Comparative whole-genome approach to identify bacterial traits for microbial interactions

Luca Zoccarato*[a], Daniel Sher*[b], Takeshi Miki[c], Daniel Segrè[d,e], Hans-Peter Grossart*[a,f,g]

(a) Department Experimental Limnology, Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), 16775 Stechlin, Germany

(b) Department of Marine Biology, Leon H. Charney School of Marine Sciences, University of Haifa, 3498838 Haifa, Israel

(c) Department of Environmental Solution Technology, Ryukoku University, 612-8577 Kyoto, Japan

(d) Departments of Biology, Biomedical Engineering, Physics, Boston University, 02215 Boston, MA

(e) Bioinformatics Program & Biological Design Center, Boston University, 02215 Boston, MA

(f) Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), 14195 Berlin, Germany

(g) Institute of Biochemistry and Biology, Potsdam University, 14476 Potsdam, Germany

Corresponding authors(*)

Luca Zoccarato, zoccarato@igb-berlin.de; Daniel Sher, dsher@univ.haifa.ac.il; Hans-Peter Grossart, hgrossart@igb-berlin.de

Abstract

Microbial interactions shape the structure and function of microbial communities with profound consequences for biogeochemical cycles and ecosystem health. Yet, most interaction mechanisms are studied only in model systems and their prevalence is unknown. To systematically explore the functional and interaction potential of sequenced marine bacteria, we developed a new trait-based approach, and applied it to 473 complete genomes (248 genera), representing a significant fraction of marine microbial communities. We identified genome functional clusters (GFCs) which condense bacterial diversity into groups with potentially common ecology and life history. Most

GFCs revealed unique combinations of interaction traits, some widely distributed (e.g., antimicrobial activity), others less common (e.g., production of siderophores and phytohormones, biosynthesis and export of specific B vitamins). Specific GFCs, comprising Alpha- and Gammaproteobacteria, are predicted to preferentially interact synergistically and/or antagonistically with bacteria and phytoplankton. Moreover, linked trait clusters (LTCs) identify traits that may have evolved together and point to specific modes of interactions. Our approach translates multidimensional genomic information into an atlas of marine bacteria and their ecosystem functions, relevant for understanding the fundamental rules that govern community assembly and dynamics.

Keywords

marine bacteria; functional clusters; genetic traits; interaction mechanisms, B vitamins; siderophores; phytohormones; motility and chemotaxis; secretion systems; antimicrobial compounds

Introduction

Interactions among aquatic microorganisms such as symbiosis, parasitism, predation and competition, greatly shape the composition and activity of microbial communities [1–3]. In particular, interactions between heterotrophic bacteria and primary producers (phytoplankton) influence the growth of both organisms [4,5] with consequences for the ecosystem functioning and the biogeochemical cycles [6,7]. For instance, heterotrophic bacteria consume up to 50% of the organic matter released by phytoplankton, significantly affecting the dynamics of the huge pool of dissolved organic carbon in the oceans [8]. Thus, if and how a bacterium can interact with other microorganisms may have important consequences for the biological carbon pump in the current and future oceans [9,10].

Studies using specific model organisms in binary co-cultures have started to elucidate mechanisms underlying marine microbial interactions (mostly between bacteria and phytoplankton) [1,2,5]. Although these results do not reflect the complexity of natural environments and the potential for higher-order effects [11], they allow to identify the chemical signals and resulting changes in gene expression and physiology that underlie these interactions. For example, bacteria associated with phytoplankton (*e.g.* within the phycosphere, [5,12]) gain access to labile organic carbon released by the primary producers, e.g. amino acids and small sulfur-containing compounds [13–18]. In return, phytoplankton benefit from an increased accessibility to nutrients via bacteria-mediated processes, e.g. nitrogen and phosphorus remineralization [19], vitamin supply [14,20] and iron scavenging via formation of siderophores [21,22]. In addition to such metabolic interactions, direct signalling may also occur between bacteria and phytoplankton, with heterotrophic bacteria directly controlling the phytoplankton cell cycle through phytohormones [13,23] or harming it via toxins [18,24]. Through such specific infochemical-mediated interactions, bacteria may also directly affect the release rate of organic carbon from phytoplankton, as well as rates of mortality and aggregation [18,23,25].

