Title

Charting the Complexity of the Marine Microbiome through Single-Cell Genomics.

Journal

Cell, 179(7)

Authors

Pachiadaki, Maria

Brown, Julia

Brown, Joseph

et al.

Publication Date

2019-12-12

DOI

10.1016/j.cell.2019.11.017

Peer reviewed

HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2020 December 12.

Published in final edited form as:

Cell. 2019 December 12; 179(7): 1623–1635.e11. doi:10.1016/j.cell.2019.11.017.

Charting the complexity of the marine microbiome through single cell genomics

Maria G. Pachiadaki[1,2], **Julia M. Brown**[1], **Joseph Brown**[1], **Oliver Bezuidt**[1], **Paul M. Berube**[3], **Steven J. Biller**[3,6], **Nicole J. Poulton**[1], **Michael D. Burkart**[4], **James J. La Clair**[4], **Sallie W. Chisholm**[3,5], **Ramunas Stepanauskas**[1,7,*]

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

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

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

[4]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, U.S.A.

[5]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, U.S.A.

[6]Current address: Department of Biological Sciences, Wellesley College, Wellesley, Massachusetts 02481, U.S.A.

[7]Lead contact.

Summary:

Marine bacteria and archaea play key roles in global biogeochemistry. To improve our understanding of this complex microbiome, we employed single cell genomics and a randomized, hypothesis-agnostic cell selection strategy to recover 12,715 partial genomes from the tropical and subtropical euphotic ocean. A substantial fraction of known prokaryoplankton coding potential was recovered from a single, 0.4 mL ocean sample, which indicates that genomic information disperses effectively across the globe. Yet, we found each genome to be unique, implying limited clonality within prokaryoplankton populations. Light harvesting and secondary metabolite biosynthetic pathways were numerous across lineages, highlighting the value of single cell genomics to advance the identification of ecological roles and biotechnology potential of uncultured microbial groups. This genome collection enabled functional annotation and genus-

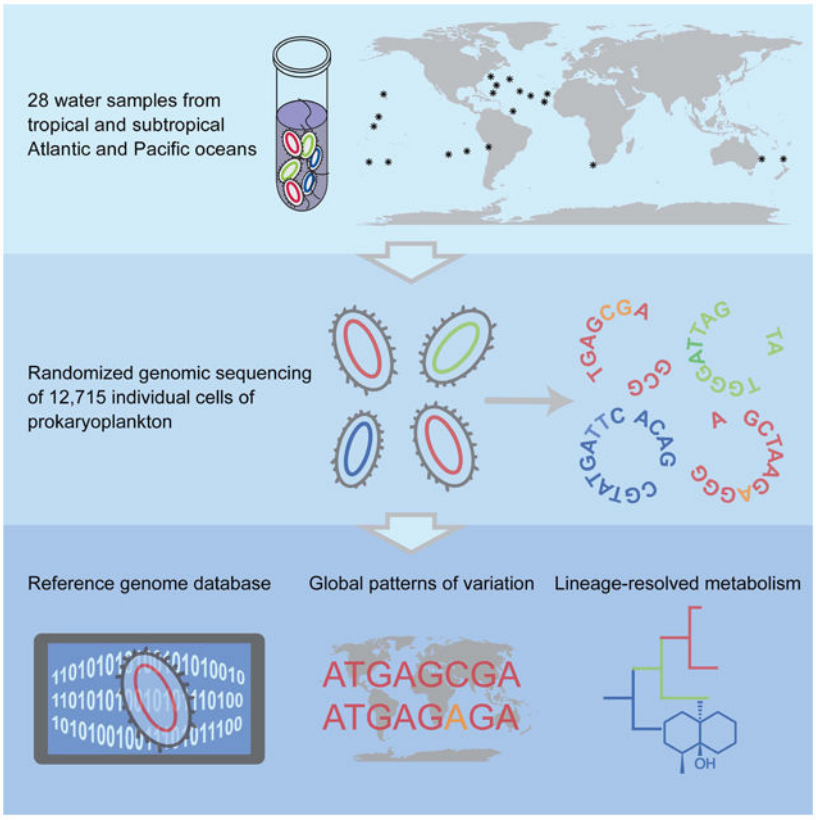
*Correspondence: rstepanauskas@bigelow.org.

Author contributions

R.S. developed the concept, managed the project and led manuscript preparation. M.G.P., J.M.B., J.B., O.B., M.D.B. and J.J.L. performed data analyses and produced figures and tables. M.G.P. and J.M.B. led data quality control. J.B. developed GORG Classifier. M.G.P. and J.M.B. wrote sections on carbon and nitrogen metabolisms and open reading frame clusters, respectively. J.M.B., M.D.B. and J.J.L performed biosynthetic gene clusters analyses. P.M.B., S.J.B. and S.W.C. oversaw field sample collection and selection, and helped manage the project. N.J.P. performed cell sorting and size calibration. All authors contributed to data interpretation and manuscript preparation.

level taxonomic assignments for >80% of individual metagenome reads from the tropical and subtropical surface ocean, thus offering a model to improve reference genome databases for complex microbiomes.

Graphical Abstract



Blurb:

The analysis of single cell genomics of marine microorganisms reveals a very high degree of uniqueness between individual cells, implying limited clonality within populations and establishes that a large fraction of global genetic diversity can be recovered from a single sample, suggesting effective global dispersal or prokaryoplankton.

Keywords

Genomics; reference database; microbial ecology; plankton; oceanography; omics; single cell genomics; biogeography; biodiversity; bioprospecting

Introduction

Unicellular, microbial life has been playing a central role in global biogeochemical processes, ecosystem functioning, and the health of multicellular organisms since its emergence >3.5 Gy ago (Falkowski et al., 2008). Traditional, pure culture-based

microbiology techniques cannot represent the staggering degree of microbial diversity that fills every imaginable life-sustaining niche in the biosphere (Locey and Lennon, 2016; Rappé and Giovannoni, 2003). Thus, studies of natural microbiomes increasingly rely on cultivation-independent tools, in particular the comparative sequence analyses of DNA, RNA and protein that are bulk-extracted from the environment (Handelsman, 2004; Sunagawa et al., 2015). The taxonomic and functional annotation of such metagenomic, metatranscriptomic and metaproteomic data (collectively referred to as “meta-omics”) depends heavily on the availability of suitable reference genomes. Unfortunately, recent studies found that existing reference genomes represent only 5% and 0.4% of gene clusters in human gut (Li et al., 2014) and marine (Sunagawa et al., 2015) metagenomes, respectively. The fraction of individual metagenomic reads that can be recruited on reference genomes ranges from <10% in the ocean to <1% in soils when using a ≥95% average nucleotide identity (ANI) threshold (Nayfach et al., 2016), where ANI in the range of 94-96% is commonly used as an operational delineator of microbial species (Ciufo et al., 2018; Konstantinidis and Tiedje, 2005; Konstantinidis et al., 2006). The paucity of adequate reference genomes remains a major limiting factor in our ability to fully interpret the majority of meta-omics data from most microbiomes.

Novel analytical approaches are being continuously developed to enhance the interpretation of meta-omics data. Improved computational tools for *de novo* assembly and binning of metagenomic reads into discernable units have revealed the coding potential of many deep lineages of Bacteria and Archaea from increasingly complex microbiomes (Anantharaman et al., 2016; Tyson et al., 2004). However, the representation and accuracy of metagenome bins deteriorates at family and lower taxonomic levels, resulting in frequent chimerism (Sczyrba et al., 2017), likely due to a combination of technical constraints and a high degree of cell-to-cell genomic diversity within the environment. For example, no medium-to-high quality bins could be produced for *Candidatus* Pelagibacter (SAR11), the most abundant lineage of marine planktonic bacteria, from global sets of shotgun metagenomes (Delmont et al., 2018; Tully et al., 2018). Furthermore, genomic bins from metagenomes often lack rRNA operons, impairing their taxonomic positioning in the context of rRNA-based phylogeny (Anantharaman et al., 2016; Delmont et al., 2018; Tyson et al., 2004).

Single cell genomics is an alternative approach for cultivation-independent recovery of microbial genomes (Ishoey et al., 2008; Kashtan et al., 2014; Stepanauskas, 2012; Woyke et al., 2017). In contrast to metagenome assembly and binning, single cell genomics does not rely on the assumption of microbial population clonality and instead produces genomic sequences of individual cells. Earlier studies demonstrated that relatively small single cell genomics datasets, consisting of tens of partial genomes, can substantially improve the recruitment of meta-omics data from the ocean (Swan et al., 2013), soil (Choi et al., 2017) and other environments (Garcia et al., 2018; Rinke et al., 2013).

Here we evaluate the capacity of large-scale single cell genomics to represent the genomic makeup of a complex, global microbiome, the surface (epipelagic) ocean in tropical and subtropical latitudes from 40°S to 40°N. Marine microorganisms are of essential importance in geochemical cycling, nutrient remineralization, and climate formation; they comprise one of the largest microbiomes on Earth and have been extensively explored by meta-omics

approaches (Falkowski et al., 2008; Giovannoni et al., 1990; Rusch et al., 2007; Sunagawa et al., 2015; Venter et al., 2004). Using 28 epipelagic samples from the tropical and subtropical Atlantic and Pacific oceans, for which complementary metagenomics and targeted single cell genomics data have been reported previously (Berube et al., 2018; Biller et al., 2018; Hewson et al., 2009; Malmstrom et al., 2012), we generated and sequenced an untargeted library of single amplified genomes (SAGs) of planktonic bacteria and archaea - prokaryoplankton. This dataset differs from earlier single cell genomics projects in both its large scale and randomized, unbiased cell selection strategy, making it suitable for quantitative data mining that is agnostic to the original hypotheses of the study. Additionally, we employed an improved flow cytometry technique to measure physical sizes of the sequenced cells (Stepanauskas et al., 2017), thus adding a new layer of information about the analyzed, uncultured microorganisms.

Results

Prokaryoplankton genomic diversity

Of the 20,288 SAGs generated from the 28 environmental samples (Table S1), 12,715 SAGs (Table S2) produced >20 kbp genome assemblies with no detectable contamination, resulting in a cumulative assembly size of 8.1 Gbp (Table 1). We named this dataset the Global Ocean Reference Genomes Tropics, or GORG-Tropics, database. A subset of 6,236 GORG-Tropics genomes, which we call GORG-BATS248, was obtained from a single, 0.4 mL seawater sample aliquot from the Bermuda Atlantic Time-series Study (BATS) station in the Sargasso Sea to assess the coding potential of prokaryoplankton on local versus global scales. On average, we estimate that 38% of each cell's genome was recovered. The GORG-Tropics database is over an order of magnitude larger than previously reported microbial single cell genomics datasets (Berube et al., 2018; Kashtan et al., 2014; Pachiadaki et al., 2017; Rinke et al., 2013; Swan et al., 2013). By processing each genome individually, the risk of errors in *de novo* genome assembly and the computational cost were minimized relative to metagenome assembly and binning. This study required small sample volumes (0.1 to 0.4 mL) and little processing in the field, which could facilitate the automation of sample collection in the future.

In order to assess the genome-level diversity of marine prokaryoplankton, we calculated the pairwise ANI among the 4,741 GORG-Tropics genomes with ≥50% estimated completion. Most ANI values were <80%, indicating that few of the prokaryoplankton cells were closely related (Fig. 1A). Only ~9,500 (0.08%) of the >11 million genome pairs were found to belong to the same, nominal "species", as defined by the >96% ANI cutoff that was recently adopted by the National Center of Biotechnology Information (NCBI) (Ciufo et al., 2018). This highlights a disconnect between the current, nominal definitions of microbial species on the one hand and the natural, yet poorly understood patterns of genomic variability and microevolution in marine prokaryoplankton and other microbiomes.

