

1 Supplementary Information for

2 **Cryptic Speciation of a Pelagic *Roseobacter* Population Varying at a Few**
3 **Thousand Nucleotide Sites**

4
5 Xiaojun Wang, Yao Zhang, Minglei Ren, Tingying Xia, Xiao Chu, Chang Liu, Xingqin Lin,
6 Yongjie Huang, Zhuoyu Chen, Aixin Yan, Haiwei Luo*

7
8
9 *Corresponding author. Email: hluo2006@gmail.com
10

11 **This PDF file includes:**

12 Supplementary Text1. Materials and Methods
13 Supplementary Text2. Results
14 Figures S1 to S13
15 Tables S1 to S12
16 Supplementary References
17
18

19 **Text 1. Materials and Methods**

20 *1.1 Sample collection and bacterial isolation*

21 *1.2 Genome sequencing, assembly, and annotation*

22 *1.3 Ortholog prediction and phylogenomic construction*

23 *1.4 Recombination inference and population structure analysis*

24 *1.5 Divergent allele replacement inference based on the outlier d_S*

25 *1.6 Mobile genetic element prediction*

26 *1.7 Substrate utilization assays*

27 *1.8 Motility tests*

28 *1.9 Oxidative and osmotic stress sensitivity assay*

29 **Text 2. Supplementary Results**

30 *2.1 Speciation does not occur in a sympatric *Marinobacterium* population*

31 *2.2 Utilization of polyamines*

32 *2.3 Inference of the history and pattern of novel allele replacements*

33 *2.4 The pattern of *Roseobacter* population differentiation fits an existing microbial*
34 *speciation model*

35

36

37 **Text 1. Materials and Methods**

38 1.1 Sample collection and bacterial isolation

39 A 10 L and a 1 L sample of surface seawater (the upper one meter) was collected at a
40 coastal site of the South China Sea near Xiamen in China. The samples were stored at 4 °C in the
41 dark and immediately returned to the laboratory for bacterial isolation. The 1 L sample was kept
42 from shaking during transportation and was not subjected to filtration for the preservation of the
43 microenvironments. The 10 L sample was subjected to filtration for the seawater medium
44 preparation and sample dilution. The physiochemical data about the samples were shown in
45 Table S1. Seawater subsampled from the 10 L sample for flow cytometry was pre-filtered
46 through the 20 µm-pore size mesh to remove large particles and zooplankton, added to
47 glutaraldehyde (0.5% final concentration), incubated at 4 °C for 15 min in the dark, flash-frozen
48 in liquid nitrogen and then stored at -80 °C until analysis. The prokaryotic abundance was
49 estimated to be $\sim 10^6$ cells ml⁻¹ using a BD Accuri C6 flow cytometer (BD Biosciences, CA, USA)
50 by staining with 1×10^{-4} SYBR Green I (v/v, final concentration, Molecular Probes) [1].

51 Next, we employed a dilution-to-extinction method (Fig. S1) for bacterial cultivation.
52 Briefly, the 10 L sample was filtered through 0.22 µm-pore size polycarbonate filters (47 mm
53 diameter, Millipore) and autoclaved, and then was used as the seawater medium for subsequent
54 isolation. Using the *in situ* seawater rather than artificial seawater can better simulate the *in situ*
55 conditions providing multiple necessary elements to diverse bacterial species. To increase the
56 chance of obtaining diverse slow-growing species, six different types of seawater culture media
57 were prepared by supplementation of ultralow concentrations of different substrates (Table S2).
58 Among the six types of seawater media, the simplest one was supplemented with glucose only,
59 with the resulting isolates named by a prefix of “xm-g”. Three more complex ones were

60 supplemented with an amino acid mixture (xm-a), a vitamin mixture (xm-v), and a combination
61 of glucose, the amino acid mixture and the vitamin mixture (xm-m), respectively. The remaining
62 two seawater media are the most complex, as they were supplemented with the SAR11 medium
63 ingredients [2] with (xm-D) or without (xm-d) DMSP. Three different magnitudes of dilutions
64 were performed in our study, including a final concentration of ~ 10 cells ml^{-1} , ~ 5 cells ml^{-1} , and
65 ~ 1 cell ml^{-1} . Take the final concentration of ~ 1 cell ml^{-1} as an example. A volume of 1 ml
66 seawater inoculum from the 1 L sample was used to make 1:10 serial dilutions with the
67 autoclaved filtered seawater (Fig. S1). At the fourth dilution, an aliquot of 0.1 ml was dispensed
68 into a glass tube, which has already contained a 9.9 ml seawater medium along with one type of
69 the nutrient supplements mentioned above (Fig. S1). For each of the six different types of
70 nutrient supplements, 10 replicates were used, which makes a total of 60 glass tubes. After
71 growing for one month at *in situ* temperature (~ 24 °C) in the laboratory, a 200 μl subsample
72 containing bacteria mixture from the above liquid medium was spread onto a single agar plate
73 (DifcoTM Marine agar 2216). These plates were incubated at 30 °C under the dark condition.
74 Only a single colony was isolated from each plate and was subsequently purified using the
75 streaking plate technique. This would avoid collecting identical strains replicated during the
76 laboratory cultivation.

77 Genomic DNA was extracted using the TIANamp Bacteria DNA Kit (OSR-M502,
78 TIANGEN Biotech). The 16S rRNA gene sequence was amplified using the Polymerase Chain
79 Reaction (PCR). The reagents for the PCR include 25 μl Premix Taq (TAKARA, Version 2.0), 1
80 μl forward primer (final concentration 0.4 μM , 27_F: 5'-AGAGTTTGATCCTGGCTCAG-3')
81 and 1 μl reverse primer (final concentration 0.4 μM , 1492_R: 5'-
82 TACGGYTACCTTGTTACGACTT-3'), 22 μl the nuclease-free water, and 1 μl template DNA.

83 The thermocycling condition for the PCR includes the initial denaturation at 95 °C for 5 minutes,
84 followed by 30 cycles (95 °C for 45 seconds, 55 °C for 45 seconds and 72 °C for 90 seconds),
85 final extension at 72 °C for 10 minutes. After determining the 16S rRNA gene sequence, a
86 population related to *Roseovarius* in the *Roseobacter* group of Alphaproteobacteria and a second
87 population related to *Marinobacterium* of Gammaproteobacteria were identified, each consisting
88 of 16 strains showing (nearly) identical 16S rRNA gene sequences.

89

90 1.2 Genome sequencing, assembly, and annotation

91 The quality of the genomic DNA of the above 32 isolates was required to pass the
92 following criteria: $A_{260\text{nm}}/A_{280\text{nm}} > 1.8$, $A_{260\text{nm}}/A_{230\text{nm}} > 2.0$, and $A_{260\text{nm}} > A_{270\text{nm}}$, which was
93 measured using NanoDrop™ 2000 Spectrophotometer (Thermo). Genome sequencing with
94 Illumina HiSeq 2500 was used to generate paired 251 bp reads and performed at the Hubbard
95 Center for Genome Studies in the University of New Hampshire (NH, USA). The resulting raw
96 reads were trimmed using Trimmomatic v0.36 [3]. Nextera adaptors were removed, and the three
97 beginning and trailing base pairs (bps) of each read were also trimmed if the quality score is
98 lower than three. The trimmed reads each with less than 50 bp were discarded. Next, FastQC
99 v0.11.5 [4] was used to check the quality of the remaining reads. The *de novo* assembly of the
100 clean reads sequenced from each genome was performed using SPAdes assembler v3.9.1 [5]
101 with the default parameters, and contigs shorter than 1,000 bp were not used for the downstream
102 analyses.

103 To facilitate the population genomic analyses of the *Roseobacter* population, the isolate
104 xm-d-517 was additionally sequenced with a long-insert (20 kb) library using the RSII platform
105 of PacBio sequencing technology. A complete and closed genome consisting of a chromosome

106 and a plasmid of the strain was assembled based on the Illumina short reads and the PacBio long
107 reads using Unicycler v0.4.6 [6], which follows a new hybrid assembly pipeline to resolve
108 bacterial genome from a combination of short and long reads. The completeness and
109 contamination of the scaffolded assembly were evaluated for each genome using CheckM v1.0.7
110 [7]. The gene calling of each genome assembly was performed using Prokka v1.11[8], and the
111 functional annotation of each protein-coding gene was performed using Prokka, the RAST server
112 v2.0 [9], the KEGG database v82.0 [10, 11] and the CDD database v3.16 [12].

113

114 1.3 Ortholog prediction and phylogenomic construction

115 The orthologous gene families among strains in each population were identified using
116 OrthoFinder 2.2.7 [13]. For each gene family, the amino acid sequences were aligned using
117 MAFFT v7.215[14], and gaps in the alignment were trimmed using TrimAl v1.4.rev15 [15].
118 Next, the trimmed alignments were concatenated and used to construct the phylogenomic tree.
119 Considering the potentially heterogeneous evolutionary rate among different gene families, the
120 data partition model was implemented using PartitionFinder2 [16], and the estimated partition
121 scheme was incorporated in the maximum likelihood phylogenomic construction using RAxML
122 v8.1.22 [17]. The phylogeny of the *Roseobacter* population was rooted with *Aliiroseovarius*
123 *crassostreae* CV919-312 [18] and *A. crassostreae* DSM16950 (RefSeq assembly accession
124 number: GCA_001307765.1 and GCA_900116725.1), and the phylogeny of the
125 *Marinobacterium* population was rooted with *Marinobacterium* sp. AK27 (RefSeq assembly
126 accession number: GCA_000705555.1). These outgroup species were chosen because they are
127 phylogenetically distinct from, but most closely related to, the two populations under study,
128 respectively. The phylogeny of *Marinobacterium* genus (Fig. S12) was constructed with

129 additional 12 *Marinobacterium* strains, which can be retrieved from RefSeq assembly accession
130 numbers: GCA_000220545.2, GCA_000378045.1, GCA_000428985.1, GCA_000620085.1,
131 GCA_001528745.1, GCA_001651805.1, GCA_003014615.1, GCA_003250495.1,
132 GCA_004339595.1, GCA_900107855.1, GCA_900108065.1, GCA_900155945.1.

133

134 1.4 Recombination inference and population structure analysis

135 The whole genome sequences of the strains within each population were aligned using
136 progressiveMauve v2.3.1[19] with the default settings. The core genomic regions, which are
137 shared by all strains of a population and longer than 500 bp, were extracted using the
138 stripSubsetLCB module provided by Mauve [20]. With the core genome alignment and the
139 phylogenomic tree as inputs, the recombination events occurring in each population were
140 inferred using ClonalFrameML v1.1 [21], which uses maximum likelihood inference to detect
141 recombination in a computationally efficient way. The shared ancestry among the strains in the
142 population was inferred with ChromoPainter and FineStructure [22]. The inputs for the
143 ChromoPainter, including haplotype data formatted as ‘phase’ files and the recombination map
144 files, were prepared following the instructions. After generating the chunk count data, the GUI
145 version of the FineStructure was used to perform a model-based clustering using the Markov
146 Chain Monte Carlo (MCMC) approach with the default settings. Two independent runs with
147 random seed yielded consistent assignments of individuals to co-ancestral populations, indicating
148 the convergence as described in the manual. The coancestry plot was visualized using the R
149 script ‘fineRADstructure.R’ [23].

150

151 1.5 Divergent allele replacement inference based on the outlier ds

152 Allelic replacements with divergent species via homologous recombination in the core
153 gene families of the *Roseobacter* population were identified through the detection of orthologs
154 with anomalously large between-clade synonymous substitution rate (d_s), which was described
155 earlier [24–26]. Briefly, synonymous mutations are often considered neutral as they do not
156 change the amino acid sequences, and most variations in synonymous divergence among loci are
157 mainly caused by the stochastic nature of mutations across the whole genome. However, if a
158 divergent allele was acquired via recombination, the recombined loci would expectedly show
159 unusually large synonymous substitution rate (d_s), compared to the remaining loci in the genome.

160 There have been a few arguments that synonymous changes are under selection at other
161 levels. For example, nitrogen limitation and carbon limitation each were demonstrated to act as
162 selective pressures in the pelagic marine environment, which drives genomic G+C content to
163 decrease and increase, respectively, in marine bacterial populations [27, 28]. In this case,
164 mutations at all genomic sites, including synonymous sites, are under selection. However, these
165 pressures indiscriminately affect synonymous sites of all genes in the genome, which is unlikely
166 to result in a small subset of gene families with unusually large d_s values. Another potential
167 selective source at synonymous sites is codon usage bias. Alternative synonymous codons are
168 generally not used in equal frequencies, and the codon usage bias is correlated with gene
169 expression levels in fast-growing microorganisms [29, 30]. A recent study showed that codon
170 usage bias in highly expressed genes is driven by selection to maximize translation speed or
171 accuracy [31], whereas the codon usage in weakly expressed genes is thought to reflect mutation
172 pressure in the absence of selection [32, 33]. Therefore, strong translational selection reduces
173 synonymous substitution rate in highly expressed genes, while synonymous changes in weakly

174 expressed genes are randomly affected by mutation. Therefore, this mechanism cannot lead to a
175 small subset of outlier gene families with unusually large d_S values.

176 Based on the above principles, if a gene family shows that pairwise d_S values between
177 Clade R-I and Clade R-II are enormously large but pairwise d_S values within each clade are
178 extremely small, it can be inferred that the allelic replacement in this family occurred at either
179 the last common ancestor (LCA) of Clade R-I or the LCA of Clade R-II. In practice, for each
180 single-copy protein-coding gene family shared by all strains in the *Roseobacter* population, d_S
181 was estimated for all possible pairs of the strains using the yn00 module in PAML [34]. To
182 detect the core gene families showing anomalous patterns of d_S , all pairwise d_S values of all
183 single-copy core gene families were clustered using the k-means clustering algorithm. The
184 number of optimal clusters (k=2, Fig. S2A) was determined using the R package ‘NbClust’ [35],
185 which provides a variety of indices for cluster validity.

186 Next, gene trees were constructed to determine the potential phylogenetic sources for the
187 genes displaying anomalous patterns of d_S . The potential gene donors were searched against 89
188 available *Roseobacter* genomes closely related to the population under study. For each gene
189 family displaying anomalous patterns of d_S , their putative orthologous genes in the above 89
190 genomes were identified using BLASTP program with an e-value of 1e-5 [36], aligned with
191 MAFFT [14] at the amino acid sequence level, and the alignments were trimmed with TrimAl
192 [15]. The maximum likelihood phylogenetic trees were constructed using IQ-Tree [37] with the
193 parameters “-mset WAG,LG,JTT,JTTDCMut -mrate E,I,G,I+G -mfreq FU -bb 1000”.

194

195 *1.6 Mobile genetic element prediction*

196 The mobile genetic elements (MGEs) including plasmids, genomic islands, prophages,
197 insertion sequences (IS), and integrons were predicted for the 16 genomes of the *Roseobacter*
198 population. Among these, the potential plasmid sequences were predicted using plasmidSPAdes
199 [38] and Recycler [39], the potential genomic islands (GIs) were predicted using the online
200 IslandViewer 4 [40] by summarizing the results from all four methods hosted by this service, the
201 potential prophages were predicted using the PHASTER web server [41], the potential IS
202 families were predicted using ISEScan [42], and the potential integrons were predicted using
203 IntegronFinder [43] and benchmarked using *Vibrio cholerae* N16961 with known integrons [44].
204 All of these programs were implemented with default parameters.

205

206 1.7 Substrate utilization assays

207 The phenotypic microarray (PM) technology from BiOLOGTM is a high-throughput
208 method for measuring a large number of cellular properties simultaneously [45]. The technology
209 uses the color change of its patented redox as a reporter of active metabolism [46]. In this study,
210 two types of microplates including PM01 and PM02A covering 190 carbon sources were used to
211 assay the phenotypic differentiation between Clade R-I and Clade R-II of the *Roseobacter*
212 population, each represented by two strains (xm-d-517 and xm-m-339-2 for Clade R-I; xm-m-
213 314 and xm-v-204 for Clade R-II) each with three replicates. The experiment was performed
214 following the procedures recommended by the manufacturer. The strains were incubated on the
215 petri dish plate of Marine Agar 2216 (BD DifcoTM) at 30 °C overnight. Bacterial cells were
216 collected and suspended using a mixture consisting of the IF-0a inoculation medium, sterile
217 distilled water, and NaCl solution. The final concentration of NaCl was 2% (m/w), which was
218 determined from a pilot growth experiment. After adjusting the turbidity of cell suspension to 60%

219 using the BiOLOG turbidity meter, the redox Dye Mixes D (100X) was added to the mixture and
220 the final concentration was adjusted to 1.5X. Then the bacterial suspension was homogenized
221 and inoculated in each well of the plates (100 μ l). All PM plates were incubated on the OmniLog
222 instrument under 30 °C. After incubating the PM plates for five days, the raw data were collected,
223 and substrate curves were generated with the programs provided by the manufacturer. A
224 substrate curve represents the bacterium's respiration activity, a useful proxy for the traditional
225 bacterial growth curve [45]. Most substrate curves either resemble bacterial growth curves
226 indicating active utilization of these substrates or locate at the baseline indicating little utilization.
227 Next, the assayed compounds in the PM plates were linked to the KEGG metabolic pathways
228 with the *opm* package [47]. The package analyzes and compares the respiration-based growth
229 data of the four strains generated from the OmniLog platform, and identifies compounds that
230 were utilized significantly differently between strains. Based on the KEGG compound ID
231 corresponding to the compounds in the PM microplate, the program proposes candidate
232 metabolic pathways by which these compounds are utilized. Through manual inspection, we
233 established the potential link between the phenotypic variation and the genotypic variation
234 among the tested strains.

