Open access • Posted Content • DOI:10.1101/2021.03.10.434754

Metapopulation Structure of Diatom-associated Marine Bacteria — Source link

Liping Qu, Liping Qu, Xiaoyuan Feng, Yuerong Chen ...+5 more authors

Institutions: The Chinese University of Hong Kong, Shantou University

Published on: 10 Mar 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Population, Roseobacter and Metapopulation

Metapopulation Structure of Diatom-associated Marine Bacteria

Liping Qu[1,2^], Xiaoyuan Feng[3,2^], Yuerong Chen[1], Lingyu Li[1], Xiaojun Wang[3,2], Zhong Hu[1,4], Hui Wang[1,4*], Haiwei Luo[3,2*]

[1]Biology Department, College of Science, and Guangdong Provincial Key Laboratory of Marine Biotechnology, and Institute of Marine Sciences, Shantou University, Shantou, China
[2]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
[3]Simon F. S. Li Marine Science Laboratory, School of Life Sciences and State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong SAR
[4]Southern Marine Science and Engineering Guangdong Laboratory, Guangzhou, China

[^]These authors contributed equally to this study.

[*]**Corresponding author:**
Haiwei Luo
School of Life Sciences, The Chinese University of Hong Kong
Shatin, Hong Kong SAR
Phone: (+852) 39436121
E-mail: hluo2006@gmail.com

Hui Wang
Biology Department, College of Science, Shantou University
Shantou, China
Phone: (+86) 754 86502721
E-mail: wanghui@stu.edu.cn

Running Title:

Key words: Roseobacter, *Sulfitobacter*, diatom, phycosphere, population structure

Abstract

Marine bacteria-phytoplankton interaction ultimately shapes ecosystem productivity. The biochemical mechanisms underlying their interactions become increasingly known, yet how these ubiquitous interactions drive bacterial evolution has not been illustrated. Here, we sequenced genomes of 294 bacterial isolates associated with 19 coexisting diatom cells. These bacteria constitute eight genetically monomorphic populations of the globally abundant Roseobacter group. Six of these populations are members of *Sulfitobacter*, arguably the most prevalent bacteria associated with marine diatoms. A key finding is that populations varying at the intra-specific level have been differentiated and each are either associated with a single diatom host or with multiple hosts not overlapping with those of other populations. These closely related populations further show functional differentiation; they differ in motility phenotype and they harbor distinct types of secretion systems with implication for mediating organismal interactions. This interesting host-dependent population structure is even evident for demes within a genetically monomorphic population but each associated with a distinct diatom cell, as shown by a greater similarity in genome content between isolates from the same host compared to those from different hosts. Importantly, the intra- and inter-population differentiation pattern remains when the analyses are restricted to isolates from intra-specific diatom hosts, ruling out distinct selective pressures and instead suggesting coexisting microalgal cells as physical barriers of bacterial gene flow. Taken together, microalgae-associated bacteria display a unique microscale metapopulation structure, which consists of numerous small populations whose evolution is driven by random genetic drift.

Since marine phytoplankton contribute one-half of global primary production (*1*) and since heterotrophic bacterioplankton process 40-50% of the carbon fixed by marine phytoplankton (*2, 3*), bacteria-phytoplankton interaction is an important process that ultimately drives carbon cycling and regulates ecosystem productivity. The physical interface mediating these ubiquitous interactions is a microzone of a few cell diameters immediately surrounding an individual phytoplankton cell, which is termed as 'phycosphere' (*4*). In the phycosphere of eukaryotic marine phytoplankton lineages (e.g., diatoms, dinoflagellates, coccolithophores, and green pico-algae), bacterial communities are consistently dominated by a handful of taxa including *Rhodobacteraceae* (mostly the Roseobacter group), *Alteromonadaceae*, and *Flavobacteriaceae* (*4, 5*), and the bacterial community assembly at the phycosphere of a given phytoplankton species is reproducible (*6*). These recurrent patterns in part result from the innate ability of phytoplankton to modulate their bacteria consortia by secreting secondary metabolites such as rosmarinic acid and azelaic acid released by a diatom species, which promote the attachment and growth of certain roseobacters but suppress opportunistic bacteria (*7*). Another important mechanism is the resource-based niche partitioning among these major bacterial associates at nutrient-enhanced phycosphere (*8, 9*). For example, diatoms may use their abundant metabolites such as 2,3-dihydroxypropane-1-sulfonate (DHPS) for targeted feeding of beneficial symbionts among which roseobacters represent a dominant group (*10*).

Recent studies have revealed a greater diversity of the mechanisms underlying the symbiosis between Roseobacter lineages and phytoplankton species than previously appreciated. Some roseobacters such as *Sulfitobacter* spp., *Ruegeria* spp. and *Dinoroseobacter* spp. establish mutualistic interactions with diatoms and microscopic green algae by providing growth factors such as vitamins and indole-3-acetic acid (IAA) to phytoplankton hosts in exchange for labile

78 organic matter (*10-12*), whereas others such as *Sulfitobacter* spp. are virulent to coccolithophores
79 by releasing algicides (*13*). In another type of interaction, some roseobacters including
80 *Phaeobacter* spp. and *Dinoroseobacter* spp. each act initially as a mutualist and later as a
81 parasite of coccolithophores and dinoflagellates, respectively (*14, 15*). An important implication
82 from these studies is that closely related roseobacters (e.g., members of *Sulfitobacter*) may
83 employ different mechanisms to interact with phytoplankton. This high diversity of roseobacter-
84 phytoplankton interaction therefore suggests that the phycosphere may act as an effective barrier
85 of gene flow among symbiotic roseobacters associated with different phytoplankton cells,
86 leading to independent evolution of even closely related roseobacter populations in the
87 seemingly well mixed seawater.

88 To test this hypothesis, we sought to determine the population structure of roseobacters
89 colonizing the phycosphere of coexisting microalgal cells. A few environmental factors
90 including nutrient availability (*16*), interactions with phages (*17*) and phytoplankton or particles
91 (*18*) are known to drive roseobacter population differentiation. To single out diatom phycosphere
92 from other confounding factors that may drive roseobacter evolution, populations associated with
93 diatoms isolated from a single seawater sample were analyzed. We collected 1 L of seawater
94 from the Pearl River Estuary located at the northern boundary of the South China Sea, isolated
95 45 diatom cells varying at the level of phylogenetic relatedness (Fig. 1A, Table S1), cultivated
96 over 850 roseobacters associated with 19 of these diatoms (Fig. 1A, Table S2), and sequenced
97 294 genomes of these roseobacters (Table S3), among which six genomes are complete and
98 closed (Table S4) by additional sequencing with Nanopore (Supplemental Text 1.1-1.3). These
99 newly sequenced roseobacters comprise eight clades (Fig. 1B; see methods in Supplemental Text
100 1.4), among which six are related to three species of *Sulfitobacter*, a roseobacter genus most

commonly found on diatoms (*19*). These include clade-2a and clade-2b related to *Sulfitobacter pseudonitzschiae*, clade-2c related to *S. geojensis*, clade-2d, clade-2e1 and clade-2e2 related to *S. mediterraneus* (Fig. 1B). The remaining clade-1 and clade-3 are related to *Marivita cryptomonadis* and *Ponticoccus* sp. LZ-14, respectively, which are distantly related to *Sulfitobacter*. Members within each clade share identical 16S rRNA genes, display whole-genome average nucleotide identity (ANI) over 99.99% (Fig. S1), and vary up to 45 non-singleton single nucleotide polymorphisms (SNPs) in which the rare variant occurs in at least two genomes (Table S5). Despite this genetic monomorphism, there is ample evidence that the within-clade members are predominantly from the environment rather than a result of clonal replications during the laboratory cultivation of the diatom cells, a required process prior to roseobacter isolation (see Supplemental Text 2, Fig. S2 and Table S5).

A key finding is that closely related roseobacter populations associated with different diatom cells are often genetically differentiated (see methods in Supplemental Text 1.6). This is clearly supported by populations diverged at different phylogenetic depths. The two clades (clade-2a and clade-2b) related to *S. pseudonitzschiae* share 94.16% ANI (Fig. S1), which is at the boundary (95% ANI) delineating a distinct species (*20*). The SNP density within each clade is extremely low across the whole genome but becomes very high between the clades (Fig. S3), indicating that these two clades each have fixed distinct alleles and are thus genetically isolated. The clade-2a is composed solely of members isolated from *Skeletonema menzelii* 26 (abbreviated as ‘SM26’), whereas the clade-2b comprise members from five diatom cells of the class *Coscinodiscophyceae* affiliated with three species including *Skeletonema costatum* (SC1 & SC33), *Mediolabrus comicus* (MC36 & MC52), and *Thalassiosira tenera* (TT37), and one diatom cell of a distantly related class *Mediophyceae* affiliated with the species *Trieres chinensis*

(TC12) (Fig. 1B, Table S2). Among the three clades related to *S. mediterraneus*, clade-2e1 and clade-2e2 share 97.78% ANI, suggesting that they have not yet separated into two distinct species. However, a dramatic increase in SNP density of between-clade comparisons compared to the within-clade comparisons across the whole genome (Fig. S4A,B,C) indicates that these two clades are under ongoing speciation. The clade-2d is more divergent, showing 89.34% ANI to clade-2e1 and 89.51% ANI to clade-2e2, though it shares with the latter two clades very high similarities (99.86% and 99.93%, respectively) at the 16S rRNA gene sequence. The SNP density derived from the comparison between clade-2e1 and clade-2e2 is much lower than when each is compared to clade-2d (Fig. S4C,D,E), suggesting that the genomes are differentiated to a greater extent when the phylogenetic depth becomes larger for these overall very closely related clades. Importantly, these clades each have a distinct host range. Members of clade-2e1 are exclusively associated with a single diatom host (SC5), whereas members of clade-2e2 are from three hosts of the same diatom species (SC2, SC4 and SC7) and members of clade-2d are from hosts of two different families (SC1 and MC36) (Fig. 1B, Table S2). Population differentiation of the sampled clades is further supported by genome rearrangement in both chromosome and plasmids, as shown by the more conserved gene order of within-clade members (Fig. S2) compared to that of between-clade members (Fig. S5) for clade-2a and clade-2b. Multiple genome rearrangement events were also observed between clade-2d and clade-2e2 (Fig. S6), though within-clade comparisons cannot be made because no more closed genomes are available in both clades.

Diatom-dependent differentiation is also evident from the more closely related roseobacter demes associated with different hosts but sharing membership of the same clade. Because of the genetic monomorphism at the core genomes (Table S5), the within-clade

147 members do not show a reliable phylogenetic structure. We therefore turned to explore the
148 accessory genes which are shared by a subset of the genomes under comparison. A simple
149 clustering based on the presence and absence pattern of the accessory genes identified clusters
150 corresponding to distinct diatom hosts. In clade-2b, for example, members associated with SC33
151 largely constitute an independent cluster separated from members associated with other hosts
152 (Fig. 2A). Likewise, in clade-2c, members from *Minutocellus polymorphus* 20 (MP20) are
153 overall well separated from those associated with *Thalassiosira rotula* 60 (TR60) (Fig. 2B); in
154 clade-2e2, most members associated with SC2 are separated from those with other hosts (Fig.
155 2C); and in clade-3, members from SC2 and those from SC6 are generally clustered into two
156 separate groups (Fig. 2D). For the remaining clades (clade-2d & clade-1) with members
157 associated with multiple hosts, host-dependent clustering is not obvious (Fig. S7).
158 We further identified important evidence for population differentiation at the functional
159 level. The secretion systems are well known to mediate bacteria-bacteria and bacteria-host
160 interactions. Interestingly, the presence and absence pattern of three secretion systems
161 differentiates the clade-2e1, clade-2e2 and clade-2d related to *S. mediterraneus* (Table S6).
162 Specifically, the type VI secretion system (T6SS) transports effector proteins into both
163 prokaryotic and eukaryotic cells in a contact-dependent manner (*21, 22*). This system was
164 reported in only a few roseobacter lineages (*23, 24*). We showed an exclusive presence of a
165 T6SS gene cluster on the chromosome of clade-2d members. Another uncommon secretion
166 system in roseobacters, thus far only reported in the roseobacter species *Marinovum algicola*
167 (*25*), is the type II secretion system (T2SS), which promotes the release of folded proteins,
168 mainly extracellular enzymes such as proteases, lipases, phosphatases, and polysaccharide
169 hydrolases, to the extracellular milieu or displayed on the cell surface (*26*). We found an

exclusive occurrence of a T2SS cluster on a plasmid of the clade-2d members. In terms of the type IV secretion system (T4SS), the *virB/D4* type secretes effector proteins and plasmid DNA to target both bacteria and hosts (*27-29*), whereas the *trb* type transports plasmid DNA between bacteria (*29*). While the *virB/D4* is commonly found among roseobacters (*30*), the *trb* is rare in these bacteria. Consistent with this pattern, a *virB/D4* gene cluster was found in all three clades, but a *trb* gene cluster was exclusively identified on the chromosome of clade-2e2 members. In the case of the *S. pseudonitzschiae* related clades, both clade-2a and clade-2b carry the *virB/D4*-based T4SS, but they differ in copy numbers. The *virB/D4* copy number difference was similarly found between the three *S. mediterraneus* related clades, but a unique observation was that the clade-2d members possess an additional copy on their chromosomes instead of the plasmids where this type of secretion system usually locates. No other secretion systems were found in clade-2a and clade-2b. Gene clusters encoding all secretion systems locate within the genomic islands except the T6SS of clade-2d (Table S6), suggesting that roseobacter-diatom and/or roseobacter-bacteria interactions are highly dynamic.

