Title:

Gene exchange networks define species-like units in marine prokaryotes

Authors:

R. Stepanauskas[1*], J.M. Brown[1], U. Mai[2], O. Bezuidt[1], M. Pachiadaki[3], J. Brown[1], S.J. Biller[4], P.M. Berube[5], N.R. Record[1], S. Mirarab[6].

Affiliations:

[1] Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine, 04544, U.S.A.

[2] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92093, U.S.A.

[3] Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, U.S.A.

[4] Department of Biological Sciences, Wellesley College, Wellesley, Massachusetts 02481, U.S.A.

[5] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, U.S.A.

[6] Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, California 92093, U.S.A.

*Correspondence to: rstepanauskas@bigelow.org

SUMMARY

Post-submission note. Since the original submission of this manuscript to bioRxiv, we discovered that some of our results may be impacted by the limitations of some of the comparative genomics tools used in this study. We are working on a revised version of the manuscript.

Although horizontal gene transfer is recognized as a major evolutionary process in Bacteria and Archaea, its general patterns remain elusive, due to difficulties tracking genes at relevant resolution and scale within complex microbiomes. To circumvent these challenges, we analyzed a randomized sample of >12,000 genomes of individual cells of Bacteria and Archaea in the tropical and subtropical ocean - a well-mixed, global environment. We found that marine microorganisms form gene exchange networks (GENs) within which transfers of both flexible and core genes are frequent, including the rRNA operon that is commonly used as a conservative taxonomic marker. The data revealed efficient gene exchange among genomes with <28% nucleotide difference, indicating that GENs are much broader lineages than the nominal microbial species, which are currently delineated at 4-6% nucleotide difference. The 42 largest GENs accounted for 90% of cells in the tropical ocean microbiome. Frequent gene exchange within GENs helps explain how marine microorganisms maintain millions of rare genes and adapt to a dynamic environment despite extreme genome streamlining of their individual cells. Our study suggests that sharing of pangenomes through horizontal gene transfer is a defining feature of fundamental evolutionary units in marine planktonic microorganisms and, potentially, other microbiomes.

MAIN TEXT

45 Horizontal gene transfer (HGT) is a key evolutionary process in *Bacteria* and *Archaea*, with
46 impacts ranging from the spread of antibiotic resistance in human pathogens to the emergence of
47 new players in planetary biogeochemical cycles [1-5]. While many HGT studies have focused on
48 remarkable yet rare gene transfers between evolutionarily distant microorganisms, routine HGT
49 among close relatives has also been reported in the ocean [6-9] and other environments [10-14]. The
50 majority of microbial genes are flexible, i.e. they are not present in all members of a given
51 lineage, which suggests frequent exchange [15,16]. However, the identification of HGT
52 phylogenetic boundaries and predominant sources, as well as the estimation of HGT rates in
53 nature have proven difficult [1-5,10,11]. Most flexible genes are rare [17,18], thus studies of their
54 distribution in specific organisms may require bias-free genome sampling at scales that remain
55 hard to achieve. Furthermore, homologous recombination, one of the key steps in HGT, is the
56 most frequent among near-identical genomes [10,19], where it is also the hardest to detect using
57 computational tools that rely on DNA sequence discrepancies [12,20]. Yet another challenge is
58 complex and poorly understood microbial biogeography, which acts from microscopic to global
59 scales and obscures the discrimination of the effects of evolutionary distance and encounter rates
60 on the distribution of genetic material in a population [21,22]. Consequently, ecosystem-wide and
61 global patterns of HGT remain largely unknown. Likewise, although many theoretical concepts
62 of microbial evolution view HGT as essential to speciation [21], such concepts have not replaced
63 arbitrary species demarcations in contemporary microbial taxonomy [23-25], due to limited
64 empirical support.

65 Marine planktonic *Bacteria* and *Archaea* (prokaryoplankton) play essential roles in the
66 global carbon cycle, nutrient remineralization and climate formation and constitute two thirds of
67 ocean's biomass [26-29]. Prokaryoplankton inhabiting epipelagic environments in tropical and

68 subtropical latitudes are efficiently dispersed around the globe by ocean currents and differ
69 genomically from prokaryoplankton inhabiting higher latitudes and greater depths [17,30]. Thus,
70 they form a global microbiome with clear geographic boundaries that has been evolving over
71 geological time scales. Here, we analyze 12,715 randomly sampled single amplified genomes
72 (SAGs), called GORG-Tropics, recovered from 28 globally distributed samples of tropical and
73 subtropical, euphotic ocean water [17]. Initial analyses found, remarkably, that no two of these cells
74 had identical gene repertoire [17], indicating that HGT may play a major role in shaping these
75 genomes. We leverage the large size and randomized genome sampling approach of GORG-
76 Tropics to analyze how HGT relates to microbial evolutionary distance in this global, well-mixed
77 microbiome.
78
79 **Flexible genes**
80 The distribution of flexible genes is expected to be randomized in a population undergoing
81 frequent HGT [15,16]. In search for genealogical boundaries of HGT, we analyzed how differences
82 in gene content relate to genome nucleotide difference (GND) – a commonly used metric of
83 evolutionary distance calculated as 1-ANI (average nucleotide identity) [10,19,31,32] – in pairs of
84 GORG-Tropics genomes. Surprisingly, we observed a non-linear relationship, with a distinct
85 breakpoint at 27.6% GND **(Fig. 1A).** Below this threshold, the fraction of non-orthologous genes
86 increased only slightly with increasing GND, while a strong, positive relationship emerged at
87 evolutionary distances above this GND value**.** The saturation of codon wobble positions could
88 not explain this discontinuity, because a similar, non-linear pattern emerged in the relationship
89 between genome-wide amino acid difference (AAD) and the fraction of non-orthologous genes,
90 with an inflection point of 39.3% AAD (**Fig. 1B**). To confirm that these patterns were not an

91 artifact of our computational tools, we searched for and found no discontinuity in the GND-AAD
92 relationship at these values (**Fig. S1).** Next, we generated a set of mock genomes with
93 randomized gene content, which failed to produce discontinuities in the relationship between
94 GND and fraction of non-orthologous genes (**Fig. S2A**). Furthermore, changing the amino acid
95 identity threshold in the CompareM [33] ortholog identification algorithm from 30% (default) to
96 50% had only a minor impact on the breakpoint in GORG-Tropics (**Fig. S2A**). Finally, we
97 replaced the CompareM *de novo* ortholog search with identification of shared hidden Markov
98 model-based Prokka [34] annotations, which produced a similar pattern with a near-identical
99 breakpoint (**Fig. S2B**). This indicates that discontinuities at ~28% GND and ~39% AAD are
100 intrinsic features of prokaryoplankton in the tropical and subtropical ocean. Below these
101 thresholds, GND has little influence on the fraction of non-orthologous genes, which averaged at
102 29% and was similar to the average of 28% of strain-specific genes found in well-characterized,
103 nominal bacterial species [35]. Similar discontinuities, although at slightly higher GND values,
104 were observed in pairwise differences in genome size and G+C content (**Fig. 1E**). We propose
105 that the most plausible explanation for the observed discontinuity is HGT frequency being
106 sufficient to maintain shared pools of flexible genes (pangenomes) among bacterioplankton cells
107 below ~28% GND. Above this approximate threshold, insufficient frequency of HGT fixation
108 leads to the separation of pangenomes and accumulation of gene content differences in
109 individual cells.

110 The distribution of GORG-Tropics SAG pairs along the GND and AAD gradients exhibited
111 multiple sharp peaks averaging ~ 19.5%, 22.5%, 25%, 28% and 30% GND (**Fig. 1 C-D**). These
112 peaks did not coincide with breakpoints in relationships between GND, AAD and gene content
113 (**Fig. 1 A-B**), indicating that mechanisms causing these peaks had no major impact on our

114 estimated HGT boundaries. The low frequency of genome pairs with near-0% GND suggests that
115 GND distribution peaks were not caused by multiple lineages forming blooms of near-clonal
116 populations during the GORG-Tropics sample collection. Based on the finding of the same peaks
117 in geographically and temporally distant samples (**Fig. 1F**) and their mixed taxonomic
118 composition (**Fig. 1A-B**), we speculate that these peaks reflect concurrent population expansions
119 of multiple lineages during ancient, ecosystem-wide perturbations, and deserve further
120 investigation. As reported previously [17], fewer than 0.1% of GORG-Tropics SAG pairs fell
121 below the 4% GND threshold for the contemporary, nominal definition of microbial species [24].
122 This contrasts with several recent reports of elevated abundance of microorganisms with near-
123 0% GND, which were interpreted as empirical support of the contemporary species delineations
124 [36,37]. Although differences between marine and non-marine microbiomes may play a role, this
125 discrepancy among studies also underlines the potential importance of genome selection
126 strategies: a) randomized sample of individual cells [17], b) public databases dominated by specific
127 taxa [36], and c) metagenome bins defined by assembly and binning algorithms [37].
128