235 According to functional annotation of the core genes underlying the differentiation of the
236 *Roseobacter* population, we further tested the utilization of polyamines (putrescine and
237 spermidine) as a sole carbon source and a sole nitrogen source, respectively. The four strains
238 were inoculated in liquid medium (Difco™ Marine broth 2216) overnight at 28 °C. Cell cultures
239 were centrifuged at 4,000 rpm under room temperature for three minutes, and the pellets were
240 resuspended and washed with the basal salt medium without any carbon or nitrogen sources (See
241 Table S11 for the ingredients). Next, about 60 μ l of the cell culture was added to test tubes

242 containing 3 ml of intended media with equal initial inoculation concentration. The tubes were
243 incubated at 28 °C with shaking, and the growth of each strain was monitored by measuring
244 OD₆₀₀ every two hours. The intended media consisted of the basal medium and the supplemented
245 carbon and nitrogen source (Table S11). When spermidine and putrescine each were tested as a
246 sole carbon source, pyruvate was used as a comparison and distilled water as a negative control.
247 When each were tested as a sole nitrogen source, NH₄Cl was used as a comparison. For each test,
248 the growth of cells in marine broth 2216 (Difco™) was monitored simultaneously to ensure that
249 the growth rate of the four stains are similar given proper conditions. The statistical comparisons
250 of the between-clade growth rate were conducted with one-way ANOVA.

251

252 1.8 Motility tests

253 The swimming motility of the four strains (xm-d-517 and xm-m-339-2 for Clade R-I;
254 xm-m-314 and xm-v-204 for Clade R-II) was tested on 0.18% (w/v) soft agar marine broth plates.
255 Overnight culture of each of the strains was sub-cultured (1:200 dilution) into the marine broth
256 2216 (Difco™) at 28 °C for 4-5 hours with shaking. When OD₆₀₀ reached 0.6-0.8, a suspension
257 of cells (3 µl) was spotted at the center of a freshly prepared semi-solid swimming plate. After
258 incubation for 11 days at 28 °C, the distance of colony migration around the inoculation site was
259 evaluated by the diameter of the covered areas. The capacity of swimming motility was indicated
260 by the longest distance bacterial cells could swim on the plate. All assays were conducted in
261 triplicates. For sedimentation assays, the overnight culture of each strain was diluted (1:1) into
262 fresh marine broth 2216. 2 ml of the suspension was transferred to a 14-ml falcon test tube and
263 was allowed to sediment at room temperature for up to 24 hours without shaking. Three
264 replicates were conducted for each sample, and each experiment was performed at least twice.

265

266 1.9 Oxidative and osmotic stress sensitivity assay

267 To examine bacterial susceptibility to oxidative stress (H₂O₂) and osmotic stress (NaCl)
268 inducing agents, freshly streaked colonies of the four strains (xm-d-517 and xm-m-339-2 for
269 Clade R-I; xm-m-314 and xm-v-204 for Clade R-II) were inoculated into 2 ml marine broth 2216
270 (Difco™). After incubation at 28 °C for 24 hours with shaking, each bacterial culture was
271 diluted 1:20 in marine broth 2216 and grown at 28 °C with shaking until OD₆₀₀ 0.6~0.8. Then the
272 cell number of each strain was normalized to OD₆₀₀ as 0.5 and serially diluted with sterile
273 phosphate-buffered saline (PBS), and 5 µl cell suspensions with 10⁻³-10⁻⁶ dilution fold were
274 spotted on marine broth 2216 agar plates without or with 0.1 mM H₂O₂ or 200 mM NaCl. Plates
275 were incubated at 28 °C for 2 days, and the images of plates were recorded. Each experiment
276 was performed at least twice.

277

278 **Text 2. Supplementary results**

279 2.1 Speciation does not occur in a sympatric *Marinobacterium* population

280 The analyses in the main text demonstrated that the *Roseobacter* population is under
281 ongoing speciation and the results indicate that phycosphere is one of the likely ecological niches
282 that drove its speciation. A natural next question is whether this is a general pattern in other
283 bacterial populations in the sympatric environment. In our culture collection resulted from the
284 same sampling and bacterial isolation campaign, there is another bacterial species also consisting
285 of 16 strains related to the *Marinobacterium* genus in Gammaproteobacteria. No other species
286 have sufficient amount of closely related isolates amenable for our population genomic analyses.
287 Isolates in the *Marinobacterium* population share the 16S rRNA gene sequence identity of

288 99.90±0.13% and ANI of 98.12±4.57%. In terms of the intraspecific diversity, despite that the
289 *Marinobacterium* species and the *Roseobacter* species is nearly indistinguishable when these two
290 criteria were compared, they differ surprisingly by a factor of over 10 in the density of SNPs in
291 their core genomes (57,045 per Mbp versus 4,242 per Mbp; Table S4).

292 Although the 16 strains in the *Marinobacterium* population are similarly grouped into
293 two phylogenetic clusters (hereafter Clade M-I and Clade M-II) based on all single-copy core
294 genes (the tree shown on the left of Fig. S10), the results of ClonalFrameML show that
295 recombination occurs nearly as frequently as point mutations ($\rho/\theta=0.76$), indicating a sexual
296 population structure. In addition, seven fineSTRUCTURE populations were identified in the 16
297 strains of the *Marinobacterium* population (Fig. S10). Among these, one population consists of
298 five strains (xm-d-530, xm-g-59, xm-d-579, xm-d-564, and xm-a-152) spanning all over the
299 phylogenomic tree, indicating that frequent gene flow occurs between distinct phylogenetic
300 groups. For the remaining fineSTRUCTURE populations that share the membership of the
301 monophyletic groups shown in the phylogeny, the clustering order of the fineSTRUCTURE
302 populations shown in the dendrogram is completely different from the branching order of the
303 corresponding monophyletic groups. Therefore, it is difficult to observe any fineSTRUCTURE
304 population or monophyletic group in *Marinobacterium* that is well differentiated from the
305 remaining fineSTRUCTURE populations. Furthermore, the SNP density distribution of the
306 whole *Marinobacterium* population is similar to that of the Clade M-I in this population (Fig.
307 S11), suggesting that genetic diversity of the whole population is well represented by members
308 of Clade M-I. This pattern lends further support for its panmictic population structure with no
309 signature for genetic differentiation between Clade M-I and Clade M-II.

310 Our analyses indicate that the micro-evolutionary pattern of the *Marinobacterium* species does
311 not correlate with phycosphere or other microscale ecological niches. A simple genome content
312 analysis showed that none of the genomes in this population contain any motility and chemotaxis
313 genes (Fig. S13), which are required to establish symbiosis with phytoplankton [48–50] or
314 explore other microenvironments. Moreover, the *Marinobacterium* species under study may have
315 undergone genome reduction during its evolutionary history, as members of this species are
316 equipped with streamlined genomes (~2.3 Mbp, Table S12) and phylogenetically imbedded into
317 other *Marinobacterium* species with larger genomes (3.5-5.6 Mbp, Fig. S12). Reduction of
318 genomic G+C content, a signature of genome streamlining in marine bacteria [51], was also
319 observed (~48% versus 54%-64%; Fig. S12). These genomic features indicate that members of
320 the isolated *Marinobacterium* species are oligotrophic bacteria with streamlined genomes, which
321 are not known to actively explore microenvironments including phycospheres [51–53]. Thus, the
322 observed population differentiation potentially driven by phycosphere may be restricted to
323 bacteria equipped with more versatile metabolism like the *Roseobacter* population studied here.

324

325 2.2 Utilization of polyamines

326 A few other substrates included in the PM microplates were differentially utilized by the
327 four strains, but the pattern disagrees with the phylogenetic divide of these strains. For instance,
328 only two strains, xm-d-517 (R-I) and xm-m-314 (R-II), could use putrescine as a sole carbon
329 source (Table S10 and Fig. S5-H08), which was inconsistent with the phylogenetic grouping of
330 the four strains. Putrescine and spermidine (the latter not included in the two PM microplates)
331 are the most important short-chain polyamines, and they are mainly produced by phytoplankton
332 and other planktonic organisms and consumed as carbon and nitrogen sources by the

333 *Roseobacter* group and a few other bacterial lineages in the ocean [54]. Their uptake is thought
334 to occur through the ABC-type transporter genes encoded by the *pot* gene cluster *potABCD* [55],
335 which (from xm-d-517_03110 to xm-d-517_03113) are part of the 200 core genes subjected to
336 recombination with external species. While the protein products of *potA* and *potD* are both
337 essential to spermidine and putrescine uptake together with channel forming proteins encoded by
338 *potB* and *potC* [56], the novel allele replacement for *potA* and *potD* occurred at the LCA of
339 Clade R-II and the LCA of Clade R-I, respectively, suggesting that functional innovation may
340 have occurred in both Clade R-I and Clades R-II members. We therefore performed simple
341 growth assays and confirmed that the differential utilizations by the four representative strains of
342 putrescine and spermidine each as a sole carbon source (Fig. S6) do not match the phylogenetic
343 divide of these strains. The two polyamines each was further tested as a sole nitrogen source.
344 While the putrescine did not show differential utilization between the two clades, the spermidine
345 was utilized significantly better by Clade II members than by Clade I members (One-way
346 ANOVA, $p < 0.001$, Fig. 3G).

347

348 2.3 Inference of the history and pattern of novel allele replacements

349 Among the 200 core families, 183 have orthologs in the two genomes of *Aliiroseovarius*
350 *crassostreae*, the available lineage most closely related to the *Roseobacter* population (Fig. S8A).
351 These 183 gene trees each show close relationship between Clade R-I, Clade R-II and *A.*
352 *crassostreae*, but they differ in the branching order of the three lineages (see examples of gene
353 trees in Fig. S7A, B, C). This indicates that *A. crassostreae* or lineages related to it were the
354 sources of the novel alleles. We noted that, however, all gene trees summarizing the evolutionary
355 relationship between Clade R-I, Clade R-II and *A. crassostreae* are rooted with very distant

356 outgroups (again see Fig. S7A, B, C as examples of gene trees). It is well known that appropriate
357 outgroups allow for reliable inference of ancestral state of the lineages under comparison and
358 thus of the rooted tree topology (Fig. S8B), whereas too distant outgroups weaken the confidence
359 of the ancestral inference and the reliability of the resulting topology (Fig. S8C). Hence, the
360 evolutionary relationship of the three lineages in the gene trees inferred with the regular tree
361 building method is not reliable due to the lack of an appropriate outgroup.

362 As such, an alternative approach was employed to infer the gene tree topology, which
363 calculates the pairwise neutral genetic distances (approximated by pairwise d_S values) among the
364 three lineages and sets the root between the two lineages showing the greatest distance (i.e., the
365 midpoint rooting). Among the 183 (out of 200) core gene families that have orthologs in *A.*
366 *crassostreae*, 148 each show the smallest d_S in the comparison of Clade R-II and *A. crassostreae*,
367 suggesting that Clade R-I branches off first in the gene tree (Fig. S9-i-a) and that recombination
368 may have occurred from unsampled lineages that branched earlier than *A. crassostreae* to the
369 LCA of Clade R-I on the rooted species tree (the red arrow in Fig. S9-i-b). An alternate
370 explanation is that these gene families underwent recombination between the LCA of *A.*
371 *crassostreae* and the LCA of Clade R-II (the grey arrow in Fig. S9-i-b). The latter mechanism
372 homogenizes the genetic materials at these gene families, which is expected to lead to unusually
373 low d_S values between recombined lineages [25]. Because the d_S values between *A. crassostreae*
374 and Clade R-II are large and comparable to other pairwise comparisons in each gene family
375 (Table S9), the second mechanism was ruled out.

376 Likewise, four gene families each were inferred to show a gene tree topology by
377 clustering Clade R-I and *A. crassostreae* (Fig. S9-ii-a) and to have undergone recombination
378 from the unsampled lineages to the LCA of Clade R-II (the red arrow in Fig. S9-ii-b). In another

379 31 gene families, the gene tree topologies (Fig. S9-iii-a) were inferred to be identical to the
380 species tree in which Clade R-I and Clade R-II are clustered, suggesting that unsampled lineages
381 that branched off following *A. crassostreae* donated alleles to either the LCA of Clade R-I or the
382 LCA of Clade R-II (Fig. S9-iii-b). The remaining 17 gene families do not have orthologs in *A.*
383 *crassostreae*, and their most closely related genes are sampled from more distant lineages (see an
384 example of gene tree in Fig. S7D). In these cases, the gene tree topologies (Fig. S9-iv-a) were
385 inferred to be congruent with the species tree, and novel alleles were likely acquired from
386 unsampled lineages that are closely related to the *Roseobacter* population under study (Fig. S9-
387 iv-b).

388

389 2.4 The pattern of *Roseobacter* population differentiation fits an existing microbial speciation 390 model

391 The ecological differentiation could be driven by either genome-wide selective sweeps
392 [57] or gene-specific selective sweeps [58]. The first model states that acquisition of an adaptive
393 genetic trait enables a genotype to outcompete all others in the population, and recombination is
394 rare enough to sustain period selection which purges genetic diversity to near zero across the
395 genome. This theory finds its first and strong support only recently from a natural lake bacterial
396 population, which showed genome-wide genotype succession over eight years [59]. Alternatively,
397 acquisitions of new alleles may enable the cell to explore a new ecological niche and form a new
398 ecotype, and genome-wide selective sweeps can effectively separate the gene pool of the new
399 ecotype from that of the old one and lead to new species formation [57]. Whether this second
400 mechanism occurs in nature remains contentious [60, 61].

401 In contrast, the second model proposes that high frequency of recombination results in
402 the spread of adaptive alleles in the population while preventing the elimination of the genome-
403 wide diversity. In the *Roseobacter* population under study, the very low frequency of
404 recombination ($\rho/\theta=0.076$) prevents adaptive alleles acquired at the three core genomic regions
405 from unlinking to the rest of the genome, thus gene-specific selective sweeps were less likely to
406 occur. Instead, the genome-wide diversity within the derived clade (Clade R-I) is extremely low
407 (Fig. 1C), which was likely a result of the genome-wide selective sweep. The present study
408 therefore provides the first evidence that genome-wide selective sweep drives bacterial
409 speciation in nature. One caveat is the limited number of available isolates affiliated with Clade
410 R-I may discount the use of the genome-wide selective sweep as an exclusive mechanism to
411 explain the data. If genetically diverse members of Clade R-I do exist but remain unsampled,
412 which carry the same adaptive alleles but are more diverged in the rest of the genome, the
413 mechanisms driving the speciation process may become more complicated.
414

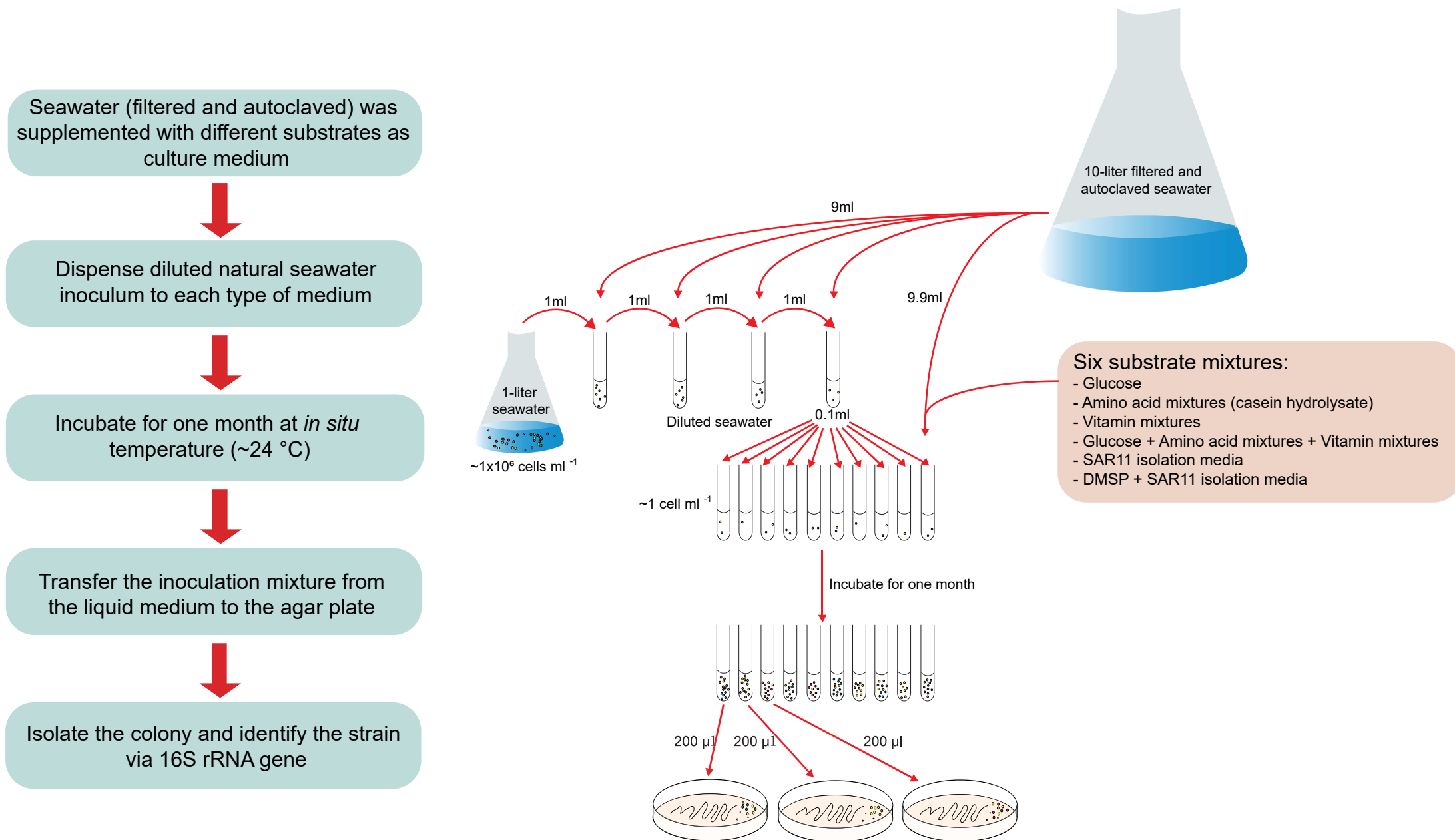


Figure S1. The overview of a dilution-to-extinction cultivation approach used for *Roseobacter* isolation. The flowchart on the left describes the general procedure of the approach, and the schematic plot on the right provides the details of the dilution strategy. More information can be found in SI Text 1.1.