None of the genomes were identical to each other at the nucleotide level. Only 121 genome pairs (0.001%), 119 of which came from GORG-BATS248, had an ANI >99.9%. In these 121 pairs, the rate of nucleotide substitutions exceeded the rate of methodological errors (Stepanauskas et al., 2017) by an order of magnitude, and the ratio of their non-synonymous

versus synonymous substitutions averaged ~0.1, which is indicative of purifying selection and cannot be explained by sequencing or assembly errors. These 121 genome pairs also contained non-syntenic regions that encompassed entire operons and putative prophages (Fig. 1B). We found the same regions with ≥80% nucleotide identity in multiple other SAGs of the corresponding lineages SAR11 surface group 1 (e.g. AG-913-O18) and S25-593 (e.g. AG-457-K09), but not in other microbial groups, which provides further support for biological origins of these non-syntenic regions rather than methodological artifacts. This vast microdiversity across all lineages of marine prokaryoplankton expands on the prior studies of *Prochlorococcus*, the most abundant phototroph in the ocean (Kashtan et al., 2014, 2017), as well as other studies of large genome libraries (Good et al., 2017; Shapiro et al., 2012; Wolf et al., 2016). Such genomic variability, which includes both point mutations and gene content variation, likely plays a major role in the collective functioning of complex microbiomes, their compositional dynamics in time and space, and resilience to environmental change.

A recent survey of all prokaryote genomes in NCBI databases identified a pronounced discontinuity between >95% ANI values found within named, nominal species and <83% ANI values found in interspecies comparisons, which was interpreted as an indication of evolutionary forces sustaining biological species-like cohorts of Bacteria and Archaea (Jain et al., 2018). We found no evidence for such discontinuity in GORG-Tropics (Fig. 1A). This may indicate different patterns of microbial diversification in the ocean as compared to other environments. Alternatively, the elevated frequency of ANI >95% among genomes currently held in NCBI may reflect non-random genome sampling, with an overrepresentation of a small number of medically relevant lineages that were selected for sequencing based on the current, nominal species definitions and isolation techniques. The randomized cell selection approach used in our study offers an unbiased view of the genomic composition and evolutionary dynamics of the analyzed microbiomes.

Representation of global prokaryoplankton by GORG-Tropics

We recruited individual reads of 119 publicly available metagenomes from the tropical and subtropical epipelagic (Table S3) using a 95% nucleotide identity threshold to gauge how much of global prokaryoplankton diversity is represented in GORG-Tropics. This recruited 6.3-72.6% (mean=40.0%) of metagenome reads, indicating that GORG-Tropics contains a substantial fraction of the global prokaryoplankton coding potential at the nominal species resolution (Fig. 2A). An average of 58% recruitment could be achieved by relaxing the nucleotide identity threshold, demonstrating that inter-study comparisons require uniform methods (Fig. 2B). We observed strong metagenome fragment recruitment in the Indian Ocean despite using only samples from the Atlantic and Pacific to generate GORG-Tropics (Fig. 2C). Metagenome recruitment was substantially lower, averaging 11%, in temperate and polar waters (Fig. S1). These patterns are consistent with prior reports of water temperature and latitude being the primary drivers of the global distribution of marine planktonic bacteria, archaea and protists, and support the hypothesis that microbes can be dispersed longitudinally over long distances (Seeleuthner et al., 2018; Sunagawa et al., 2015; Swan et al., 2013).

GORG-Tropics substantially improved the recruitment of prokaryoplankton metagenomes as compared to other available reference genome databases, including the recently published genomes of 3,726 marine prokaryotes (Klemetsen et al., 2018) and the complete sets of metagenome assembly bins produced from either exclusively marine metagenomes (Delmont et al., 2018; Tully et al., 2018) or all public metagenomes from diverse environments (Parks et al., 2017) (Fig. 2A). Even GORG-BATS248, which consists of SAGs generated from a single, 0.4 mL water sample, outperformed metagenome bins that were produced from the global set of metagenome reads that was used in recruitment. This indicates that a substantial fraction of the global coding potential of marine prokaryoplankton resides in each tiny parcel of ocean water, due to effective mixing on a global scale. The observed improvement of prokaryoplankton representation by GORG-Tropics likely depends on the ability of single cell genomics to recover variable genome regions in non-clonal populations. This corroborates our finding that close relatives are rare among randomly sampled cells (Fig. 1A).

Next, we analyzed how well GORG-Tropics represents the global pool of prokaryoplankton coding sequences (CDS), which was recently estimated to comprise approximately 40 million clusters, based on the TARA Oceans shotgun metagenomics data (Sunagawa et al., 2015). The co-clustering of CDS from GORG-Tropics genomes and TARA Oceans tropical and subtropical metagenomes produced ~35 million clusters. About 1 million of these clusters, comprising 29% of all CDS, were shared between GORG and TARA, including >99.9% of clusters containing >1,000 CDS (Fig. 3A). Thus, GORG-Tropics offers genomic context to the majority of abundant CDS clusters encoded by surface ocean prokaryoplankton in the tropics and subtropics.

Notably, 97% of all clusters had <10 CDS, while 12% of GORG-Tropics CDS and 24% of TARA CDS were singletons, consistent with the prevalence of rare genes in microorganisms from other environments (Wolf et al., 2016). We found that 56% of GORG-Tropics singletons were also present in the TARA dataset, indicating that they are not artifacts of techniques employed in single cell genomics (GORG) or shotgun metagenomics (TARA). Rare genes were not enriched in eukaryotic or viral sequences (Table S4), implying that most of them do not originate from phage infections or contaminants from Eukarya. While the overall length of CDS peaked at 200 amino acids (aa), singleton length exhibited a bimodal distribution, with local maxima at around 50 aa and 200 aa length (Fig. 3B). Examining GORG CDS specifically, we found no major differences in functional composition between the annotated singletons and clusters (Fig. 3C). However, singletons were enriched in hypothetical proteins (Fig. 3B). On average, the hypothetical CDS were shorter than CDS for which functional annotations could be obtained in both singletons and clusters (Fig. 3D). Annotated singletons were not enriched in short CDS, providing no evidence for incomplete or degrading genes of known function to comprise a major part of rare genes in prokaryoplankton, although difficulties in annotating degraded genes may play a role. The distinct peak in abundance at 50 aa among unannotated singletons suggests that rare prokaryoplankton CDS may be enriched in short sequences of unknown function.

Lineage-resolved genome features of marine prokaryoplankton

Complete or near-complete 16S rRNA gene sequences were recovered from 5,536 GORG-Tropics SAGs, enabling their taxonomic assignment to 20 phyla, 31 classes, 43 orders, 55 families, 49 genera and 1 species of Bacteria and Archaea. The general taxonomic composition of GORG-Tropics is consistent with prior explorations of marine prokaryoplankton using 16S rRNA surveys and shotgun metagenomics (DeLong et al., 2006; Giovannoni et al., 1990), with the predominance of Proteobacteria, Bacteroidetes and Cyanobacteria phyla, and with more than one third of the cells belonging to the lineage SAR11 Surface 1 (Fig. 4, Table S5). Our results also highlight the numeric abundance of lineages such as AEGEAN-169 (4.8% of prokaryoplankton in the analyzed samples) that have received limited attention so far (Reintjes et al., 2019). Importantly, many complete and near-complete 16S rRNA genes from SAGs could not be assigned to SILVA database's taxonomic ranks: classes (1.2%), orders (2.3%), families (11%), genera (72%), and species (99.98%). This demonstrates that a large fraction of marine prokaryoplankton remains taxonomically uncharted.

Over 40 distinct, previously defined prokaryoplankton lineages were represented by at least 10 members in GORG-Tropics (Figs. 4-5, Table S5). Many of these lineages have no or few cultured representatives and no previously published genomes. Our data indicate that most of the prevalent lineages have small genomes (1-2 Mbp), low G+C content (29-35%), and small cell diameters (0.2-0.5 μm) (Figs. 4, S2). These findings are consistent with previous reports of genome streamlining and small cell sizes of the cultured isolates of SAR11, the most abundant lineage of marine prokaryoplankton (Giovannoni, 2017). However, some lineages did not conform to this predominant pattern. For example, Arctic 97B-4 (Verrucomicrobia), OM60 (Gammaproteobacteria), KI89 (Gammaproteobacteria), E01-9C-26 (Gammaproteobacteria), and the Roseobacter cluster (Alphaproteobacteria) exhibited average genome sizes >3 Mbp, G+C content >45% and cell diameters >0.4 μm. This indicates specialized ecological niches and divergent adaptations among lineages with streamlined and non-streamlined genomes.

There was a positive correlation between cell size and genome size among the prokaryoplankton lineages (Fig. 4A), which is in agreement with prior reports that examined non-marine environments and used different methods (Sorensen et al., 2019). This may be caused by both variables being constrained by selective pressures toward streamlining in the pelagic environment (Giovannoni et al., 2014). *Prochlorococcus* and unclassified Synechococcaceae cyanobacteria formed some of the most pronounced outliers in the relationship between cell size and genome size, as they have small genomes (1.65±0.12 and 1.64±0.13 Mbp) despite comparatively large diameters (0.55±0.20 and 0.70±0.23 μm), which may be required to accommodate the photosynthetic machinery. On the opposite end of the spectrum, Verrucomicrobia lineage Arctic 97B-4 and Alphaproteobacteria lineage Roseobacter had similar or smaller cell size estimates than *Prochlorococcus* while possessing >4 Mbp genomes, with their larger genomes likely reflecting elevated metabolic versatility.

The number of CDS clusters encoded by specific lineages (pangenome size) correlated positively with the number of SAGs in a lineage, indicating that we have not exhausted

pangenomes of any of these lineages (Fig. 4B). The largest pangenome (~100,000 clusters) was recovered from the most abundant lineage SAR11 Surface 1, despite their individual genome size averaging only 1.3 Mbp (Table S5). However, lineages containing larger genomes tended to have a greater slope in the pangenome size relative to each new genome added to the analysis (Fig. 4B-C). Most of the rare clusters are lineage-specific, displaying narrow phylogenetic distributions, and most of the genes in large pangenomes are rare (Table S5). Remarkably, we observed no signs of exhausting the pangenome pools of these lineages, with an average of ~45 new clusters added with each new SAG sequenced in lineages represented by >200 SAGs (Fig. 4C-D).

Coding potential for carbon and nitrogen fixation

Microbial fixation of C and N into reduced, biologically accessible forms is essential to the productivity of marine ecosystems. By screening GORG-Tropics for key genetic markers, we identified the presence of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) form I genes, indicative of CO_2 fixation via the Calvin-Benson-Bassham cycle, in Cyanobacteria and in several lineages of Proteobacteria (Fig 5, Fig S3, Table S5). We identified RuBisCO form IC/D, the thiosulfate-induced cytochrome *soxAX*, and bacteriochlorophyll genes in the Alphaproteobacteria lineage Ca. Luxescamonaceae, confirming the recent findings by Graham et al. (2018). This suggests that *Ca.* Luxescamonaceae may be capable of anoxygenic photosynthesis using reduced sulfur compounds as electron donors, in contrast to the well-documented process of anoxygenic phototrophy that does not result in net CO_2 fixation (discussed below). The composition of GORG-Tropics SAGs implies that *Ca.* Luxescamonaceae comprise ~1.1% of prokaryoplankton in tropical and subtropical epipelagic samples (Fig. S2, Table S5) and are an order of magnitude more abundant than initially proposed (Graham et al., 2018). Additionally, both RuBisCO and bacteriochlorophyll genes were detected in two *Parahaliaea* (Gammaproteobacteria) SAGs. The discovery of two potential photolithotrophic lineages, one of which is rather abundant, is unexpected because anoxygenic photosynthesis is thought to rely on reduced sulfur compounds, which are found in low concentrations in oxygenated ocean (Ksionzek et al., 2016). Further studies will be required to confirm and quantify the genomics-predicted involvement of *Ca.* Luxescamonaceae and *Parahaliaea* in anoxygenic photosynthesis.