While much is known about microbial interactions involving model organisms such as specific strains of *Roseobacter* [13,18–20,24], *Alteromonas* [26–28], Cyanobacteria [19,29] or *Vibrio* [30,31], relatively little is known regarding how widely distributed such interactions are across microbial diversity. The few experimental studies that measure microbial interactions across diversity (e.g. [32–34]) are usually limited in their phylogenetic scope and are performed under conditions different from natural marine environments. However, the knowledge obtained from model organisms on the molecular mechanisms underlying microbial interactions and the increasing availability of high-quality genomes present an opportunity to map known interaction mechanisms to a large set of bacterial species from various taxa. Here, we re-analyse 421 previously published genomes of diverse marine bacteria that represent a significant fraction of marine microbial communities (213 genera), providing an “atlas” of their functional metabolic capacity. The atlas includes also 52 bacteria isolated from extreme marine habitats, humans and plant roots which serve as functional out-groups and represent well known symbiotic bacteria. Several previous studies have aimed to characterize and cluster genomes based on their predicted functional similarity, defined usually using individual genes (e.g. [35–38]) or coarse

functional categories (e.g. COGs, [37,39,40])(Supplementary Table 1, Supplementary text). We chose to take a trait-based approach rather than a gene-based one, which is an intermediate level of resolution between individual genes and coarse functions. Trait-based approaches offer a new perspective to investigating microbial diversity with a more mechanistic understanding [41] but have been used only in few specific cases to highlight putative bacterial interactions (e.g., [42]). We focused on the following traits: 1) the metabolic capacity to synthesize or degrade specific biomolecules and to regulate such processes (KEGG modules); 2) specific functions related to microbial interactions, such as motility, chemotaxis and the ability to produce molecules such as siderophores, toxins and antibiotics. This allows to classify genomes into coherent functional units, some of which recapitulate known bacterial groups with well-defined ecological roles, while others refer to potential new groups. Furthermore, genetic traits can be grouped into linked trait clusters, representing functions that likely evolved together and may be functionally connected (i.e. participating in the same process). Our approach maps the mechanisms of microbial interactions elucidated in model organisms across bacterial diversity, suggests specific groups of bacteria likely to interact in a similar manner, and identifies trait combinations that likely mediate these interactions.

Results and Discussion

Genome functional clusters (GFCs) group genomes with similar ecology

To obtain an overview on the functional capabilities of marine bacteria, we re-annotated a set of 473 high-quality genomes and analysed them using a trait-based workflow, which focuses on the detection of complete genetic traits rather than on the presence of individual genes (Supplementary Fig. 1 and Supplementary text). Genetic traits were represented by metabolic KEGG modules, secondary metabolite pathways, transporters, phytohormones and siderophores production, as well as the degradation of specific sulfur metabolites. Among the identified genetic traits, those known to mediate cell-cell interactions in bacterial model systems (see [4,5,43]) were flagged as “interaction traits” (e.g. production of certain vitamins, vibrioferrin or a specific secretion systems; Supplementary Table 2).