Since these secretion systems may mediate either pathogenic, or commensal, or mutualistic relationships with hosts and/or other bacteria (*21, 26, 28, 29, 31, 32*), their differential presence among the clades suggests that distinct clades may exert different and even opposite physiological effects on the diatom hosts. This motivated us to set up experimental assays (Supplemental Text 1.7) to compare the effects of co-culture of diatom and roseobacter, the latter represented by each of the *S. mediterraneus* and *S. pseudonitzschiae* clades, on the growth of the diatom. Among the 11 tested roseobacter isolates, eight significantly promoted the growth of the diatom, whereas the remaining three did not significantly change the growth rate of the diatom (Fig. S8). We did not observe consistent differences between closely related clades

193 regarding their effects on the diatom growth (Fig. S8). While this assay was motivated by the
194 observation of clade-specific secretion systems, there is no direct link between the algal growth
195 change and the differential presence of the secretion systems in the bacterial symbionts.
196 Another important metabolic trait relevant to roseobacter-phytoplankton interaction is the
197 bacterial motility (*33*), which is differentially present among these related roseobacter clades.
198 Three phylogenetically distinct flagellar gene clusters (FGCs) designated as *fla1*, *fla2* and *fla3*
199 (Fig. 3A) have been identified in the Roseobacter group, and carrying any of them may enable
200 motility (*34, 35*). Among these, *fla2* is present in a plasmid of the *S. pseudonitzschiae* related
201 clade-2e1, clade-2e2 and clade 2d, whereas *fla1* was exclusively found on the chromosome of
202 clade-2e2 (Fig. 3A, Table S6; see methods in Supplemental Text 1.8). Despite the presence of
203 the flagellum-encoding gene clusters, the flagella were not detected by transmission electron
204 microscopy (Fig. 3B) and the motility phenotype was not observed under the experimental
205 condition (Fig. 3B; Supplemental Text 1.9). The lack of flagella and motility in these related
206 clades is likely due to inappropriate physicochemical conditions set in the laboratory experiment,
207 as temperature (*36*), pH (*37*), salinity (*38*) and metal ions (*39*) were demonstrated to induce the
208 expression of the FGC genes in other bacteria. In terms of the *S. pseudonitzschiae* related clade-
209 2a and clade-2b, they did not possess any type of FGC (Table S6), but both instead carry
210 homologs of two candidate gene clusters (i.e., type-IVb tight adherence pilus gene cluster; Table
211 S6) recently hypothesized to be responsible for dendritic motility (*34*). We showed that the
212 clade-2b members possess an additional copy located within a genomic island compared to
213 clade-2a members, consistent with the greater swimming and dendritic motilities observed in the
214 former (Fig. 3C).

Previous studies demonstrated that selection for niche adaptation drives sympatric population differentiation in free-living prokaryotic lineages (*18, 40-42*). In the present study, we provided evidence that closely related but genetically discrete populations of several sympatric Roseobacter lineages each have a distinct diatom host range. The pattern of *Sulfitobacter mediterraneus* related populations (clade-2e1, clade-2e2, clade-2d) is of particular interest. This is because populations of this species varying at different stages of differentiation, including those at the very beginning (i.e., within-clade demes differentiated only by accessory gene content), at the middle (i.e., closely related clade under ongoing speciation), and at the completion of speciation, were captured and each found to be associated with different hosts of the same diatom species *Skeletonema costatum*. Since members of the same microalgal species likely release a similar set of organic compounds to the phycosphere and impose other physicochemical parameters (e.g., reactive oxygen species) at similar levels, the observed differentiation of the symbiotic bacterial populations is less likely driven by ecological selection imposed by differential exposure to different microalgal exudates. Instead, symbiotic bacteria may be trapped in the phycosphere (*43*), leading to a reduced opportunity of recombination between bacteria associated with different diatom cells compared to those within the same phycosphere. This is a new mechanism of bacterial population differentiation in the pelagic ocean and represents one of the few examples of population differentiation at the sympatric scale due to physical barriers of gene flow. To put it in context, previous cases of bacterial population differentiation in a sympatric pelagic environment were linked to ecological barriers of gene flow, such as differentiated populations colonizing organic particles of different sizes (*41*) or those inhabiting the bulk seawater versus phycosphere/particles (*18*).

Our observation of subdivision of highly closely related populations each showing genetic monomorphism has important implications for understanding the population structure of the diatom-associated symbiotic roseobacters. Similar population structure was previously demonstrated in obligately host-dependent bacteria such as endosymbionts subjected to repeated bottlenecks during transmissions to new hosts in small numbers of bacterial cells (*44*), and also proposed for generalist marine bacteria such as *Vibrio* spp. which experience short bursts as a result of intensive use of ephemeral resources like organic particles followed by dispersal and colonization of new particles with low numbers of cells (*45*). These two known mechanisms lead to the formation of "metapopulation structure", in which the population is divided into subpopulations each colonizing a transient resource such as hosts and particles (*45*). Our results suggest that the population structure of the diatom-associated roseobacters aligns well with the metapopulation structure. Hence, phycosphere colonization represents a new mechanism leading to bacterial metapopulation structure in the pelagic ocean. The exact processes leading to the formation of metapopulation structure of these diatom-associated roseobacters remains unknown, however. It could be a result of short burst owing to intensive use of the organic substrates enriched in the phycosphere. It is also possible that bacteria-diatom associations have been maintained at the evolutionary timescale, such that diatom host-dependent population differentiation is evident even at the completion of speciation.

Formation of metapopulation structure in a bacterial species leads to a reduced effective population size (N_e) of the species (*45*), a key parameter in understanding the population genetic mechanism underpinning biological evolution and defined as the size of an ideal population carrying the same amount of the neutral genetic diversity as is observed in the real population (*45, 46*). Because N_e is the inverse of the power of random genetic drift (*45*), the reduced N_e of a

bacterial species owing to the formation of metapopulation structure suggests an increased power of genetic drift in driving the evolution of diatom-association roseobacters. As a consequence, the diatom-associated roseobacter populations are predicted to more readily accept the horizontally transferred genetic elements, which are often mildly deleterious owing to the selfish propagation of most mobile genetic elements at the expense of cellular fitness (*47*) but may also carry functional traits such as antimicrobial genes that increase competitive advantages of the bacteria at the phycosphere. Given that diatoms and roseobacters are among the most abundant phytoplankton and bacterial groups, respectively, in today's ocean (*19*), our findings have important implications for biogeochemical cycles mediated by bacteria-phytoplankton interactions.

It is important to clarify that while the concept of phycosphere has been adopted in this and many other studies, there has been no direct experimental evidence for its occurrence because of technological challenges in separating this microenvironment from the bulk seawater (*4*). Prior studies established microalgal phycosphere as a hotspot of carbon and nutrient cycling in the pelagic ocean. Here, we revealed that phycosphere of diatom cells may act as an effective physical barrier of gene flow between nearly identical symbiotic roseobacters, thereby conferring a new role of phycosphere in driving the evolution of pelagic marine bacteria.

References

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237-240 (1998).
2. F. Azam, T. Fenchel, J. G. Field, J. S. Gray, L. A. Meyer-Reil, F. Thingstad, The ecological role of water-column microbes in the sea. *Marine Ecology Progress Series* **10**, 257-263 (1983).
3. J. J. Cole, S. Findlay, M. L. Pace, Bacterial production in fresh and saltwater ecosystems: A cross-system overview. *Marine Ecology Progress Series* **43**, 1-10 (1988).
4. J. R. Seymour, S. A. Amin, J.-B. Raina, R. Stocker, Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology* **2**, 17065 (2017).
5. A. Buchan, G. R. LeCleir, C. A. Gulvik, J. M. Gonzalez, Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nature Reviews Microbiology* **12**, 686–698 (2014).
6. J. Mönnich, J. Tebben, J. Bergemann, R. Case, S. Wohlrab, T. Harder, Niche-based assembly of bacterial consortia on the diatom *Thalassiosira rotula* is stable and reproducible. *The ISME Journal* **14**, 1614-1625 (2020).
7. A. A. Shibl, A. Isaac, M. A. Ochsenkühn, A. Cárdenas, C. Fei, G. Behringer, M. Arnoux, N. Drou, M. P. Santos, K. C. Gunsalus, C. R. Voolstra, S. A. Amin, Diatom modulation of select bacteria through use of two unique secondary metabolites. *Proceedings of the National Academy of Sciences* **117**, 27445-27455 (2020).
8. H. Fu, M. Uchimiya, J. Gore, M. Moran, Ecological drivers of bacterial community assembly in synthetic phycospheres. *Proceedings of the National Academy of Sciences* **117**, 201917265 (2020).
9. F. X. Ferrer-González, B. Widner, N. R. Holderman, J. Glushka, A. S. Edison, E. B. Kujawinski, M. A. Moran, Resource partitioning of phytoplankton metabolites that support bacterial heterotrophy. *The ISME Journal* **15**, 762-773 (2021).
10. B. P. Durham, S. Sharma, H. Luo, C. B. Smith, S. A. Amin, S. J. Bender, S. P. Dearth, B. A. S. Van Mooy, S. R. Campagna, E. B. Kujawinski, E. V. Armbrust, M. A. Moran, Cryptic carbon and sulfur cycling between surface ocean plankton. *Proceedings of the National Academy of Sciences* **112**, 453-457 (2015).
11. S. A. Amin, L. R. Hmelo, H. M. van Tol, B. P. Durham, L. T. Carlson, K. R. Heal, R. L. Morales, C. T. Berthiaume, M. S. Parker, B. Djunaedi, A. E. Ingalls, M. R. Parsek, M. A. Moran, E. V. Armbrust, Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**, 98-101 (2015).
12. M. B. Cooper, E. Kazamia, K. E. Helliwell, U. J. Kudahl, A. Sayer, G. L. Wheeler, A. G. Smith, Cross-exchange of B-vitamins underpins a mutualistic interaction between *Ostreococcus tauri* and *Dinoroseobacter shibae*. *The ISME Journal* **13**, 334-345 (2019).
13. N. Barak-Gavish, M. J. Frada, C. Ku, P. A. Lee, G. R. DiTullio, S. Malitsky, A. Aharoni, S. J. Green, R. Rotkopf, E. Kartvelishvily, U. Sheyn, D. Schatz, A. Vardi, Bacterial virulence against an oceanic bloom-forming phytoplankter is mediated by algal DMSP. *Science Advances* **4**, eaau5716 (2018).
14. I. Wagner-Dobler, B. Ballhausen, M. Berger, T. Brinkhoff, I. Buchholz, B. Bunk, H. Cypionka, R. Daniel, T. Drepper, G. Gerdts, S. Hahnke, C. Han, D. Jahn, D. Kalhoefer, H. Kiss, H.-P. Klenk, N. Kyrpides, W. Liebl, H. Liesegang, L. Meincke, A. Pati, J.

Petersen, T. Piekarski, C. Pommerenke, S. Pradella, R. Pukall, R. Rabus, E. Stackebrandt, S. Thole, L. Thompson, P. Tielen, J. Tomasch, M. von Jan, N. Wanphrut, A. Wichels, H. Zech, M. Simon, The complete genome sequence of the algal symbiont *Dinoroseobacter shibae*: a hitchhiker's guide to life in the sea. *The ISME Journal* **4**, 61-77 (2010).