The presence of genes for sulfur oxidation and RuBisCO indicated the potential for chemoautotrophy in lineages SAR324, Litoricola and ZD405 (Table S5). This is consistent with previous observations that sulfur oxidation-based chemoautotrophy is prevalent in the oxygenated ocean below the epipelagic region (Swan et al., 2011). Intriguingly, SAR11, which constitute a major fraction of marine prokaryoplankton, require reduced sulfur compounds for heterotrophic growth (Tripp et al., 2008). Collectively, this indicates a greater role for reduced sulfur in the biogeochemistry of oxygenated ocean than is currently assumed, potentially operating through cryptic cycles similar to those in hypoxic zones (Canfield et al., 2010).

Photoheterotrophic light harvesting via rhodopsin and bacteriochlorophyll, which does not involve net carbon fixation, is utilized by aquatic microorganisms as a supplementary source

of energy (Béjà et al., 2000; Koblížek, 2015; Pinhassi et al., 2016). We found rhodopsin genes in 58% of all SAGs. Considering the 38% average genome recovery in GORG-Tropics SAGs (Table 1), this finding suggests that most prokaryoplankton cells in the analyzed samples had the potential for photoheterotrophy. Among lineages with >10 SAGs, these genes were only absent from Nitrosopumilales (Thaumarchaeota), Arctic97B-4 (Verrucomicrobia) and Cyanobacteria. Furthermore, bacteriochlorophyll and type-II photochemical reaction centers, but not CO_2 fixation pathways, were identified in Roseobacter (Alphaproteobacteria), OM60 (NOR5) (Gammaproteobacteria) and some rare lineages (Table S5). This reinforces the prevalence of non-photosynthetic harvesting of solar energy in prokaryoplankton, which was recently suggested to absorb a similar amount of solar energy as chlorophyll-a-based phototrophy (Gómez-Consarnau et al., 2019).

We found no evidence for N_2 fixation pathways in any of the analyzed SAGs, including 17 members of the Planctomycetes phylum. This stands in contrast to a recent report of planktonic Planctomycetes being involved in nitrogen fixation (Delmont et al., 2018) and suggests that the capacity for nitrogen fixation among free-living prokaryoplankton in the oxygenated epipelagic waters of the tropical and subtropical ocean is rare, as might be expected from the high demand for energy and Fe and the sensitivity to O2 of this process. In agreement with the established role of Thaumarchaeota in ammonium oxidation (Francis et al., 2005; Könneke et al., 2005; Wuchter et al., 2006), we found ammonia monooxygenase genes in Nitrosopumilales (Table S5). No SAGs contained the genes required for commamox, the complete oxidation of ammonia to nitrate (Daims et al., 2015), suggesting that commamox is likely not significant in the euphotic, oxygenated, tropical ocean. Similarly, no SAGs were found to contain nitrite oxidoreductase, in agreement with previous studies reporting that nitrite oxidizing bacteria are scarce in the euphotic ocean, since they are outcompeted by phytoplankton (Smith et al., 2014; Zakem et al., 2018).

Respiratory nitrate reductase *(narG)* and nitrous oxide reductase *(norZ)*, indicative of denitrification, were found in a small number of SAGs. *narG* was detected in genomes belonging to Alphaproteobacterial lineage Roseobacter and Gammaproteobacteria lineages SAR92 and ZD405, while *nosZ* was recovered in Bacteroidetes lineages Marinoscillum, NS2b, NS4, NS5 and NS9. All but one of the genomes encoding denitrification genes originated from a single, oxygen-depleted sample in the East Tropical South Pacific (Table S1) suggesting a localized distribution. Cosmopolitan lineages likely adapt to local low oxygen conditions by acquiring genes that enable the use of alternative electron acceptors, as shown recently for SAR11 (Tsementzi et al., 2016). These findings highlight the utility of large-scale, randomized single cell genomics to identify the potential of specific microbial lineages to contribute to biogeochemically important processes.

Lineage-resolved biosynthetic gene clusters

Secondary metabolites are important in microbial ecology and are utilized by humans as sources of antibiotics, anti-cancer drugs and other therapeutic compounds (Fenical and Jensen, 2006; Gerwick and Moore, 2012). To date, secondary metabolites in bacteria associated with marine sediments, corals, tunicates, and sponges have received the most attention, while studies of prokaryoplankton have been limited in phylogenetic scope, and

primarily focused on cultivated isolates (Fenical and Jensen, 2006; Gerwick and Moore, 2012). In an effort to bridge this knowledge gap, we applied the genome mining tool antiSMASH (Blin et al., 2017) on the GORG-Tropics dataset. This uncovered, in a quantitative and phylogenetically resolved manner, a remarkably diverse suite of predicted gene clusters for the biosynthesis of terpenes, bacteriocins, polyketides, arylpolyenes, phosphonates, lassopeptides, microcins, ectoines, non-ribosomal peptides, N-Acyl homoserine lactones and other secondary metabolites (Fig. 5, Table S5).

Although Actinobacteria from soils and marine sediments have served as the primary microbial source of bioactive compounds in biotechnology (Rigali et al., 2018), we found that the two most abundant Actinobacteria lineages in marine prokaryoplankton are among the most deplete in biosynthetic gene clusters (Fig. 5, Table S5). Only terpene synthesis clusters were found in the Sva0996 lineage, while no recognizable biosynthetic clusters were found in *Actinomarina*. This is consistent with the genome sizes of *Actinomarina* and Sva0996 being some of the smallest among prokaryoplankton lineages (Figs. 4 and S2, Table S5), although we cannot exclude the possibility that some secondary metabolite clusters escaped detection. Interestingly, some of the uncultured lineages, such as SAR324 (Deltaproteobacteria), Arctic97B-4 (Verrucomicrobia) and Marinamargulisbacteria (Margulisbacteria) encoded among the most diverse sets of biosynthetic clusters, which suggests potential targets for future studies.

Terpene clusters were found in most prokaryoplankton lineages (Fig. 5, Table S5), in agreement with a recent report identifying them in many bacterial genomes in public databases (Yamada et al., 2015). Of particular relevance in terms of therapeutic potential was the observed diversity of polyketide synthase genes (PKSs), constituting markers of one of the major classes of natural products (Helfrich et al., 2019; Hertweck, 2009). Many of the Type I PKS systems shared >80% identity with known PKS-type polyunsaturated fatty acid (PUFA) synthases, which have commercial markets for both prescription drug and nutraceutical applications (Calder, 2015). Several Type I PKS pathways contained conserved 4’-phosphopantetheinyltransferases (PPTases), particularly those from the Sfp superfamily specific for secondary metabolism (Beld et al., 2014). An interesting example of a modular type I PKS cluster was found in SAG AG-912-B08, where the presence of trans-acyltransferase (AT) domains suggested the potential biosynthesis of macrolides, a class of natural products well known for therapeutic utility (Karpiński, 2019) but with unknown function in the oceans. The GORG-Tropics SAGs also contained multiple hybrid, non-ribosomal peptide synthase (NRPS)-Type I PKS and trans-AT PKS systems, natural product classes that have demonstrated utility as antibiotics and chemotherapeutics (Amoutzias et al., 2016; Helfrich et al., 2019; Hertweck, 2009). Several of the NRPS pathways, e.g. in the Bacteroidetes SAG AG-313-C05, displayed biosynthetic elements for siderophores, small molecule iron chelators secreted to scavenge growth-limiting metals (Hider and Kong, 2010). This is just a small selection of the thousands of biosynthetic gene clusters identified in the GORG-Tropics SAGs.

The observed abundance and diversity of biosynthetic clusters in marine prokaryoplankton is surprising, considering their generally small genomes (Figs. 4 and S2, Table S5) and dilute environment, where intercellular communication and warfare may be less effective than in

biofilms and other, more crowded settings. Thus, consideration should be given to the potential for the products of these biosynthetic clusters to play yet unknown, intracellular and intercellular roles. The general patterns highlight how large-scale single cell genomics enables a methodical exploration of biosynthetic capabilities of uncultured microorganisms. Our findings may facilitate the generation of new hypotheses leading to novel insights into the roles of secondary metabolites in microbial interactions in nature as well as translate practical applications for biotechnological and medicinal applications. Cultivation-independent research tools are becoming essential in studies of chemical ecology and bioprospecting (Gerwick and Moore, 2012; Harvey et al., 2015). In contrast to meta-omics, single cell genomics recovers complex biosynthetic clusters from an individual cell, which may improve the characterization of variable regions of these clusters, ensure the compatibility of co-dependent genes and help selecting suitable heterologous expression systems as well as aid in the design of more effective methods for laboratory culturing.

GORG-Tropics as a reference database for prokaryoplankton meta-omics

To improve the utility of this dataset, we created a computational pipeline - the GORG Classifier - which facilitates interpretation of meta-omics data using the GORG-Tropics SAGs as a reference. This tool integrates GORG-Tropics into Kaiju (Menzel et al., 2016) to produce taxonomic and functional annotations of shotgun metagenomes, metatranscriptomes and metaproteome peptide sequences. Evaluation of the performance of the GORG Classifier was conducted by analyzing pre-annotated, mock metagenomes of prokaryoplankton from the tropical versus temperate epipelagic ocean. Mock metagenomes were produced by generating new, randomized SAG datasets separate from GORG-Tropics, and then computationally shredding them to imitate Illumina shotgun reads. These mock metagenomes were analyzed with the GORG Classifier, with either GORG-Tropics or the NCBI non-redundant database (nr) serving as a reference database. The taxonomic and functional assignments obtained were compared to the values expected from the source SAG annotations.

We found that the GORG-Tropics database substantially improved the sensitivity and accuracy of both taxonomic and functional assignments of the tropical epipelagic mock metagenome reads (Fig. 6A). The accuracy of taxonomic assignments was improved at all levels, with lower taxonomic levels showing the greatest improvement. For example, GORG-Tropics enabled correct genus-level classification of 83% reads while keeping the error rate at <0.1%, as compared to only 28% reads accurately classified with the NCBI nr database. The functional assignments showed an even more dramatic improvement, where GORG-Tropics enabled accurate annotation of 86% reads, as compared to only 0.15% reads being correctly annotated with Prokka (Seemann, 2014). This workflow enables both functional and taxonomic annotation of individual, short reads without the computationally expensive and error-prone assembly and binning steps, while retaining the quantitative aspect of raw read data. The annotation improvements offered by GORG-Tropics were limited to mock metagenomes from the tropical and subtropical epipelagic ocean and did not extend into temperate regions, where erroneous taxonomic assignments were prevalent with both nr and GORG-Tropics databases (Fig. 6B). This is consistent with the global patterns of metagenome fragment recruitment in this study (Fig. 2) and earlier findings of

prokaryoplankton differing among tropical, temperate and polar regions, as well as the deep ocean (DeLong et al., 2006; Mende et al., 2017; Swan et al., 2011, 2013).

Discussion

Our expansive, randomized single cell genomics approach enabled quantitative analyses of the distribution of hereditary material in prokaryoplankton of tropical and subtropical epipelagic ecosystem, a complex microbiome that plays a key role in global biogeochemical cycles, in unprecedented detail. We found all 12,715 sequenced cells to be genomically unique and a large fraction of them taxonomically uncharted. We also observed a substantial portion of the global prokaryoplankton pangenome in a single, 0.4 mL ocean water sample. These findings provide a new perspective on the genomic complexity and organization of microbiomes in nature. In particular, each cell's genomic uniqueness offers possible explanations for the large pangenomes of marine microbial lineages and challenges in their separation into metagenome bins.