129 **Core genes**

130 Frequent HGT of core genes (genes that are present in all members of a lineage) is expected to
131 randomize the distribution of their alleles in a population. We utilized GORG-Tropics to study
132 how GND relates to the divergence of two highly conserved, universal genes that are broadly
133 utilized in phylogenetic analyses: 16S rRNA [38] and *rpsC* [39]. In various bacterioplankton lineages,
134 divergences of both genes correlated with GND only weakly at low GND values, indicating
135 frequent HGT **(Fig. 2 A-B)**. We found completely identical 16S rRNA nucleotides and predicted
136 RpsC proteins in cell pairs with up to 14% and 19% GND. In contrast, more evolutionarily

137 distant SAG pairs produced strong correlations of GND versus 16S rRNA and RpsC
138 divergences, with inflection points of these non-linear relationships approximating 23% and 22%
139 GND. Uncorrelated variable rates of evolution among lineages could not explain this pattern,
140 because an opposite trend (loss of correlation with increasing GND) was observed in a
141 simulation with randomly varied evolution rates **(Fig. S3)**. This suggests that HGT frequency of
142 16S rRNA and *rpsC* genes is sufficient to counteract the accumulation of nucleotide substitutions
143 among cells that fall below these approximate divergence thresholds. Further support for this
144 conclusion is provided by the high levels of discordance between 16S and 23S rRNA gene
145 phylogenies at low GND and low levels of discordance at high GND, with the breakpoint at
146 ~22% GND (**Fig. 2C-D**). This implies frequent recombination of rRNA operon segments among
147 cells with <22% GND.

148 Interestingly, a discernible subset of phylogenetically diverse genome pairs formed
149 correlation patterns that ran parallel to the bulk of data points, with unusually low 16S rRNA and
150 RpsC divergences relative to their GND (marked as "outliers" in **Fig. 2 A-B).** We hypothesize
151 that such outliers form as a result of divergent GND thresholds for the frequent HGT of highly
152 conserved genes (**Fig. 2**) versus genome average (**Fig. 1**), leading to differences in the
153 evolutionary breadth of populations that share pangenomes of various genes.

154 The rRNA operon is generally considered one of the most stable and slowest-evolving
155 genome elements and therefore has served as the foundation for microbial genealogy inferences
156 since the 1970s [40]. This assumption was challenged by several studies that detected HGT of
157 rRNA genes in closely related strains and produced functional ribosomes from mosaic rRNA *in*
158 *vitro* [12]. Our results help reconcile these conflicting findings by identifying the specific GND
159 range in which the fixation of horizontally transferred rRNA genes may be frequent in marine

160 prokaryoplankton. The results of GND comparisons with 16S rRNA and *rpsC* divergences
161 indicate that even core genes are involved in frequent HGT among close relatives. This
162 conclusion corroborates with previous reports of recombination of core genes is SAR11, the
163 most abundant lineage of marine prokaryoplankton [6]. We hypothesize that the consistent
164 presence of core genes in closely related microorganisms is not caused by the exclusion of these
165 genes from HGT, as often assumed [33], but rather by a dynamic process that involves HGT and
166 selection against cells without compatible variants of these genes.
167

168 **Recent HGT**

169 We counted protein-coding genes with 0%, ≤1%, and ≤5% nucleotide difference in genome
170 pairs along the GND gradient as a proxy for putative, recent HGT events **(Fig. 3A)**. Observations
171 of such genes declined exponentially from 0% to 27% GND range. The slope of the exponential
172 decline was similar among the various COG categories, although the breakpoint varied, with
173 categories P (inorganic ion transport and metabolism) and F (nucleotide metabolism and
174 transport) dominating putative HGT counts at high GND **(Fig. 3B).** Accordingly, prior studies
175 have reported exponential declines of recombination rates among closely related bacterial [10] and
176 archaeal [19] isolates along the GND gradient. Our findings of near-identical genes in genome pairs
177 may be a result of both HGT and variable rates of nucleotide substitution among genes, as the
178 two factors are difficult to disentangle. Nevertheless, it is interesting that the GORG-Tropics
179 dataset, which contains a wide spectrum of microbial diversity, exhibited a non-linear decline in
180 near-identical gene counts at GND > 28%. This discontinuity was observed at the same GND
181 value when using either 0%, 1% or 5% gene divergence thresholds, indicating a robust,
182 biological feature rather than a methodological artifact**.** The accelerated decline in counts of

183 near-identical genes at 28% GND coincided with the discontinuity in gene content similarity at
184 the same GND value (**Fig. 1**), in support of infrequent HGT among cells with >28% GND.
185

186 **Gene exchange networks**

187 Using the 28% GND threshold, we clustered GORG-Tropics SAGs into groups that we term
188 gene exchange networks (GENs) according to single-linkage clustering. These GENs vary in the
189 extent of internal linkage, ranging from 11% to 100% (mean 78%). Of the 12,715 GORG-
190 Tropics SAGs, 9,176 SAGs formed 168 GENs of ≥2 members (**Figs. 4, S4**). Of the 3,539 SAGs
191 remaining as singletons, 94% had assembly sizes <0.5 Mbp (**Fig. S5**), suggesting that their
192 exclusion from GENs is mostly due to limited genome recovery. The 42 most populous GENs
193 accounted for >90% of the 7,208 SAGs with ≥0.5 Mbp assemblies. Given the good agreement in
194 taxonomic and genomic compositions of GORG-Tropics SAGs and other marine
195 prokaryoplankton 'omics datasets [17], these 42 most populous GENs constitute a realistic
196 representation of the majority of prokaryoplankton in the tropical and subtropical ocean.
197 Most GENs contained only one of the previously described lineages of marine
198 prokaryoplankton (**Figs. 4, S4**), demonstrating good agreement between GENs and the earlier,
199 primarily 16S rRNA-based phylogenetic demarcations, with a few notable exceptions. The
200 largest GEN, comprised of 3,625 SAGs, encompassed all SAGs of SAR11 (Alphaproteobacteria)
201 ecotypes surface-1, surface-2, surface-3, deep-1 and "unclassified", indicating that these lineages
202 share a single pangenome, although with some internal sub-clustering (**Fig. S4**). This observation
203 helps explain an earlier finding of rhodopsin, a light-harvesting system, being encoded in
204 ecotype deep-1, which is found primarily below the euphotic depths [41]. Meanwhile, SAR11
205 surface-4 formed a separate GEN, which agrees with the recent findings of divergent

206 biosynthetic capabilities [17] and rRNA operon structure [42] of this lineage as compared to other
207 SAR11. Our network analysis also indicated an incomplete separation of Flavobacteriaceae
208 lineages NS2b, NS4 and NS5. By contrast, we found that many of the previously described
209 lineages, which span genus to phylum taxonomic units, form multiple GENs. Notably,
210 *Prochlorococcus* (Cyanobacteria) formed two large and several small GENs, indicating a
211 separation of gene pools between the high-light adapted clades (e.g. HLI, HLII, HLIII, and
212 HLIV) and the low-light adapted LLI clade. These data suggest that within the world's most
213 abundant lineage of phototrophs, there exist multiple GENs, each with likely separate
214 evolutionary trajectories. The most abundant Gammaproteobacteria lineage SAR86 formed four
215 GENs with >90 members, three of which had low linkage and substructures, indicating a high
216 degree of genomic and evolutionary heterogeneity. Multiple GENs were also formed by lineages
217 SAR116, PS1 (Alphaproteobacteria), OM60 (Gammaproteobacteria), SAR324
218 (Deltaproteobacteria), MGII (Crenarchaeota) and others. In general, the topology of GENs
219 ranged from high linkage (e.g. the two largest GENs of *Prochlorococcus*) to a complex internal
220 structure (e.g. the largest GENs of Flavobacteriaceae and SAR86), which is consistent with the
221 gradual separation of sister GENs along the GND gradient (**Figs. 1, 2**). Thus, GENs offer a
222 biologically meaningful and practical paradigm for studies of environmental microorganisms.
223

224 **Discussion**

225 Our findings suggest that bacterioplankton in the tropical euphotic ocean form gene exchange
226 networks that are characterized by shared pangenomes, with 28% GND as their approximate
227 boundary (**Fig. 5**). Extensive HGT within GENs helps explain how the ~40 million gene types,
228 most of which rare [17,18], are maintained by bacterioplankton lineages that are characterized by

229 extreme genome streamlining [43], with an average genome size of 1.6 Mbp [17]. Access to a large
230 pool of readily exchanged genes within GENs likely facilitates adaptations of marine
231 bacterioplankton to a dynamic environment despite the limited coding potential of their
232 individual cells.

233 The observed GEN boundaries are soft, indicating that the transition from HGT-driven GEN
234 cohesion to pangenome separation between sister GENs takes place over considerable
235 evolutionary time (**Fig. 1**) and is not uniform across gene types (**Figs. 2, 4**). Therefore, a fixed
236 threshold in GEN delineation would not be an accurate reflection of natural evolutionary
237 processes. Nevertheless, the soft breakpoint that averages ~28% GND demonstrates that marine
238 bacterioplankton GENs exceed the evolutionary breadth of current, nominal definitions of
239 bacterial species, which range between 4-6% GND and are designed to maintain the continuity
240 of historical taxonomies [23-25]. Assuming an average amino acid substitution rate of ~0.05% per
241 Myr [44], the age of bacterioplankton GENs likely ranges in hundreds of Myr. This time scale is
242 consistent with the estimated age of *Prochlorococcus*, which forms two of the largest
243 bacterioplankton GENs, and human commensal lineages for which reliable calibration of the
244 molecular clock is available [44,45]. The longevity of GENs helps explain the efficient, global
245 exchange of genes within marine GENs, because the entire ocean water volume is overturned by
246 the global thermohaline circulation every 1-2 kyr, while surface ocean currents operate at
247 substantially shorter timeframes [46].