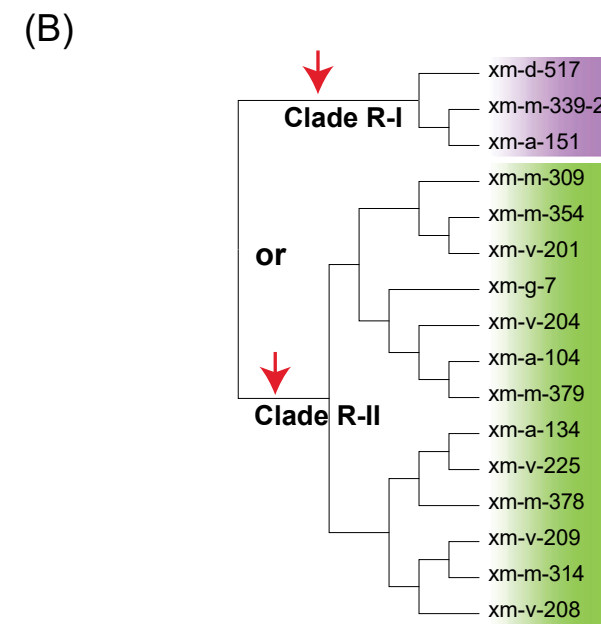
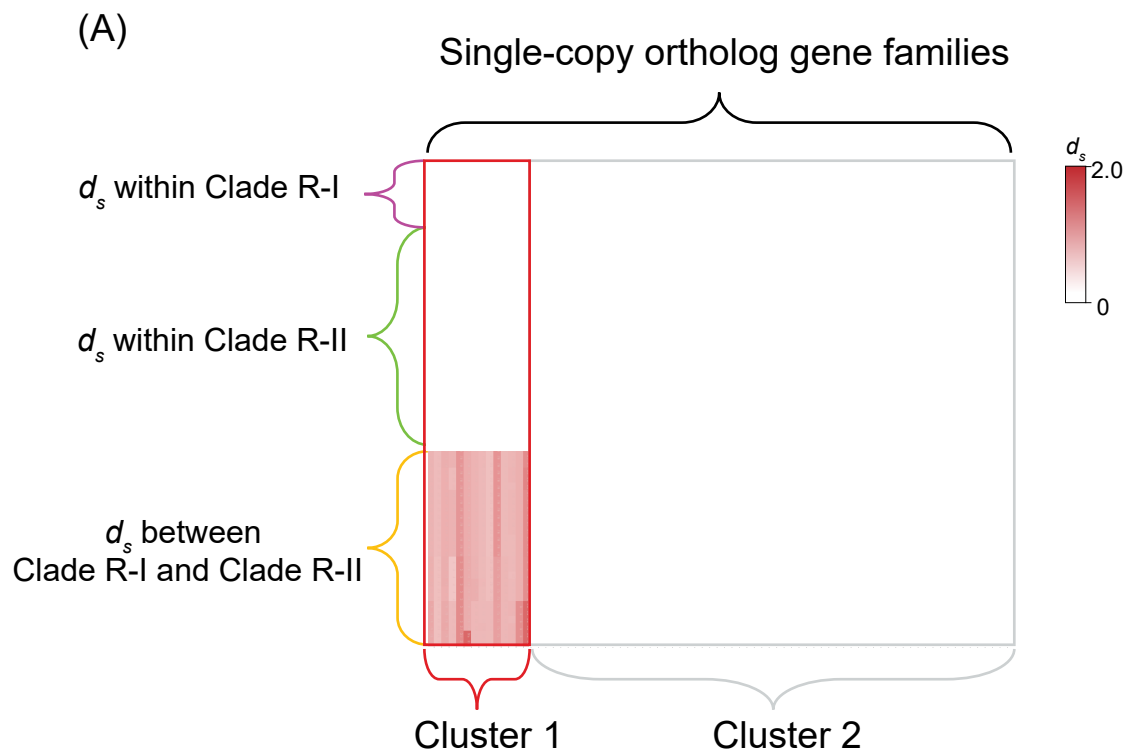
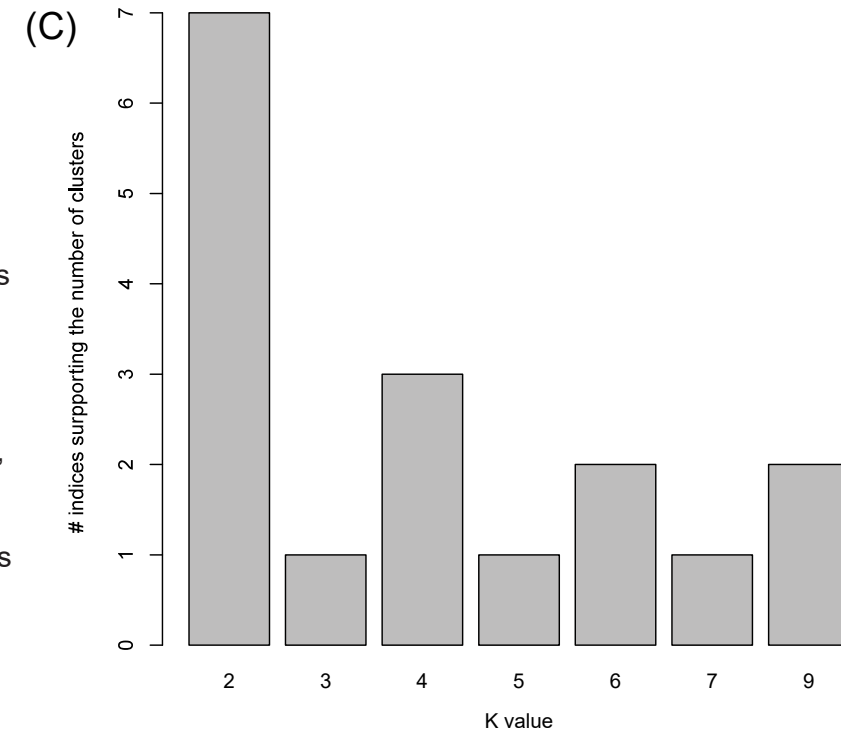


Figure S2. Illustration of allelic replacement inference using d_s values. (A) The demo heatmap of the d_s values calculated for every possible pair of genomes across all single-copy orthologous gene families, with warmer colour indicating higher d_s values. The gene families are grouped into two clusters according to (C). Cluster 1 shows unusually large d_s values between Clade R-I and Clade R-II but small d_s values within each clade, whereas both between- and within-clade d_s values are small in Cluster 2. (B) Evolutionary history of an example gene from Cluster 1 mapped to the genome tree. Due to the unusually large between-clade d_s values and little diversity within each clade, the allelic replacement with distant lineages is inferred to have occurred at the LCA of Clade R-I or that of Clade R-II. (C) The determination of optimal number of clusters by cluster validity indices using 'NbCluster' package in R language. The best cluster number is determined to be two, which is supported by seven indices ('duda', 'pseudot2', 'beale', 'gap', 'frey', 'sindex', 'sdbw') for cluster validity. It means that the core genes in the *Roseobacter* population are grouped into two clusters based on the pairwise d_s values among all 16 strains across 2,846 single-copy orthologous gene families.



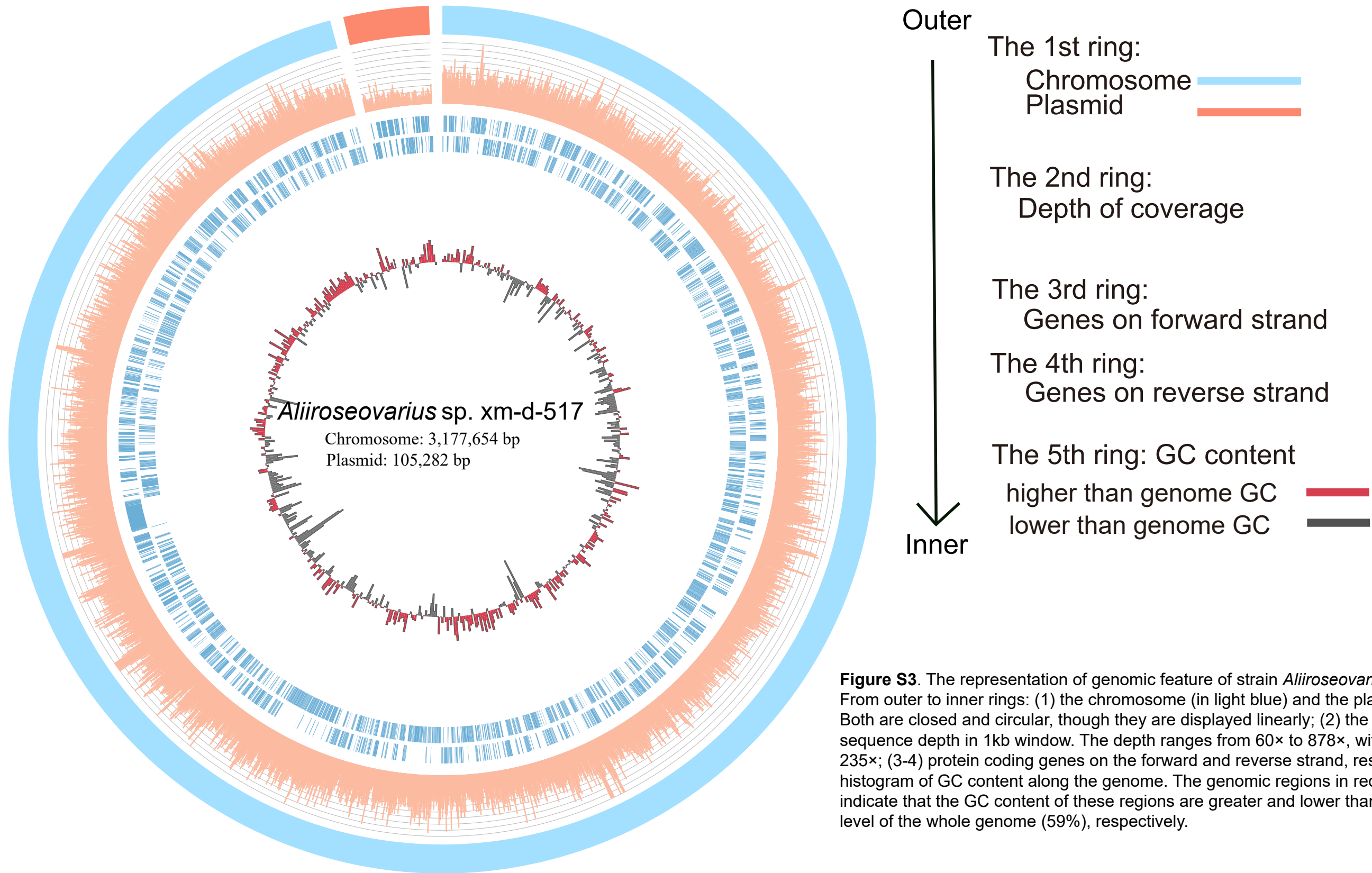


Figure S3. The representation of genomic feature of strain *Aliiroseovarius* sp. xm-d-517. From outer to inner rings: (1) the chromosome (in light blue) and the plasmid (in light red). Both are closed and circular, though they are displayed linearly; (2) the histogram of sequence depth in 1kb window. The depth ranges from 60× to 878×, with a median of 235×; (3-4) protein coding genes on the forward and reverse strand, respectively; (5) the histogram of GC content along the genome. The genomic regions in red and black indicate that the GC content of these regions are greater and lower than the average level of the whole genome (59%), respectively.

PM 01 (Carbon Sources)

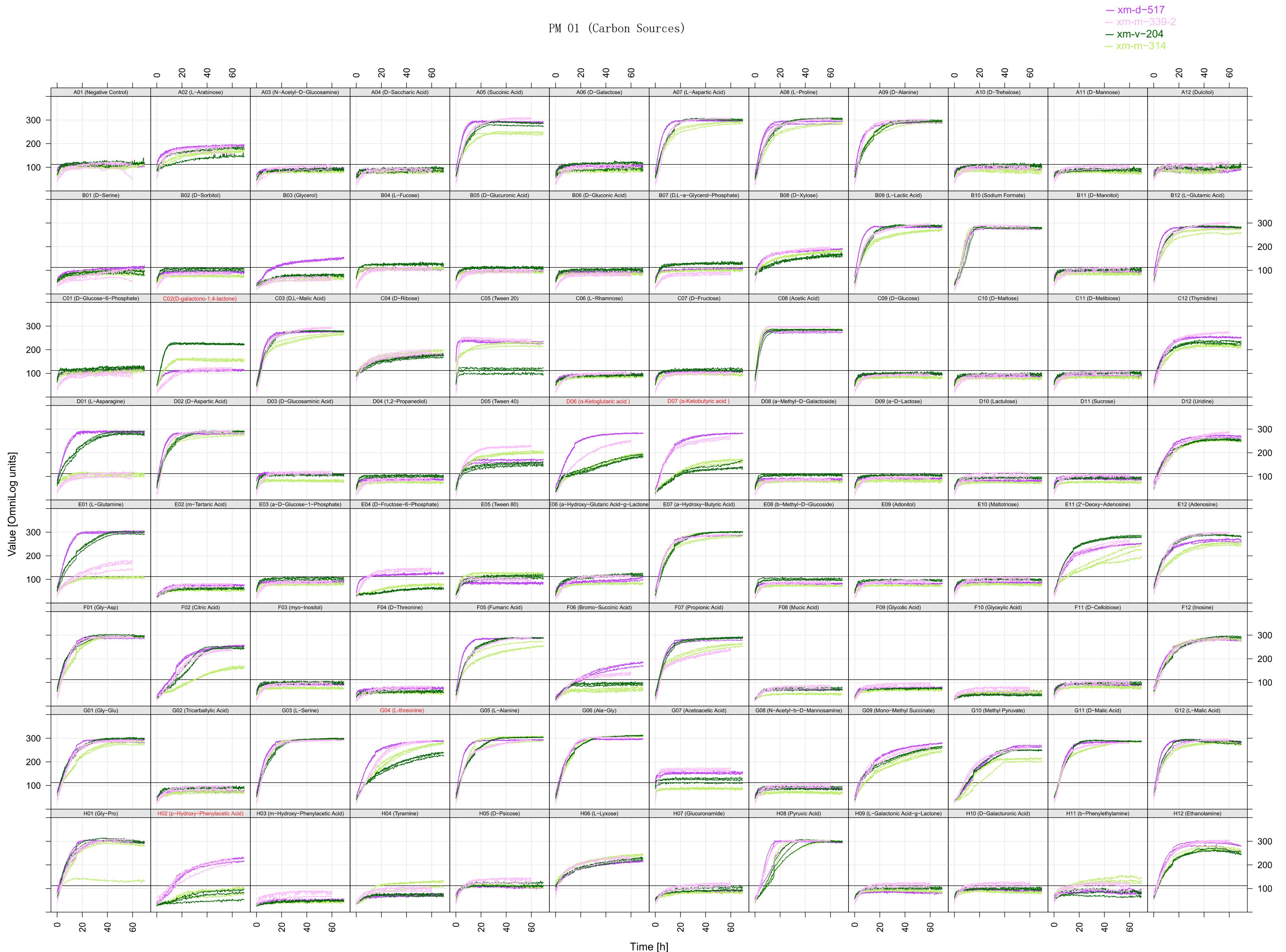


Figure S4. The utilization of 95 carbon sources in phenotypic microarray microplate PM01 by the strain xm-d-517 (purple) and xm-m-339-2 (light purple) from Clade R-I and the strain xm-v-204 (green) and xm-m-314 (light green) from Clade R-II in the *Roseobacter* population. The first well (A01) in the microplate is the negative control without any carbon source. The substrate curves show the respiration activity of bacteria as a proxy for the traditional bacterial growth curve. Most of substrate curves in the microplate either resemble bacterial growth curves or are near the baseline. The former indicates that the strain could use the corresponding substrates for growth, whereas the latter indicates that no respiration is detected in the well, namely, incapable of using the substrates. Three replicates are performed for each strain. Five significantly differentially utilized substrates by members from these two diversified clades are highlighted in red.