15. E. Segev, T. P. Wyche, K. H. Kim, J. Petersen, C. Ellebrandt, H. Vlamakis, N. Barteneva, J. N. Paulson, L. Chai, J. Clardy, R. Kolter, Dynamic metabolic exchange governs a marine algal-bacterial interaction. *eLife* **5**, e17473 (2016).
16. Y. Sun, Y. Zhang, J. T. Hollibaugh, H. Luo, Ecotype diversification of an abundant Roseobacter lineage. *Environmental Microbiology* **19**, 1625-1638 (2017).
17. E. C. Sonnenschein, K. F. Nielsen, P. D'Alvise, C. H. Porsby, J. Melchiorsen, J. Heilmann, P. G. Kalatzis, M. Lopez-Perez, B. Bunk, C. Sproer, M. Middelboe, L. Gram, Global occurrence and heterogeneity of the Roseobacter-clade species *Ruegeria mobilis*. *The ISME Journal* **11**, 569-583 (2017).
18. X. Wang, Y. Zhang, M. Ren, T. Xia, X. Chu, C. Liu, X. Lin, Y. Huang, Z. Chen, A. Yan, H. Luo, Cryptic speciation of a pelagic Roseobacter population varying at a few thousand nucleotide sites. *The ISME Journal* **14**, 3106-3119 (2020).
19. S. A. Amin, M. S. Parker, E. V. Armbrust, Interactions between diatoms and bacteria. *Microbiology and Molecular Biology Reviews* **76**, 667-684 (2012).
20. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114 (2018).
21. T. G. Sana, A. Hachani, I. Bucior, C. Soscia, S. Garvis, E. Termine, J. Engel, A. Filloux, S. Bleves, The second type VI secretion system of *Pseudomonas aeruginosa* strain PAO1 is regulated by quorum sensing and Fur and modulates internalization in epithelial cells. *Journal of Biological Chemistry* **287**, 27095-27105 (2012).
22. J. M. Silverman, Y. R. Brunet, E. Cascales, J. D. Mougous, Structure and regulation of the type VI secretion system. *Annual Review of Microbiology* **66**, 453-472 (2012).
23. Y. Zhang, Y. Sun, N. Jiao, R. Stepanauskas, H. Luo, Ecological genomics of the uncultivated marine Roseobacter lineage CHAB-I-5. *Applied and Environmental Microbiology* **82**, 2100-2111 (2016).
24. A. J. Collins, M. S. Fullmer, J. P. Gogarten, S. V. Nyholm, Comparative genomics of Roseobacter clade bacteria isolated from the accessory nidamental gland of *Euprymna scolopes*. *Frontiers in Microbiology* **6**, 123 (2015).
25. O. Frank, M. Göker, S. Pradella, J. Petersen, Ocean's Twelve: flagellar and biofilm chromids in the multipartite genome of *Marinovum algicola* DG898 exemplify functional compartmentalization. *Environmental Microbiology* **17**, 4019-4034 (2015).
26. M. Nivaskumar, O. Francetic, Type II secretion system: A magic beanstalk or a protein escalator. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1843**, 1568-1577 (2014).
27. R. Fronzes, P. J. Christie, G. Waksman, The structural biology of type IV secretion systems. *Nature Reviews Microbiology* **7**, 703-714 (2009).
28. K. Wallden, A. Rivera-Calzada, G. Waksman, Microreview: Type IV secretion systems: versatility and diversity in function. *Cellular Microbiology* **12**, 1203-1212 (2010).
29. L. Chen, Y. Chen, D. W. Wood, E. W. Nester, A New Type IV Secretion System Promotes Conjugal Transfer in Agrobacterium tumefaciens. *Journal of Bacteriology* **184**, 4838-4845 (2002).

30. H. Luo, M. A. Moran, Evolutionary ecology of the marine Roseobacter clade. *Microbiology and Molecular Biology Reviews* **78**, 573-587 (2014).
31. K. V. Korotkov, M. Sandkvist, W. G. J. Hol, The type II secretion system: biogenesis, molecular architecture and mechanism. *Nature Reviews Microbiology* **10**, 336-351 (2012).
32. M. S. Nelson, M. J. Sadowsky, Secretion systems and signal exchange between nitrogen-fixing rhizobia and legumes. *Front Plant Sci* **6**, 491-491 (2015).
33. H. Geng, R. Belas, Molecular mechanisms underlying roseobacter–phytoplankton symbioses. *Current Opinion in Biotechnology* **21**, 332-338 (2010).
34. P. Bartling, J. Vollmers, J. Petersen, The first world swimming championships of roseobacters—Phylogenomic insights into an exceptional motility phenotype. *Systematic and Applied Microbiology* **41**, 544-554 (2018).
35. V. Michael, O. Frank, P. Bartling, C. Scheuner, M. Goker, H. Brinkmann, J. Petersen, Biofilm plasmids with a rhamnose operon are widely distributed determinants of the 'swim-or-stick' lifestyle in roseobacters. *The ISME Journal* **10**, 2498-2513 (2016).
36. O. A. Soutourina, P. N. Bertin, Regulation cascade of flagellar expression in Gram-negative bacteria. *FEMS Microbiology Reviews* **27**, 505-523 (2003).
37. R. A. Alm, P. Guerry, T. J. Trust, The Campylobacter sigma 54 flaB flagellin promoter is subject to environmental regulation. *Journal of Bacteriology* **175**, 4448-4455 (1993).
38. C. Karlsen, S. M. Paulsen, H. S. Tunsjø, S. Krinner, H. Sørum, P. Haugen, N.-P. Willassen, Motility and flagellin gene expression in the fish pathogen *Vibrio salmonicida*: Effects of salinity and temperature. *Microbial Pathogenesis* **45**, 258-264 (2008).
39. A. Guzzo, C. Diorio, M. S. DuBow, Transcription of the *Escherichia coli fliC* gene is regulated by metal ions. *Applied and Environmental Microbiology* **57**, 2255-2259 (1991).
40. F. M. Cohan, Transmission in the origins of bacterial diversity, from ecotypes to phyla. *Microbiology Spectrum* **5**, (2017).
41. B. J. Shapiro, J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabó, M. F. Polz, E. J. Alm, Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48-51 (2012).
42. H. Cadillo-Quiroz, X. Didelot, N. L. Held, A. Herrera, A. Darling, M. L. Reno, D. J. Krause, R. J. Whitaker, Patterns of gene flow define species of thermophilic Archaea. *PLOS Biology* **10**, e1001265 (2012).
43. J.-B. Raina, V. Fernandez, B. Lambert, R. Stocker, J. R. Seymour, The role of microbial motility and chemotaxis in symbiosis. *Nature Reviews Microbiology* **17**, 284-294 (2019).
44. L.-M. Bobay, H. Ochman, Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evolutionary Biology* **18**, 153 (2018).
45. C. Fraser, E. J. Alm, M. F. Polz, B. G. Spratt, W. P. Hanage, The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741-746 (2009).
46. B. Charlesworth, Effective population size. *Current Biology* **12**, R716-R717 (2002).
47. R. J. Hall, F. J. Whelan, J. O. McInerney, Y. Ou, M. R. Domingo-Sananes, Horizontal Gene transfer as a source of conflict and cooperation in Prokaryotes. *Frontiers in Microbiology* **11**, (2020).

48. Y. Liu, Q. Lin, J. Feng, F. Yang, H. Du, Z. Hu, H. Wang, Differences in metabolic potential between particle-associated and free-living bacteria along Pearl River Estuary. *Science of The Total Environment* **728**, 138856 (2020).
49. P. Daste, D. Neuville, B. Victor-Baptiste, A simple procedure for obtaining clonal isolation of diatoms. *British Phycological Journal* **18**, 1-3 (1983).
50. R. R. L. Guillard, J. H. Ryther, Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (cleve) Gran. *Canadian Journal of Microbiology* **8**, 229-239 (1962).
51. M. Sapp, A. S. Schwaderer, K. H. Wiltshire, H. G. Hoppe, G. Gerdts, A. Wichels, Species-specific bacterial communities in the phycosphere of microalgae? *Microbial Ecology* **53**, 683-699 (2007).
52. J. Zhou, M. A. Bruns, J. M. Tiedje, DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology* **62**, 316 (1996).
53. L. Hu, X. Peng, J. Zhou, Y. Zhang, S. Xu, X. Mao, Y. Gao, J. Liang, Characterizing the interactions among a dinoflagellate, flagellate and bacteria in the phycosphere of *Alexandrium tamarense* (Dinophyta). *Frontiers in Marine Science* **2**, 100 (2015).
54. M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, T. L. Madden, NCBI BLAST: a better web interface. *Nucleic Acids Research* **36**, W5-W9 (2008).
55. F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539-539 (2011).
56. S. Capellagutierrez, J. M. Sillamartinez, T. Gabaldon, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
57. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268-274 (2014).
58. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587-589 (2017).
59. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**, 518-522 (2018).
60. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**, W256-W259 (2019).
61. H. Wang, H. D. t. Laughinghouse, M. A. Anderson, F. Chen, E. Willliams, A. R. Place, O. Zmora, Y. Zohar, T. Zheng, R. T. Hill, Novel bacterial isolate from Permian groundwater, capable of aggregating potential biofuel-producing microalga *Nannochloropsis oceanica* IMET1. *Applied and Environmental Microbiology* **78**, 1445-1453 (2012).
62. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**, 996-1004 (2018).
63. S. Nurk, A. Bankevich, D. Antipov, A. A. Gurevich, A. Korobeynikov, A. Lapidus, A. D. Prjibelski, A. Pyshkin, A. Sirotkin, Y. Sirotkin, R. Stepanauskas, S. R. Clingenpeel, T.

Woyke, J. S. McLean, R. Lasken, G. Tesler, M. A. Alekseyev, P. A. Pevzner, Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology* **20**, 714-737 (2013).

64. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**, 1043-1055 (2015).

65. Y. Chen, F. Nie, S.-Q. Xie, Y.-F. Zheng, Q. Dai, T. Bray, Y.-X. Wang, J.-F. Xing, Z.-J. Huang, D.-P. Wang, L.-J. He, F. Luo, J.-X. Wang, Y.-Z. Liu, C.-L. Xiao, Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications* **12**, 60 (2021).

66. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540-546 (2019).

67. R. Vaser, I. Sovic, N. Nagarajan, M. Sikic, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* **27**, 737-746 (2017).

68. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One* **9**, e112963 (2014).

69. R. R. Wick, M. B. Schultz, J. Zobel, K. E. Holt, Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350-3352 (2015).

70. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).

71. T. Brettin, J. J. Davis, T. Disz, R. A. Edwards, S. Gerdes, G. J. Olsen, R. Olson, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, J. A. Thomason, R. Stevens, V. Vonstein, A. R. Wattam, F. Xia, RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports* **5**, 8365 (2015).

72. M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, S. K. S, KEGG: kyoto encyclopaedia of genes and genomes. *Nucleic Acids Research* **28**, 27-30(24) (2000).

73. G. S. Vernikos, J. Parkhill, Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22**, 2196-2203 (2006).

74. D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, G. W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533-1542 (2017).

75. M. Simon, C. Scheuner, J. P. Meierkolthoff, T. Brinkhoff, I. Wagnerdobler, M. Ulbrich, H. Klenk, D. Schomburg, J. Petersen, M. Goker, Phylogenomics of *Rhodobacteraceae* reveals evolutionary adaptation to marine and non-marine habitats. *The ISME Journal* **11**, 1483-1499 (2017).

76. J. Rozewicki, S. Li, K. M. Amada, D. M. Standley, K. Katoh, MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research* **47**, W5-W10 (2019).

77. T. J. Treangen, B. D. Ondov, S. Koren, A. M. Phillippy, The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology* **15**, 524 (2014).

78. C. Jain, L. M. Rodriguezr, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114 (2018).
79. R. C. Team, R: A language and environment for statistical computing. *MSOR connections* **1**, 12-21 (2014).
80. A. C. Darling, B. Mau, F. R. Blattner, N. T. Perna, Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14**, 1394-1403 (2004).
81. A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).
82. C. Chen, H. Chen, Y. Zhang, H. R. Thomas, M. H. Frank, Y. He, R. Xia, TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant* **13**, 1194-1202 (2020).
83. L. Marquez, G. A. Morales, M. A. S. De Rodriganez, E. Almansa, F. J. Moyano, M. Diaz, Comments on the calculation of the specific growth rate in experiments with untagged individuals. *Scientia Marina* **79**, 505-508 (2015).
84. R. P. Trenkenshu, Calculayion of the specific growth rate of microalgae. *Marine Biological Journal* **4**, 100-108 (2019).
85. D. Pfeiffer, D. Schüler, Quantifying the benefit of a dedicated “magnetoskeleton” in bacterial magnetotaxis by live-cell motility tracking and soft agar swimming assay. *Applied and Environmental Microbiology* **86**, e01976-01919 (2020).
86. D. Krysciak, J. Grote, M. Rodriguez Orbegoso, C. Utpatel, K. U. Förstner, L. Li, C. Schmeisser, H. B. Krishnan, W. R. Streit, RNA Sequencing analysis of the broad-host-range strain *Sinorhizobium fredii* NGR234 identifies a large set of genes linked to quorum sensing-dependent regulation in the background of a *traI* and *ngrI* deletion mutant. *Applied and Environmental Microbiology* **80**, 5655 (2014).
87. E. Teira, S. Martínez-García, C. Lønborg, X. A. Alvarez-Salgado, Growth rates of different phylogenetic bacterioplankton groups in a coastal upwelling system. *Environmental microbiology reports* **1**, 545-554 (2009).
88. Y. Sun, K. E. Powell, W. Sung, M. Lynch, M. A. Moran, H. Luo, Spontaneous mutations of a model heterotrophic marine bacterium. *The ISME Journal* **11**, 1713-1718 (2017).

