiosa* Temecula1 (28058114, 28058986), *Yersinia enterocolitica* 8081 (121663664, 122087364), *Yersinia pestis* Angola (162350793, 162350914, 162354752), *Yersinia pestis* Antiqua (108777911, 108782079, 108782179, 108782189), *Yersinia pestis* CO92 (30407161, 5763810, 5832423, 5834685), *Yersinia pestis* KIM (22002119, 3883003), *Yersinia pestis* Nepal516 (108773814, 108777796, 108777901), *Yersinia pestis* Pestoides F (145209020, 145212871, 145212960), *Yersinia pestis* biovar *Microtus* 91001 (45357124, 45357210, 45357241, 45357364, 45438631), *Yersinia pseudotuberculosis* IP32953 (108860311, 51587641, 51591639), *Yersinia pseudotuberculosis* IP 31758 (152958106, 152958171, 152958308), *Yersinia pseudotuberculosis* PB1 (186696370, 186700521), *Yersinia pseudotuberculosis* YPIII (169748796), *Zymomonas mobilis* ZM4 (56542470), uncultured Termite group 1 bacterium phylotype Rs D17 (170286962, 170287724, 170287734, 170287738), uncultured methanogenic archaeon RC-I (116077928).

5. Evaluation of 16S PCR based results using GOS 16S metagenomic sequences: We identified 10,312 16S sequences in the GOS metagenomic (Sanger) data. We assessed the representation of the sequenced genomes in these data by searching their 16S sequences against the metagenomic sequences. At 97% match identity, only 27% of the GOS metagenomic 16S sequences recruited to the sequenced genomes – a number comparable to the PCR analysis; furthermore, the order of the top recruiting clades was very similar to the order observed using the PCR data. We also assessed the representation of the metagenomic sequences in the PCR data - 60% of the metagenomic sequences match at $\geq 99\%$ identity to PCR sequences; this increases to 81% at $\geq 97\%$ identity. The largest uncultured OTUs identified by the PCR analysis also recruit the largest number of metagenomic sequences (see **SM3.xls**, worksheet `rank_pcr_otu_recruit_wgs.xlsx`), thereby showing that these OTUs are not the result of PCR bias. Furthermore, when the 40% of metagenomic sequences that recruit at $< 99\%$ identity (to the PCR sequences), are clustered at 97% identity, 1,664 OTUs were produced. However, most of these OTUs are of small sizes, with only two of them containing more than 50 sequences, thus showing that there are likely not many large uncultured OTUs that are metagenomic specific (i.e. missed by PCR).

6. CRISPR/Cas analysis: Some of the most intriguing apparently absent genes involves phage resistance in bacteria: the CRISPR/Cas genes. CRISPR arrays consist of series of direct repeats in a bacterial or archaeal genome interspersed with short spacer sequences acquired from previously invading viruses or plasmids. The immune response is carried out by a system of Cas (CRISPR associated) genes⁵. In addition to the analyses reported in the section *Genomic and functional characterization*, we also compared the known lifestyle data for CRISPR-rich genomes versus the same data for genomes where no CRISPR elements were found. We observed a clear correlation of CRISPR presence with the lifestyle: among CRISPR-rich genomes, 70% belonged to spatially constrained or dense environments, such as hydrothermal vents or bacterial mats, while among CRISPR-free genomes, an estimated 90% belonged to surface picoplankton (see *Methods*). We hypothesize that it is too metabolically expensive for picoplanktonic bacteria to maintain a complex system of Cas genes, given that its efficiency depends on the exact CRISPR spacer alignment against the invading phage genome⁶. The likelihood of a planktonic microbe encountering the same phage over time is reduced due to physical dispersion and mixing mechanisms. This is in contrast with microbes inhabiting matrix-attached (mats or particles) or area-restricted (hydrothermal vents) environments where the chances of repeated encounters with the same phage are higher.

References

- 1 <http://web.camera.calit2.net/>.
- 2 Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**, e77 (2007).
- 3 Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74 (2004).
- 4 Shaw, A. K. *et al.* It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**, 2200-2210, doi:EMI1626 [pii] 10.1111/j.1462-2920.2008.01626.x (2008).
- 5 Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-170, doi:327/5962/167 [pii] 10.1126/science.1179555 (2010).
- 6 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712, doi:315/5819/1709 [pii] 10.1126/science.1138140 (2007).