PM 02 (Carbon Sources)

— xm-d-517
 — xm-m-339-2
 — xm-v-204
 — xm-m-314

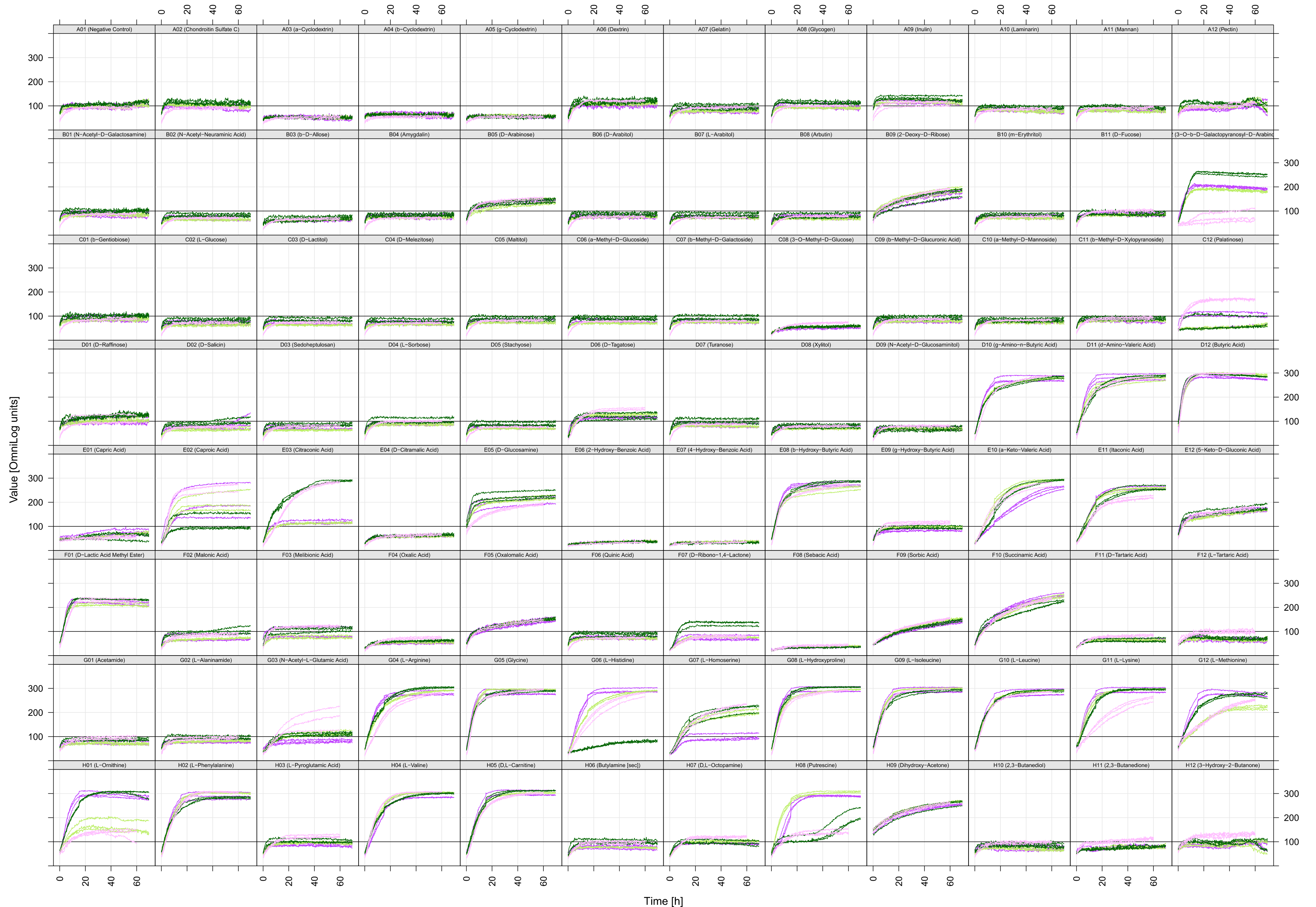


Figure S5. The utilization of 95 carbon sources in phenotypic microarray microplate PM02A by the four representative strains (same as those shown in Fig. S4) in the *Roseobacter* population. Three replicates are performed for each strain. No differentially utilized substrate from this plate by members from the two diversified clades of the *Roseobacter* population was identified.

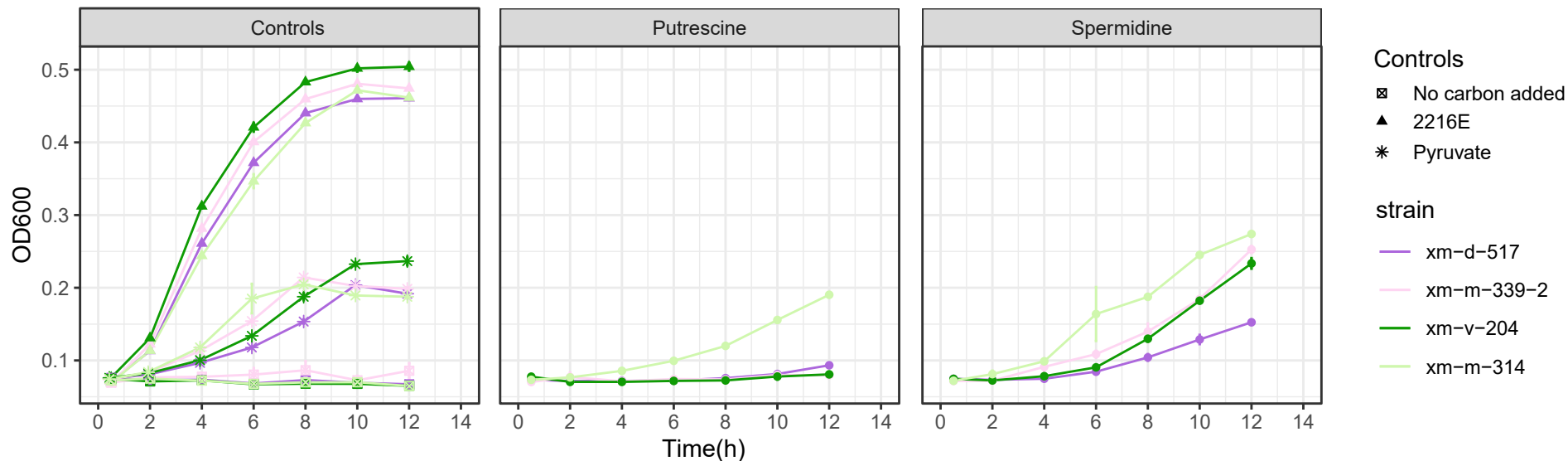
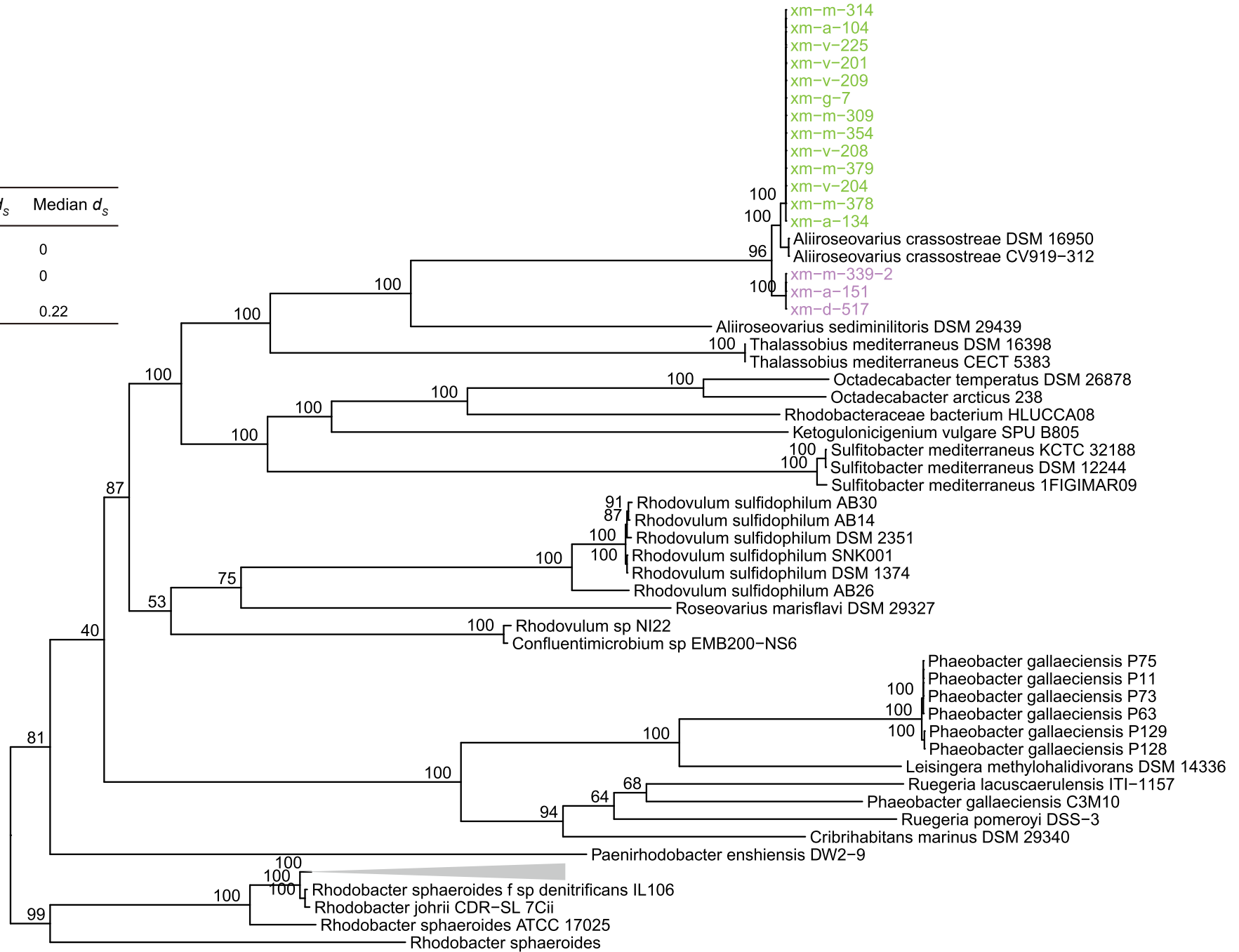


Figure S6. Growth experiments of the four representative strains with and without carbon sources, in which three replicates were performed for each strain. Left: the negative control without any carbon source, the positive control when pyruvate is used as a sole carbon source, and another control inoculated in rich medium (Difco™ Marine broth 2216). Middle and right: the growth curves of the four strains with putrescine and spermidine as a sole carbon source, respectively. The lines for the four strains are indicated in distinct colors (xm-d-517 in purple; xm-m-339-2 in light purple; xm-v-204 in green; xm-m-314 in light green).

Figure S7. Examples illustrating the inference of the evolutionary scenarios underlying novel allele replacements at the core genes each showing an unusually large between-clade d_S value in the *Roseobacter* population. The four example (A, B, C and D) core gene families (named with gene locus from strain xm-d-517) each represent a distinct evolutionary path (i, ii, iii and iv) shown in Fig. S9. The two diversified clades of the *Roseobacter* population are highlighted with distinct colors in the gene tree, consistent with the color scheme shown in their phylogenomic tree (Fig. 1A).

(A) xm-d-517_03123

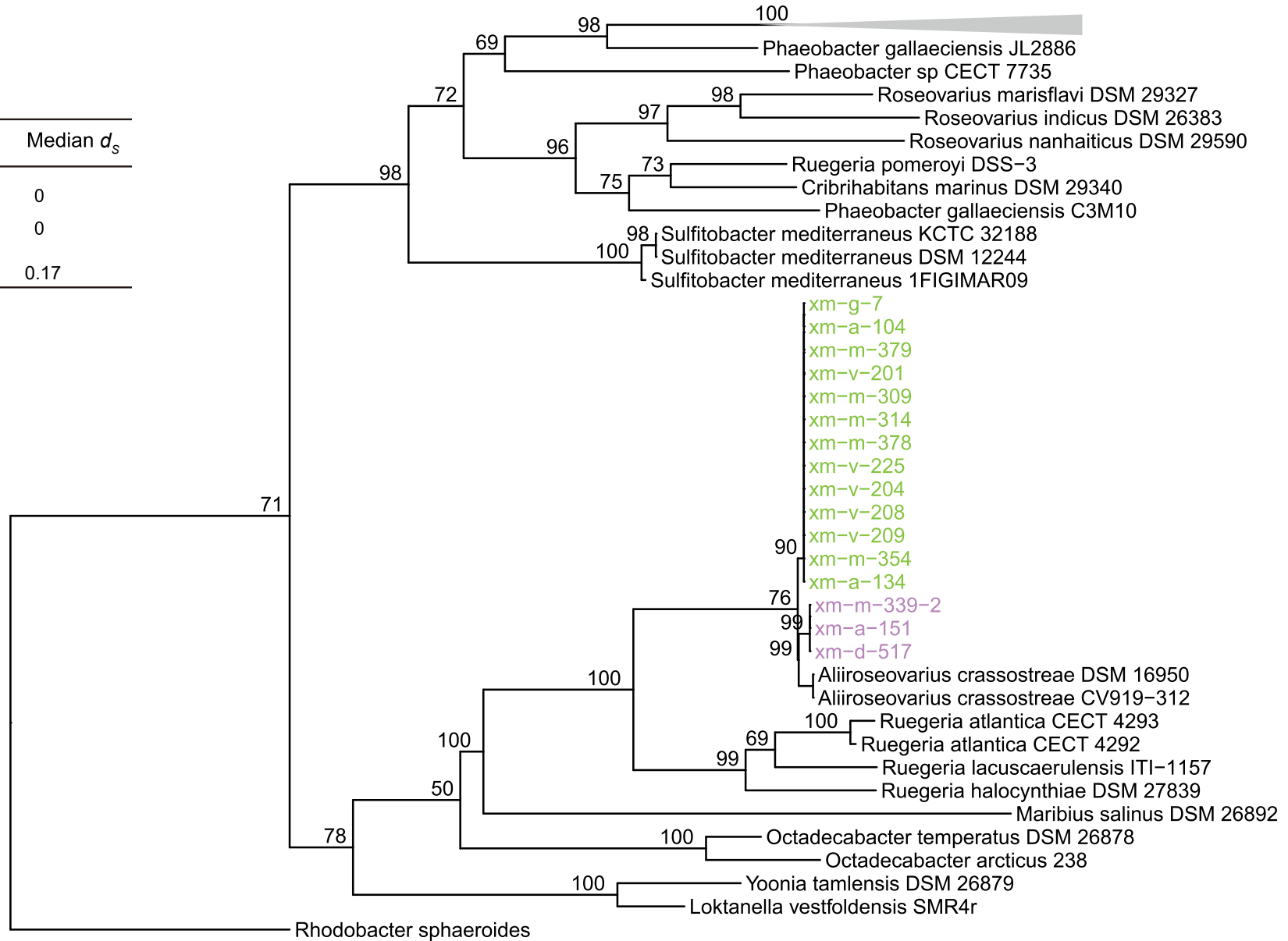
Group	Mean d_s	Median d_s
Within Clade R-I	0	0
Within Clade R-II	0	0
Between clades	0.22	0.22