Acknowledgements

We thank Xingqin Lin for helpful discussion. We also thank Qingyu Li for assisting microalgal isolation and identification, Min Yang for helping with microalgal cell counting, Fumin Yang, Qing Zhang and Huiling Chen for their help in bacteria isolation, Xiaoyu Yang for assisting a preliminary experiment of the motility assays. H.W. was supported by the National Natural Science Foundation of China (41676116), Key Special Project for Introduced Talents Team of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (GML2019ZD0606), and Guangdong Science and Technology Department (2019A1515011139). H.L. was supported by the Hong Kong Research Grants Council General Research Fund (14163917) and the Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16).

Author Contributions

H.L. conceptualized the work, designed this study, directed the bioinformatics analyses, interpreted the data, and wrote the main manuscript. H.W. directed the sampling work, diatom isolation, and roseobacter cultivation, experimental assay, and related writing, co-interpreted the data, and provided comments to the manuscript. L.Q. collected the sample, performed cultivation and characterized the cultures. X.F. and L.Q. performed all the bioinformatics, co-interpreted the data, drafted the technical details, and prepared figures and tables. L.L. contributed to bacterial isolation, and L.L. and Y.C. performed physiological assays. X.W. contributed to the bioinformatics. H.Z. contributed to the discussion and provided comments to the manuscript.

Conflict of Interest

561 The authors declare no competing commercial interests in relation to the submitted work.

562

563 **Data availability**

564 The 18S rRNA gene sequences of the diatoms are available at NCBI under the accession
565 number MW494549 - MW494585. The raw genomic sequencing data and assembled genomes of
566 the 294 isolates are available at NCBI under the accession number PRJNA691705.

567

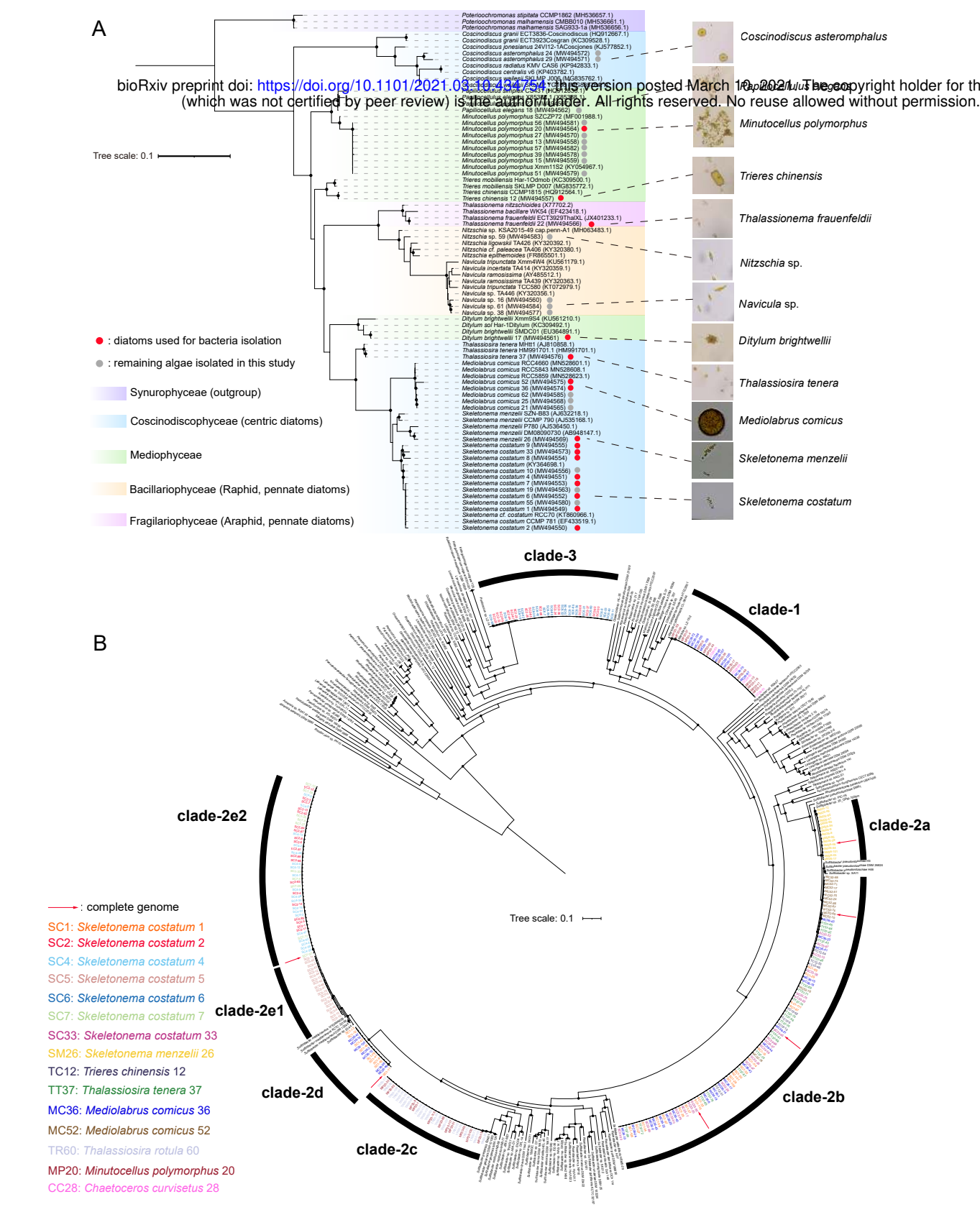

Fig 1. Phylogenetics of diatoms and Roseobacters isolated from this study. (A) Maximum likelihood (ML) tree showing the phylogenetic positions of diatoms isolated in this study. The phylogeny was inferred using IQ-TREE based on the 18S rRNA gene sequences with lengths longer than 1,600 bp. Three *Synurophyceae* strains were used as the outgroup. Solid circles in the phylogeny indicate nodes with ultrafast bootstrap (UFBoot) values > 95%. Diatom strains used for bacteria cultivation and other isolated microalgae are marked with red and gray dots, respectively. The photos of microalgae obtained under an optical microscope are shown on the right panel. (B) ML phylogenomic tree showing the phylogenetic positions of roseobacters isolated from this study. The phylogeny was inferred using IQ-TREE based on the concatenation of 120 conserved bacterial proteins (see methods). Solid circles in the phylogeny indicate nodes with UFBoot values > 95%. The associated microalgae of bacterial strains are differentiated with colors. The six complete and closed genomes are marked with red arrows.

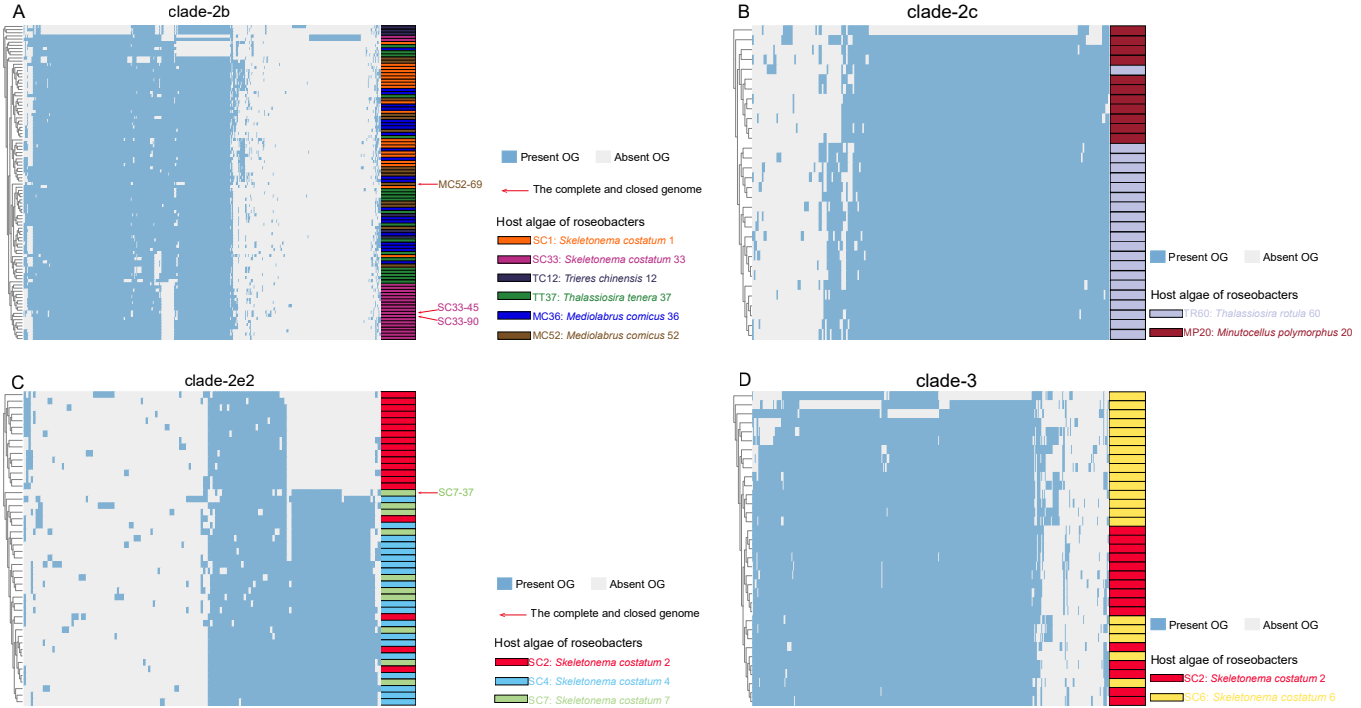

Fig. 2. The clustering of accessory genes in the genomes of (A) clade-2b, (B) clade-2c, (C) clade-2e2 and (D) clade-3. The dendrogram of genome clustering was generated based on the presence and absence of their orthologous gene families (OGs), which are colored in blue and gray, respectively. The associated microalgae of bacterial strains are differentiated with colors. The complete and closed genomes are marked with red arrows.