The approach we employed here enabled the first methodical, lineage-resolved survey of gene clusters involved in energy, nitrogen and secondary metabolisms. This confirmed an earlier finding of the genomic potential for aerobic anoxygenic photosynthesis in *Ca.* Luxescamonaceae, and showed that this lineage of Alphaproteobacteria is substantially more abundant than previously thought. The abundance and diversity of the identified biosynthetic clusters suggested an importance of secondary metabolites in the dilute environment of free-living prokaryoplankton and offered a bioprospecting roadmap for biotechnology applications.

Utilized as a reference database, GORG-Tropics enabled accurate assignment of both taxonomy and predicted functions to the majority of individual metagenome reads from the tropical and subtropical epipelagic, which was not possible before. We expect the GORG-Tropics to serve as a useful resource for marine microbiology. We also propose that randomized single cell genomics should serve as a new, instrumental approach for studies of soil, plant, mammalian and other microbiomes in order to fill our major knowledge gaps about these important microbial players in the functioning of diverse ecosystems and macroorganisms, as well as in climate change and other global processes (Cavicchioli et al., 2019).

STAR Methods

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources, reagents, and scripts should be directed to and will be fulfilled by the Lead Contact, Ramunas Stepanauskas, (rstepanauskas@bigelow.org).

This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Field sample collection—Aquatic samples were collected using Niskin bottles from the epipelagic zone at 20 tropical and subtropical locations in the Pacific and Atlantic oceans (Fig. 2A; Table S1). 1-2 mL aliquots of raw seawater were transferred to sterile cryovials, amended with 10% (final concentration) glycerol for cryoprotection, flash-frozen in liquid nitrogen, and stored at −80°C.

METHOD DETAILS

Single amplified genome (SAG) data generation—The generation, sequencing, *de novo* assembly, annotation and quality control of SAGs were performed at the Bigelow Laboratory for Ocean Sciences' Single Cell Genomics Center (scgc.bigelow.org). First, we utilized SAGs generated in an earlier study (Berube et al., 2018), where one 384-well microplate of SAGs was generated from 24 field samples (Table S1). Additional SAGs were generated from four samples collected during the cruise BULA (Table S1), one microplate per sample. The cryopreserved seawater samples were thawed and pre-filtered through a 40 μm mesh size cell strainer (Becton Dickinson). In order to discriminate heterotrophic bacteria and extracellular particles, seawater samples were incubated with the SYTO-9 DNA stain (5 μM final concentration; Thermo Fisher Scientific) for 10-60 min, after which the particle green fluorescence (proxy for nucleic acid content), light forward scatter (proxy for size), and the ratio of green versus red fluorescence (for improved discrimination of cells from detrital particles) were used to define the sort gate. Fluorescence-activated cell sorting (FACS), cell diameter determination, cells lysis and whole genome amplification with WGA-X were performed as previously described (Stepanauskas et al., 2017).

To gain a deeper understanding of prokaryoplankton coding potential within a single sample, 37 additional microplates of SAGs were generated from a single cryovial of sample BATS248. Twenty of these plates were generated by cell sorting based on the SYTO-9 stain as above, but the sort gate was inclusive of particles with fluorescence spectra typical to *Synechococcus*. Ten of the additional BATS248 microplates were produced after cell labeling with an alternative probe RedoxSensor Green (1 μM final concentration for 20-40 minutes at room temperature; Thermo Fisher Scientific), which targets viable cells (Stepanauskas et al., 2017). The final seven supplementary microplates of BATS248 were generated by sorting particles that fell below the typical prokaryote sort gate on the SYTO-9 fluorescence axis. For all but five prokaryoplankton lineages there was no statistically significant difference in the relative abundance among SAGs generated with SYTO-9 and RedoxSensor Green probes (Welsh two sample *t*-test, $p>0.5$; Fig. S5). However, to avoid potential methodological biases, only SAGs that were generated with the SYTO-9 stain and the typical prokaryoplankton sort gate were used in the quantitative analyses of prokaryoplankton lineages. All 37 supplementary BATS248 SAG plates and all four plates of equatorial SAGs from the BULA expedition were generated with an extended spectrum of index FACS size calibration (Stepanauskas et al., 2017), which included a *Pelagibacter ubique* calibration culture, allowing us to accurately size prokaryote cells in the range of 0.2-2.0 μm.

Sequencing and *de novo* assembly of SAGs—All SAGs were subject to Low Coverage Sequencing (LoCoS) (Stepanauskas et al., 2017), after which 150 SAGs with lowest Cp values from each plate were selected for deeper, post-LoCoS sequencing. While LoCoS generated a variable number of 2×150bp reads per SAG, with an average of ~300k, the post-LoCoS sequencing produced >2M reads for each selected SAG. The goal of this selection and sequencing strategy was to dedicate a deeper, post-LoCoS sequencing effort to a taxonomically unbiased set of SAGs with the highest potential for good genome recovery, based on the previously observed negative correlation between WGA-X Cp and subsequent genome recovery (Stepanauskas et al., 2017). We found no significant taxonomic differences between SAGs with high and low WGA-X Cp values, providing indications that this strategy does not introduce compositional biases during this selection process (X^2-test, p>0.05; Fig. S6). SAG paired-end libraries were created with Nextera XT kits (Illumina), sequenced with a NextSeq 500 (Illumina) and *de novo* assembled using a workflow utilizing SPAdes (Bankevich et al., 2012), as previously described (Stepanauskas et al., 2017). The quality of the sequencing reads was assessed using FastQC and the quality of the assembled genomes was assessed using checkM (Parks et al., 2015) and tetramer frequency analysis (Woyke et al., 2009). This workflow was previously evaluated for assembly errors using three bacterial benchmark cultures with diverse genome complexity and %GC, indicating no non-target and undefined bases in the assemblies and average frequencies of mis-assemblies, indels and mismatches per 100 kbp being 1.5, 3.0 and 5.0 (Stepanauskas et al., 2017).

Although the single cell genomes in this data set were screened for contamination introduced during cell sorting and DNA amplification, users should be aware that these screening procedures may not completely eliminate the potential for multiple genomes being present in the same assembly. Some SAGs, for example, may be derived from cells infected with a bacteriophage (i.e. virocells) and thus contain both host and virus genomes. Other single cells may contain multiple genomes due to a close physical association between two cells that resulted in co-sorting and co-amplification of DNA.

Taxonomic and functional annotation of SAG assemblies—16S rRNA gene regions longer than 500 bp were identified using local alignments provided by BLAST against CREST's (Lanzén et al., 2012) curated SILVA reference database SILVAMod v128 and taxonomic assignments were based on a reimplementation of CREST's last common ancestor algorithm. The taxonomic assignments were used to group SAGs into lineages. Lineage clustering was performed by grouping SAGs with the same SILVA affiliation and the name of lineages correspond to their name of lowest rank in SILVA. In this manuscript, we report lineages that have 10 or more representatives, with the exception of SAR202, Marinamargulisbacteria and NKB19. The latter lineages have fewer than 10 representatives. Marinamargulisbacteria were initially annotated as ML635J-21 Cyanobacteria (k__Bacteria;p__Cyanobacteria;c__ML635J-21;o__?;f__?;g__?;s__?) and NKB19 did not receive taxonomic annotation below Superkingdom level (k__Bacteria;p__?;c__?;o__?;f__?;g__?;s__?). For these two lineages, near full-length 16S rRNA gene sequences were aligned using the SINA aligner (Pruesse et al., 2012) with a curated Bacterial domain 16S rRNA gene phylogeny clustered at an operational taxonomic unit (OTU) threshold of ≥85% nucleotide identity. Maximum-likelihood (ML) phylogenies

were created with MEGA 6.0 (Tamura et al., 2013) using the General Time Reversible (GTR) Model, with Gamma distribution with invariable sites (G+I), and 95% partial deletion for 100 replicate bootstraps. If the SAG 16S rRNA gene sequence had ≥85% nucleotide identity to an unclassified 16S rRNA gene sequence in the database, and phylogenetically clustered with those sequences (i.e. shared a monophyletic node), it was classified as the corresponding candidate phylum.

For Unclassified Rhodobacteraceae, a phylogenetic approach to refine the taxonomy was also applied. The ≥1,300 bp 16S rRNA gene sequences of the 84 SAGs that were initially annotated as “Rhodobacteraceae Unclassified” by CREST were combined with 222 cultured representatives of various Alphaproteobacteria families, aligned using SINA (Pruesse et al., 2012), and used to construct a ML tree with RAxML (Stamatakis, 2014). 16S rRNA gene sequences of Unclassified Rhodobacteraceae SAGs that were shorter than 1,300 bp were then placed in the RAxML tree. The initially Unclassified Rhodobacteraceae SAGs formed two distinct, bootstrap-supported clades (bootstrap value >90). Metabolic gene content and whole genome trees verified that one of the clades was the recently described *Ca.* Luxescamonaceae. For the concatenated protein tree, 13 SAGs with 17 other Alphaproteobacteria genomes that were obtained from the NCBI, including 4 MAGs identified as *Ca.* Luxescamonaceae by Graham et al. (2018) were used for phylogeny. The GToTree phylogenomic workflow (Lee, 2019) was utilized to determine phylogeny of these genomes using a HMM set of 117 single copy gene targets for Alphaproteobacteria. Genomes containing at least half of the total single copy gene targets were kept for further analysis and downstream phylogenetic placement. A ML phylogenetic tree based on the final concatenated SCG sets was generated using FastTree version 2.1.10 (Price et al., 2009) with the default parameters (Fig. S4). Table S2 contains the SILVA taxonomy and the refined lineage assignment of each SAG.

Functional annotation was first performed using Prokka (Seemann, 2014) with default Swiss-Prot databases supplied by the software. Prokka was run a second time with a custom protein annotation database built from compiling Swiss-Prot (Bateman et al., 2017) entries for Archaea and Bacteria. The output of Prokka and the secondary annotation were joined into a single, tab-delimited table with headers identifying the origin of the assignment. Biosynthetic pathways were identified using AntiSMASH 4.0 (Blin et al., 2017), with KnownClusterBlast, ActiveSiteFinder and SubClusterBlast options.

To generate a RuBisCO tree, all RuBisCO sequences from the GORG-Tropics SAGs, 10 sequences of marine MAGs, and sequences from 72 cultured representatives (Fig. S3) were aligned with ClustalW (Thompson et al., 1994). The RuBisCO tree was constructed with MEGA X (Kumar et al., 2018) using the ML and the Le Gascuel 2008 model (Le and Gascuel, 2008). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (6 categories (+G, parameter = 1.4194)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 1.27% sites). The tree is drawn to scale, with branch lengths

measured in the number of substitutions per site. This analysis involved 317 amino acid sequences. All positions with less than 75% site coverage were eliminated.

All unique bacterial 16S sequences from GORG-Tropics were aligned by SINA (Pruesse et al., 2012) and a ML tree with the Kimura 2-parameter model (Kimura, 1980) was constructed using MEGA X (Kumar et al., 2018). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood approach, and then selecting the topology with the superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+G, parameter = 0.6966)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 18.34% sites). The tree is drawn to scale, with branch lengths measured as the number of substitutions per site. This analysis involved 2,898 nucleotide sequences. All positions with less than 75% site coverage were eliminated.