0.05

(B) xm-d-517_03119

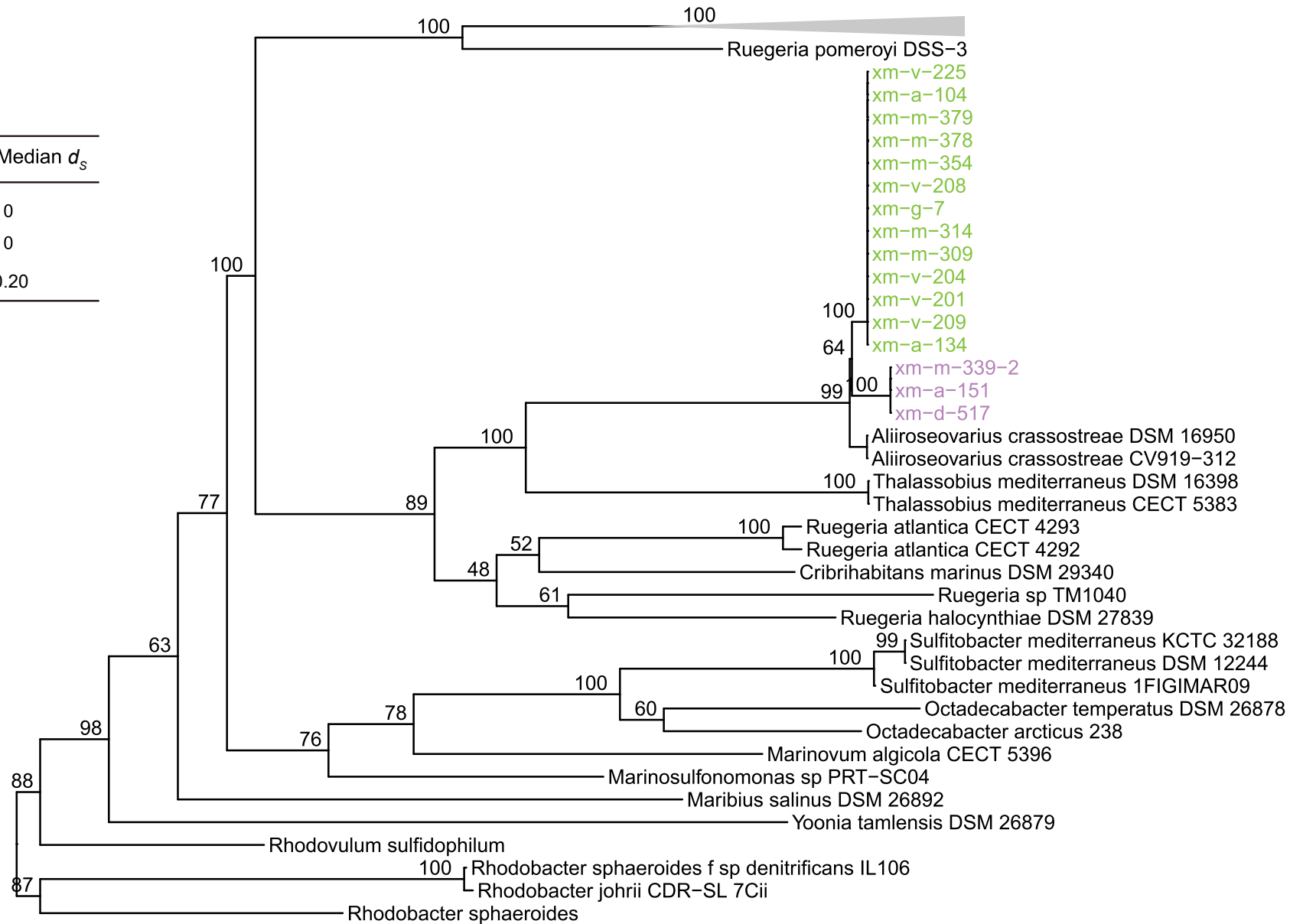
Group	Mean d_s	Median d_s
Within Clade R-I	0	0
Within Clade R-II	0	0
Between clades	0.17	0.17



0.05

(C) xm-d-517_03154

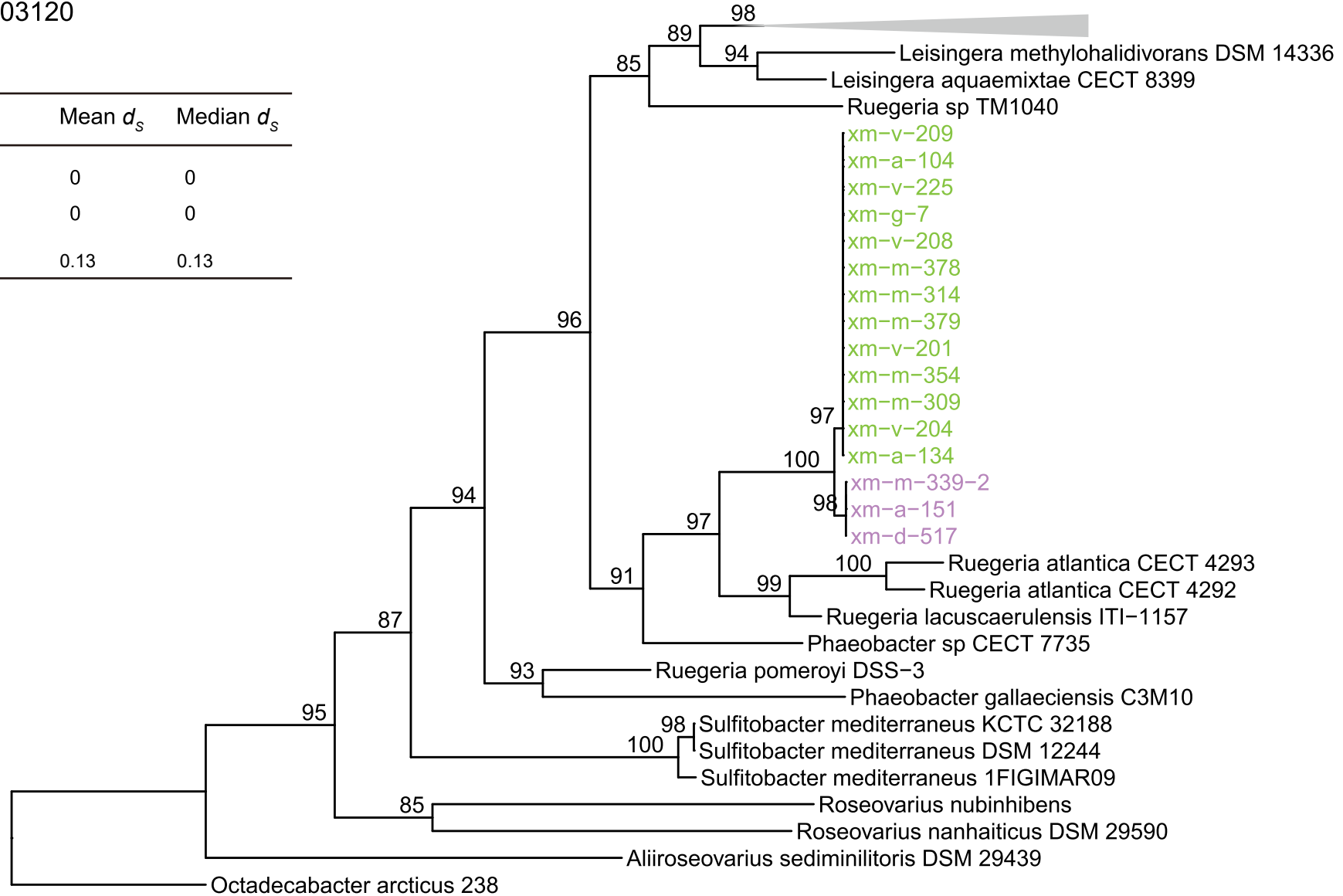
Group	Mean d_s	Median d_s
Within Clade R-I	0	0
Within Clade R-II	0	0
Between clades	0.20	0.20



0.05

(D) xm-d-517_03120

Group	Mean d_s	Median d_s
Within Clade R-I	0	0
Within Clade R-II	0	0
Between clades	0.13	0.13



0.05

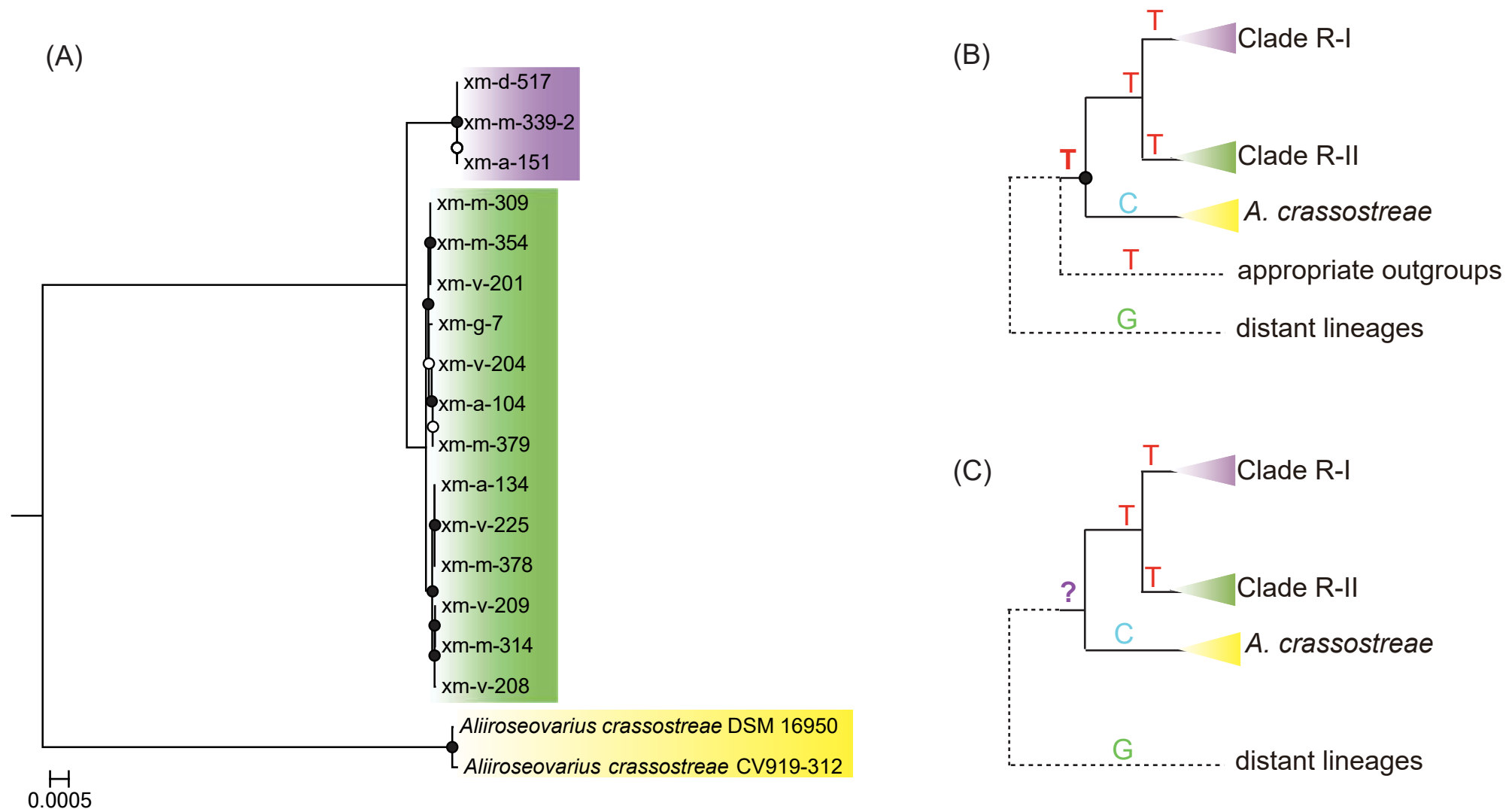


Figure S8. (A) The RAxML maximum-likelihood phylogenomic tree of the *Roseobacter* population and *A. crassostreae* based on concatenated single-copy core genes. Solid and open circles at the nodes indicate the frequency of the group defined by that node is at least 95 and 80. (B and C) Two examples of ancestral state inference. Above each branch is ancestral state of each lineage. (B) The ancestral state of the LCA of Clade R-I, Clade R-II and *A. crassostreae* can be reliably inferred when appropriate outgroups are available. The solid circles denote the root of three lineages. (C) Available outgroups are too distant to infer ancestral state of three lineages.

Gene tree topology inferred by neutral genetic distance

Inferred recombination history

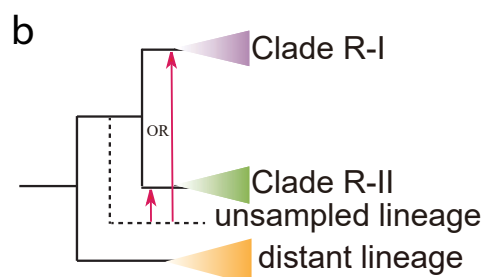
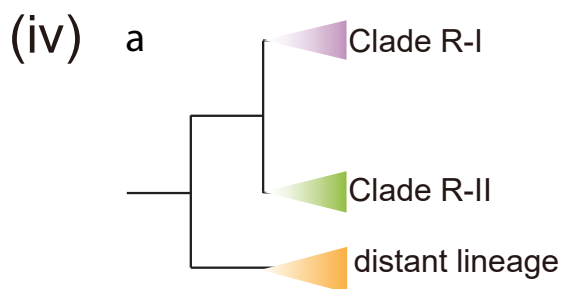
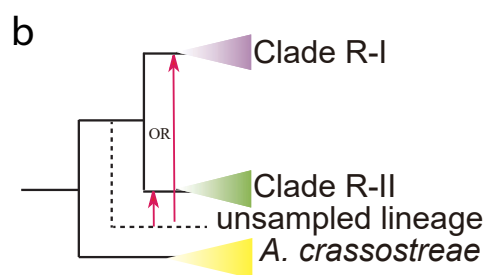
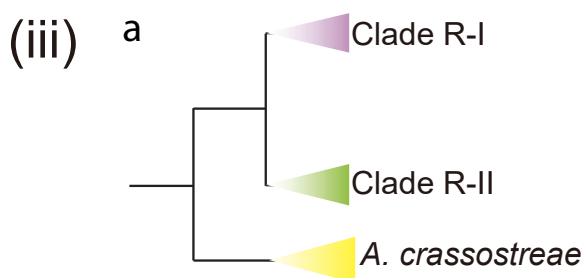
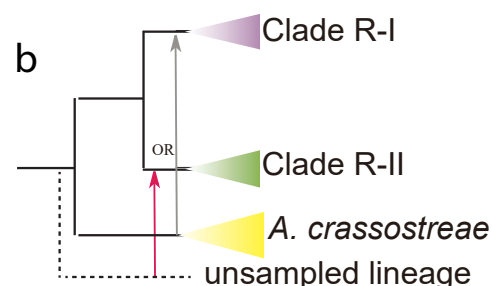
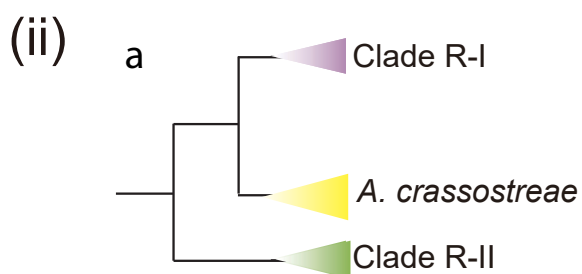
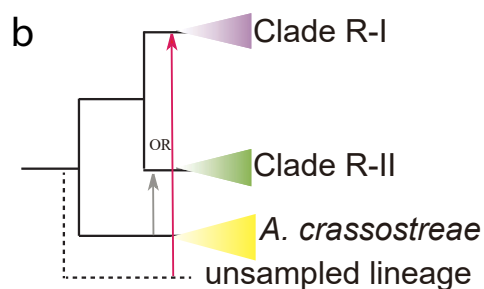
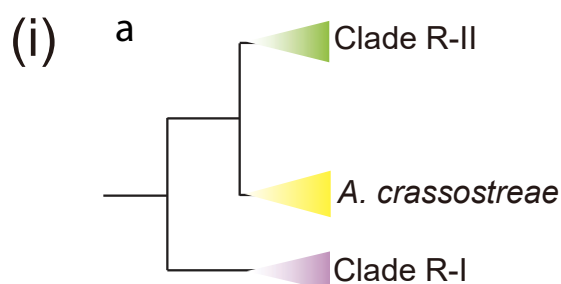


Figure S9. The inferred recombination history of the 200 core gene families that drove population differentiation. Left shows gene tree topology inferred based on pairwise neutral genetic distance (d_s) among Clade R-I, Clade R-II and *Aliiroseovarius crassostreae*, and right is the inferred recombination history mapped on the species tree of Clade R-I, Clade R-II and *A. crassostreae*. (i) Clustering between Clade R-II and *A. crassostreae* (i-a) indicates the LCA of Clade R-I replaced novel alleles from unsampled lineages that branched earlier than *A. crassostreae* (the red arrow; i-b) or recombination between the LCA of Clade R-II and that of *A. crassostreae* (the grey arrow; i-b). The history indicated by the grey arrow is rejected by the d_s analysis; (ii) Clade R-I and *A. crassostreae* share closer evolutionary relationship (ii-a), indicating unsampled lineages which branched earlier than *A. crassostreae* donated alleles to the LCA of Clade R-II (the red arrow; ii-b) or recombination between the LCA of Clade R-I and *A. crassostreae* (the grey arrow; ii-b). The history indicated by the grey arrow is rejected by the d_s analysis; (iii) gene tree shows topology consistent with species tree (iii-a), suggesting either the LCA of Clade R-I or that of Clade R-II replaced alleles from unsampled lineages that branched off following *A. crassostreae* (iii-b); (iv) gene tree is congruent with species tree, though no orthologs can be found in *A. crassostreae* (iv-a), suggesting either the LCA of Clade R-I or that of Clade R-II replaced alleles from unsampled lineages that branched off following the “distant lineage”.