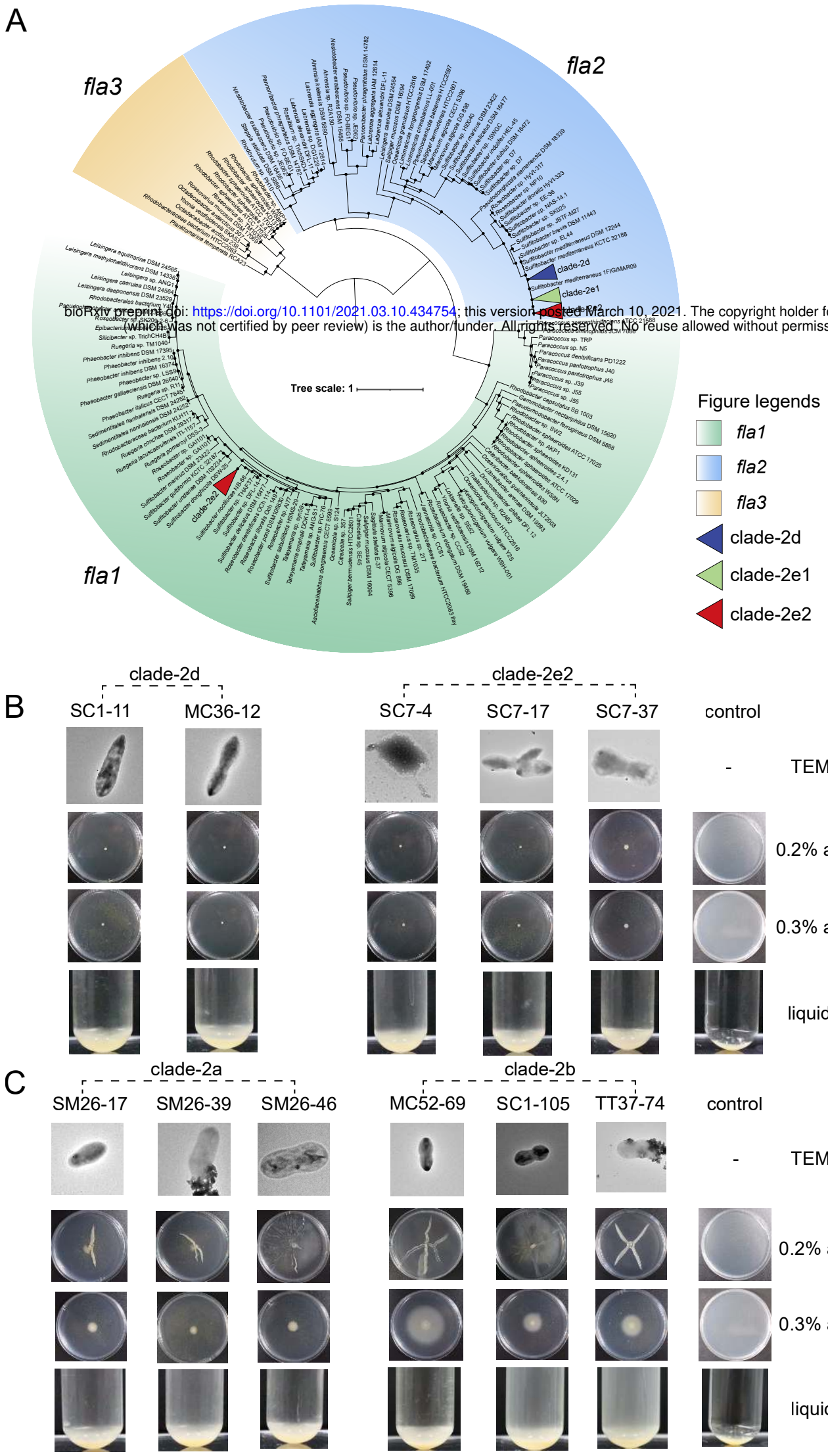

Fig. 3. The metabolic traits. (A) ML phylogenetic tree of the three homologous types of flagellar gene clusters (FGCs) found in the Roseobacter group. This phylogeny was built based on four marker flagellar proteins (FliF, FlgI, FlgH and FlhA). Solid circles in the phylogeny indicate nodes with UFBoot values > 95%. (B, C) Photos showing the cellular morphology under a transmission electron microscope (TEM), motility on the 2216E agar plates with 0.2% or 0.3% agar, and sedimentation phenotypes in the liquid 2216E medium. Eleven representative strains were used in the assay, including two from clade-2d (B), three from clade-2e2 (B), three from clade-2a (C) and three from clade-2b (C). All strains from clade-2e1 were lost.

Supplementary Materials for

Metapopulation Structure of Diatom-associated Marine Bacteria

Liping Qu, Xiaoyuan Feng, Yuerong Chen, Lingyu Li, Xiaojun Wang, Zhong Hu, Hui Wang*, Haiwei Luo*

*Hui Wang. Email: wanghui@stu.edu.cn
*Haiwei Luo. Email: hluo2006@gmail.com

This PDF file includes:

Supplementary Text
Figs. S1 to S8

Other Supplementary Materials for this manuscript include the following:

Data S1 to S6

Text 1. Methods

1.1 Isolation, identification and phylogenetic analysis of the microalgae

One-liter of seawater (0.5 m below surface) was collected from the Pearl River Estuary (113.7221° E, 21.9935° N) in January 2018 (*48*), stored in a cooler (4 °C) and transferred to the laboratory within 24 hours. Single cells of microalgae were isolated from the seawater using micropipettes under optical microscope following a previously described method (*49*). The microalgal cells were then washed repeatedly using fresh axenic F/2 medium (*50*) to remove the free-living bacteria and loosely associated bacteria around microalgal cells (*51*). The washed microalgal cells each were inoculated in a 24-well plate containing 1 mL of fresh F/2 medium to increase the cell density. After incubation at 20°C with 12 h/12 h light-dark cycle at 200 µmol photons m^{-2} s^{-1} for 3-5 days, the microalgal culture in each well was transferred into a conical flask with 30 mL of fresh F/2 medium to increase the biomass of microalgae. In the exponential growth phase, 15 mL of the cultures each were fixed with 1% Lugol's solution and sent to the Fujian Provincial Department of Ocean and Fisheries for morphological characterization and species identification. Besides, the taxonomy of microalgae was further validated using 18S rRNA genes. The DNA was extracted from another 10 mL of the microalgal cultures using the CTAB method (*52*). The 18S rRNA genes were amplified using primers (SSU-F: 5'-ACCTGGTTGATCCTGCCAGT-3' and SSU-R: 5'-TCACCTACGGAAACCTTGT-3') following a previous study (*53*) and sent for sequencing using the Sanger platform. These 18S rRNA gene sequences were subjected to a maximum likelihood (ML) phylogenetic analysis along with a few reference sequences, which were selected based on a preliminary BLASTn (*54*) result against the NCBI nt database. These sequences were aligned using Clustal Omega v1.2.4 (*55*) with default parameters and trimmed using trimAl v1.4.rev15 (*56*) with '-automated1'

641 option. The phylogenetic tree was constructed using IQTREE v1.6.12 (*57*) with the Modelfinder
642 (*58*) for model selection, and 1,000 ultrafast bootstrap replicates were sampled to assess the
643 robustness of the phylogeny (*59*). The phylogenetic tree was visualized using iTOL v5.6 (*60*).

644 *1.2 Isolation and identification of bacteria in phycosphere*

645 The 1 mL of microalgal cultures each were collected during the logarithmic growth phase
646 followed by a 10-fold serial dilution. Solid 2216E agar plates (BD Bioscience, USA) were spread
647 with 100 µL of each dilution and incubated at 25°C for one week. Single colonies were selected
648 and repeatedly streaked on 2216E agar plates to isolate and purify bacterial strains initially
649 associated with microalgae. The purified bacterial strains were subjected to the colony
650 polymerase chain reaction (colony PCR) to retrieve the 16S rRNA genes for taxonomy
651 identification. The 16S rRNA genes were amplified using the universal primers 27F and 1492R
652 as described previously (*61*) and were partially (~700 bp) sequenced using 27F on the Sanger
653 platform at Invitrogen Trading (Shanghai) Co., Ltd. Their taxonomy was inferred using BLASTn
654 on the NCBI website, and a total of 294 strains with top BLAST hit to the Roseobacter group
655 were kept for genome sequencing.

656 *1.3 Genome sequencing, assembly and annotation*

657 The genomic DNA of 294 Roseobacter genomes was extracted using the Bacterial Genome
658 DNA Rapid Extraction Kit (Guangzhou Dongsheng Biotech Co., Ltd.) and sequenced at the
659 Beijing Genomics Institute (BGI, China) using the Illumina Hiseq xten PE150 platform. Raw
660 reads were quality trimmed using Trimmomatic v0.39 (*62*) with options
661 'SLIDINGWINDOW:4:15 MAXINFO:40:0.9 MINLEN:40' and assembled using SPAdes
662 v3.10.1(*63*) with '-careful' option. Only contigs with length > 1,000 bp and sequencing depth >

663 5x were retained. Genome completeness, contamination, and strain heterogeneity (Table S3)
664 were calculated using CheckM v1.1.2 (*64*).

665 Six isolates (SM26-46, SC33-45, SC33-90, MC52-69, SC1-11 and SC7-37) were
666 additionally sequenced with the Nanopore platform (Nextomics Biosciences Co., Ltd.) to retrieve
667 complete and closed genomes. The mismatches between Nanopore and Illumina reads were
668 reconciled according to the following procedure. Raw reads of the Nanopore sequencing were
669 first corrected by Necat v0.0.1 (*65*) with 'PREP_OUTPUT_COVERAGE=100
670 CNS_OUTPUT_COVERAGE=50' parameters. The polished reads were then assembled using
671 Flye v2.6 (*66*) with the '--plasmids' parameter. The initial assemblies were corrected twice using
672 the polished Nanopore sequencing reads by Racon v1.4.13 (*67*) with '-m 8 -x -6 -g -8 -w 500'
673 options and five times using the Illumina sequencing reads by Pilon v1.23 (*68*) with default
674 parameters. The Bandage v0.8.1 (*69*) was used to check whether the final assembled
675 chromosomes and plasmids were closed, which showed that the chromosome and plasmids in all
676 six genomes are closed except the plasmid 2b_P2 in MC52-69 and plasmid 2b_P5 in SC33-90
677 (Table S4).

678 Protein-coding genes were identified using Prokka v1.14.6 (*70*) with default parameters, and
679 their functions were annotated using online RAST (*71*) and KEGG server (*72*). Genomic islands
680 were predicted using Alien_hunter v1.7 (*73*) with default parameters.

681 *1.4 Phylogenomic tree construction*

682 An ML phylogenetic tree was constructed based on 120 conserved bacterial genes (*74*) at the
683 amino acid level to identify the phylogenetic position of 294 sequenced Roseobacter genomes.
684 Other reference Roseobacter genomes included in the phylogeny were used following previous
685 studies (*35, 75*). The 120 conserved proteins each were aligned using MAFFT v7.222 (*76*) with

default parameters and trimmed using trimAl with '-resoverlap 0.55 -seqoverlap 60' options. The trimmed alignments were linked together to form a super-alignment for each genome. The phylogenetic tree was constructed using IQTREE v1.6.12 (*57*) with the Modelfinder (*58*) for model selection, and a total of 1,000 ultrafast bootstrap replicates were sampled to assess the robustness of the phylogeny (*59*). The phylogenetic tree was visualized using iTOL v5.6 (*60*).

1.5 Plasmid identification for the clade-2e1 members

Since a closed genome was not available to the clade-2e1, we used the following procedure to detect whether a contig is part of the chromosome or plasmid. Contigs of clade-2e1 genomes were aligned to the complete and closed genome of clade-2e2 (SC7-37) using Parsnp v1.2 (*77*). The contig was considered to be located on the chromosome or plasmid if > 80% region of this contig was aligned to the chromosome or plasmid of SC7-37. The remaining contigs were considered as unassigned.

1.6 Population genomics analyses

The whole-genome average nucleotide identity (ANI) was calculated using FastANI v1.3 (*78*) to assess the genomic sequence similarity within and between clades. Besides, single nucleotide polymorphisms (SNPs) were identified using Parsnp v1.2 (*77*) with default parameters, and the SNP density was calculated using 10 Kbp sliding windows and plotted with custom scripts in R v3.6.1 (*79*). To investigate the genomic structural variation between clades, two complete and closed genomes (SC1-11 in clade-2d and SC7-37 in clade-2e2) were aligned using Mauve v2015-02-26 (*80*) with default parameters and the segment arrangement of the chromosome and the plasmid was visualized using Mauve with the 'Min LCB weight' parameter around 2,000. A similar comparison was also performed for two complete and closed genomes

from clade-2a (SM26-46) and clade-2b (SC33-45), and for three complete and closed genomes within clade-2b (SC33-45, SC33-90 and MC52-69), respectively.

The clustering of accessory genes within clade was used to investigate whether the association with different hosts caused the differentiation of roseobacters at the genome content level. Orthologous gene families (OGs) were predicted using Roary v3.13.0 (*81*) with default parameters for genomes of the eight clades separately. The presence/absence pattern of accessory OGs was summarized as a binary matrix. The Euclidean distance of each pair of genomes was calculated using TBtools v1.0695 (*82*), and then genomes were clustered with the 'complete' method and visualized using TBtools.

1.7 Co-cultivation of an axenic diatom culture and the Roseobacter isolates

The 11 Roseobacter representatives each were co-cultured with an axenic diatom culture *Phaeodactylum tricornutum* CCMP2561 to verify the effect of these roseobacters on the microalgal growth. The axenic microalgal culture was obtained from the Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China. The diatom was inoculated in axenic F/2 medium at 20°C with 12 h/12 h light-dark cycle at 200 μmol photons m^{-2} s^{-1}. The microalgal cells were counted three times each day using the Beckman Coulter Z2 (Beckman Coulter Inc., America) until they reached the logarithmic growth phase. The initial cell concentration of the microalgae used for growth assay was about 10^6 cells mL^{-1}.