Metagenomic fragment recruitment—We calculated the percentage of read fragments recruited from 119 publically available marine metagenomes (Table S3; Fig. 2A) from epipelagic tropical and subtropical oceanic regions against seven genomic databases: (I) GORG-Tropics vl (current study), (II) GORG-BATS248 (current study), (III) All 98 marine isolates and SAGs sequenced before 2013 (Swan et al., 2013); (IV) 957 nonredundant metagenome assembled genomes (MAGs) from TARA Oceans (Delmont et al., 2018); (V) 2,631 nonredundant good quality (completion≥50% and contamination ≤10%) MAGs from TARA Oceans (Tully et al., 2018); (VI) 7,903 MAGs from all NCBI metagenomes (Parks et al., 2017); and (VII) all 3,726 marine prokaryotic genomes in MarDB database (Klemetsen et al., 2018). The methodological details of the recruitment are reported in detail previously (Pachiadaki et al., 2017). In brief, paired metagenomic reads were joined using flash version 1.2.11 with the following parameters: -x 0.05 -m 20 -M 150 (Magoč and Salzberg, 2011). Successfully joined metagenomic reads were subsampled to 10^6 reads using seqtk and were aligned to a concatenated file containing all genomes from each of the aforementioned databases using bwa mem with the default parameters (Li et al., 2009). Read alignment files were filtered using samtools (Li et al., 2009) and pysam to identify reads aligning at 95% percent identity over a minimum alignment length of 100 nt. Results were visualized using the ggplot2 package in R. For the visualization of the biogeographical distribution of the metagenomic recruitment, the R packages maps and mapdata were used. The percent of reads aligned against the GORG-Tropics database at various percent identity thresholds (100, 98, 95, 92, 90, 88, 85, 80 and 70) was also calculated.

Gene clustering—All coding sequences (CDS) from TARA Oceans tropical and subtropical epipelagic samples (Sunagawa et al., 2015) were downloaded from EMBL and translated into amino acid sequences (92,128,162 sequences). GORG-Tropics protein sequences (8,589,814 sequences) were called using prodigal with the '-p meta' flag to mimic protein calling used for metagenomic samples (Hyatt et al., 2010). TARA and GORG proteins were then combined and clustered using the lindclust clustering method (80% identity threshold and 80 kmers) within the MMseqs2 software package (Steinegger and Soding, 2017). Clustering parameters were selected to reflect the stringent parameters used

for previously reported cluster analyses for TARA microbial metagenomic data (Sunagawa et al., 2015). We attempted to replicate previous clustering methods exactly, but ran into computational resource limitations and instead found MMseqs2 to be a more efficient clustering tool. The amino acid length for singletons and cluster seed sequences from clusters with >10 members were used to generate Figure 3B using seaborn and matplotlib python packages for plotting.

For functional examination of singleton sequences and clusters from GORG-Tropics, and for lineage-specific pangenome analyses, all translated CDS from GORG-Tropics SAGs called by Prokka and functionally annotated as described above were combined and clustered once again, using the lindclust clustering method (80% identity threshold and 80 kmers) within the MMseqs2 software package.

To draw rarefaction curves (Fig. 3C) for each lineage, CDS were randomly sampled from the MMseqs2 output tsv file and determined to be either a member of a previously sampled cluster, or a new cluster. This was repeated until all CDS from each lineage were sampled. Results were plotted as the number of sequences sampled versus the number of clusters accumulated for each lineage using the matplotlib python plotting library. To calculate the number of clusters added per genome (Fig. 3D), rarefaction curves were drawn similarly to 3C, except that sequences were sampled per randomly selected SAG. A linear regression was calculated for the last 10 selected SAGs per lineage against the number of accumulated clusters added per SAG, and the calculated slope was recorded as the rate of CDS clusters added per genome. This was repeated 10 times per lineage, to account for variability in genome completeness among randomly sampled SAGs. The average rate of CDS clusters added per genome was used for plotting and reporting.

Average nucleotide identity (ANI) and synteny analyses—Paiwise ANI was calculated for all GORG-Tropics SAGs with greater than 50% completeness using fastANI with default parameters (Jain et al., 2018). ANI distributions were plotted using seaborn and matplotlib python packages. Synonymous and non-synonymous mutations were assessed by conducting all against all BLASTP searches between pairs of SAGs that shared ≥99% ANI using a 95% sequence identity cut-off. Selected sequence pairs were aligned using Clustal Omega (Sievers et al., 2011) with default parameters. Using the PAL2NAL tool (Suyama et al., 2006), the nucleotide sequences that correspond with each of the aligned protein sequence pairs were converted into codon alignments. The resulting codon alignment pairs were used to estimate synonymous and non-synonymous substitution ratios using the YN00 program from PAML4.8 with an implementation of the Yang and Nielsen 2000 method (Yang, 2007; Yang and Nielsen, 2000).

GORG-Tropics database for Kaiju—The annotated assemblies of SAGs from which the 16S rRNA gene was recovered were compiled into a GORG-Tropics reference database for implementation in Kaiju, a computational tool for meta-omics read annotation (Menzel et al., 2016). The database consists of contigs (GORG_v1.fasta), gene sequences (GORG_v1_<taxonomy>.faa), Kaiju indexes based on NCBI taxonomy (GORG_v1_NCBI.fmi) and SILVAmod taxonomy (GORG_v1_CREST.fmi), and a text reference file linking contig, gene, gene coordinates, and functional annotations to gene

sequence headers (GORG_v1.tsv). The link between gene sequence references and the Kaiju indexes allows both taxonomic and functional annotation. Annotation of DNA or amino acid sequences is performed using Kaiju against GORG's index (-m 11 -e 3). Using Kaiju's supplemental method addTaxonNames, the taxonomic lineage can be added based on the selected index. Names (names.dmp) and nodes (nodes.dmp) definitions per taxonomy are required by Kaiju and each are supplied. Using the GORG tabular annotation data, Kaiju hits are mapped to their respective annotated SAG assembly, from which complete functional annotations are retrieved, including enzyme commission number, gene identifiers, and gene product descriptions.

Evaluation of GORG-based metagenome annotation—We employed mock metagenome datasets to evaluate classifications based on the GORG reference database. New libraries of prokaryoplankton SAGs were generated, LoCoS-sequenced and annotated from one tropical epipelagic and one temperate epipelagic sample (Table S1), two 384-well microplates per sample, following the same procedures as described above. For this purpose, the tropical epipelagic water sample was collected from 80 m depth in the central Atlantic Ocean (22°48’36.0” N, 46°03’37.8” W) on October 14, 2017, during the AT39-01 North Pond CORKs research cruise. The temperate epipelagic sample was collected from 1 m depth in the Gulf of Maine (43°5'37.72" N, 69°34'41.25" W) on April 12, 2017. In both cases, one microplate of SAGs was generated by sorting cells in a typical prokaryote gate using the SYTO-9 stain, and one microplate of SAGs was generated using the RedoxSensor Green probe, as described above. This resulted in 211 tropical and 124 temperate SAGs from which the 16S rRNA genes were retrieved. The 16S-containing assemblies were taxonomically and functionally annotated in the same way as GORG-Tropics SAGs and then computationally shredded into 150-280 basepair shreds using the bbtools script randomreads.sh, resulting in approximately 23 and 7 million mock metagenome reads from each of the environments, corresponding to >20x coverage. The script’s default parameter “adderrors=t” introduced substitution errors in the obtained mock metagenomic reads that are typical to the Illumina sequencing technology. The reads were analyzed with Kaiju with GORG-Tropics as its underlying database, and the obtained taxonomic and functional assignments were compared to the expected values, based on the source SAG annotations.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical significance was determined through *t*-test or X^2-test as reported in the method detail section. All computational and statistical analyses were conducted using the aforementioned referenced open source software tools.

DATA AND CODE AVAILABILITY

GORG database genomes are available at NCBI under bioproject ID PRJEB33281 and at Open Science Framework under DOI 10.17605/OSF.IO/PCWJ9. Mock metagenomes are also available at Open Science Framework under DOI 10.17605/OSF.IO/PCWJ9. As previously reported (Berube et al., 2018), ancillary physical, chemical, and biological data associated with the data set can be accessed from C-MORE (http://hahana.soest.hawaii.edu/cmoreDS/), HOT (http://hahana.soest.hawaii.edu/hot/hot-dogs/), BATS (http://

bats.bios.edu/), GEOTRACES (http://www.geotraces.org/), and SCOPE (http://scope.soest.hawaii.edu/data/) using the sample metadata available in Table S1.

ADDITIONAL RESOURCES

This study did not generate additional resources.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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 appreciate the HOT, BATS, SCOPE, and AT39-01 North Pond CORKs research cruise operations teams for their assistance with sampling, as well as the captains and crew of the R/V Kilo Moana, R/V Ka`imikai-O-Kanaloa, R/V Atlantic Explorer, R/V Pelagia, R/V Southern Surveyor, R/V Tangaroa, RRS James Cook, RRS Discovery, R/V Melville, R/V Knorr and R/V Atlantis. We thank Stephen Giovannoni (Oregon State University) for supplying us with SAR11 cultures, which enabled us to extend the range of cell size calibration during fluorescence-activated cell sorting, and Beth Orcutt (NSF award OCE-1536539) for providing samples from the central Atlantic. We also thank Eric Becraft (University of North Alabama), Mary Ann Moran (University of Georgia), and William Fenical (Scripps Institution of Oceanography) for advice in data interpretation. This work was funded by the Simons Foundation (Life Sciences Project Award ID 510023, R.S.; Life Sciences Project Award ID 337262, S.W.C.; SCOPE Award ID 329108, S.W.C.), as well as the National Science Foundation (OCE-1153588, OCE-1356460, and DBI-0424599 to S.W.C. and OCE-1335810 to R.S.) and NIH grant R21AI134037 to J.J.L., M.D.B. and R.S.

References

Amoutzias DG, Chaliotis A, and Mossialos D (2016). Discovery strategies of bioactive compounds synthesized by nonribosomal peptide synthetases and type-I polyketide synthases derived from marine microbiomes. Mar. Drugs 14, 80.

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, et al. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commun 7, 13219. [PubMed: 27774985]

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol 19, 455–477. [PubMed: 22506599]

Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, et al. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169. [PubMed: 27899622]

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, et al. (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289, 1902. [PubMed: 10988064]

Beld J, Sonnenschein EC, Vickery CR, Noel JP, and Burkart MD (2014). The phosphopantetheinyl transferases: catalysis of a post-translational modification crucial for life. Nat. Prod. Rep 31, 61–108. [PubMed: 24292120]

Berube PM, Biller SJ, Hackl T, Hogle SL, Satinsky BM, Becker JW, Braakman R, Collins SB, Kelly L, Berta-Thompson J, et al. (2018). Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. Sci. Data 5, 180154. [PubMed: 30179231]

Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL, Coe A, Bergauer K, Bouman HA, et al. (2018). Marine microbial metagenomes sampled across space and time. Sci. Data 5, 180176. [PubMed: 30179232]

Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de los Santos ELC, Kim HU, Nave M, et al. (2017). antiSMASH 4.0 -improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 45, W36–W41. [PubMed: 28460038]

Calder PC (2015). Marine omega-3 fatty acids and inflammatory processes: Effects, mechanisms and clinical relevance. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1851, 469–484. [PubMed: 25149823]

Canfield DE, Stewart FJ, Thamdrup B, De Brabandere L, Dalsgaard T, Delong EF, Revsbech NP, and Ulloa O (2010). A cryptic sulfur cycle in oxygen-minimum - zone waters off the Chilean coast. Science 330, 1375. [PubMed: 21071631]

Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, Behrenfeld MJ, Boetius A, Boyd PW, Classen AT, et al. (2019). Scientists' warning to humanity: microorganisms and climate change. Nature Reviews Microbiology 17, 569–586. [PubMed: 31213707]

Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, Flater J, Tiedje JM, Hofmockel KS, Gelder B, et al. (2017). Strategies to improve reference databases for soil microbiomes. ISME J. 11, 829–834. [PubMed: 27935589]

Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, Brover S, Schoch C, Kimchi A, and DiCuccio M (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Int. J. Syst. Evol. Microbiol 68, 2386–2392. [PubMed: 29792589]

Daims H, Lebedeva EV, Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M, Vierheilig J, Bulaev A, et al. (2015). Complete nitrification by *Nitrospira* bacteria. Nature 528, 504. [PubMed: 26610024]

Delmont TO, Quince C, Shaiber A, Esen ÖC, Lee ST, Rappé MS, McLellan SL, Lücker S, and Eren AM (2018). Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat. Microbiol 3, 804–813. [PubMed: 29891866]

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, Martinez A, Sullivan MB, Edwards R, Brito BR, et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. Science 311, 496. [PubMed: 16439655]

Falkowski PG, Fenchel T, and Delong EF (2008). The microbial engines that drive earth's biogeochemical cycles. Science 320, 1034. [PubMed: 18497287]

Fenical W, and Jensen PR (2006). Developing a new resource for drug discovery: marine actinomycete bacteria. Nature Chemical Biology 2, 666–673. [PubMed: 17108984]

Francis CA, Roberts KJ, Beman JM, Santoro AE, and Oakley BB (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc Natl Acad Sci U S A. 102, 14683. [PubMed: 16186488]

Garcia SL, Stevens SLR, Crary B, Martinez-Garcia M, Stepanauskas R, Woyke T, Tringe SG, Andersson SGE, Bertilsson S, Malmstrom RR, et al. (2018). Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. ISME J. 12, 742–755. [PubMed: 29222442]

Gerwick WH, and Moore BS (2012). Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. Chemistry and Biology 19, 85–98. [PubMed: 22284357]

Giovannoni SJ (2017). SAR11 bacteria: the most abundant plankton in the oceans. Annu. Rev. Mar. Sci 9, 231–255.