248 Although GORG-Tropics did not capture putative HGT events among cells with >32% GND,
249 this does not mean that such events never occur, as exemplified by prior reports of cross-domain
250 HGT [47]. Some gene flow among evolutionarily distant lineages may also take place through
251 intermediary lineages. However, the rate of fixation of genes acquired across GEN boundaries

252 may be insufficient to obscure the predominant evolutionary patterns, akin to the rare
253 introgression events in eukarya [48]. Furthermore, some ancient HGT between now evolutionarily
254 distant lineages may have occurred while both the donor and the recipient were still part of the
255 same GEN. Likewise, multiple studies have demonstrated distinct patterns of HGT within
256 substantially more narrow lineages of marine microorganisms, variably named "ecotypes",
257 "genotypes", "backbones" and "populations"[8,9,49]. Our data suggest that that such lineages may
258 not maintain isolated pangenomes, therefore their divergence from sister lineages may be
259 temporary and reversible. We speculate that HGT manifests at multiple levels, with impacts of
260 evolutionary proximity, encounter rates and HGT mechanisms producing nested, fractal-like
261 patterns of gene flow in complex microbiomes **(Fig. 5C)**.

262 Efficient, global mixing of the euphotic, tropical ocean simplifies studies of microbiome-
263 wide HGT in this environment. Dispersal and cell encounter limitations likely have a more
264 profound impact on HGT patterns in environments such as soils and organismal microbiomes.
265 However, the lifespan of GENs ranging in hundreds of millions of years may enable sufficient
266 HGT to maintain global GENs even in such compartmentalized microbiome types. Accordingly,
267 discontinuities in genome content along nucleotide divergence gradients are apparent in several
268 earlier reports analyzing diverse microbial genome collections, although evolutionary origins of
269 these discontinuities were not discussed [31,50]. Our findings suggest that sharing of pangenomes
270 through HGT defines fundamental evolutionary units in marine prokaryoplankton and
271 demonstrates a novel approach to explore general HGT patterns in other microbiomes. Viewed
272 as cohorts of cells with dynamic, shared gene pools, GENs offer a new conceptual framework to
273 guide studies of microbial evolution, ecology and biotechnological exploration.

274

SUPPLEMENTAL INFORMATION

Materials and Methods

Figures S1-S6

Tables S1-S2

ACKNOWLEDGMENTS

We thank Elizabeth Fergusson, Brian Thompson, Corianna Mascena and Ben Tupper at the Bigelow Laboratory for Ocean Sciences' Single Cell Genomics Center for the generation of single cell genomic data. We also thank Daniel Buckley (Cornell University) and Thane Papke (University of Connecticut) for valuable advice. **Funding:** This work was funded by the Simons Foundation (Life Sciences Project Award ID 510023, RS) and the National Science Foundation (OCE-1335810 and OIA-1826734 to RS; III-1845967 to SM and UM).

AUTHOR CONTRIBUTIONS

R.S. developed the concept, managed the project and led manuscript preparation. J.M.B. performed most data analyses related to key genome characteristics and clustering. U.M. and S.M. performed phylogenetic discordance analyses and variable evolutionary rate modeling. O.B. performed putative HGT identification. N.R.R. performed change point regression analyses. M.P. and J.B. led data management and curation. P.M.B. and S.J.B. oversaw field sample collection and selection. All authors contributed to data interpretation and manuscript preparation.

297 **REFERENCES**

298 1 Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes:
299 quantification and classification. *Annual Reviews in Microbiology* **55**, 709-742 (2001).
300 2 Goldenfeld, N. & Woese, C. Biology's next revolution. *Nature* **445**, 369 (2007).
301 3 Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of
302 life. *Nat Rev Genet* **16**, 472-482, doi:10.1038/nrg3962 (2015).
303 4 Lawrence, J. G. & Retchless, A. C. The myth of bacterial species and speciation. *Biology*
304 *and Philosophy* **25**, 569-588 (2010).
305 5 Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of
306 evolutionary ecology. *Nat Rev Microbiol* **12**, 263-273, doi:10.1038/nrmicro3218 (2014).
307 6 Vergin, K. L. *et al.* High intraspecific recombination rate in a native population of
308 Candidatus Pelagibacter ubique (SAR11). *Environmental Microbiology* **9**, 2430-2440
309 (2007).
310 7 Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation
311 of bacteria. *Science* **335**, 48-51 (2012).
312 8 Kashtan, N. *et al.* Single-cell genomics reveals hundreds of coexisting subpopulations in
313 wild Prochlorococcus. *Science* **344**, 416-420, doi:10.1126/science.1248575 (2014).
314 9 Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J. & Polz, M. F. A Reverse Ecology
315 Approach Based on a Biological Definition of Microbial Populations. *Cell* **178**, 820-834
316 e814, doi:10.1016/j.cell.2019.06.033 (2019).
317 10 Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial
318 speciation. *Science* **315**, 476-480 (2007).

319 11 Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the
320 human microbiome. *Nature* **480**, 241-244, doi:10.1038/nature10571 (2011).
321 12 Gogarten, J., Doolittle, W. & Lawrence, J. Prokaryotic evolution in light of gene transfer.
322 *Molecular biology and evolution* **19**, 2226-2238 (2002).
323 13 Whitaker, R. J., Grogan, D. W. & Taylor, J. W. Recombination shapes the natural
324 population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol Biol*
325 *Evol* **22**, 2354-2361 (2005).
326 14 Doroghazi, J. R. & Buckley, D. H. Widespread homologous recombination within and
327 between Streptomyces species. *ISME Journal* **4**, 1136-1143, doi:10.1038/ismej.2010.45
328 (2010).
329 15 McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes.
330 *Nat Microbiol* **2**, 17040, doi:10.1038/nmicrobiol.2017.40 (2017).
331 16 Vos, M. & Eyre-Walker, A. Are pangenomes adaptive or not? *Nat Microbiol* **2**, 1576,
332 doi:10.1038/s41564-017-0067-5 (2017).
333 17 Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through
334 Single-Cell Genomics. *Cell* **179**, 1623-1635.e1611, doi:10.1016/j.cell.2019.11.017
335 (2019).
336 18 Sunagawa, S., Karsenti, E., Bowler, C. & Bork, P. Computational eco-systems biology in
337 Tara Oceans: Translating data into knowledge. *Molecular Systems Biology* **11**,
338 doi:10.15252/msb.20156272 (2015).
339 19 Williams, D., Gogarten, J. P. & Papke, R. T. Quantifying homologous replacement of
340 loci between haloarchaeal species. *Genome Biology and Evolution* **4**, 1223-1244,
341 doi:10.1093/gbe/evs098 (2012).

342 20 Beiko, R. G., Harlow, T. J. & Ragan, M. A. Highways of gene sharing in prokaryotes.
343 *Proceedings Of The National Academy Of Sciences Of The United States Of America*
344 **102**, 14332-14337, doi:10.1073/pnas.0504068102 (2005).
345 21 Palmer, M., Venter, S. N., Coetzee, M. P. A. & Steenkamp, E. T. Prokaryotic species are
346 sui generis evolutionary units. *Systematic and Applied Microbiology* **42**, 145-158,
347 doi:10.1016/j.syapm.2018.10.002 (2019).
348 22 Andam, C. P., Choudoir, M. J., Vinh Nguyen, A., Sol Park, H. & Buckley, D. H.
349 Contributions of ancestral inter-species recombination to the genetic diversity of extant
350 Streptomyces lineages. *ISME J* **10**, 1731-1741, doi:10.1038/ismej.2015.230 (2016).
351 23 Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species
352 definition for prokaryotes. *Proceedings Of The National Academy Of Sciences Of The*
353 *United States Of America* **102**, 2567-2572 (2005).
354 24 Ciufo, S. *et al.* Using average nucleotide identity to improve taxonomic assignments in
355 prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol* **68**, 2386-2392,
356 doi:10.1099/ijsem.0.002809 (2018).
357 25 Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny
358 substantially revises the tree of life. *Nat Biotechnol* **36**, 996-1004, doi:10.1038/nbt.4229
359 (2018).
360 26 Falkowski, P. G., Fenchel, T. & DeLong, E. F. The Microbial Engines that Drive Earth's
361 Biogeochemical Cycles. *Science* **320**, 1034-1039 (2008).
362 27 Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic Diversity in
363 Sargasso Sea Bacterioplankton. *Nature* **345**, 60-63 (1990).