The 11 Roseobacter strains each were inoculated in 5 mL of 2216E liquid medium and incubated at 25°C with 150 rpm shaking until the OD_{600} reached 0.6 - 0.7. For each strain, 3 mL of bacterial culture was centrifuged for 1 min at 12,000 rpm, and pellets were washed three times with sterile seawater and re-suspended in 100 μL of sterile seawater.

730 The 100 µL of bacterial suspension was inoculated in 30 mL of axenic microalgal culture as
731 the experimental group, and another 30 mL of the axenic microalgal culture without bacterial
732 inoculations were used as the negative control. Three replicates were set for all control and
733 experimental groups. The microalgal cells were counted in triplicate at the third day of the co-
734 culture experiment. The specific growth rate of the microalgae (μ_c for the control group and μ_e
735 for the experimental groups) over three days was calculated as:

736 $$\mu = (\ln B_2 - \ln B_1)/(t_2 - t_1)$$

737 where B_1 and B_2 were the cellular density (concentration) in the culture at t_1 (0 d) and t_2 (3
738 d), respectively (*83, 84*). The significance level between μ_c and μ_e was estimated using the
739 Student's t-test.

740 *1.8 Phylogenetic analysis for flagellar gene clusters*

741 Three homologous and phylogenetically distinguishable types of flagellar gene cluster
742 (FGC) have been identified in the Roseobacter group (*35*). While none of them were identified in
743 closely related clade-2a and clade-2b members, its homologs were found in clade-2d, clade-2e1
744 and clade-2e2 members. Next, phylogenetic analysis was performed for all the 83 genomes from
745 the latter three clades to identify their FGC type. Four flagellar marker genes (*fliF, flgH, flgI, and*
746 *flhA*) (*35*) were aligned at the amino acid level using MAFFT v7.222 (*76*) with default
747 parameters and trimmed using trimAl with '-resoverlap 0.55 -seqoverlap 60' options. The
748 trimmed alignments were concatenated and a phylogenetic tree was constructed using IQTREE
749 v1.6.12 (*57*) with the Modelfinder (*58*) function. A total of 1,000 ultrafast bootstrap replicates
750 were sampled to assess the robustness of the phylogeny (*59*), and the phylogenetic tree was
751 visualized using iTOL v5.6 (*60*).

752 *1.9 Motility assay*

753 Eleven roseobacters were used in the motility assay, including three strains in clade-2a, three
754 in clade-2b, two in clade-2d, and three in clade-2e2. The clade-2e1 was not included because all
755 strains in this clade were lost. The 11 strains each were inoculated in 5 mL of 2216E liquid
756 medium and incubated at 25°C with 150 rpm shaking until the optical density value at 600 nm
757 (OD_{600}) reached 0.6-0.7. The flagella and pili were observed under a transmission electron
758 microscope (TEM; JEM-F200, Japan), and the motility of representative strains was tested on
759 solid and liquid medium.

760 Aliquots (6 mL) of liquid culture for each bacterial strain was centrifuged at 1,000 rpm for
761 10 min. The pellets were rinsed with axenic water 2-3 times. Pellets were fixed overnight with 1
762 mL of 2.5% glutaraldehyde solution at 4°C. Fixed cells were resuspended with axenic water and
763 stabilized on the 200-mesh copper grid (Beijing Zhongjingkeyi Technology Co., Ltd) with the
764 carbon support film. Stabilized cells were stained with 2% sodium phosphotungstate solution for
765 1-2 min and observed by TEM.

766 The motility of the 11 strains were tested on soft agar plates with 0.2% and 0.3% agar (w/v)
767 following previously study (*34, 85*). Soft agar plates were point inoculated with 3 µL of bacterial
768 pre-cultures and incubated at 25°C for 6 d. The motility was also tested using sedimentation
769 assays by inoculating 100 µL of bacterial cultures in 5 mL of fresh 2216E liquid medium and
770 incubated at room temperature for 24 h without shaking (*86*).

771

Text 2. Supplemental result

High genetic similarities of within-clade members are not due to clonal replication during laboratory cultivation

Considering that all members of each clade showed extremely high genetic identity at the DNA sequence level, it is possible that the collected isolates can be a mix of cells each representing a distinct genotype originally inhabiting the wild environment and cells replicated clonally during the laboratory cultivation of the microalgae. The latter situation needs to be considered since roseobacter isolation was performed after the one-month incubation of microalgae and since roseobacters divide approximately once per day (*87*). It is important to differentiate these two scenarios because the inference of population structure could be biased if laboratory clones dominated our collections. Here, we provide several lines of evidence against the laboratory replication hypothesis.

If all members of a clade would have been laboratory replicates from a single ancestor, we asked whether the observed SNPs could be explained by mutations during the period of laboratory cultivation of the diatoms. If a simple assumption is made that each SNP site was caused by a mutation, then the expected number of mutations (S) can be estimated based on the spontaneous mutation rate (μ), bacterial growth rate or number of cell divisions per day (G), the alignment length (L), laboratory cultivation time (T) and number of genomes under comparison (N), according to the following equation:

$$S = \mu * G * L * T * N$$

To make it simple, we used the mutation rate of the model Roseobacter strain, *Ruegeria pomeroyi* DSS-3 ($\mu = 1.39 \times 10^{-10}$ per nucleotide per generation) (*88*) and the doubling time of wild roseobacters (G = *approximately one generation per day*) (*87*). Given that L = 4.34, 4.84,

795 4.37, 3.89, 3.56, 3.87, 3.74 and 4.31 Mb, and N = 29, 15, 100, 32, 16, 19, 48 and 35 genomes, for
796 clade-1, clade-2a, clade-2b, clade-2c, clade-2d, clade-2e1, clade-2e2 and clade-3, respectively,
797 the expected number of mutations of these clades are 0.53, 0.30, 1.82, 0.52, 0.24, 0.31, 0.75 and
798 0.63, respectively, under 30 days of laboratory cultivation of the diatoms. Since the expected
799 numbers of mutations are all below one except clade-2b but there are a few dozen SNPs in six of
800 the eight clades (Table S5), the observed SNPs cannot arise from mutations during the laboratory
801 cultivation period.

802 One may argue that some of the SNPs could result from sequencing error. While it is
803 difficult to distinguish between mutations and sequencing errors for singleton SNPs, presence of
804 non-singleton SNPs (the rare variant found in at least two genomes) is convincing evidence
805 against sequencing errors owing to the extremely low chance of having sequencing errors at the
806 same sites. We identified non-singleton SNPs in six out of the eight clades (Table S5). In
807 particular, all sequenced genomes from clade-1 and clade-2a each exhibited a unique
808 combination of the non-singleton SNPs, suggesting that none of them represent laboratory clones
809 (Table S5). For members from clade-2b, clade-2e1, clade-2e2 and clade-3, we identified 26, 14,
810 9 and 14 unique combinations of the non-singleton SNPs in 100, 19, 48 and 35 genomes,
811 respectively, suggesting that at least a sizable number of the members in these clades represent
812 the wild genotypes. In the case of the remaining two clades, clade-2c and clade-2d, no non-
813 singleton SNPs were found in the 32 and 16 members, respectively.

814 It is important to note that the SNP analysis is restricted to the core DNA shared by all
815 members of a clade, and members without displaying biologically meaningful SNPs in their core
816 genomes may differ in their accessory genomes. We therefore asked whether genomes from each
817 clade each harbor a unique combination of accessory genes (Table S5). However, this analysis is

818 compromised by the incomplete genome assembly of reads derived from Illumina sequencing.
819 This issue is mitigated, but not eradicated, when the genes missing in the complete and closed
820 genome are used. For the four clades (clade-2a, clade-2b, clade-2d and clade-2e2) with at least
821 one complete and closed genome, all of genomes each exhibit a unique combination of accessory
822 genes (Table S5), which is evidence against the laboratory clonal replication hypothesis.
823 A higher-level marker to discriminate between genomes is gene order and genome
824 rearrangement, which is best characterized using complete and closed genomes. We therefore
825 closed three genomes (SC33-45, SC33-90 and MC52-69) from the same clade (clade-2b) each
826 with additional sequencing by Nanopore (Table S4). Particularly, two of them (SC33-45 and
827 SC33-90) are barely differentiated at the accessory genome content level (Fig. 2A), further
828 highlighting the potential value of the genome rearrangement analysis. These three genomes
829 (SC33-45, SC33-90 and MC52-69) differ in chromosomal DNA length (3,836,086 bp, 3,834,406
830 bp and 3,850,659 bp, respectively), and one DNA inversion event has occurred on their
831 chromosome involving an orthologous DNA segment varying in their length (4644 bp, 4648 bp
832 and 1164 bp, respectively; Fig. S2). They possess seven homologous plasmids, of which five are
833 closed for all three genomes and thus are useful for comparison. While these five closed
834 plasmids each have conserved gene order, they differ slightly in their lengths. Here are
835 assembled lengths of these five closed plasmids in the three isolates: 2b_P1 (82,238 bp, 82,229
836 bp and 82,397 bp), 2b_P3 (290,592 bp, 290,600 bp and 290,590 bp), 2b_P4 (281,159 bp,
837 281,160 bp and 281,166 bp), 2b_P6 (210,913 bp, 210,913 bp and 210,915 bp), and 2b_P7
838 (110,566 bp, 110,566 bp and 110,565 bp).
839 Taken together, the available patterns of non-singleton SNPs in the core genomes, the strain-
840 specific gene combination in the accessory genomes, the length of chromosomal DNA and

841 plasmid DNA, and strain-specific order of genomic fragments strongly favor the hypothesis that
842 most of the analyzed roseobacters each represents a distinct genotype originally from the wild
843 environment.
844

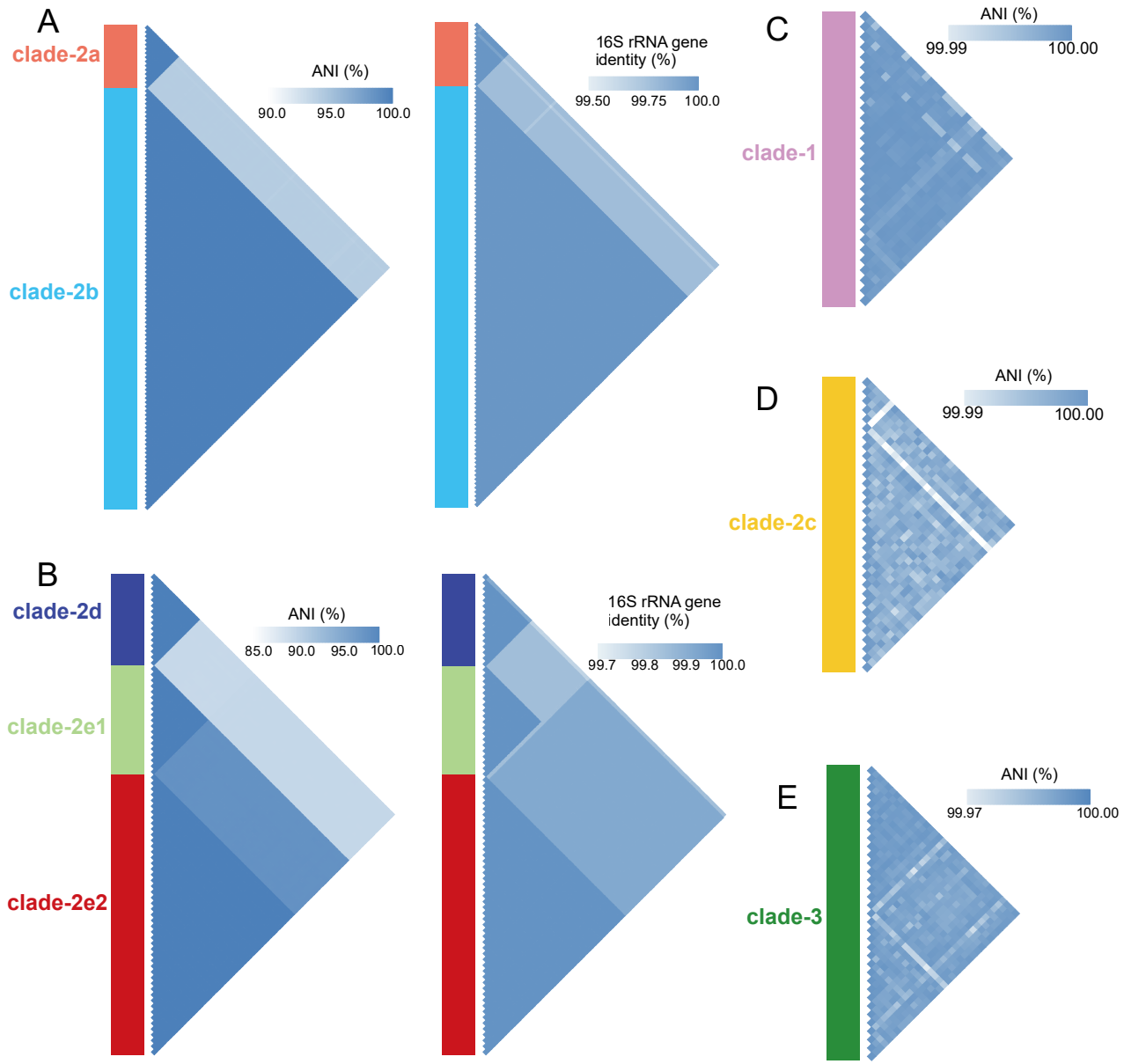

Fig. S1. The whole-genome average nucleotide identity (ANI) and/or 16S rRNA gene identity for (A) clade-2a and clade-2b, (B) clade-2d, clade-2e1, and clade-2e2, (C) clade-1, (D) clade-2c, and (E) clade-3. The 16S rRNA gene identity is not shown for clade-1, clade-2c, and clade-3 because the 16S rRNA genes are identical within each clade.