Giovannoni SJ, Britschgi TB, Moyer CL, and Field KG (1990). Genetic diversity in Sargasso Sea bacterioplankton. Nature 345, 60–63. [PubMed: 2330053]

Giovannoni SJ, Cameron Thrash J, and Temperton B (2014). Implications of streamlining theory for microbial ecology. ISME J. 8, 1553. [PubMed: 24739623]

Gish W, and States DJ (1993). Identification of protein coding regions by database similarity search. Nature Genetics 3, 266–272. [PubMed: 8485583]

Gomez-Consarnau L, Raven JA, Levine NM, Cutter LS, Wang D, Seegers B, Aristegui J, Fuhrman JA, Gasol JM, and Sanudo-Wilhelmy SA (2019). Microbial rhodopsins are major contributors to the solar energy captured in the sea. Sci Adv 5, eaaw8855. [PubMed: 31457093]

Good BH, McDonald MJ, Barrick JE, Lenski RE, and Desai MM (2017). The dynamics of molecular evolution over 60,000 generations. Nature 551, 45. [PubMed: 29045390]

Graham ED, Heidelberg JF, and Tully BJ (2018). Potential for primary productivity in a globally-distributed bacterial phototroph. ISME J. 12, 1861–1866. [PubMed: 29523891]

Handelsman J (2004). Metagenomics: application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev 68, 669. [PubMed: 15590779]

Harvey AL, Edrada-Ebel R, and Quinn RJ (2015). The re-emergence of natural products for drug discovery in the genomics era. Nature Reviews Drug Discovery 14, 111. [PubMed: 25614221]

Helfrich EJN, Ueoka R, Dolev A, Rust M, Meoded RA, Bhushan A, Califano G, Costa R, Gugger M, Steinbeck C, et al. (2019). Automated structure prediction of trans-acyltransferase polyketide synthase products. Nature Chemical Biology 15, 813–821. [PubMed: 31308532]

Hertweck C (2009). The Biosynthetic Logic of Polyketide Diversity. Angewandte Chemie International Edition 48, 4688–4716. [PubMed: 19514004]

Hewson I, Paerl RW, Tripp HJ, Zehr JP, and Karl DM (2009). Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. Limnol. Oceanogr 54, 1981–1994.

Hider RC, and Kong X (2010). Chemistry and biology of siderophores. Nat. Prod. Rep 27, 637–657. [PubMed: 20376388]

Hunter JD (2007). Matplotlib: A 2D Graphics Environment. Computing in Science and Engineering 9, 90–95.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, and Hauser LJ (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. [PubMed: 20211023]

Ishoey T, Woyke T, Stepanauskas R, Novotny M, and Lasken RS (2008). Genomic sequencing of single microbial cells from environmental samples. Current Opinion in Microbiology 11, 198–204. [PubMed: 18550420]

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, and Aluru S (2018). High throughput ANI analysis of 90k prokaryotic genomes reveals clear species boundaries. Nat. Commun 9, 5114. [PubMed: 30504855]

Karpiński TM (2019). Marine Macrolides with Antibacterial and/or Antifungal Activity. Mar Drugs 17, 241.

Kashtan N, Roggensack SE, Berta-Thompson JW, Grinberg M, Stepanauskas R, and Chisholm SW (2017). Fundamental differences in diversity and genomic population structure between Atlantic and Pacific *Prochlorococcus*. ISME J. 11, 1997. [PubMed: 28524867]

Kimura M (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol 16, 111–120. [PubMed: 7463489]

Klemetsen T, Raknes IA, Fu J, Agafonov A, Balasundaram SV, Tartari G, Robertsen E, and Willassen NP (2018). The MAR databases: development and implementation of databases specific for marine metagenomics. Nucleic Acids Res. 46, D692–D699. [PubMed: 29106641]

Koblížek M (2015). Ecology of aerobic anoxygenic phototrophs in aquatic environments. FEMS Microbiol. Rev 39, 854–870. [PubMed: 26139241]

Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, and Stahl DA (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437, 543–546. [PubMed: 16177789]

Konstantinidis KT, and Tiedje JM (2005). Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102, 2567. [PubMed: 15701695]

Konstantinidis KT, Ramette A, and Tiedje JM (2006). The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 1929–1940.

Ksionzek KB, Lechtenfeld OJ, McCallister SL, Schmitt-Kopplin P, Geuer JK, Geibert W, and Koch BP (2016). Dissolved organic sulfur in the ocean: Biogeochemistry of a petagram inventory. Science 354, 456. [PubMed: 27789839]

Kumar S, Stecher G, Li M, Knyaz C, and Tamura K (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol 35, 1547–1549. [PubMed: 29722887]

Lanzén A, Jørgensen SL, Huson DH, Gorfer M, Grindhaug SH, Jonassen I, Øvreås L, and Urich T (2012). CREST – Classification resources for environmental sequence tags. PLoS One 7, e49334. [PubMed: 23145153]

Le SQ, and Gascuel O (2008). An improved general amino acid replacement matrix. Mol. Biol. Evol 25, 1307–1320. [PubMed: 18367465]

Lee MD (2019). Applications and Considerations of GToTree: a user-friendly workflow for phylogenomics. Evolutionary Bioinformatics 15, 1.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079. [PubMed: 19505943]

Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, et al. (2014). An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol 32, 834. [PubMed: 24997786]

Locey KJ, and Lennon JT (2016). Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A. 113, 5970. [PubMed: 27140646]

Magoč T, and Salzberg SL (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinforma. Oxf. Engl 27, 2957–2963.

Malmstrom RR, Rodrigue S, Huang KH, Kelly L, Kern SE, Thompson A, Roggensack S, Berube PM, Henn MR, and Chisholm SW (2012). Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. ISME J. 7, 184. [PubMed: 22895163]

Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, and DeLong EF (2017). Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat. Microbiol 2, 1367–1373. [PubMed: 28808230]

Menzel P, Ng KL, and Krogh A (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun 7, 11257. [PubMed: 27071849]

Nayfach S, Rodriguez-Mueller B, Garud N, and Pollard KS (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 26, 1612–1625. [PubMed: 27803195]

Oliveira PH, Touchon M, Cury J, and Rocha EPC (2017). The chromosomal organization of horizontal gene transfer in bacteria. Nat. Commun 8, 841. [PubMed: 29018197]

Pachiadaki MG, Sintes E, Bergauer K, Brown JM, Record NR, Swan BK, Mathyer ME, Hallam SJ, Lopez-Garcia P, Takaki Y, et al. (2017). Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. Science 358, 1046. [PubMed: 29170234]

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, and Tyson GW (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055. [PubMed: 25977477]

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, and Tyson GW (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol 2, 1533–1542. [PubMed: 28894102]

Pinhassi J, DeLong EF, Beja O, Gonzalez J, and Pedrós-Alió C (2016). Marine bacterial and archaeal ion-pumping rhodopsins: genetic diversity, physiology, and ecology. Microbiol. Mol. Biol. Rev 80, 929. [PubMed: 27630250]

Price MN, Dehal PS, and Arkin AP (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol 26, 1641–1650. [PubMed: 19377059]

Pruesse E, Peplies J, and Glöckner FO (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28, 1823–1829. [PubMed: 22556368]

Rappé MS, and Giovannoni SJ (2003). The Uncultured Microbial Majority. Annu. Rev. Microbiol 57, 369–394. [PubMed: 14527284]

Reintjes G, Tegetmeyer HE, Bürgisser M, Orlić S, Tews I, Zubkov M, Voβ D, Zielinski O, Quasi C, Glöckner FO, et al. (2019). On-site analysis of bacterial communities of the ultraoligotrophic South Pacific Gyre. Appl. Environ. Microbiol 85, e00184–19. [PubMed: 31076426]

Rigali S, Anderssen S, Naômé A, and van Wezel GP (2018). Cracking the regulatory code of biosynthetic gene clusters as a strategy for natural product discovery. Biochem. Pharmacol 153, 24–34. [PubMed: 29309762]

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431. [PubMed: 23851394]

Rusch DB, Halpem AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, et al. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLOS Biology 5, e77. [PubMed: 17355176]

Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droge J, Gregor I, Majda S, Fiedler J, Dahms E, et al. (2017). Critical Assessment of Metagenome Interpretation - a benchmark of metagenomics software. Nat. Methods 14, 1063. [PubMed: 28967888]

Seeleuthner Y, Mondy S, Lombard V, Carradec Q, Pelletier E, Wessner M, Leconte J, Mangot J-F, Poulain J, Labadie K, et al. (2018). Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. Nat. Commun 9, 310. [PubMed: 29358710]

Seemann T (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. [PubMed: 24642063]

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, and Aim EJ (2012). Population genomics of early events in the ecological differentiation of bacteria. Science 336, 48. [PubMed: 22491847]

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol 7, 539. [PubMed: 21988835]

Smith JM, Chavez FP, and Francis CA (2014). Ammonium uptake by phytoplankton regulates nitrification in the sunlit ocean. PLoS One 9, e108173. [PubMed: 25251022]

Sorensen JW, Dunivin TK, Tobin TC, and Shade A (2019). Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. Nature Microbiology 4, 55–61.

Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. [PubMed: 24451623]

Steinegger M, and Söding J (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol 35, 1026–1028. [PubMed: 29035372]

Stepanauskas R (2012). Single cell genomics: an individual look at microbes. Current Opinion in Microbiology 15, 613–620. [PubMed: 23026140]

Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, Becraft ED, Brown JM, Pachiadaki MG, Povilaitis T, et al. (2017). Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. Nat. Commun 8, 84. [PubMed: 28729688]

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al. (2015). Structure and function of the global ocean microbiome. Science 348, 1261359. [PubMed: 25999513]

Suyama M, Torrents D, and Bork P (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609–612. [PubMed: 16845082]

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland EDP, Gomez ML, et al. (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science 333, 1296. [PubMed: 21885783]

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, Luo H, Wright JJ, Landry ZC, Hanson NW, et al. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc Natl Acad Sci U S A. 110, 11463. [PubMed: 23801761]

Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol 30, 2725–2729. [PubMed: 24132122]

Thompson JD, Higgins DG, and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680. [PubMed: 7984417]

Tripp HJ, Kitner JB, Schwalbach MS, Dacey JWH, Wilhelm LJ, and Giovannoni SJ (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. Nature 452, 741. [PubMed: 18337719]

Tsementzi D, Wu J, Deutsch S, Nath S, Rodriguez-R LM, Burns AS, Ranjan P, Sarode N, Malmstrom RR, Padilla CC, et al. (2016). SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature 536, 179. [PubMed: 27487207]

Tully BJ, Graham ED, and Heidelberg JF (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci. Data 5, 170203. [PubMed: 29337314]

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, and Banfield JF (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37–43. [PubMed: 14961025]

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al. (2004). Environmental genome shotgun sequencing of the sargasso sea. Science 304, 66. [PubMed: 15001713]

Wickham H (2011). ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics 3, 180–185.