364 28 Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest
365 Atlantic through Eastern Tropical Pacific. *PLoS Biology* **5**, e77 (2007).
366 29 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**,
367 doi:10.1126/science.1261359 (2015).
368 30 Swan, B. K. *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic
369 bacteria in the surface ocean. *Proceedings of the National Academy of Sciences of the*
370 *United States of America* **110**, 11463-11468, doi:10.1073/pnas.1304246110 (2013).
371 31 Konstantinidis, K. T. & Tiedje, J. M. Towards a genome-based taxonomy for
372 prokaryotes. *Journal Of Bacteriology* **187**, 6258-6264 (2005).
373 32 Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the
374 genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**,
375 1929-1940 (2006).
376 33 Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes
377 substantially expands the tree of life. *Nature Microbiology* **2**, 1533-1542,
378 doi:10.1038/s41564-017-0012-7 (2017).
379 34 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-
380 2069, doi:10.1093/bioinformatics/btu153 (2014).
381 35 Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses.
382 *Current Opinion in Microbiology* **23**, 148-154, doi:10.1016/j.mib.2014.11.016 (2015).
383 36 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
384 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
385 *Nat Commun* **9**, doi:10.1038/s41467-018-07641-9 (2018).

386 37 Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate
387 Bacterial Species Boundaries. *mSystems* **5**, doi:10.1128/mSystems.00731-19 (2020).
388 38 Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The
389 primary kingdoms. *Proceedings Of The National Academy Of Sciences Of The United*
390 *States Of America* **74**, 5088-5090 (1977).
391 39 Wu, D., Jospin, G. & Eisen, J. A. Systematic Identification of Gene Families for Use as
392 "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and
393 Archaea and Their Major Subgroups. *PLoS ONE* **8** (2013).
394 40 Woese, C. R. *et al.* Conservation of primary structure in 16S ribosomal RNA. *Nature*
395 **254**, 83-85 (1975).
396 41 Cameron, T. J. *et al.* Single-cell enabled comparative genomics of a deep ocean SAR11
397 bathytype. *ISME Journal* **8**, 1440-1451, doi:10.1038/ismej.2013.243 (2014).
398 42 Haro-Moreno, J. M. *et al.* Ecogenomics of the SAR11 clade. *Environ Microbiol* **22**,
399 1748-1763, doi:10.1111/1462-2920.14896 (2020).
400 43 Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining
401 theory for microbial ecology. *ISME J* **8**, 1553-1565, doi:10.1038/ismej.2014.60 (2014).
402 44 Ochman, H. & Wilson, A. C. Evolution in bacteria: Evidence for a universal substitution
403 rate in cellular genomes. *Journal of Molecular Evolution* **26**, 74-86 (1987).
404 45 Lebreton, F. *et al.* Tracing the Enterococci from Paleozoic Origins to the Hospital. *Cell*
405 **169**, 849-861 e813, doi:10.1016/j.cell.2017.04.027 (2017).
406 46 Doos, K., Nilsson, J., Nycander, J., Brodeau, L. & Ballarotta, M. The World Ocean
407 Thermohaline Circulation. *Journal of Physical Oceanography* **42**, 1445-1460 (2012).

408 47 Nelson-Sathi, S. *et al.* Acquisition of 1,000 eubacterial genes physiologically transformed
409 a methanogen at the origin of Haloarchaea. *Proceedings of the National Academy of*
410 *Sciences of the United States of America* **109**, 20537-20542 (2012).
411 48 Green, R. E. *et al.* A draft sequence of the neandertal genome. *Science* **328**, 710-722,
412 doi:10.1126/science.1188021 (2010).
413 49 Thompson, J. R. *et al.* Genotypic diversity within a natural coastal bacterioplankton
414 population. *Science* **307**, 1311-1313 (2005).
415 50 Barco, R. A. *et al.* A Genus Definition for Bacteria and Archaea Based on a Standard
416 Genome Relatedness Index. *mBio* **11**, e02475-02419, doi:10.1128/mBio.02475-19
417 (2020).
418 51 Fong, Y., Huang, Y., Gilbert, P. B. & Permar, S. R. chngpt: Threshold regression model
419 estimation and inference. *BMC Bioinformatics* **18**, doi:10.1186/s12859-017-1863-x
420 (2017).

421

422

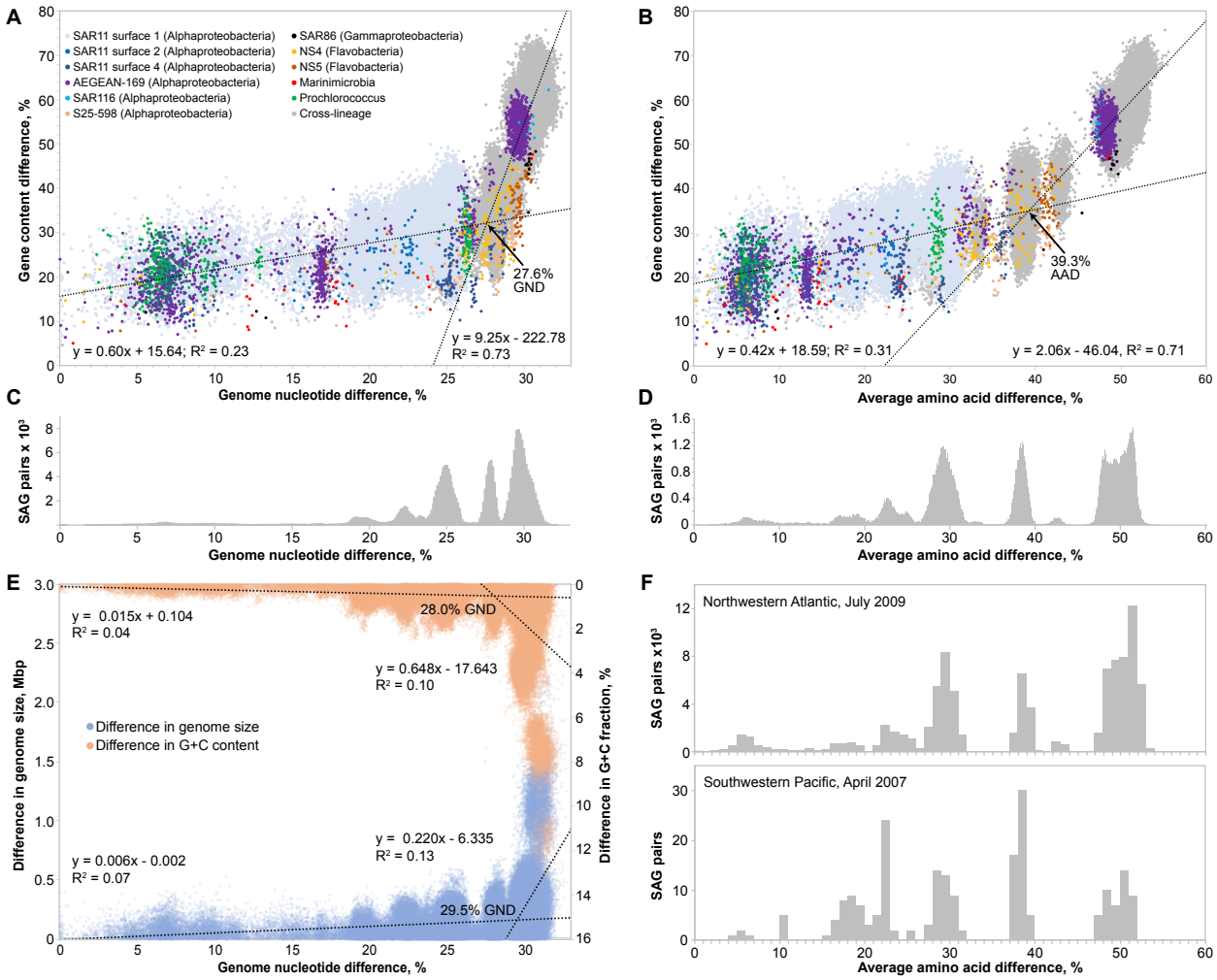

Fig. 1. Patterns of genome divergence of marine bacterioplankton**.** Fraction of non-orthologous genes in SAG pairs in relation to their genome nucleotide difference (GND) (A) and amino acid difference (AAD) (B). Frequency distribution of GND (C) and AAD (D), calculated at 0.1% GND and AAD intervals. (E) Differences in estimated genome size and G+C content in SAG pairs in relation to GND. (F) Frequency distribution of AAD in SAGs from two locations and sample collection times, calculated at 1% AAD intervals. To minimize analytical uncertainties, we sub-sampled GORG-Tropics to 819 bacterial genomes with ≥80% estimated genome completion that contain 16S rRNA genes. In panels A and B, colors of data points designate taxonomic lineages

represented by ≥30 genome pairs. Orthologs were identified using a 30% amino acid identity threshold with CompareM [33]. In panel F, Northwestern Atlantic is represented by field samples SWC-09 and SWC-10, while Southwestern Pacific is represented by field samples 117-Fiji-Hawaii-2007, 121-Fiji-Hawaii-2007, 133-Fiji-Hawaii-2007 and 137-Fiji-Hawaii-2007 [17]. Regression models and points of inflection were estimated by change point regression and segmented models [51].