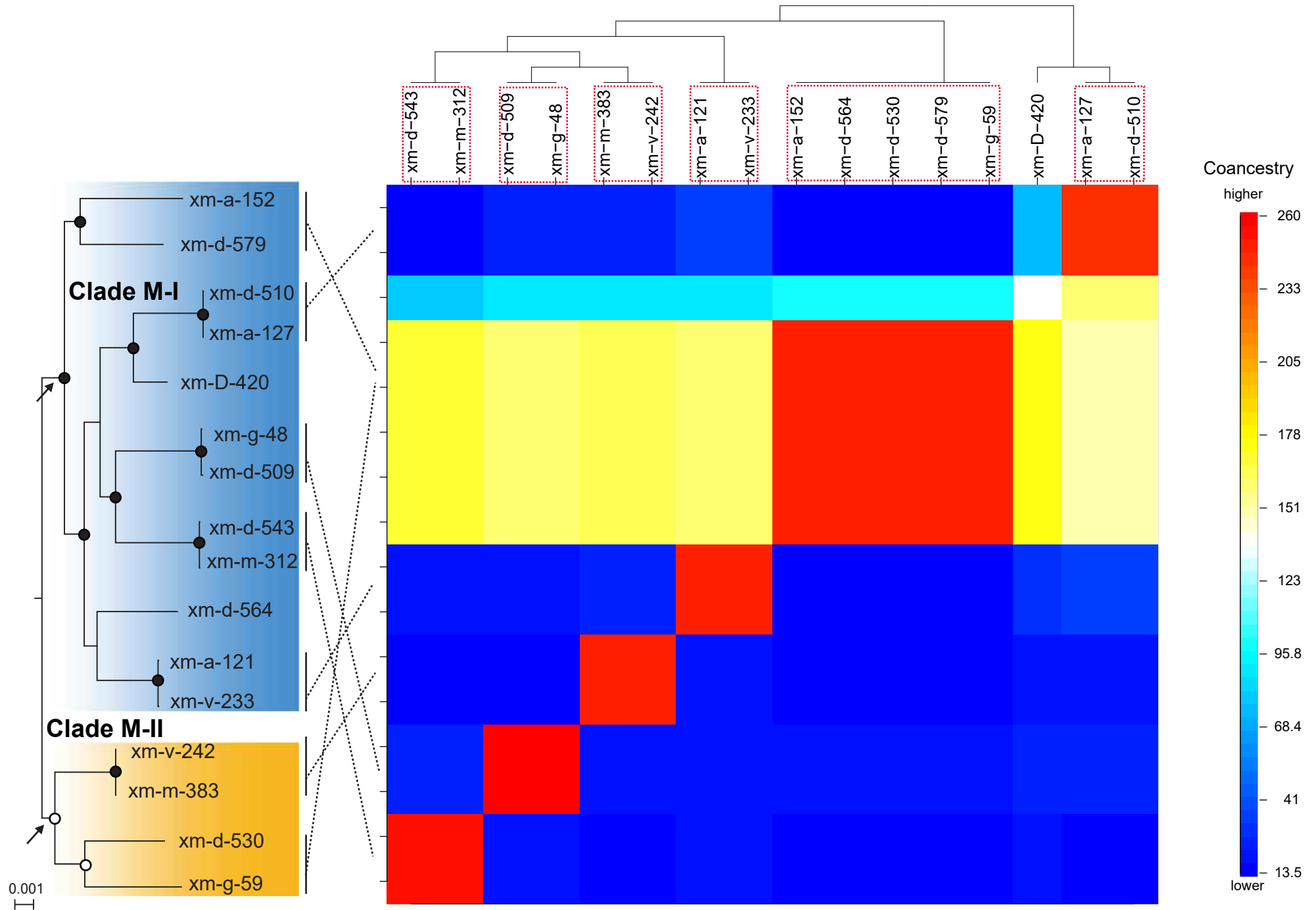


Figure S10. The RAxML maximum-likelihood phylogenomic tree and the fineSTRUCTURE coancestry matrix of the *Marinobacterium* population. The rooted phylogenomic tree is shown on the left (the outgroup is not shown). Solid and open circles at the nodes indicate the frequency of the group defined by that node is at least 95 and 80, respectively, in the 100 bootstrapped replicates. The scale bar indicates the number of substitutions per site. The two most deeply branching clades in the *Marinobacterium* population are highlighted in blue and orange. The last common ancestors of Clade M-I and Clade M-II each are marked with an arrow. The coancestry matrix is shown on the right, the legend of the matrix is same as that in Fig.1A.

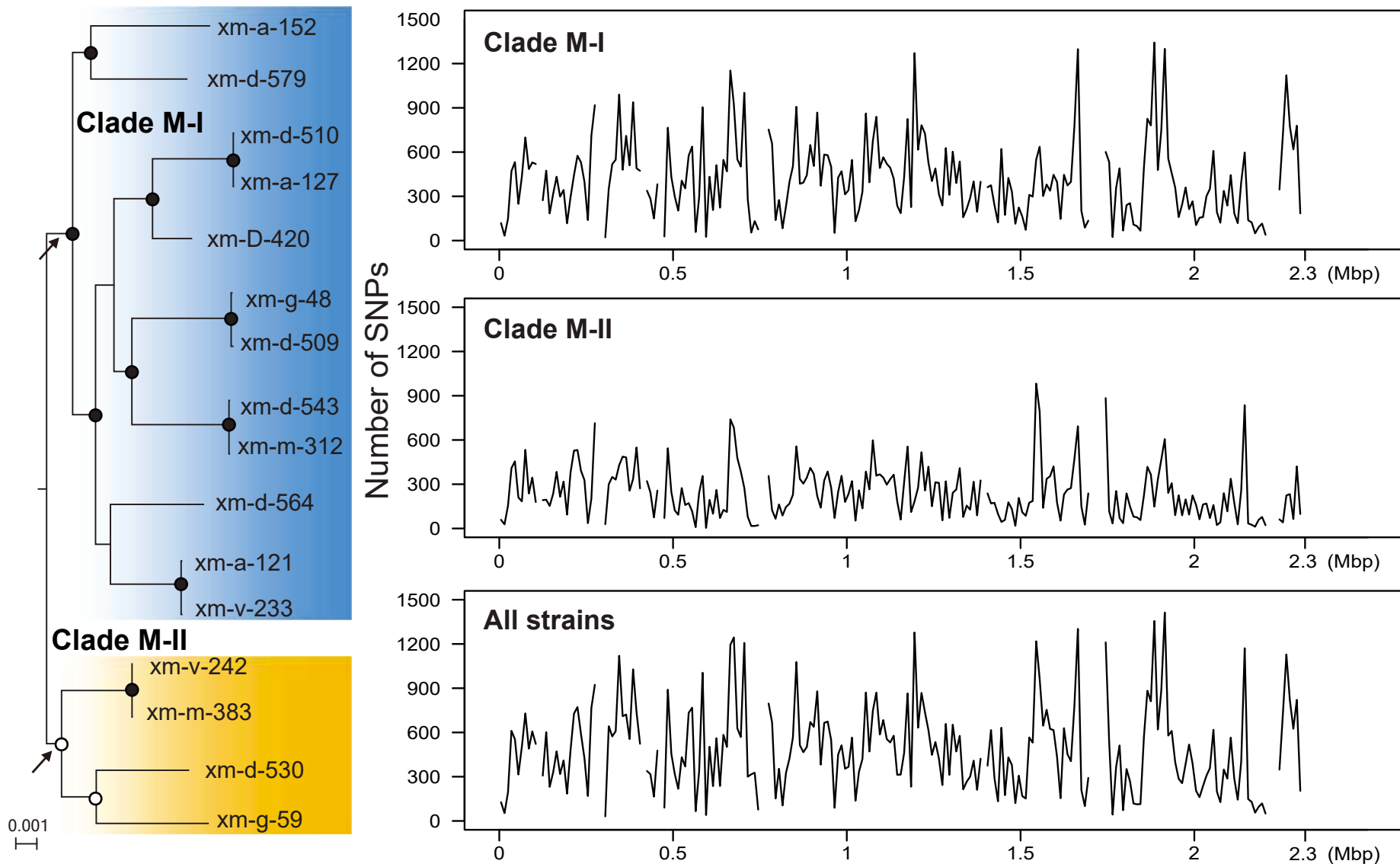


Figure S11. The distribution of SNPs along a representative genome in the *Marinobacterium* population. The SNPs are counted within 10-kb sliding windows across the reference genome. As the *Marinobacterium* population in which a closed genome sequence is not available, the strain xm-d-579 is used as a reference since it consists of the fewest contigs, and contigs are ordered decreasingly in length. The SNP density is counted within each of the two most deeply-branching clades, respectively, as well as among all strains pooled together. The phylogenomic tree of each population is on the left of the plot, where the last common ancestors of Clade M-I and Clade M-II each are marked with an arrow.

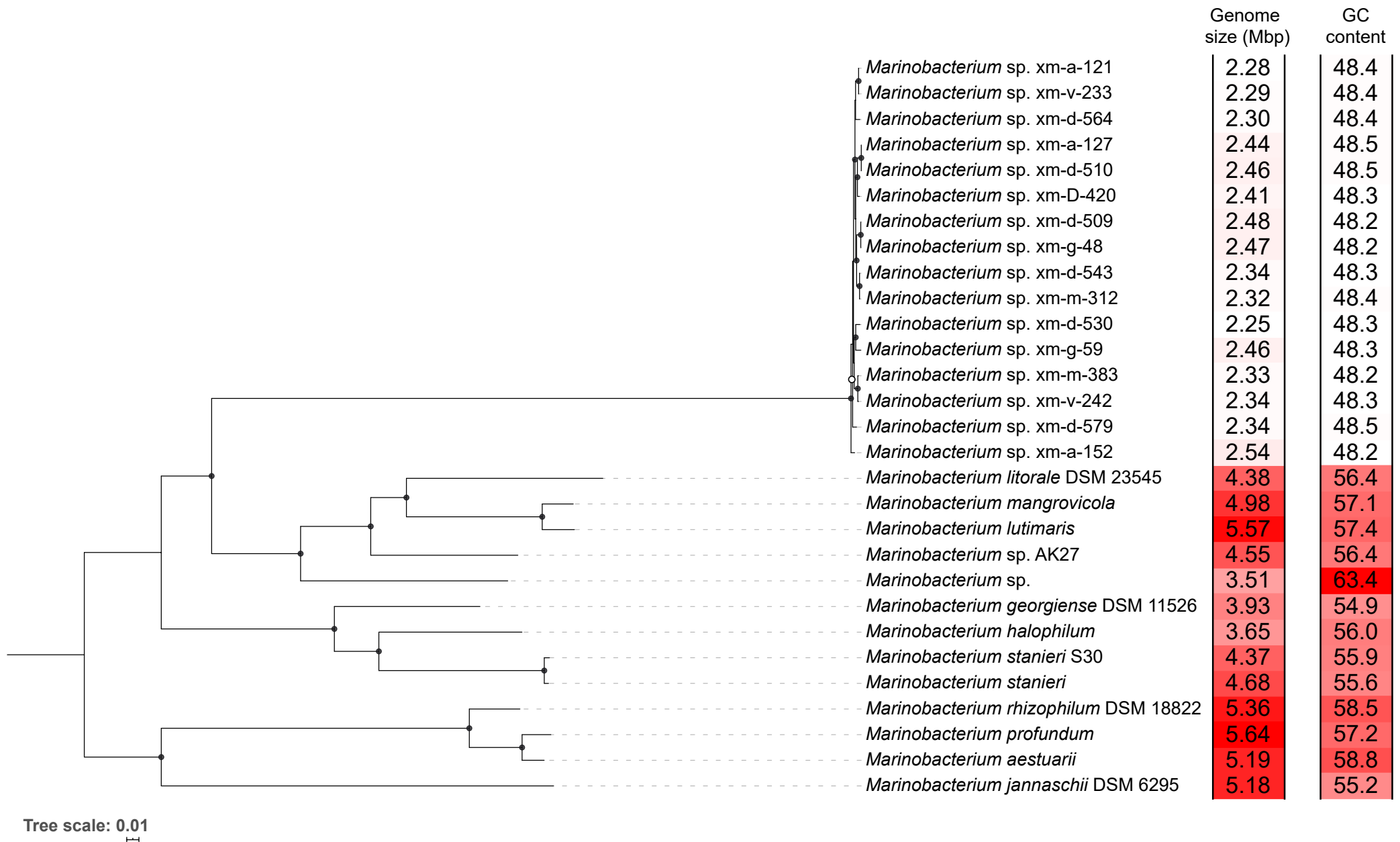




Figure S12. The RAxML maximum-likelihood phylogenomic tree of the *Marinobacterium* genus and genomic features. The rooted phylogenomic tree is shown on the left (the outgroup is not shown. The genome sequences of published *Marinobacterium* strains were downloaded from NCBI). Solid and open circles at the nodes indicate the frequency of the group defined by that node is at least 95 and 80, respectively, in the 100 bootstrapped replicates. The scale bar indicates the number of substitutions per site. The genome size and GC content are shown on the right, with warmer color representing the larger genome size or higher GC content.

RAST Subsystem 1st level

 *Roseobacter* sp.
 *Marinobacterium* sp.

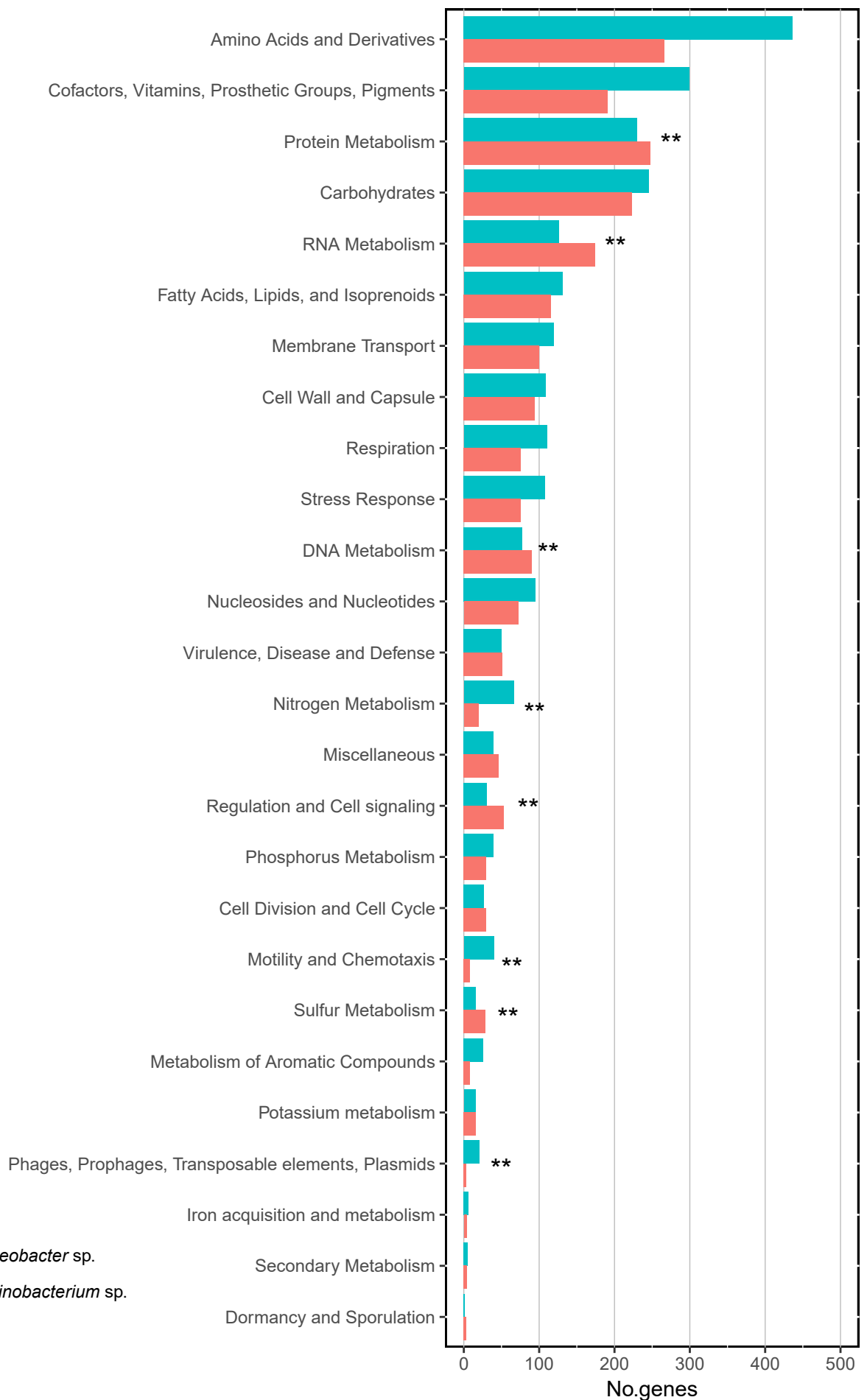


Figure S13. Comparison of functional gene abundance between the *Roseobacter* population and the *Marinobacterium* population, and genes are categorized based on the first level of subsystems from the RAST server for all genes within each population. The statistical enrichment of each functional group was assessed using χ^2 test. Two stars on the right of the bars denote significant difference ($p < 0.01$) in the relative abundance of genes in these functional categories between the two populations.

Table S1. The physicochemical data of the seawater sample from which the *Roseobacter* population and the *Marinobacterium* population were isolated.

Parameter	Value
Sampling date	2015-10-22 11:43 am
Weather	Sunny
Latitude/Longitude	24°29'12.59'' N/118°14'05.86'' E
Volume (L)	10
Temperature (°C)	23.70
Salinity (%)	29.96
pH	8.13
DO (%)	96.50
Conductivity (SPC)	45,018.00
NO ₂ ⁻ (μmol/L)	6.421
NO ₂ ⁻ +NO ₃ ⁻ (μmol/L)	35.366
Total phosphate (μmol/L)	1.645
Si (μmol/L)	34.499

Table S2. List of nutrient supplements for six types of seawater media used for bacterial isolation. These nutrient supplements each are suspended using the autoclaved and filtered (0.22 μm syringe) seawater in the preparation of the cultivation media.