Wolf YI, Makarova KS, Lobkovsky AE, and Koonin EV (2016). Two fundamentally different classes of microbial genes. Nat. Microbiol 2, 16208. [PubMed: 27819663]

Woyke T, Doud DFR, and Schulz F (2017). The trajectory of microbial single-cell sequencing. Nat. Methods 14, 1045. [PubMed: 29088131]

Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, Kiss H, Saw JH, Senin P, Yang C, and Chatterji S (2009). Assembling the marine metagenome, one cell at a time. PloS One 4, e5299. [PubMed: 19390573]

Wuchter C, Abbas B, Coolen MJL, Herfort L, van Bleijswijk J, Timmers P, Strous M, Teira E, Herndl GJ, Middelburg JJ, et al. (2006). Archaeal nitrification in the ocean. Proc Natl Acad Sci USA 103, 12317. [PubMed: 16894176]

Yamada Y, Kuzuyama T, Komatsu M, Shin-Ya K, Omura S, Cane DE, and Ikeda H (2015). Terpene synthases are widely distributed in bacteria. Proc Natl Acad Sci U S A 112, 857–862. [PubMed: 25535391]

Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol 24, 1586–1591. [PubMed: 17483113]

Yang Z, and Nielsen R (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol 17, 32–43. [PubMed: 10666704]

Zakem EJ, Al-Haj A, Church MJ, van Dijken GL, Dutkiewicz S, Foster SQ, Fulweiler RW, Mills MM, and Follows MJ (2018). Ecological control of nitrite in the upper ocean. Nat. Commun 8, 1206.

Highlights:

1. Each drop of seawater contains much of the global prokaryoplankton pangenome.
2. Individual cells' genomic uniqueness limits separation into metagenome bins.
3. Methodical survey highlights lineage-resolved energy and secondary metabolism.
4. Randomized sampling of individual genomes offers a new model to study microbiomes.

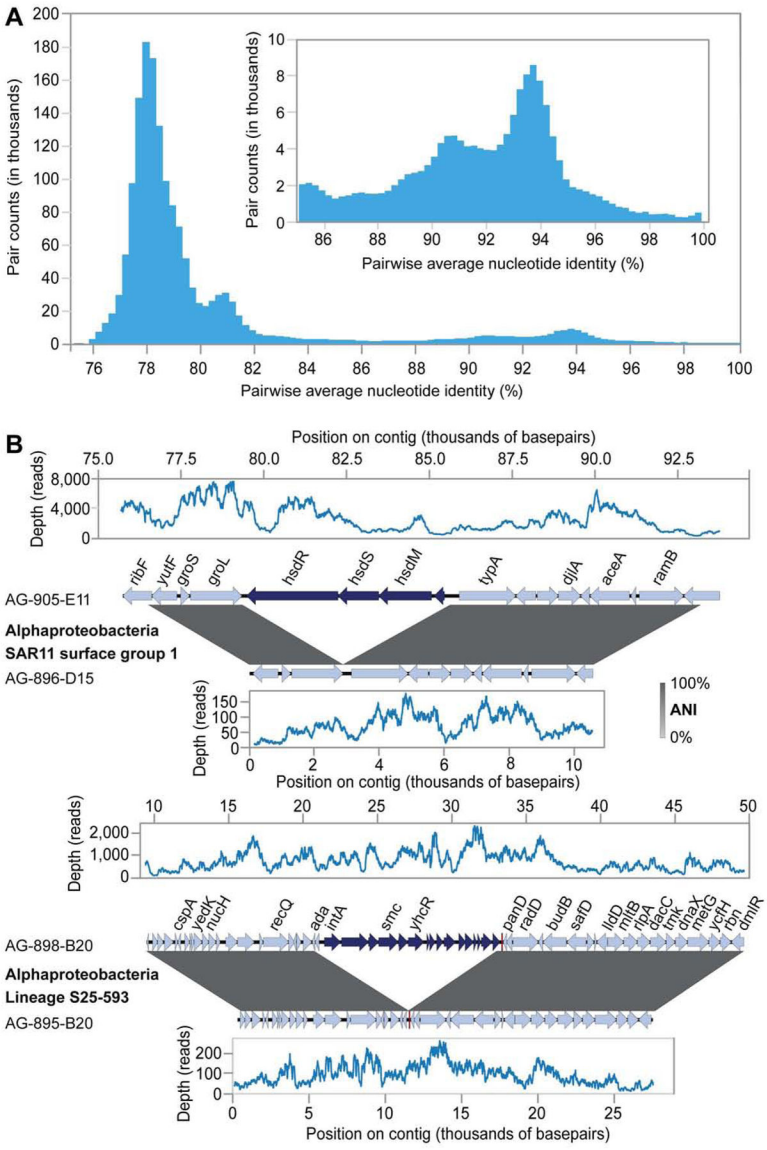


Figure 1.

Genomic diversity among GORG-Tropics SAGs. (A) Pairwise Average Nucleotide Identity (ANI) of SAGs with estimated completeness of ≥50%. The inset is an enlarged region of 85-100% ANI. (B) Examples of gene content differences among SAGs with ANI >99.9%. Red bars indicate tRNA genes. Sequence coverage depth is provided for the aligned regions and ranges from 6 to >7,000.

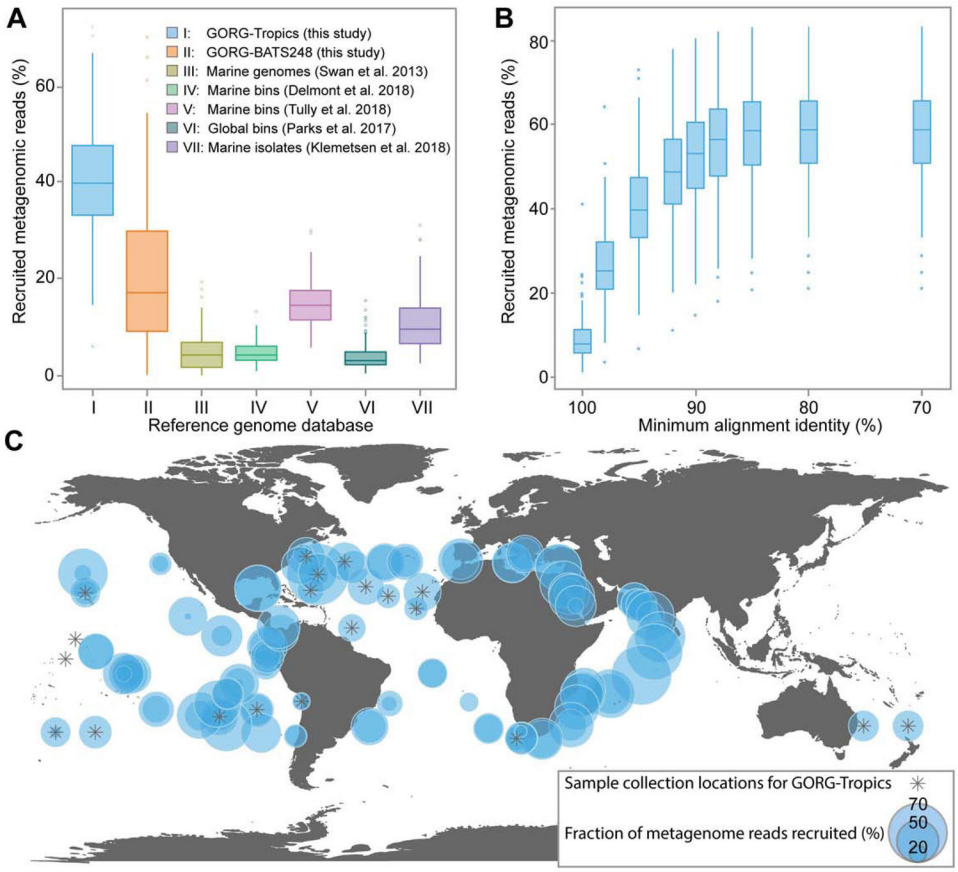


Figure 2.

Recruitment of reads from tropical and subtropical epipelagic metagenomes. (A) Fraction of reads recruited from each of the 119 public metagenomes against the following genome databases: (I) GORG-Tropics (current study), (II) GORG-BATS248 (current study), (III) all 98 marine isolates and SAGs sequenced by 2012 (Swan et al., 2013); (IV) 957 non-redundant metagenome bins from TARA Oceans (Delmont et al., 2018); (V) 2,631 non-redundant bins from TARA Oceans with estimated ≥50% completion and ≤10% contamination (Tully et al., 2018); (VI) 7,903 bins generated from all NCBI metagenomes (Parks et al., 2017); and (VII) all 3,726 marine prokaryotic genomes in the MarDB database (Klemetsen et al., 2018). Thresholds of ≥95% nucleotide identity and 100 bp alignment length were used in these analyses. (B) Fraction of reads recruited from each of the 119 public metagenomes against GORG-Tropics using various nucleotide identity thresholds and a minimum of 100 bp alignment length. (C) Geographic distribution of recruitment against GORG-Tropics at nucleotide identity ≥95%. Circle centers correspond to metagenome collection location. Geographic coordinates can be found in Table S1.

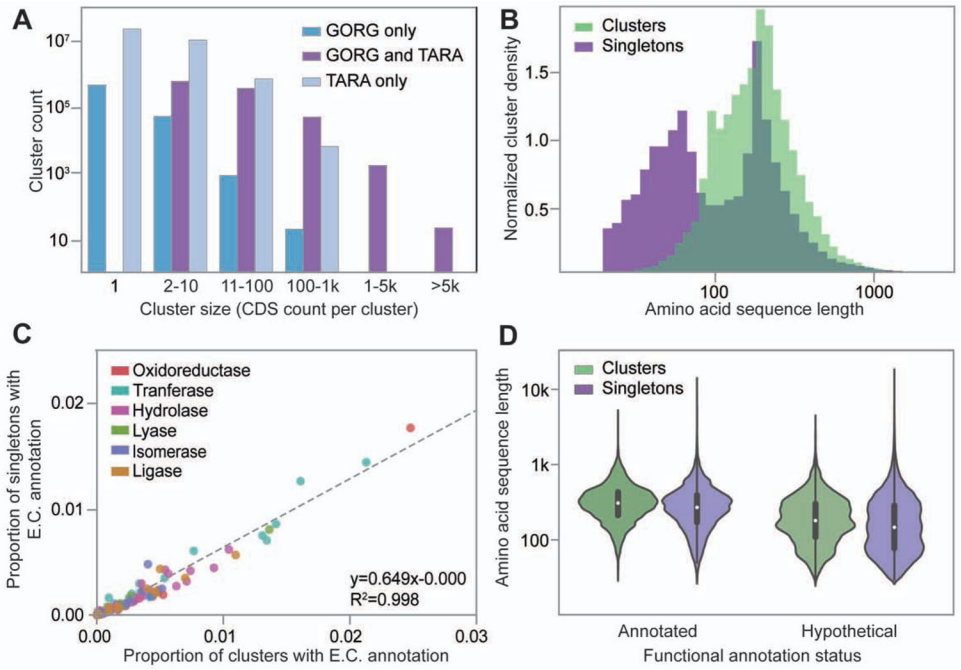


Figure 3.