423

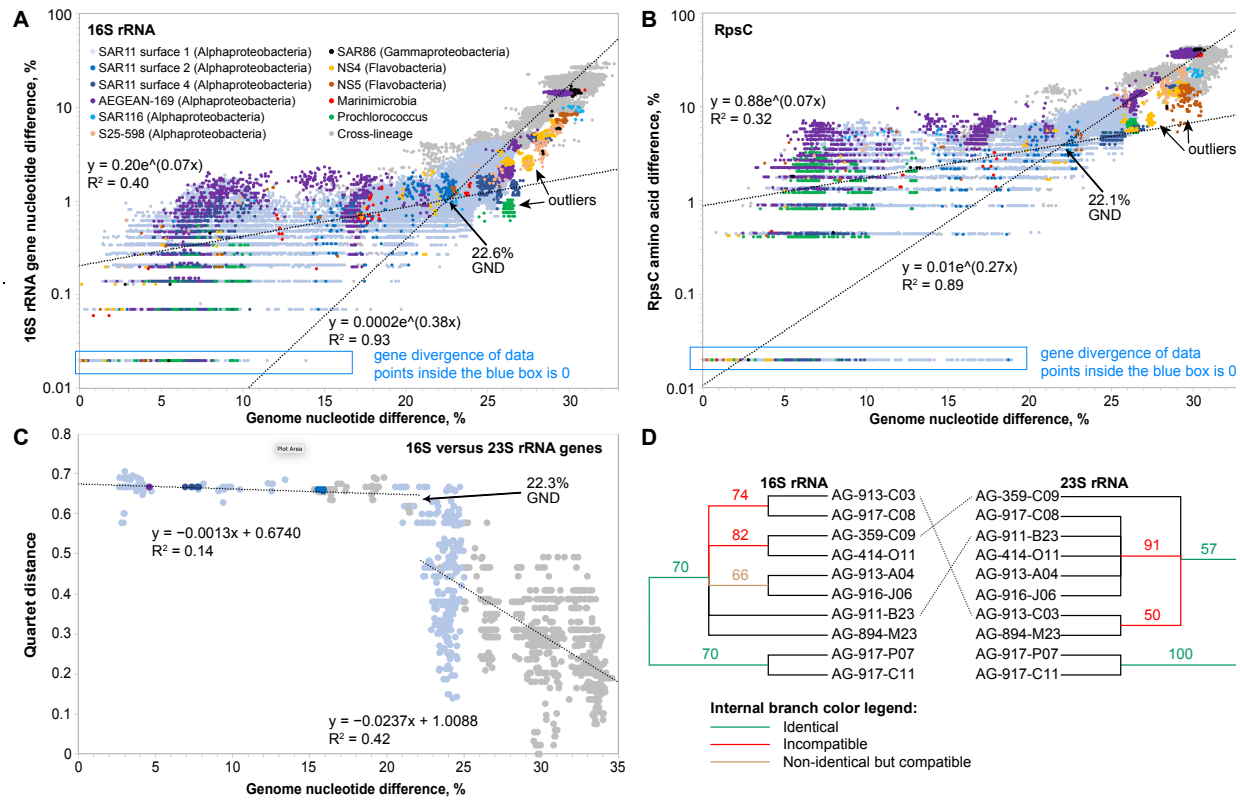

Fig. 2. Divergence of selected core genes in relation to genome divergence. Relationships between genome nucleotide difference (GND) and sequence divergences of 16S rRNA (A) and RpsC (B) genes. (C) Phylogenetic discordance of 16S and 23S rRNA gene trees along GND gradient. Each dot represents a subset of 10 SAGs, characterized by its mean pairwise GND. Tree discordance is measured using normalized quartet distance after collapsing branches with <70% bootstrap support. (D) Example of discordance of 16S and 23S rRNA gene phylogenies among SAR11 surface-1 lineage SAGs with mean 2.9% GND. Numbers by nodes indicate bootstrap support. Branches with <50% bootstrap support were removed. All genome labels are aligned except those indicated by dotted lines. Regression models and points of inflection were estimated by change point regression [51] using segmented (A and B) and stegmented (C) models. Included are 819 GORG-Tropics SAGs with ≥80% estimated

genome completion and the recovery of the 16S rRNA gene. In panels A and B, colors of data points designate taxonomic lineages represented by ≥30 genome pairs.

424

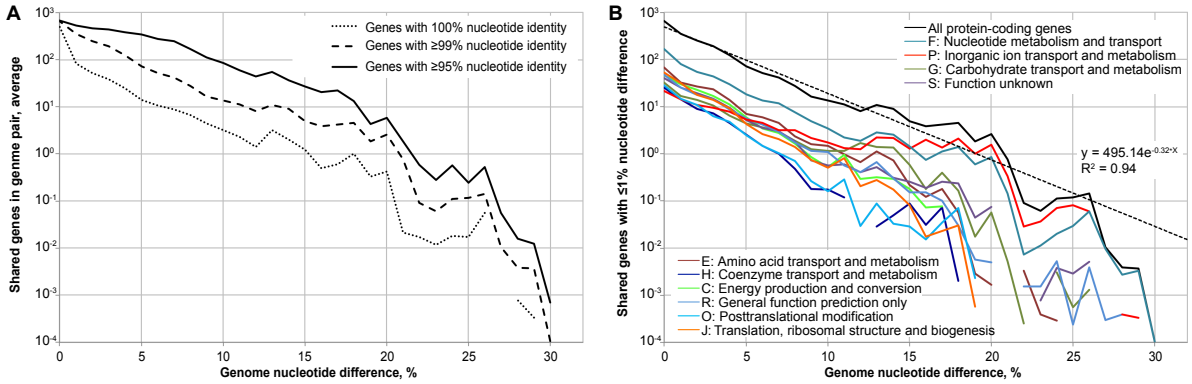

Fig. 3. Average counts of genes with high nucleotide identity in pairs of GORG-Tropics SAGs as a function of GND. Included are 819 GORG-Tropics SAGs with ≥80% estimated genome completion and the recovery of the 16S rRNA genes. Counted are only gene alignments that cover ≥80% of query sequence length. **(A)** Average counts of protein-coding genes with 95-100% nucleotide identities in genome pairs. **(B)** Average counts of protein-coding genes with ≥99% nucleotide identity in genome pairs. Colored datasets correspond to 10 most abundant COGs. Regression is based on the 0-27% GND range. Data discontinuities correspond to values equal to zero.

425

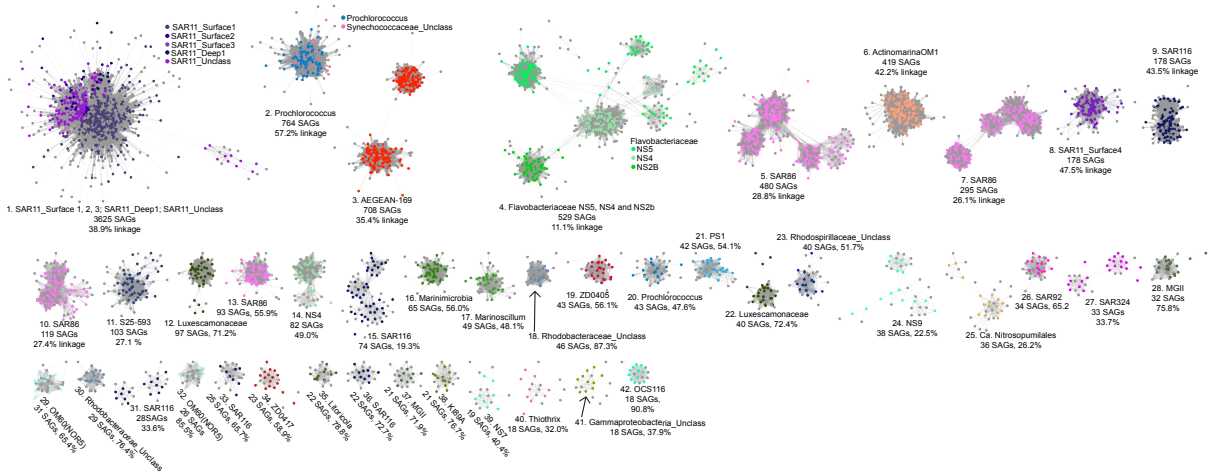

Fig. 4. The 42 GENs that account for 90% of GORG-Tropics SAGs. Nodes represent SAGs, edges represent GND ≤ 28%, node colors and titles represent literature-derived phylogenetic-ecological lineages. Text indicates GEN ID, 16S rRNA-based lineage, count of member SAGs and network linkage (ratio of actual versus possible link counts). Additional information on a complete set of GENs can be found in Fig. S4 and Data S1 and S2.