Supplements	Six types of seawater media					
	xm-g	xm-a	xm-v	xm-m	xm-d	xm-D
Glucose	0.01 g/L			0.01 g/L	0.01 g/L	0.01 g/L
An amino acid mixture (casein hydrolysate)		2 μM		2 μM		
A vitamin mixture (Centrum, Wyeth)			0.6 g/L	0.6 g/L		
DMSP						100 nM
NH_4Cl					1 μM	1 μM
KH_2PO_4					0.1 μM	0.1 μM
D-ribose					0.001% (w/v)	0.001% (w/v)
Glycerol					0.001% (w/v)	0.001% (w/v)
N-acetyl-D-glucosamine					0.001% (w/v)	0.001% (w/v)
Methylamine					0.001% (w/v)	0.001% (w/v)
Pyruvic acid					0.001% (w/v)	0.001% (w/v)
Ethanol					0.002% (v/v)	0.002% (v/v)
Thiamine HCl					20 ng/L	20 ng/L
Biotin					0.1 ng/L	0.1 ng/L
Vitamin B12					0.1 ng/L	0.1 ng/L
Folic acid					0.2 ng/L	0.2 ng/L
Para-Aminobenzoic Acid					1 ng/L	1 ng/L
Nicotinic acid					10 ng/L	10 ng/L
Inositol					100 ng/L	100 ng/L
Calcium pantothenate					20 ng/L	20 ng/L
Pyridoxine HCl					10 ng/L	10 ng/L

Table S3. Genome statistics of the 16 isolates in the *Roseobacter* population. The completeness and contamination are estimated with CheckM [7], and the remaining statistics are calculated with QUAST [62].

Strain	GC	Completeness	Contamination	# Contigs	Longest contig (bp)	Total size (bp)	N50 (bp)
xm-a-104	59.1	99.7	0.4	20	689,243	3,266,724	358,925
xm-a-134	59.1	99.7	0.4	23	821,604	3,283,027	353,983
xm-a-151	59	99.7	0.4	26	434,522	3,257,283	277,436
xm-d-517	59	99.7	0.4	2	3,177,654	3,282,936	3,177,654
xm-g-7	59	99.7	0.46	30	770,039	3,320,255	277,720
xm-m-309	59.1	99.7	0.46	15	989,758	3,332,128	823,298
xm-m-314	59.2	99.7	0.53	19	915,353	3,439,089	832,426
xm-m-339-2	59.1	99.7	0.4	33	434,570	3,387,700	277,436
xm-m-354	59.1	99.7	0.46	86	989,760	3,417,944	823,257
xm-m-378	59.1	99.7	0.4	22	821,586	3,282,870	354,031
xm-m-379	59.1	99.7	0.4	23	689,243	3,270,267	358,925
xm-v-201	59.1	99.7	0.46	15	989,759	3,332,057	823,217
xm-v-204	59.1	99.7	0.4	19	689,243	3,265,630	358,924
xm-v-208	59.2	99.7	0.46	15	1,003,587	3,441,315	832,374
xm-v-209	59.2	99.7	0.53	19	915,353	3,439,089	832,426
xm-v-225	59.1	99.7	0.4	22	821,586	3,283,060	354,023

Table S4. Genetic diversity of the *Roseobacter* population and the *Marinobacterium* population. For the purpose of comparison, published prokaryotic populations are also included which show evidence of population differentiation. The number of single nucleotide polymorphisms (SNPs) is calculated based on the core genomes of each population, and normalized by the core genome length and the number of genomes within each population. The 16S rRNA gene identity and whole-genome average nucleotide identity (ANI) each are the mean value of all pairwise comparisons. In the case of *Clostridium difficile* and *Mycobacterium tuberculosis* data sets, genome assemblies are not available and therefore their 16S rRNA gene identity and pairwise ANI are not available. As the evolution of the *Roseobacter* population is heavily affected by the three large genomic regions showing evidence of novel allele acquisitions, the genetic diversity of this population is calculated with and without these regions. In the ‘Lifestyle’ column, ‘F’ and ‘H’ represent free-living and host-associated lifestyles, respectively, whereas ‘MIX’ means both lifestyles are present in the organism.

Population	Taxonomy	Lifestyle	#Genomes	SNPs	SNPs/Mbp	SNPs/Mbp/ 10 genomes	16S rRNA gene identity (%)	Pairwise ANI (%)	Reference
<i>Roseobacter</i> sp.	Bacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae	F	16	12,860	4,242	2,651	100.00	99.76	This study
				1,277	455	283		-	This study (without the three recombination affected regions)
<i>Marinobacterium</i> sp.	Bacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae	F	16	107,547	57,045	35,653	99.90	98.12	This study
<i>Sulfolobus islandicus</i>	Archaea; Crenarchaeota; Thermoprotei; Sulfolobales; Sulfolobaceae	F	12	17,388	7,560	6,300	99.83	99.46	[63]
<i>Vibrio cyclotrophicus</i>	Bacteria; Gammaproteobacteria; Vibrionales; Vibrionaceae	F	20	111,012	31,717	15,859	100.00	99.12	[52]
<i>Myxococcus xanthus</i>	Bacteria; Deltaproteobacteria; Myxococcales; Cystobacterineae; Myxococcaceae	F	22	83,780	11,024	5,011	99.99	99.10	[64]
<i>Mesorhizobium</i> sp.	Bacteria; Alphaproteobacteria; Rhizobiales; Phyllobacteriaceae	F	38	381,344	77,826	20,481	99.80	97.88	[65]
<i>Pseudomonas koreensis</i>	Bacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae	F	53	936,976	227,974	43,014	99.28	91.82	[66]
<i>Alteromonas macleodii</i>	Bacteria; Gammaproteobacteria; Alteromonadales; Alteromonadaceae	F	9	118,001	35,436	39,373	99.28	98.66	[67]
<i>Methanosarcina mazei</i>	Archaea; Euryarchaeota; Methanomicrobia; Methanosarcinales; Methanosarcinaceae	F	56	39,660	19,929	3,559	99.93	99.16	[68]
<i>Prochlorococcus</i>	Bacteria; Terrabacteria; Cyanobacteria; Synechococcales; Prochloraceae	F	66	52,885	33,053	5,008	99.92	97.37	[69]
<i>Polynucleobacter asymbioticus</i>	Bacteria; Betaproteobacteria; Burkholderiales; Burkholderiaceae	F	9	50,477	30,778	34,198	100.00	98.81	[70]
<i>Bacillus subtilis</i>	Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae	F	6	273,619	77,076	128,460	99.68	97.09	[71]
<i>Salinibacter ruber</i>	Bacteria; Bacteroidetes; Rhodothermaceae	F	8	90,784	39,992	49,990	99.79	98.42	[72]

Population	Taxonomy	Lifestyle	#Genomes	SNPs	SNPs/Mbp	SNPs/Mbp/ 10 genomes	16S rRNA gene identity (%)	Pairwise ANI (%)	Reference
<i>Streptomyces</i>	Bacteria; Actinobacteria; Streptomycetales; Streptomycetaceae	F	24	572,708	184,150	76,729	99.86	93.49	[73]
<i>Ruegeria mobilis</i>	Bacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae	F	42	210,119	62,535	14,889	99.90	98.00	[74]
<i>Streptomyces albidoflavus</i>	Bacteria; Actinobacteria; Streptomycetales; Streptomycetaceae	MIX	30	169,594	40,003	13,334	99.97	98.82	[75]
<i>Vibrio parahaemolyticus</i>	Bacteria; Gammaproteobacteria; Vibrionales; Vibrionaceae	MIX	157	327,904	80,566	5,132	99.94	98.41	[76]
<i>Mycobacterium leprae</i>	Bacteria; Actinobacteria; Corynebacteriales; Mycobacteriaceae	H	4	215	65	162	100.00	99.99	[77]
<i>Xanthomonas campestris pathovar musacearum</i> [Xcm]	Bacteria; Gammaproteobacteria; Xanthomonadales; Xanthomonadaceae	H	14	272	84	60	100.00	99.99	[78]
<i>Bordetella pertussis</i>	Bacteria; Betaproteobacteria; Burkholderiales; Alcaligenaceae	H	7	1,437	534	763	99.86	99.89	[79]
<i>Yersinia pestis</i>	Bacteria; Gammaproteobacteria; Enterobacterales; Yersiniaceae	H	17	1,364	284	167	99.91	99.88	[80]
<i>Salmonella enterica</i> serovar Typhi	Bacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae	H	19	1,964	446	235	99.92	98.96	[81]
<i>Bacillus anthracis</i>	Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae	H	19	2,965	509	268	99.68	99.92	[82]
<i>Clostridium difficile</i>	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridiaceae	H	151	3,686	1,018	67	-	-	[83]
<i>Salmonella enterica</i> serovar Paratyphi A	Bacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae	H	149	4,584	1,126	76	99.24	99.96	[84]
<i>Mycobacterium tuberculosis</i>	Bacteria; Actinobacteria; Corynebacteriales; Mycobacteriaceae	H	22	9,037	2,031	923	-	-	[85]
<i>Streptococcus pyogenes</i>	Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;	H	44	71,558	43,000	9,773	99.95	98.78	[86]
<i>Streptococcus mutans</i>	Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;	H	57	51,802	48,073	8,434	99.93	98.97	[87]
<i>Neisseria meningitidis</i>	Bacteria; Betaproteobacteria; Neisseriales; Neisseriaceae	H	20	80,279	76,602	38,301	99.58	97.84	[88]

Table S5. The number of the bi-allelic single nucleotide polymorphisms (BiPs), the relative frequency (ρ/θ) and relative effect (r/m) of recombination to point mutation in the *Roseobacter* population calculated from the whole genome alignment (WGA) generated by progressiveMauve v2.3.1 [19]. These ratios are also calculated from the same WGA but without the three long recombination segments inferred by ClonalFrameML (Fig. 1B). Within these three recombined segments, the BiPs are categorized into homoplasious and non-homoplasious BiPs.

	# Homoplasious SNPs	# Non-homoplasious SNPs	SNP density (SNPs/Mbp)	ρ/θ	r/m
WGA	270	12,576	4,238	0.076	18.10
WGA excluding three long recombined segments	45	1,228	453	0.052	1.13
WGA excluding homoplasious SNPs in three long recombined segments	45	12,576	4,163	0.068	17.12
WGA excluding non-homoplasious SNPs in three long recombined segments	270	1,228	496	0.084	1.83

Table S6. Statistics of between-clade (i.e., Clade R-I vs. Clade R-II in the *Roseobacter* population) and within-clade d_s values for the two clusters across 2,846 shared single-copy gene families identified with the K-means clustering algorithm. The optimal number (K=2) of the clusters was determined with the R package ‘NbClust’ (Fig. S2A). The number in the parenthesis denotes the number of core gene families assigned to each cluster.

	Cluster 1 (176)		Cluster 2 (2,670)	
	Median (d_s)	Mean (d_s)	Median (d_s)	Mean (d_s)
Clade R-I vs Clade R-II	0.166	0.197	0.000	0.968E-3
Within Clade R-I	0.000	0.000	0.000	0.000
Within Clade R-II	0.000	0.011	0.000	0.448E-3

Table S7. Functional annotation of the 53 genes exclusively and universally found in either Clade R-I or Clade R-II of the *Roseobacter* population. Clade R-I and Clade R-II specific genes are represented by loci from strains xm-d-517 and xm-a-104, respectively.

Locus id	Group	Function annotation
xm-d-517_00055	Clade R-I specific	hypothetical protein
xm-d-517_00274	Clade R-I specific	hypothetical protein
xm-d-517_00799	Clade R-I specific	Oligopeptide transport ATP-binding protein OppD (TC 3.A.1.5.1)
xm-d-517_00928	Clade R-I specific	hypothetical protein
xm-d-517_00929	Clade R-I specific	hypothetical protein
xm-d-517_00930	Clade R-I specific	hypothetical protein
xm-d-517_00931	Clade R-I specific	hypothetical protein
xm-d-517_00950	Clade R-I specific	possible Bacterial Ig-like domain (group 1)
xm-d-517_01363	Clade R-I specific	hypothetical protein
xm-d-517_01364	Clade R-I specific	hypothetical protein
xm-d-517_01365	Clade R-I specific	hypothetical protein
xm-d-517_01366	Clade R-I specific	hypothetical protein
xm-d-517_01366	Clade R-I specific	hypothetical protein
xm-d-517_01367	Clade R-I specific	hypothetical protein
xm-d-517_01368	Clade R-I specific	hypothetical protein
xm-d-517_01369	Clade R-I specific	hypothetical protein
xm-d-517_01370	Clade R-I specific	hypothetical protein
xm-d-517_02126	Clade R-I specific	hypothetical protein
xm-d-517_02127	Clade R-I specific	hypothetical protein
xm-d-517_02128	Clade R-I specific	hypothetical protein
xm-d-517_02129	Clade R-I specific	hypothetical protein
xm-d-517_02131	Clade R-I specific	Type I restriction-modification system, restriction subunit R (EC 3.1.21.3)
xm-d-517_02132	Clade R-I specific	Type I restriction-modification system, specificity subunit S (EC 3.1.21.3)
xm-d-517_02133	Clade R-I specific	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72)
xm-d-517_02134	Clade R-I specific	hypothetical protein
xm-d-517_02137	Clade R-I specific	hypothetical protein
xm-d-517_02139	Clade R-I specific	hypothetical protein
xm-d-517_02231	Clade R-I specific	hypothetical protein
xm-d-517_02234	Clade R-I specific	Integrase
xm-d-517_02235	Clade R-I specific	hypothetical protein
xm-d-517_02236	Clade R-I specific	hypothetical protein
xm-d-517_02238	Clade R-I specific	hypothetical protein
xm-d-517_02240	Clade R-I specific	HigA protein (antitoxin to HigB)
xm-d-517_02241	Clade R-I specific	hypothetical protein
xm-d-517_02242	Clade R-I specific	hypothetical protein
xm-d-517_02243	Clade R-I specific	hypothetical protein
xm-d-517_02244	Clade R-I specific	hypothetical protein
xm-d-517_02275	Clade R-I specific	hypothetical protein
xm-d-517_02326	Clade R-I specific	Endoribonuclease L-PSP family protein
xm-d-517_02332	Clade R-I specific	Opine oxidase subunit C
xm-d-517_02333	Clade R-I specific	Opine oxidase subunit B
xm-d-517_02334	Clade R-I specific	Transcriptional activator protein LysR
xm-d-517_02788	Clade R-I specific	hypothetical protein
xm-d-517_03128	Clade R-I specific	Mobile element protein
xm-a-104_00048	Clade R-II specific	hypothetical protein
xm-a-104_00050	Clade R-II specific	Transcriptional regulator, MecI family
xm-a-104_00060	Clade R-II specific	hypothetical protein
xm-a-104_00074	Clade R-II specific	hypothetical protein
xm-a-104_01262	Clade R-II specific	Putative bacteriophage-related protein
xm-a-104_01977	Clade R-II specific	Bll0064 protein
xm-a-104_03010	Clade R-II specific	Histone acetyltransferase HPA2 and related acetyltransferases
xm-a-104_03018	Clade R-II specific	hypothetical protein
xm-a-104_03024	Clade R-II specific	hypothetical protein

Table S8. The number of single-copy core gene families showing amino acid substitutions in the *Roseobacter* population. The pattern of amino acid substitution is considered to be congruent with the speciation pattern of the *Roseobacter* population if the amino acid state is identical within a clade (Clade R-I or Clade R-II) but different between the two diversified clades.

	#Gene families with amino acid substitutions	#Gene families with amino acid substitutions consistent with speciation pattern	#Gene families with amino acid substitutions inconsistent with speciation pattern
194 genes families	179	149	30
The remaining 2,652 gene families	309	42	267
All 2846 genes	488	191	297

Table S10. The list of substrates that are utilized differently among the four representative strains of the *Roseobacter* population. These substrates are collected from two Biolog phenotypic microplates (PM01 and PM02A), within which the compound in each well (except the negative control) is used by the bacteria as a sole carbon source. In the column ‘Utilization pattern’, isolates are clustered with parentheses if they show similar utilization pattern, but are grouped to different parentheses if they show significantly different utilization patterns. The substrates with similar utilization patterns are grouped and separated by colored shadings.