General patterns in coding sequence (CDS) clusters. (A) CDS clusters shared between the GORG-Tropics and TARA Oceans datasets, as a function of cluster size. (B) Normalized density histogram of protein sequence length in clusters with >10 CDS (green, n = 1,035,323) and singletons (blue, n = 22,263,710). Clustering was performed on combined GORG and TARA CDS. (C) Correlation of the counts of sequences with three levels of EC annotation in clustered CDS from GORG-Tropics in clusters containing >10 members compared to corresponding annotations within GORG-Tropics singleton sequences. (D) Predicted protein size distributions for annotated and unannotated gene clusters from GORG-Tropics containing >10 sequences compared to GORG-Tropics singletons. Displayed values represent means for the last 10 SAGs sampled in each lineage.

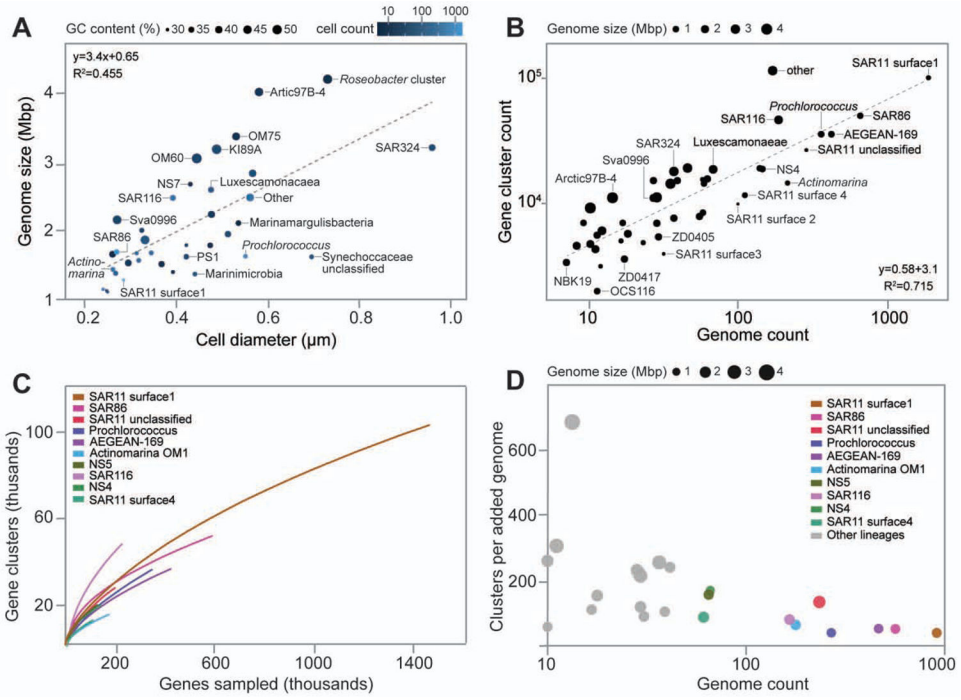


Figure 4.

General characteristics of prokaryoplankton lineages represented by ≥10 SAGs in GORG-Tropics. Phyla Marinamargulisbacteria, NBK19 and Chloroflexi were also included as individual lineages, although they contained <10 SAGs. (A) Relationships among average cell diameter, average genome size and average G+C content. (B) Relationships among SAG count, pangenome size, and average genome size. (C) Accumulation of gene clusters in prokaryoplankton lineages as a function of genes added. Included are lineages with >100k genes in GORG-Tropics. (D) Number of new gene clusters per each new SAG added to the database. Displayed are means for last 10 SAGs sampled for each lineage.

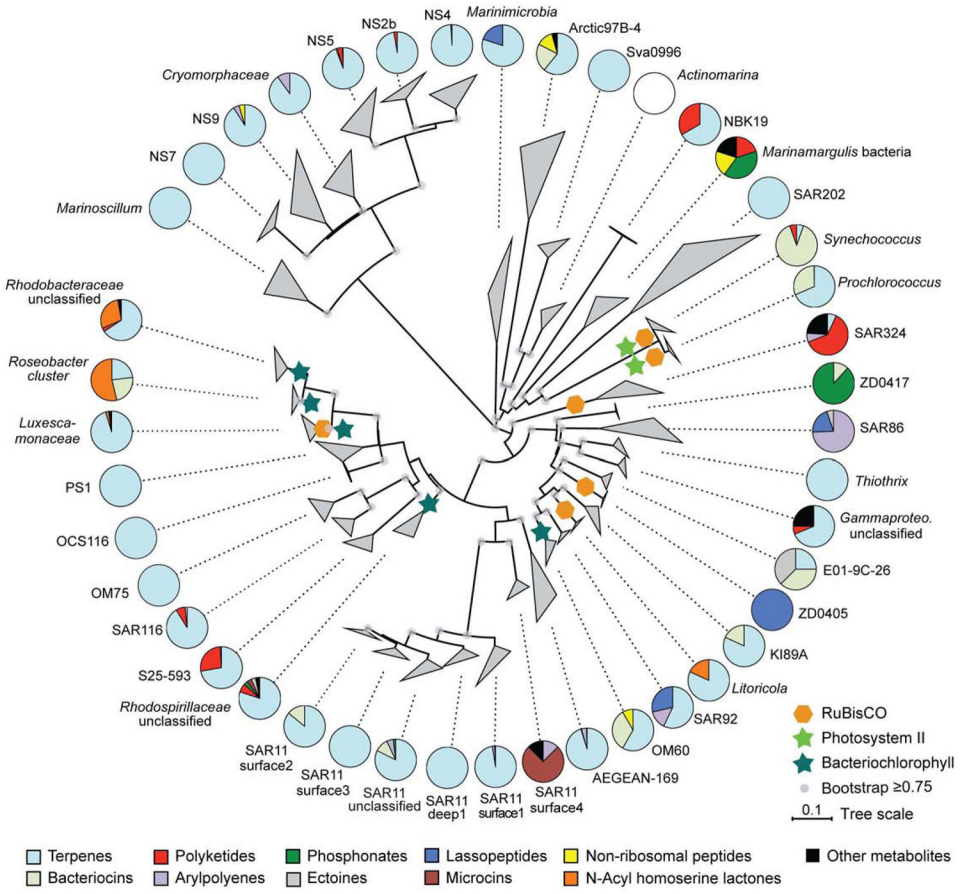


Figure 5.

Lineage-resolved genomic potential for RuBisCO, bacteriochlorophyll and secondary metabolites in the context of the 16S rRNA gene phylogeny. The phylogenetic tree was constructed using MEGA X (Kumar et al., 2018). Secondary metabolite gene clusters were predicted with antiSMASH 4.2.0 (Blin et al., 2017). Pie charts indicate relative abundances of metabolite clusters among genomes with at least one cluster within each lineage. The type of biosynthetic system is provided by color-coding and reflects a binning of antiSMASH biosynthetic gene cluster types in parentheses: Terpenes (terpene); Bacteriocins (bacteriocin and bacteriocin-terpene); Polyketides (T1pks, T1pks-nrps, T1pks-PUFA, T1pks-PUFA-otherks, T1pks-otherks, T3pks, phosphonate-T3pks-terpene, otherks-butyrolactone-nrps, transatpks, and otherks); Arylpolyenes (arylpolyene); Phosphonates (phosphonate and phosphonate-terpene); Ecotines (ectoine); Lassopeptide (lassopeptide); Microcin (microcin); Non-ribosomal peptides (nrps, bacteriocin-nrps and lantipeptide-nrps); N-Acyl homoserine lactones (hserlactone); and Other metabolites (acyl amino acids, ladderane, lantipeptide, nucleoside, PUFA, resorcinol, siderophore, and other). Expanded analyses of secondary metabolite biosynthetic potential are provided in Tables S2 and S5.

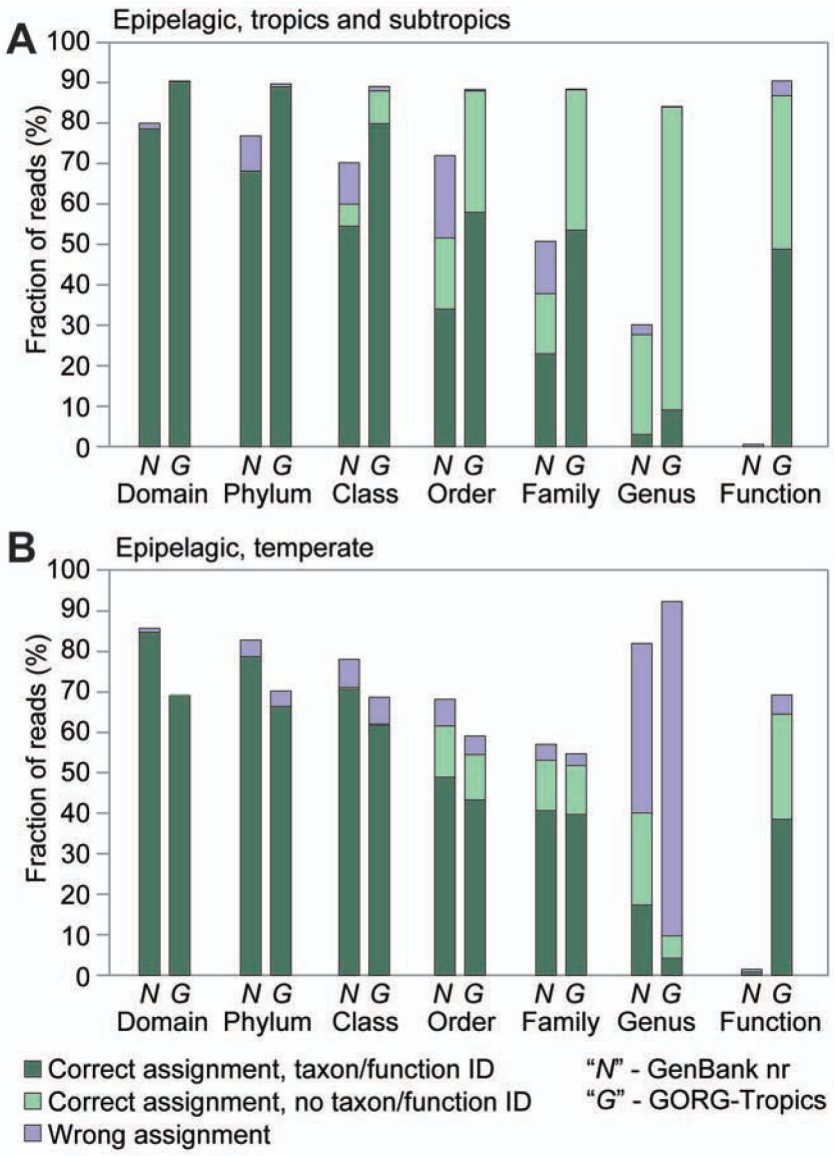


Figure 6.

Fraction of independently derived mock metagenome reads with correct and incorrect taxonomy and function assignments using GenBank nr versus GORG-Tropics databases. Kaiju (Menzel et al., 2016) was used for all taxonomy assignments and for the assignment of functions based on the GORG-Tropics database. Prokka (Seemann, 2014) was used for the functional annotation of GORG-Tropics and mock metagenome reads.

Table 1.

Overview of the GORG-Tropics database. GORG-BATS248 is a subset of GORG-Tropics and originates from a single sample from the Sargasso Sea.

Metric	GORG-Tropics	GORG-BATS248
Field samples	28	1
Sample volume analyzed, mL	3.1	0.4
SAGs sequenced	20,288	11,729
SAG assemblies ≥20 kbp	12,715	6,236
SAG assemblies ≥50% completion	4,741	2,533
SAG assemblies ≥80% completion	1,040	637
SAGs with 16S rRNA recovery	5,536	2,442
Cumulative assembly, Mbp	8,122	4,094
Average genome recovery, %	38	39
Phyla of Bacteria and Archaea	20	12
Classes of Bacteria and Archaea	31	16
Orders of Bacteria and Archaea	43	23
Families of Bacteria and Archaea	55	33
Genera of Bacteria and Archaea	49	26