426

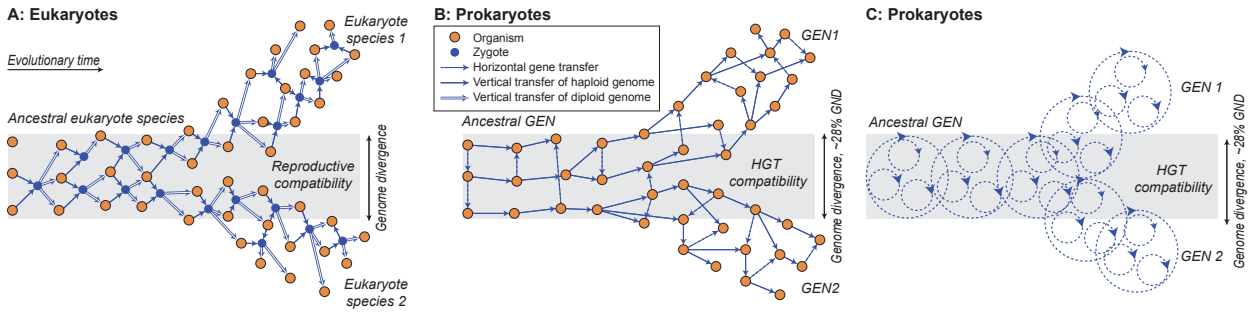

Fig. 5. Conceptualized gene flow and divergence in sexually reproducing eukaryotes (A) and prokaryoplankton GENs (B, C). In contrast to the formation of a zygote in eukaryotes, prokaryote HGT is not coupled to reproduction and is unidirectional. Furthermore, since HGT involves both homologous and non-homologous recombination, individual members of GENs contain only a fraction of GEN's pangenome. A and B depict gene flow among cells. C depicts a broader, nested/fractal pattern of gene flow within GENs.

427

MATERIALS AND METHODS

Data source

The main dataset utilized in this study is Global Oceans Reference Genomes Tropics (GORG-Tropics), which consists of 12,715 partial genomes of individual cells of bacteria and archaea obtained through a random selection from 28 globally distributed samples of tropical and subtropical, photic ocean water column [1].

Genome-wide comparisons

Average nucleotide identity (ANI) of genome pairs was calculated using the ANIb method within pyani [2]. Although computationally more demanding than alternative approaches, this technique offers superior sensitivity in a broader range of GND values [3]. Due to high computational demands of this software, we limited this analysis to those 819 GORG-Tropics SAG assemblies that contained 16S rRNA gene sequences and were estimated to have ≥80% completion. ANI comparisons with >3% overlap in both directions were kept for analysis, self-comparisons were removed. Genome nucleotide difference (GND) was calculated as GND = 1-ANI. Detection of orthologous genes and amino acid identity were determined with CompareM, which uses reciprocal best hits to identify orthologous pairs [4]. Search for nearly identical genes was done using reciprocal BLASTn [5] between the open reading frames of genome pairs. Mock genomes were generated by randomly sampling 10kb fragments from 16S containing SAGs that were ≥ 80% complete, to construct a set of 1,000 genomes that followed genome length distributions similar to the real data. Inflection points and model fits were estimated using change point regression using the segmented model, except for the regression analyses of

24 phylogenetic quartet distance, where the "stegmented" model allowing for a jump between the
25 two lines was used [6].
26
27 **Phylogenetic analyses**
28 Multiple alignment of the 16S and 23S genes was performed using SINA [7] and followed by
29 alignment trimming with trimAl version 1.4rev15 [8]. Maximum likelihood trees were inferred
30 using RAxML version 8.2.10 [9] with default settings and 1,000 bootstrap replicates. To compute
31 discordance of 16S and 23S rRNA genes, we first created 3,000 genome subsets, each containing
32 10 SAGs. We used the single-linkage agglomerative clustering algorithm implemented in
33 TreeN93 [10] to build a hierarchy of the genomes. At any threshold of GND, all genomes within
34 that threshold belong to a sub tree of the single-linkage tree. We then cut the single tree at
35 different thresholds of GND and chose 10 individual genomes at random from different subtrees.
36 This allowed us to control the pairwise GND of the 10 genomes within a certain range. By
37 varying the threshold between 0.5% and 29.5% (steps: 0.5), we obtained 60 groups of subtrees
38 and randomly sampled 50 subsets of size 10 from each group. This procedure produced 3,000
39 subsets that covered the range between 2.5% and 34.6% mean pairwise GND (average: 18.9%).
40 Once the subsets of genomes were selected, we restricted the original 16S and 23S trees to
41 include only the 10 selected SAGs. We computed the bootstrap support for the restricted trees by
42 restricting the bootstrap replicate trees to the 10 selected SAGS and computing bipartition
43 support. We then collapsed all branches with BS<75%. We discarded all pairs of subtrees that
44 had more than 2/3 quartet trees (140 out of 210 possible) as unresolved. This left us with 1,422
45 pairs of subtrees. For each pair, we computed the normalized weighted quartet distance. A
46 quartet match score was computed where each quartet was counted as 1 if resolved identically in

both trees, as 0 if resolved differently in the two trees, and 1/3 if unresolved in one or both trees (because there is a 1/3 chance for trees to resolve the quartet identically at random). The weighted quartet distance was computed by subtracting the total match score from the total quartets and then normalizing by the total number of quartets. To compute quartet matches, we used tqDist[11].

To test if a simple model of rate variation can describe patterns of GND discordance that we observed, we used a simulation. We first inferred a concatenation tree by utilizing the GToTree phylogenomic workflow [12] using an HMM set of 74 single copy gene targets. A Maximum likelihood tree was generated using FastTree version 2.1.10 [13] with the default parameters. We then rooted the tree using MV rooting [14] and made the tree ultrametric (i.e., clock-like) using wLogDate [15], fixing the tree height to 1. We then multiplied each branch length by a rate multiplier drawn randomly from the Gamma distribution with shape and scale set to 1 (thus, mean: 1; variance: 1). This model is used in many existing simulation methods for capturing rate variation across genes [16,17]. The rate multipliers represent changes in evolutionary rate and are modelled here as being independent. We repeated this process 5 times, obtaining 5 potential "gene trees" that differ from the main SAG tree only in terms of rates of evolutionary change. We then compared all pairwise distances in all pairs of these simulated gene trees and used these values to draw **Fig. S3**.

Network analysis

To construct GND-based network clusters for GORG-Tropics, we first used the results from the pyani ANIb comparison [2] computed for GORG-Tropics SAG assemblies that had ≥80% completion estimates and contained 16S rRNA gene sequences. This provided us with GND

70 estimates ranging ~0-32%. In order to complement these results with additional SAGs, we took
71 advantage of the computational efficiency of fastANI [3], which we ran on all GORG-Tropics to
72 estimate GND up to ~ 25% between SAG pairs. ANI estimates for the two calculators was linear
73 for values below 20% GND (R^2 = 0.992, fastani = ani_b * 0.977 + 2.4). Above 20% GND,
74 fastani values leveled out at GND = 25%, while ANIb values calculated by pyani extended above
75 25% GND. The pyani and FastANI results were combined and networks were identified as
76 single-linkage network graphs. Sub-modules were identified using the "community" network
77 clustering algorithm within the Louvain Python package [18] and visualized with cytoscape [19].
78

79 **References**

80 1 Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through
81 Single-Cell Genomics. *Cell* **179**, 1623-1635.e1611, doi:10.1016/j.cell.2019.11.017
82 (2019).
83 2 Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G. & Toth, I. K. Genomics and
84 taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens.
85 *Analytical Methods* **8**, 12-24, doi:10.1039/c5ay02550h (2016).
86 3 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
87 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
88 *Nat Commun* **9**, doi:10.1038/s41467-018-07641-9 (2018).
89 4 Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes
90 substantially expands the tree of life. *Nature Microbiology* **2**, 1533-1542,
91 doi:10.1038/s41564-017-0012-7 (2017).

92 5 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
93 database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).
94 6 Fong, Y., Huang, Y., Gilbert, P. B. & Permar, S. R. chngpt: Threshold regression model
95 estimation and inference. *BMC Bioinformatics* **18**, doi:10.1186/s12859-017-1863-x
96 (2017).
97 7 Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple
98 sequence alignment of ribosomal RNA genes. *Bioinformatics (Oxford, England)* **28**,
99 1823-1829, doi:10.1093/bioinformatics/bts252 (2012).
100 8 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated
101 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973,
102 doi:10.1093/bioinformatics/btp348 (2009).
103 9 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
104 large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033
105 (2014).
106 10 Moshiri, N. TreeN93: a non-parametric distance-based method for inferring viral
107 transmission clusters. *bioRxiv*, 383190, doi:10.1101/383190 (2018).
108 11 Sand, A. *et al.* TqDist: A library for computing the quartet and triplet distances between
109 binary or general trees. *Bioinformatics* **30**, 2079-2080, doi:10.1093/bioinformatics/btu157
110 (2014).
111 12 Lee, M. D. & Ponty, Y. GToTree: A user-friendly workflow for phylogenomics.
112 *Bioinformatics* **35**, 4162-4164, doi:10.1093/bioinformatics/btz188 (2019).