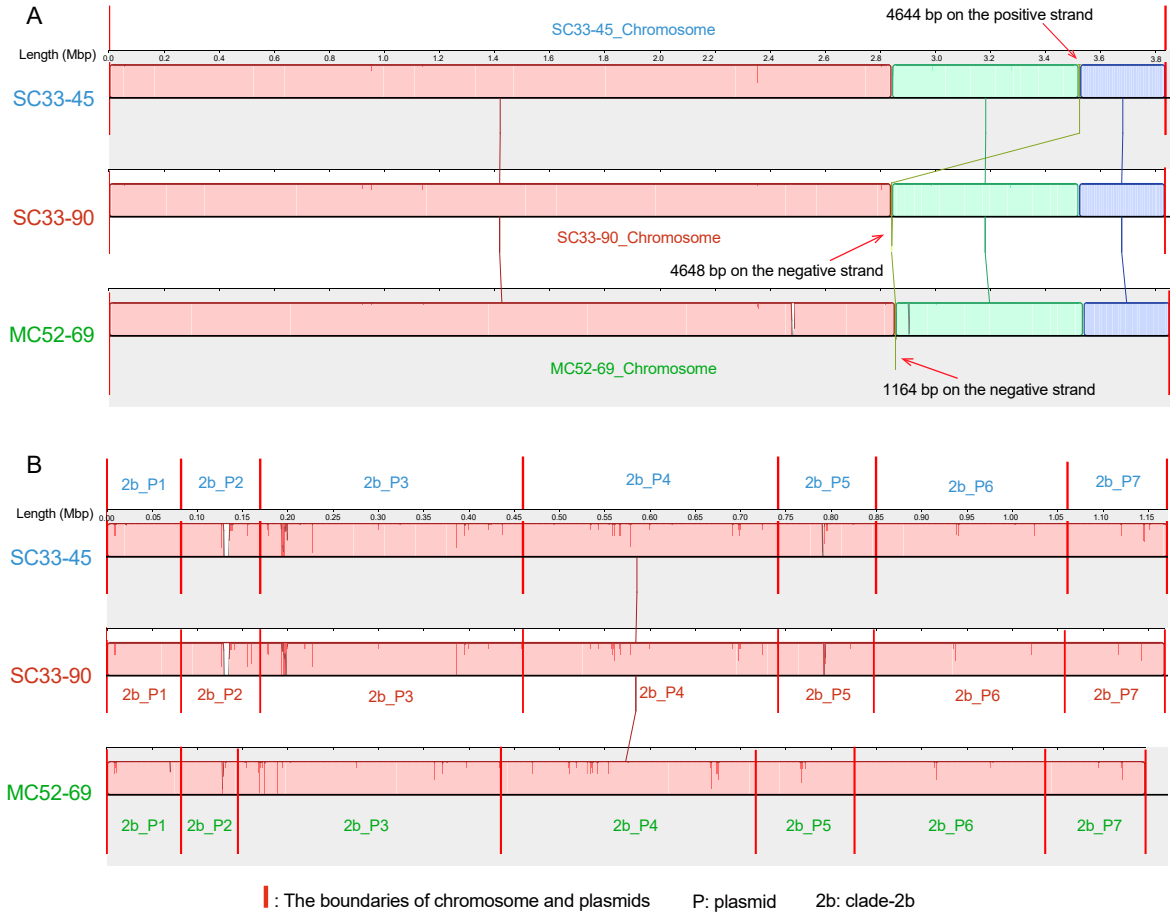

Fig. S2. The genome arrangement of (A) chromosome and (B) plasmids of the three complete and closed genomes (MC52-69, SC33-45 and SC33-90) in clade-2b. Homologous regions shared by the three genomes are represented using locally collinear blocks (LCBs) with connected lines. The minimum LCB weight is 3485, which represents the minimum number of matching nucleotides identified in all LCBs. A similarity profile is shown within each LCB, and the height of the similarity profile represents the conservation level of the alignment. LCBs above and below the centerline represent genomic regions on the forward and reverse strand, respectively. The boundaries of replicons (chromosome and plasmids) are represented by red vertical lines.

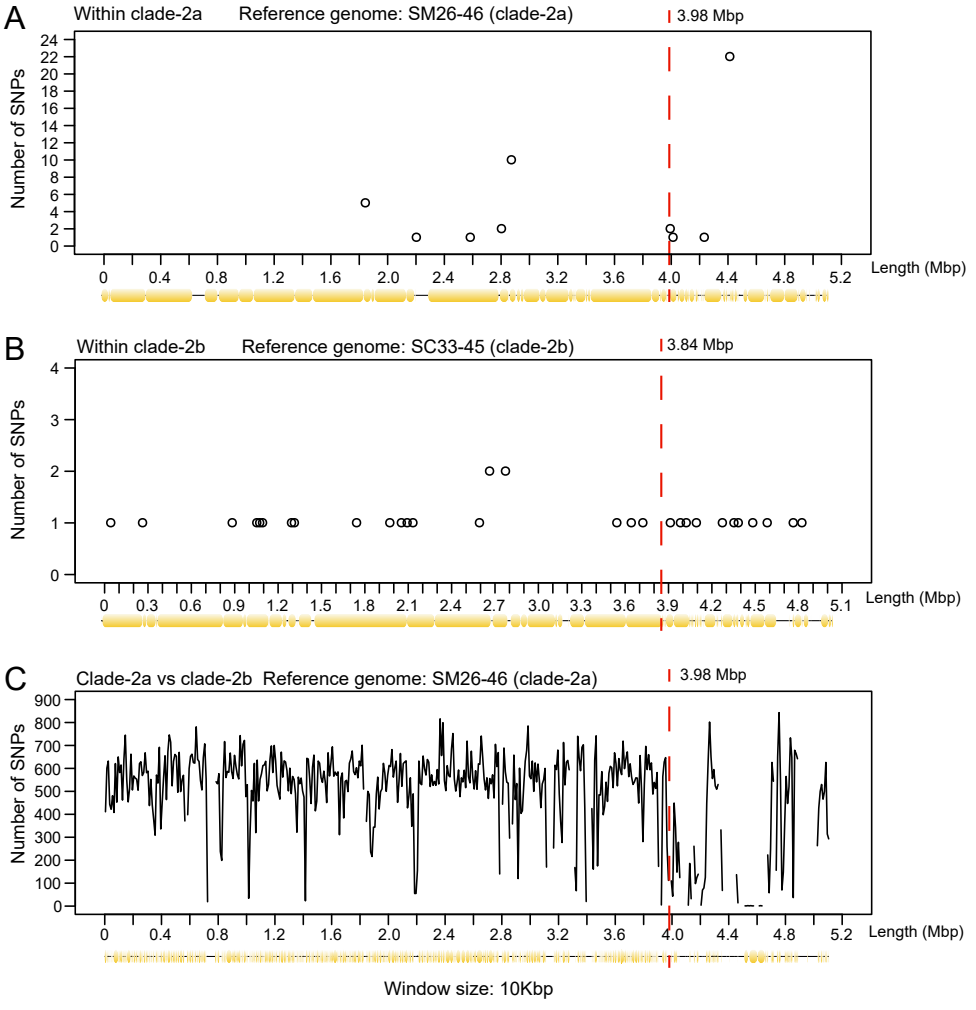

Fig. S3. The density of non-singleton SNPs (A) within clade-2a, (B) within clade-2b, and (C) between these two clades. The SNP density was calculated with a window size of 10 Kbp. The aligned region is shown at the bottom of each panel. The boundary between chromosome and plasmid is shown using a red vertical dotted line.

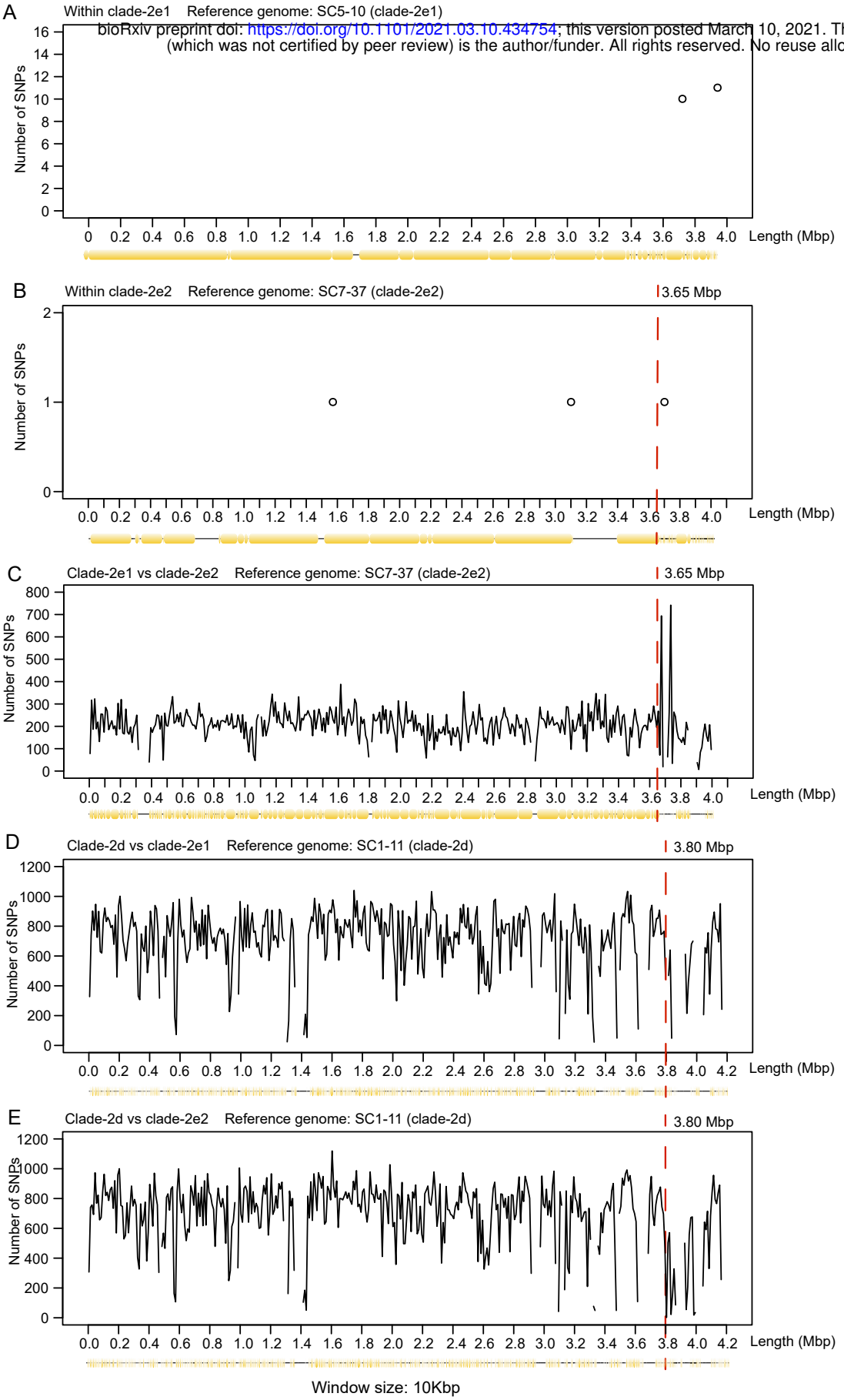

Fig. S4. The density of non-singleton SNPs (A) within clade-2e1, (B) within clade-2e2, (C) between clade-2e1 and clade-2e2, (D) between clade-2d and clade-2e1, and (E) between clade-2d and clade-2e2. The SNP density of clade-2d is not shown because no SNP was found within this clade. The SNPs density was calculated with a window size of 10 Kbp. The aligned region was shown at the bottom of each panel. The boundary between chromosome and plasmid is shown using a red vertical dotted line.