Substrate	Well No. of microplates	KEGG compound ID	Utilization pattern	Gene locus	Predicted function	Relevant pathway
D-Galactono-1,4-lactone	PM01-C02	C03383	(xm-m-339-2, xm-d-517) (xm-v-204, xm-m-314)			Galactose metabolism
2-Ketoglutaric acid	PM01-D06	C00026	(xm-m-339-2, xm-d-517) (xm-v-204, xm-m-314)	xm-d-517_02192, xm-d-517_02193	Isocitrate dehydrogenase (NADP); Isocitrate dehydrogenase (NAD)	TCA cycle
2-Ketobutyric acid	PM01-D07	C00109	(xm-m-339-2, xm-d-517) (xm-v-204, xm-m-314)			Glycine, serine and threonine metabolism
L-Threonine	PM01-G04	C00188	(xm-m-339-2, xm-d-517) (xm-v-204, xm-m-314)	xm-d-517_02199	L-Threonine dehydratase biosynthetic IIvA	Glycine, serine and threonine metabolism
p-Hydroxyphenylacetic acid	PM01-H02	C00642	(xm-m-339-2, xm-d-517) (xm-v-204, xm-m-314)			Tyrosine metabolism; Phenylalanine metabolism
L-Asparagine	PM01-D01	C00152	(xm-d-517, xm-v-204) (xm-m-339-2, xm-m-314)			Alanine, aspartate and glutamate metabolism
L-Glutamine	PM01-E01	C00064	(xm-d-517, xm-v-204) (xm-m-339-2, xm-m-314)			Alanine, aspartate and glutamate metabolism; Arginine biosynthesis
Propionic acid	PM01-F07	C00163	(xm-d-517, xm-v-204) (xm-m-339-2, xm-m-314)			Propanoate metabolism
L-Methionine	PM02A-G12	C00073	(xm-d-517, xm-v-204) (xm-m-339-2, xm-m-314)	xm-d-517_02286	Free methionine-R-sulfoxide reductase	
2'-Deoxyadenosine	PM01-E11	C00559	(xm-d-517, xm-m-339-2, xm-v-204) (xm-m-314)			Purine metabolism
Adenosine	PM01-E12	C00212	(xm-d-517, xm-m-339-2, xm-v-204) (xm-m-314)			Purine metabolism
Citric acid	PM01-F02	C00158	(xm-d-517, xm-m-339-2, xm-v-204) (xm-m-314)	xm-d-517_02192, xm-d-517_02193	Isocitrate dehydrogenase (NADP); Isocitrate dehydrogenase (NAD)	TCA cycle
Fumaric acid	PM01-F05	C00122	(xm-d-517, xm-m-339-2, xm-v-204) (xm-m-314)			TCA cycle
L-Lysine	PM02A-G11	C00047	(xm-d-517, xm-m-314, xm-v-204) (xm-m-339-2)			Lysine biosynthesis

Substrate	Well No. of microplates	KEGG compound ID	Utilization pattern	Gene locus	Predicted function	Relevant pathway
Citraconic acid	PM02A-E03	C02226	(xm-d-517, xm-m-314) (xm-m-339-2, xm-v-204)			
Putrescine	PM02A-H08	C00134	(xm-d-517, xm-m-314) (xm-m-339-2, xm-v-204)	xm-d-517_03110, xm-d-517_03111, xm-d-517_03112, xm-d-517_03113	Spermidine/putrescine-ABC transporter system	Putrescine degradation and utilization
3-0-beta-D-Galactopyranosyl-D-Arabinose	PM02A-B12	NA	(xm-d-517, xm-m-314) (xm-m-339-2) (xm-v-204)			
Tween 20	PM01-C05	C11624	(xm-d-517, xm-m-339-2, xm-m-314) (xm-v-204)			
Palatinose	PM02A-C12	C01742	(xm-m-314, xm-v-204) (xm-d-517) (xm-m-339-2)			
L-Histidine	PM02A-G06	C00135	(xm-d-517, xm-m-339-2, xm-m-314) (xm-v-204)	xm-d-517_02015	Histidine ammonia-lyase	Histidine degradation
L-Homoserine	PM02A-G07	C00263	(xm-m-339-2, xm-m-314, xm-v-204) (xm-d-517)	xm-d-517_02289	Homoserine/homoserine lactone efflux protein	Glycine, serine and threonine metabolism

Table S11. The recipe of the medium used to test the polyamine (putrescine and spermidine) utilization by the *Roseobacter* population. The recipe was modified from a medium used for isolating *Sulfurimonas autotrophica* from deep sea sediment [89]. When testing a polyamine compound as a sole carbon source, NH₄Cl is used as the nitrogen source. When testing a polyamine compound as a sole nitrogen source, pyruvate is used as the carbon source.

Ingredient	Final Concentration (w/v)	Reference for detailed medium
Basal salt		https://www.jcm.riken.jp/cgi-bin/jcm/jcm_grmd?GRMD=268
Vitamin mixtures		https://www.jcm.riken.jp/cgi-bin/jcm/jcm_grmd?GRMD=197
NaHCO ₃	0.004%	
Carbon source (polyamine or pyruvate)	0.1%	
Nitrogen source (polyamine or NH ₄ Cl)	0.1% for polyamine	
	0.025% for NH ₄ Cl	

Table S12. Genome statistics of the 16 isolates in the *Marinobacterium* population. The completeness and contamination are estimated with CheckM [7], and the remaining statistics are calculated with QUAST [62].

Species	GC	Completeness	Contamination	# Contigs	Longest contig (bp)	Total size (bp)	N50 (bp)
xm-a-121	48.4	97.93	0	17	604,343	2,278,837	341,648
xm-a-127	48.5	97.93	0	12	828,098	2,443,845	439,417
xm-a-152	48.2	97.93	0	22	325,262	2,536,603	236,555
xm-D-420	48.3	97.93	0	24	401,746	2,408,541	183,877
xm-d-509	48.2	97.5	0.43	34	754,379	2,484,536	339,335
xm-d-510	48.5	97.93	0	26	828,098	2,463,016	357,651
xm-d-530	48.3	97.93	0	10	653,235	2,249,011	370,837
xm-d-543	48.3	97.93	0.43	20	502,695	2,341,321	312,073
xm-d-564	48.4	97.93	0.43	22	449,423	2,297,663	211,158
xm-d-579	48.5	97.5	0	8	742,741	2,337,691	732,788
xm-g-48	48.2	97.93	0.43	21	754,379	2,465,319	332,500
xm-g-59	48.3	97.88	0.86	32	511,951	2,463,421	267,627
xm-m-312	48.4	97.93	0.43	19	502,173	2,316,931	280,901
xm-m-383	48.2	97.93	0.43	15	740,817	2,329,252	364,046
xm-v-233	48.4	97.93	0	25	389,131	2,288,368	211,806
xm-v-242	48.3	97.93	0.43	18	740,817	2,335,185	364,254

Supplementary References

1. Marie D, Partensky F, Vaulot D, Brussaard C. Enumeration of phytoplankton, bacteria, and viruses in marine samples. *Curr Protoc Cytom* 2001; **10**: 11.11.1-11.11.15.
2. Song J, Oh H-M, Cho J-C. Improved culturability of SAR11 strains in dilution-to-extinction culturing from the East Sea, West Pacific Ocean. *FEMS Microbiol Lett* 2009; **295**: 141–147.
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**: 2114–2120.
4. Andrews S. FastQC: a quality control tool for high throughput sequence data. *Bioinformatics* 2010; **1**.
5. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19**: 455–477.
6. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017; **13**: e1005595.
7. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; **25**: 1043–1055.
8. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; **30**: 2068–2069.
9. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genom* 2008; **9**: 75.
10. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016; **44**: D457-462.
11. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004; **32**: D277-280.
12. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, et al. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 2007; **35**: D237-240.

13. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015; **16**: 157.
14. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 2013; **30**: 772–780.
15. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; **25**: 1972–1973.
16. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol* 2016; **34**: 772–773.
17. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; **30**: 1312–1313.
18. Kessner L, Spinard E, Gomez-Chiarri M, Rowley DC, Nelson DR. Draft genome sequence of *Aliiroseovarius crassostreae* CV919-312, the causative agent of *Roseovarius* oyster disease (formerly juvenile oyster disease). *Genome Announc* 2016; **4**: e00148-16.
19. Darling AE, Mau B, Perna NT. Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010; **5**: e11147.
20. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004; **14**: 1394–1403.
21. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 2015; **11**: e1004041.
22. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS Genet* 2012; **8**: e1002453.
23. Malinsky M, Trucchi E, Lawson DJ, Falush D. RADpainter and fineRADstructure: population inference from RADseq data. *Mol Biol Evol* 2018; **35**: 1284–1290.
24. Hughes AL, Friedman R. Nucleotide substitution and recombination at orthologous loci in *Staphylococcus aureus*. *J Bacteriol* 2005; **187**: 2698–2704.
25. Hughes AL, French JO. Homologous recombination and the pattern of nucleotide substitution in *Ehrlichia ruminantium*. *Gene* 2007; **387**: 31–37.

26. Sun Y, Luo H. Homologous recombination in core genomes facilitates marine bacterial adaptation. *Appl Environ Microbiol* 2018; **84**: e02545-17.
27. Luo H, Thompson LR, Stingl U, Hughes AL. Selection maintains low genomic GC content in marine SAR11 lineages. *Mol Biol Evol* 2015; **32**: 2738–2748.
28. Hellweger FL, Huang Y, Luo H. Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model. *ISME J* 2018; **12**: 1180–1187.
29. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res* 1988; **16**: 8207–8211.
30. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res* 2005; **33**: 1141–1153.
31. Brandis G, Hughes D. The selective advantage of synonymous codon usage bias in *Salmonella*. *PLoS Genet* 2016; **12**: e1005926.
32. Shields DC, Sharp PM. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res* 1987; **15**: 8023–8040.
33. Bulmer M. The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res* 1990; **18**: 2869–2873.
34. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007; **24**: 1586–1591.
35. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. *J Stat Softw* 2014; **61**: 1–36.
36. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* 2013; **41**: W29–W33.
37. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; **32**: 268–274.
38. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. PlasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* 2016; btw493.

39. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, et al. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* 2017; **33**: 475–482.
40. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, et al. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 2017; **45**: W30–W35.
41. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016; **44**: W16–W21.
42. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* 2017; **33**: 3340–3347.
43. Cury J, Jové T, Touchon M, Néron B, Rocha EP. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* 2016; **44**: 4539–4550.
44. Rowe-Magnus DA, Guerout A-M, Biskri L, Bouige P, Mazel D. Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. *Genome Res* 2003; **13**: 428–442.
45. Bochner BR, Gadzinski P, Panomitros E. Phenotype microArrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* 2001; **11**: 1246–1255.
46. Bochner BR. New technologies to assess genotype–phenotype relationships. *Nat Rev Genet* 2003; **4**: 309–314.
47. Vaas LAI, Sikorski J, Hofner B, Fiebig A, Buddruhs N, Klenk H-P, et al. opm: an R package for analysing OmniLog(R) phenotype microarray data. *Bioinformatics* 2013; **29**: 1823–1824.
48. Smriga S, Fernandez VI, Mitchell JG, Stocker R. Chemotaxis toward phytoplankton drives organic matter partitioning among marine bacteria. *Proc Natl Acad Sci USA* 2016; **113**: 1576–1581.
49. Raina J-B, Fernandez V, Lambert B, Stocker R, Seymour JR. The role of microbial motility and chemotaxis in symbiosis. *Nat Rev Microbiol* 2019; **17**: 284–294.

50. Sonnenschein EC, Syit DA, Grossart H-P, Ullrich MS. Chemotaxis of *Marinobacter adhaerens* and its impact on attachment to the diatom *Thalassiosira weissflogii*. *Appl Environ Microbiol* 2012; **78**: 6900–6907.
51. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014; **8**: 1553–1565.
52. Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, et al. The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* 2009; **106**: 15527–15533.
53. Luo H, Moran MA. How do divergent ecological strategies emerge among marine bacterioplankton lineages? *Trends Microbiol* 2015; **23**: 577–584.
54. Mou X, Vila-Costa M, Sun S, Zhao W, Sharma S, Moran MA. Metatranscriptomic signature of exogenous polyamine utilization by coastal bacterioplankton. *Environmental Microbiology Reports* 2011; **3**: 798–806.
55. Mou X, Sun S, Rayapati P, Moran MA. Genes for transport and metabolism of spermidine in *Ruegeria pomeroyi* DSS-3 and other marine bacteria. *Aquat Microb Ecol* 2010; **58**: 311–321.
56. Kashiwagi K, Miyamoto S, Nukui E, Kobayashi H, Igarashi K. Functions of potA and potD proteins in spermidine-preferential uptake system in *Escherichia coli*. *J Biol Chem* 1993; **268**: 19358–19363.
57. Cohan FM. Transmission in the origins of bacterial diversity, from ecotypes to phyla. *Microbiol Spectr* 2017; **5**: MTBP-0014-2016.
58. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, et al. Population genomics of early events in the ecological differentiation of bacteria. *Science* 2012; **336**: 48–51.
59. Bendall ML, Stevens SLR, Chan LK, Malfatti S, Schwientek P, Tremblay J, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J* 2016; **10**: 1589–1601.
60. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol* 2014; **22**: 235–247.
61. Cohan FM. Prokaryotic species concepts. R.M. Kliman, ed. *The Encyclopedia of Evolutionary Biology*, 1st ed. 2016. pp 119–129.

62. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUASt: quality assessment tool for genome assemblies. *Bioinformatics* 2013; **29**: 1072–1075.
63. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, et al. Patterns of gene flow define species of thermophilic Archaea. *PLOS Biol* 2012; **10**.
64. Wielgoss S, Didelot X, Chaudhuri RR, Liu X, Weedall GD, Velicer GJ, et al. A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium *Myxococcus xanthus*. *ISME J* 2016; **10**: 2468–77.
65. Porter SS, Chang PL, Conow CA, Dunham JP, Friesen ML. Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic *Mesorhizobium*. *ISME J* 2017; **11**: 248–262.
66. Lopes LD, Pereira E Silva M de C, Weisberg AJ, Davis EW, Yan Q, Varize C de S, et al. Genome variations between rhizosphere and bulk soil ecotypes of a *Pseudomonas koreensis* population. *Environ Microbiol* 2018; **20**: 4401–4414.
67. López-Pérez M, Gonzaga A, Rodríguez-Valera F. Genomic diversity of “deep ecotype” *Alteromonas macleodii* isolates: evidence for pan-Mediterranean clonal frames. *Genome Biol Evol* 2013; **5**: 1220–1232.
68. Youngblut ND, Wirth JS, Henriksen JR, Smith M, Simon H, Metcalf WW, et al. Genomic and phenotypic differentiation among *Methanosarcina mazei* populations from Columbia River sediment. *ISME J* 2015; **9**: 1–15.
69. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* 2014; **344**: 416–420.
70. Hoetzing M, Hahn MW. Genomic divergence and cohesion in a species of pelagic freshwater bacteria. *BMC Genomics* 2017; **18**: 794.
71. Kopac S, Wang Z, Wiedenbeck J, Sherry J, Wu M, Cohan FM. Genomic heterogeneity and ecological speciation within one subspecies of *Bacillus subtilis*. *Appl Environ Microbiol* 2014; **80**: 4842–4853.

72. González-Torres P, Gabaldón T. Genome variation in the model halophilic bacterium *Salinibacter ruber*. *Front Microbiol* 2018; **9**: 1499.
73. Choudoir MJ, Buckley DH. Phylogenetic conservatism of thermal traits explains dispersal limitation and genomic differentiation of *Streptomyces* sister-taxa. *ISME J* 2018; **12**: 2176–2186.
74. Sonnenschein EC, Nielsen KF, D’Alvise P, Porsby CH, Melchiorsen J, Heilmann J, et al. Global occurrence and heterogeneity of the Roseobacter-clade species *Ruegeria mobilis*. *ISME J* 2017; **11**: 569–583.
75. Li Y, Pinto-Tomás AA, Rong X, Cheng K, Liu M, Huang Y. Population genomics insights into adaptive evolution and ecological differentiation in streptomycetes. *Appl Environ Microbiol* 2019; AEM.02555-18.
76. Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, et al. Epidemic clones, oceanic gene pools, and eco-LD in the free living marine pathogen *Vibrio parahaemolyticus*. *Mol Biol Evol* 2015; **32**: 1396–1410.
77. Monot M, Honore N, Garnier T, Zidane N, Sherafi D, Paniz-Mondolfi a, et al. Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. 2009; **41**: 1–24.
78. Wasukira A, Tayebwa J, Thwaites R, Paszkiewicz K, Aritua V, Kubiriba J, et al. Genome-wide sequencing reveals two major sub-lineages in the genetically monomorphic pathogen *Xanthomonas campestris* pathovar *musacearum*. *Genes* 2012; **3**: 361–377.
79. Bart MJ, van Gent M, van der Heide HG, Boekhorst J, Hermans P, Parkhill J, et al. Comparative genomics of prevaccination and modern *Bordetella pertussis* strains. *BMC Genomics* 2010; **11**: 627.
80. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, et al. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* 2010; **42**: 1140–1143.
81. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, et al. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet* 2008; **40**: 987–993.

82. Achtman M. Insights from genomic comparisons of genetically monomorphic bacterial pathogens. *Philos Trans R Soc Lond, B, Biol Sci* 2012; **367**: 860–867.
83. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat Genet* 2013; **45**: 109–13.
84. Zhou Z, McCann A, Weill F-X, Blin C, Nair S, Wain J, et al. Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc Natl Acad Sci USA* 2014; **111**: 12199–204.
85. Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* 2010; **42**: 498–503.
86. Bao Y-J, Shapiro BJ, Lee SW, Ploplis VA, Castellino FJ. Phenotypic differentiation of *Streptococcus pyogenes* populations is induced by recombination-driven gene-specific sweeps. *Sci Rep* 2016; **6**: 36644.
87. Cornejo OE, Lefébure T, Pavinski Bitar PD, Lang P, Richards VP, Eilertson K, et al. Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. *Mol Biol Evol* 2013; **30**: 881–893.
88. Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, et al. *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci USA* 2011; **108**: 4494–9.
89. Inagaki F, Takai K, Kobayashi H, Nealson KH, Horikoshi K. *Sulfurimonas autotrophica* gen. nov., sp. nov., a novel sulfur-oxidizing epsilon-proteobacterium isolated from hydrothermal sediments in the Mid-Okinawa Trough. *Int J Syst Evol Microbiol* 2003; **53**: 1801–1805.