113 13 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-
114 likelihood trees for large alignments. *PLoS ONE* **5**, doi:10.1371/journal.pone.0009490
115 (2010).
116 14 Mai, U., Sayyari, E. & Mirarab, S. Minimum variance rooting of phylogenetic trees and
117 implications for species tree reconstruction. *PLoS ONE* **12**,
118 doi:10.1371/journal.pone.0182238 (2017).
119 15 Mai, U. & Mirarab, S. Log Transformation Improves Dating of Phylogenies. *bioRxiv*,
120 2019.2012.2020.885582, doi:10.1101/2019.12.20.885582 (2020).
121 16 Rasmussen, M. D. & Kellis, M. Unified modeling of gene duplication, loss, and
122 coalescence using a locus tree. *Genome Research* **22**, 755-765,
123 doi:10.1101/gr.123901.111 (2012).
124 17 Mallo, D., De Oliveira Martins, L. & Posada, D. SimPhy: Phylogenomic Simulation of
125 Gene, Locus, and Species Trees. *Systematic Biology* **65**, 334-344,
126 doi:10.1093/sysbio/syv082 (2016).
127 18 Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of
128 communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*
129 **2008**, doi:10.1088/1742-5468/2008/10/P10008 (2008).
130 19 Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of
131 biomolecular interaction networks. *Genome Research* **13**, 2498-2504,
132 doi:10.1101/gr.1239303 (2003).
133

1 **SUPPLEMENTAL INFORMATION**

2

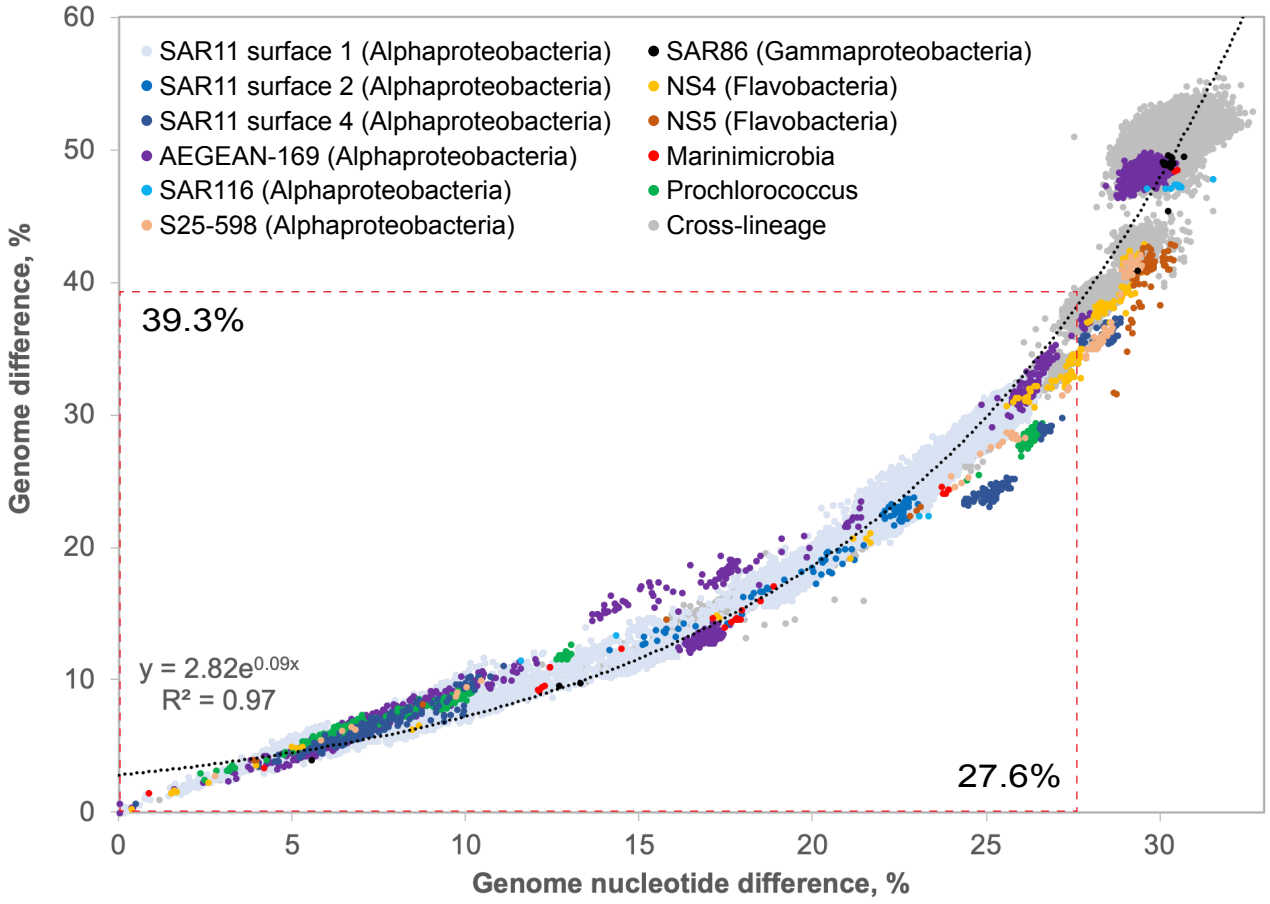

Fig. S1. Relationship between GND and AAD. Red-dotted lines indicate inflection points observed in Fig. 1.

3

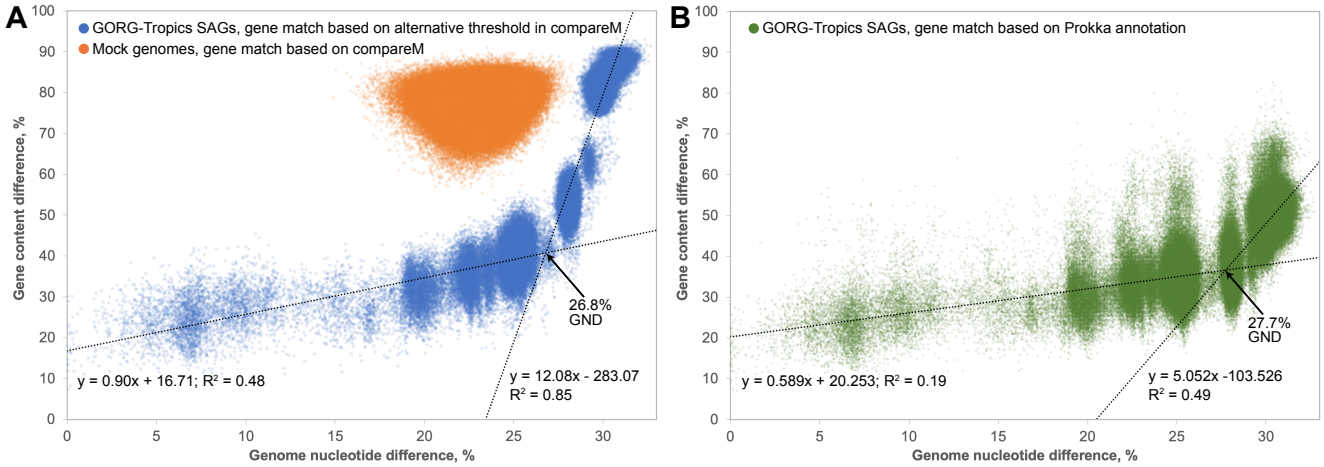

Fig. S2. Relationship between GND and the difference in protein-coding gene content. (A) CompareM-based gene content differences: mock genomes analyzed using default settings (orange) and GORG-Tropics SAGs analyzed using a 50% amino acid identity threshold in ortholog detection. (B) GORG-Tropics SAGs analyzed with Prokka.