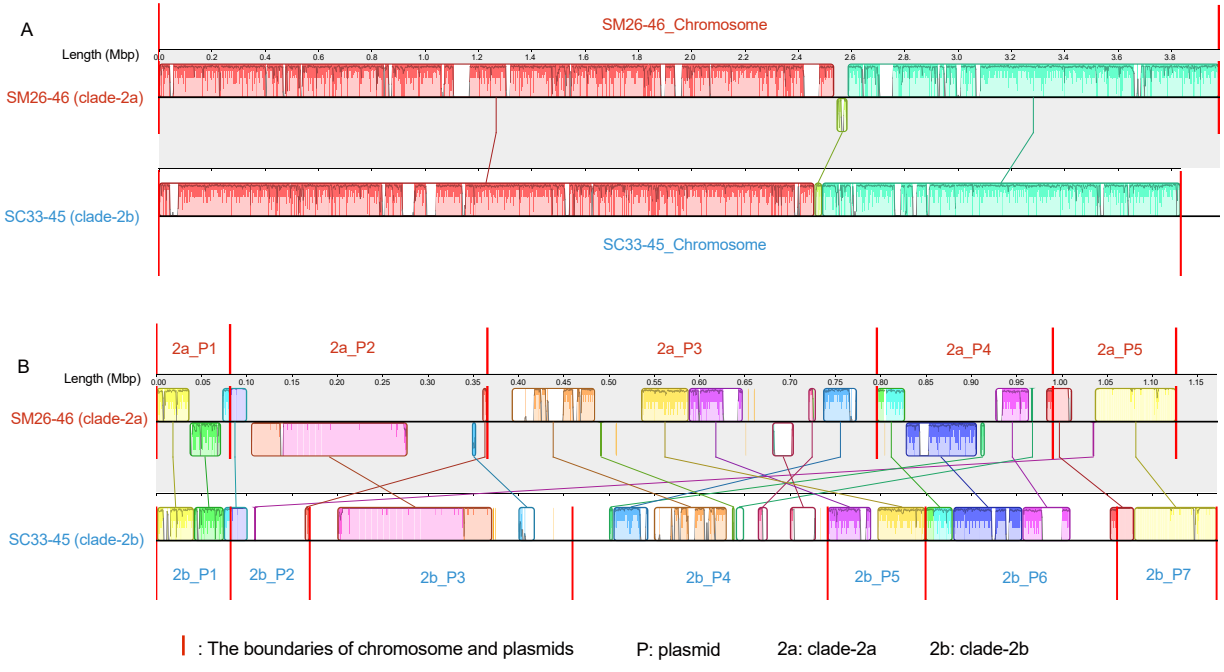

Fig. S5. The genome arrangement of (A) chromosome and (B) plasmids of the two complete genomes in clade-2a (SM26-46) and clade-2b (SC33-45). Homologous regions shared by the two genomes are represented using locally collinear blocks (LCBs) with connected lines. The minimum LCB weight is 2249, which represents the minimum number of matching nucleotides identified in all LCBs. A similarity profile is shown within each LCB, and the height of the similarity profile represents the conservation level of the alignment. LCBs above and below the centerline represent genomic regions on the forward and reverse strand, respectively. The boundaries of replicons (chromosome and plasmids) are represented by red vertical lines.

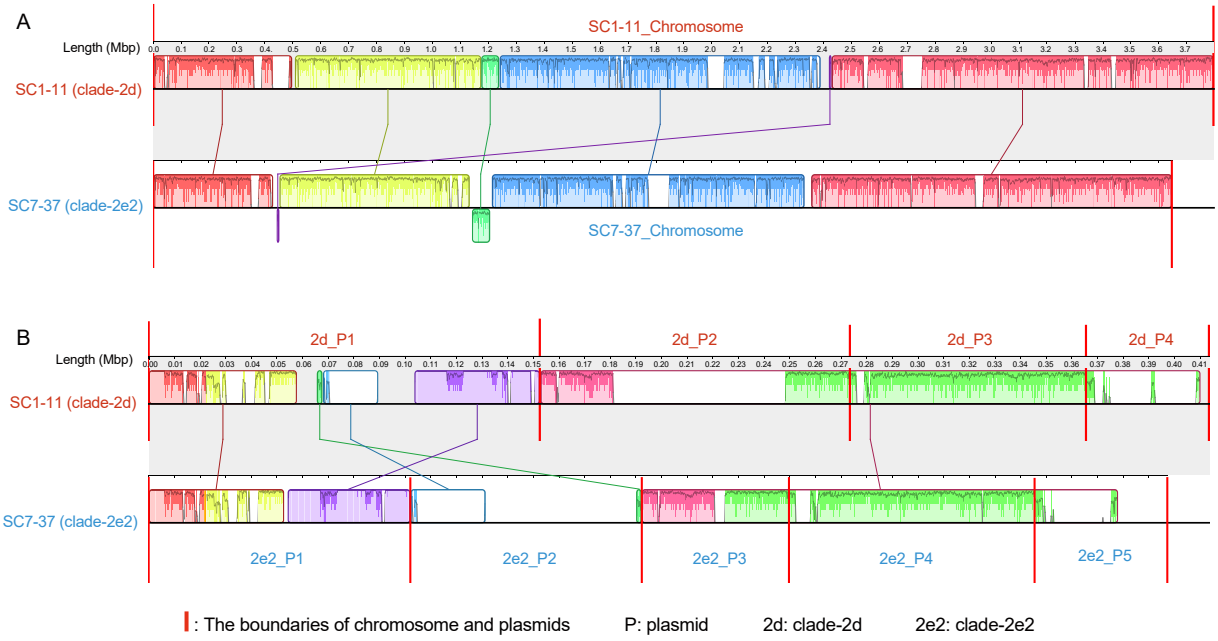

Fig. S6. The genome arrangement of (A) chromosome and (B) plasmids of the two complete genomes from clade-2d (SC1-11) and clade-2e2 (SC7-37). Homologous regions shared by the two genomes are represented using locally collinear blocks (LCBs) with connected lines. The minimum LCB weight is 2209, which represents the minimum number of matching nucleotides identified in all LCBs. A similarity profile is shown within each LCB, and the height of the similarity profile represents the conservation level of the alignment. LCBs above and below the centerline represent genomic regions on the forward and reverse strand, respectively. The boundaries of replicons (chromosome and plasmids) are represented by red vertical lines.

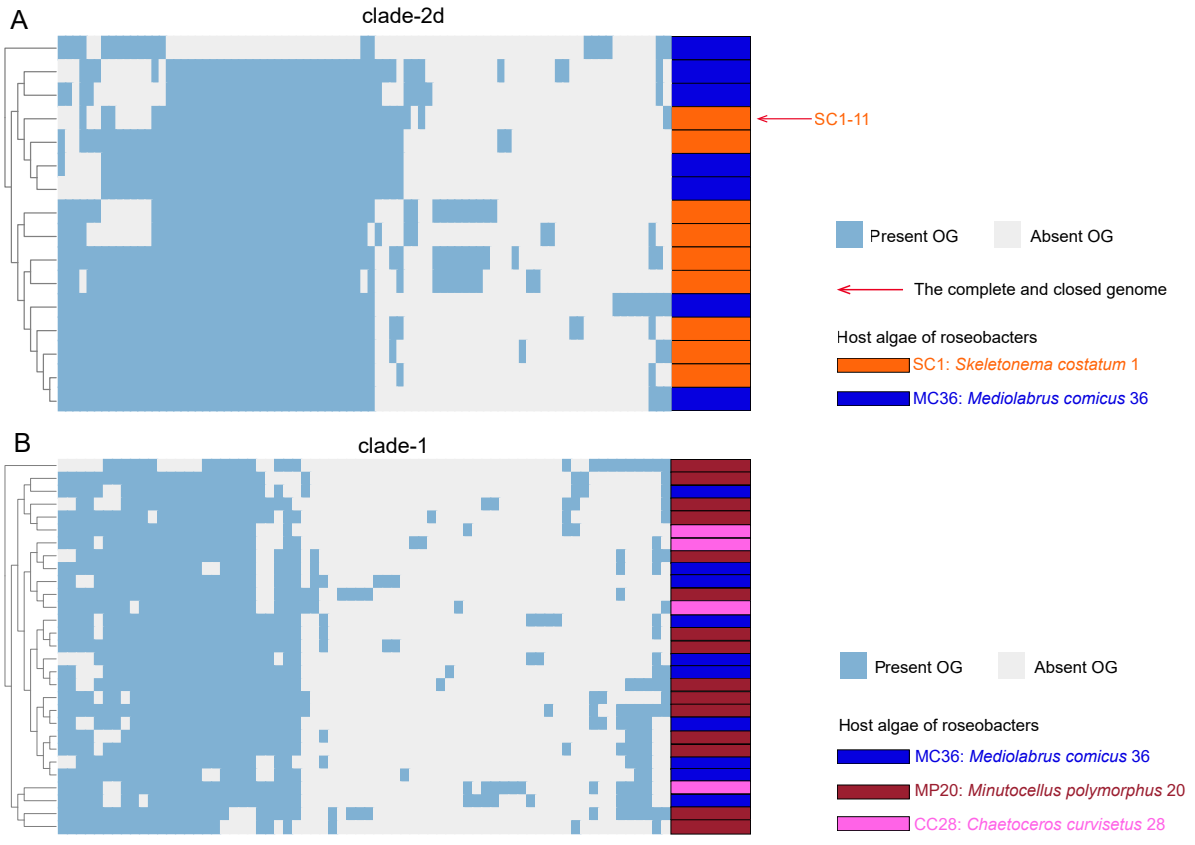

Fig. S7. The clustering of accessory genes in the genomes of clade-2d (A) and clade-1 (B). The dendrogram of genome clustering was generated based on the presence and absence of their orthologous gene families (OGs), which are colored in blue and gray, respectively. The associated microalgae of bacterial strains are differentiated with colors. The complete and closed genome is marked with a red arrow.

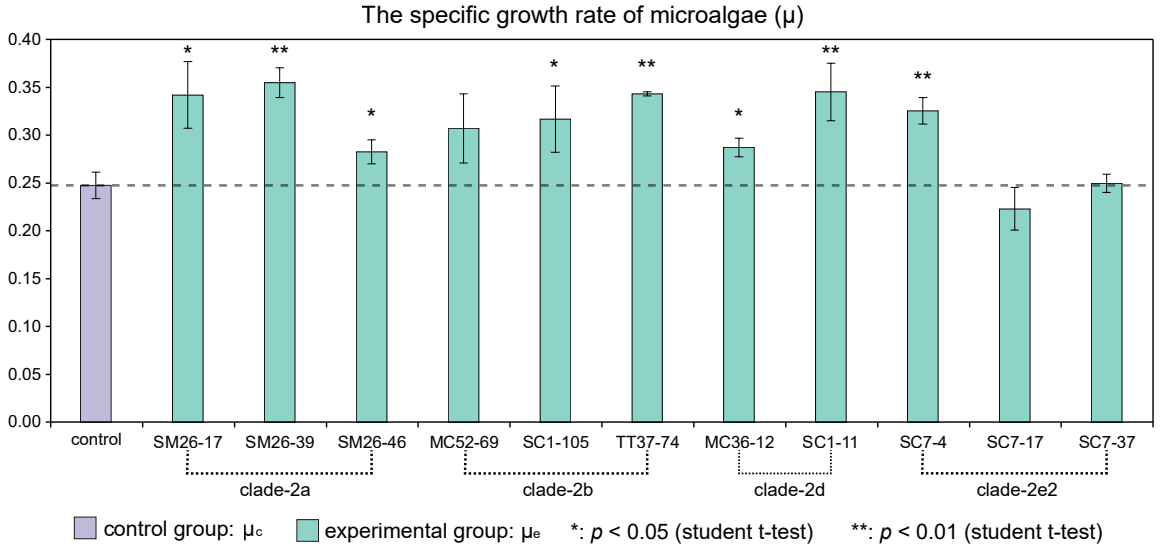

Fig. S8. The specific growth rates of microalgal strains after three days of co-culture. The microalgal specific growth rate in the control groups (μ_c) and experimental groups (μ_e) are shown in purple and green columns, respectively. The significance level $p < 0.05$ and $p < 0.01$ compared to the control group is shown using * and **, respectively.