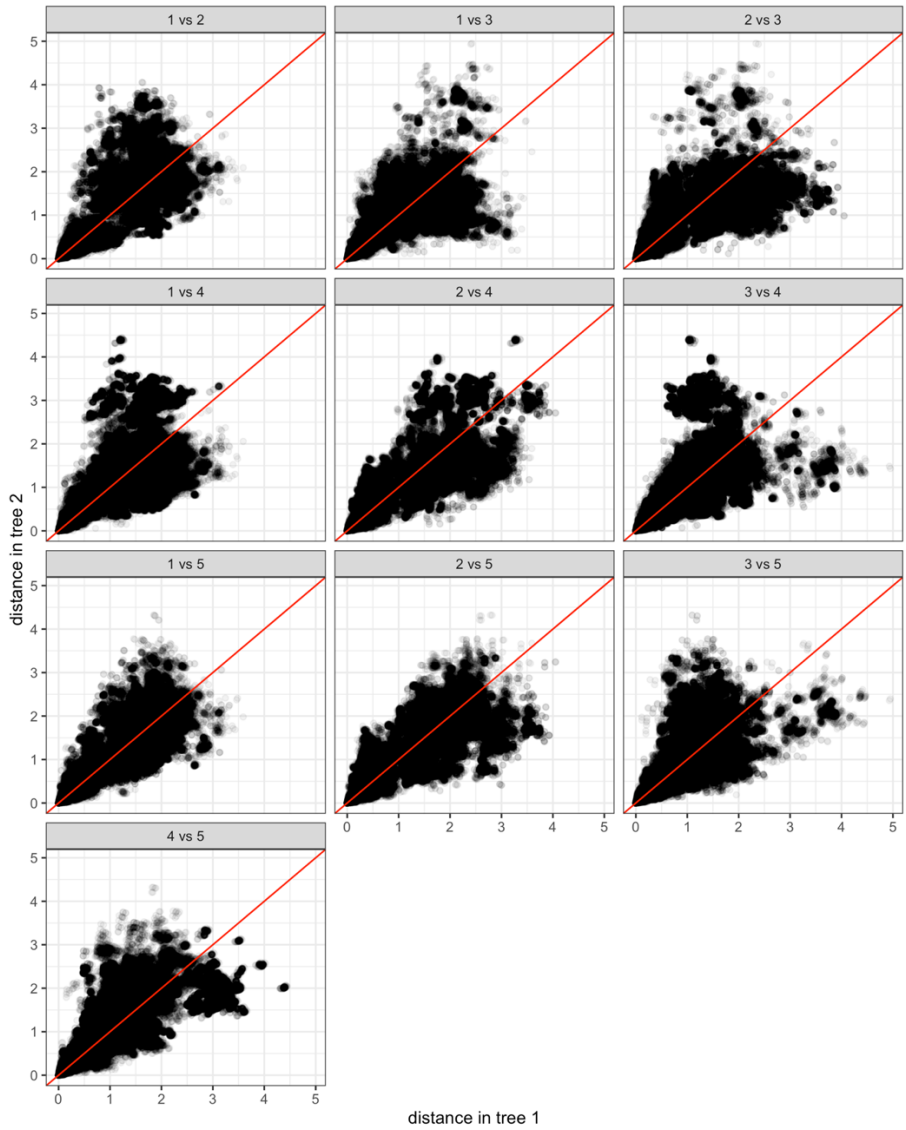

Fig. S3. Simulation of varied evolution rates. We show the correlation between branch lengths between two simulated gene trees, showing the total tree distance for all pairs of leaves, each represented as one dot. The gene trees are simulated by starting from a common tree, keeping the topology fixed, but simply multiplying all branch lengths by a rate multiplier drawn randomly from the Gamma distribution (shape and scale both set to 1). This model is used in many existing methods for capturing rate variation across genes and is found to be a good match to empirical data [1,2].

4

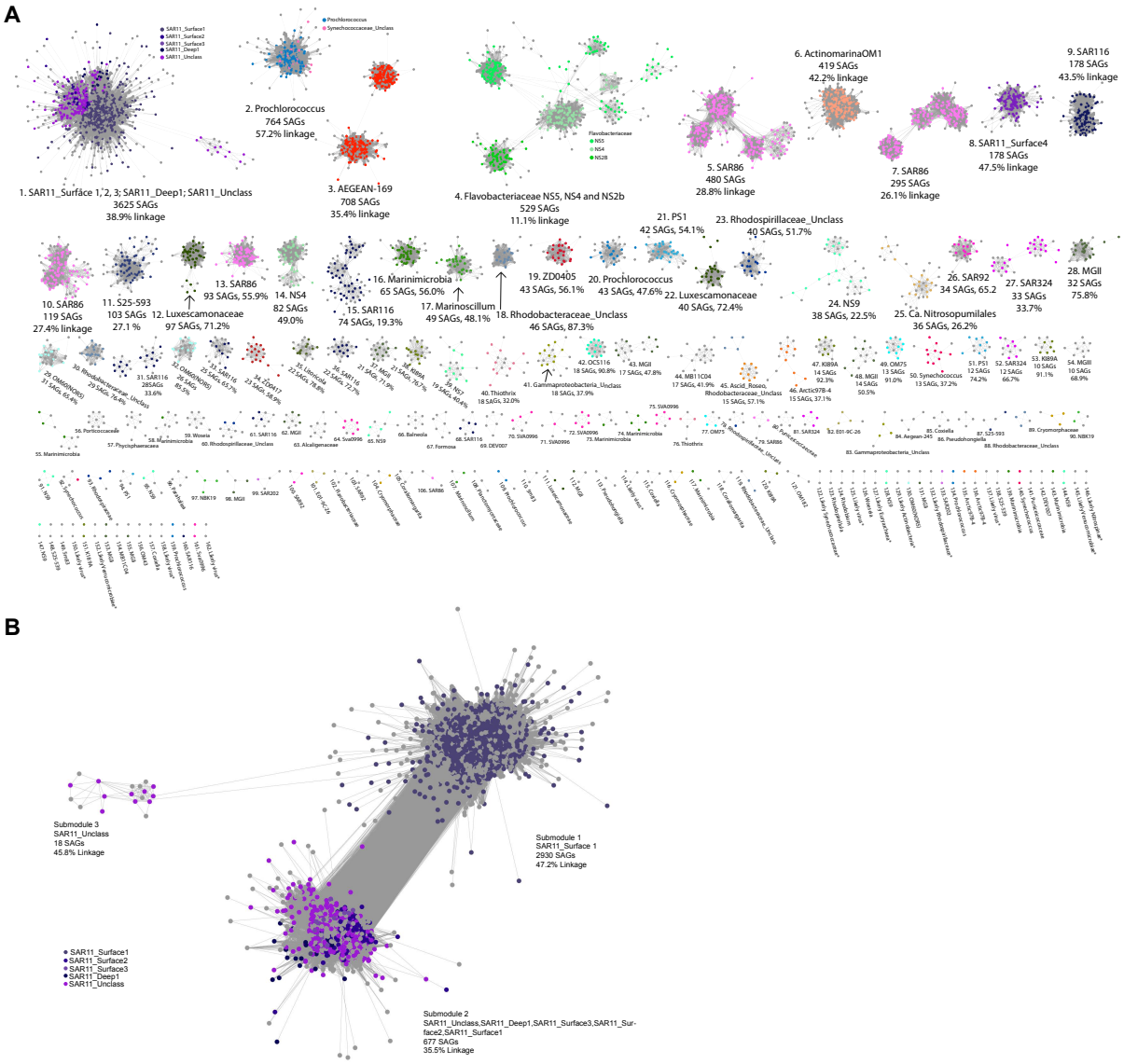

Fig. S4. Gene exchange networks (GENs). Nodes represent SAGs; edges represent GND ≤ 28%; node colors and titles represent 16S rRNA phylogenetic lineages. Text indicates GEN ID, 16S rRNA-based lineage, count of member SAGs and network linkage (ratio of actual versus possible link counts). Pane A contains all resolved GENs, pane B displays GEN #1 in greater detail. Note "*" indicates tentative identities of GENs that lack SAGs with 16S rRNA genes recovered.

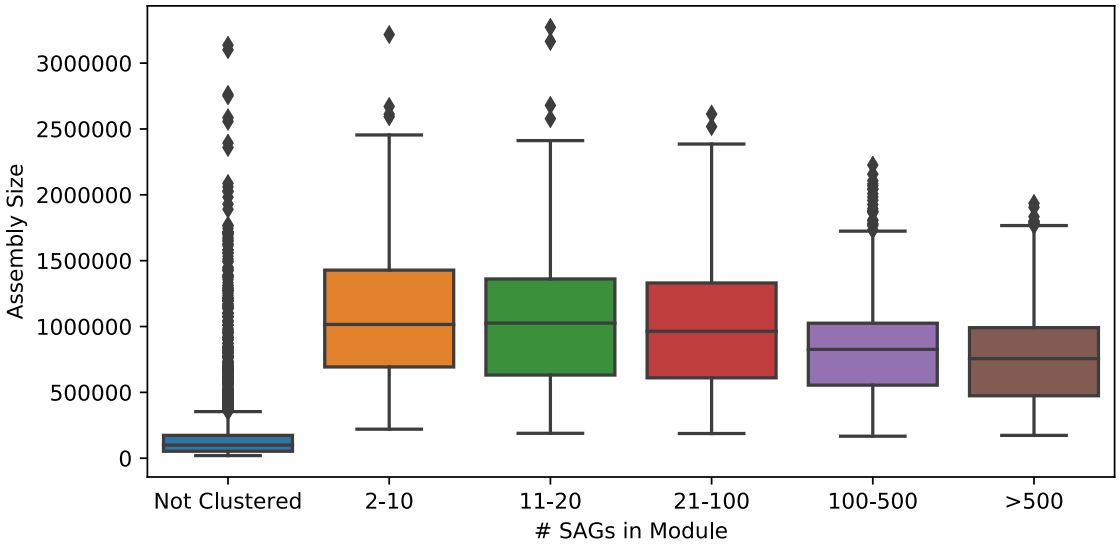

Fig. S5. Genome assembly size distribution in GENs with diverse member count.

5

Table S1. (separate file)

Assignments of GORG-Tropics SAGs to GENs illustrated in Figs. 4 and S4.

Table S2. (separate file)

General characteristics of GENs illustrated in Figs. 4 and S4. Note "*" indicates tentative identities of GENs that lack SAGs with 16S rRNA genes recovered.

References

1 Rasmussen, M. D. & Kellis, M. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research* **22**, 755-765, doi:10.1101/gr.123901.111 (2012).

2 Mallo, D., De Oliveira Martins, L. & Posada, D. SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic Biology* **65**, 334-344, doi:10.1093/sysbio/syv082 (2016).