Based on the occurrence patterns of all traits (Fig. 1), we could cluster the genomes into 47 genome functional clusters (GFCs; Supplementary Table 3). In each GFC, genomes encode similar genetic traits, and thus GFCs are expected to be coherent in terms of functional and metabolic capacity, including the ways that the grouped bacteria respond to abiotic cues and interact with other microbes. Previous genome comparison approaches have identified genome clusters that match to ecologically relevant groups (e.g. ecotypes, as defined for *Bacillus pumilus* [36] and *Prochlorococcus* [29]) or lifestyles (e.g. oligotrophic and copiotrophic species [39]). Similarly, in our analysis we found GFCs that represent group of organisms with a defined ecology and life history, such as the Pelagibacterales group (GFC 2), different ecotypes of Cyanobacteria (GFCs 15 and 36), or Vibrio groups, characterized by different host-specificity and pathogenicity (GFCs 25 and 47)(see Supplementary text). Specific GFCs were also identified for each of three groups of Gammaproteobacteria (*Alteromonas*, *Marinobacter* and *Pseudoalteromonas*) which are typically considered as "copiotrophs", often associated with organic particles or phytoplankton [44–47]. A detailed analysis of the traits found in each of the respective GFCs (Supplementary Fig. 2) suggested that *Pseudoalteromonas* and *Alteromonas* bore more genetic traits involved in the resistance against antimicrobial compounds, as well as regulation for osmotic and redox stresses. They also had similar vitamin B1 and siderophore transporters, which are different from those encoded by *Marinobacter*. *Marinobacter* possessed several more transporters for phosphonate and amino acids, as well as specific regulatory systems for adhesion (e.g. alginate and type 4 fimbriae production) and chemotaxis. These patterns advocated that there might be coherent physiological and/or ecological differences between these three groups. Overall, our GFC framework recapitulates previous knowledge on bacterial groups with defined ecology and life history, and provides a way to delineate and characterize potentially new ecological groups.

The correlation between functional and taxonomic classifications still represents an open question and, in marine environments, recent studies have provided both supporting [35,48] and disproving argumentations about the strength of such correlation [37,49]. Therefore, we sought to understand the extent to which our retrieved GFCs overlapped with the genomes' taxonomy. In our analysis, we defined a GFC as taxonomically coherent (i.e. monophyletic) when all grouped genomes belonged to the same taxon and all genomes of

that taxon were grouped in that GFC (see Supplementary Fig. 3 and Supplementary text for more information). Monophyletic GFCs imply that the taxonomic affiliation of these bacteria can predict the traits encoded in their genomes. According to this metric, 23 out of the 47 GFCs were taxonomically coherent, most of them at the genus level including all Firmicutes and half of the Alpha- and Gammaproteobacteria GFCs. Of the remaining 24 GFCs, 5 were polyphyletic (i.e. include genomes of different phyla) and 19 were paraphyletic. Paraphyletic GFCs contained genomes of multiple taxa (differing at genus, family or even order rank) whereby these taxa were also partitioned among multiple GFCs. These GFC comprised genomes of Cyanobacteria, Bacteroidota and the remaining half of the Alpha- and Gammaproteobacteria.

Overall, half of the detected GFCs were monophyletic and support the existence of a strong correlation between taxonomy and functionality in marine bacteria, while the remaining paraphyletic GFCs highlight that, in some cases, the taxonomic partitioning (based on the Genome Database Taxonomy [50]) do not completely reflect the functional differentiation. At the whole-community level, it has been shown that taxonomically distinct communities exhibit similar functional profiles, which led to the suggestion that some bacterial clades have similar genetic capacity, and can replace each other while maintaining unchanged the community functioning [51,52]. The polyphyletic GFCs may group such taxonomically different but functionally similar organisms, and this is supported by examples in GFCs 33 and 41 (grouping thermo- and halotolerant bacteria), or GFC 17 (grouping sulfur-oxidizing and facultative anaerobe bacteria). One of the main biological processes that mediates such functional “homogenisation” is horizontal gene transfer [53,54] and, indeed, most of the paraphyletic GFCs grouped with each other known copiotrophs or particle/host associated bacteria (e.g. Rhodobacteraceae in GFCs 9, 30 and 40, Vibrionaceae in GFCs 25 and 47, Alteromonadaceae in GFCs 21 and 24, Oleiphilaceae in GFC 35, or Halomonadaceae in GFC 43; [13,18,45,55,56]), whose lifestyles favour events of genetic exchange (e.g. [57]). Conversely, another study suggested that this perceived similarity in community function reflects only known metabolic pathways, and it is therefore possible that adding to our analysis also unknown genes might separate these GFCs into monophyletic ones [58]. Although hypothetical genes would be of no use (not informative) in a trait-based approach, we hypothesize that they could com-

plement the horizontal gene transfer hypothesis in explaining the blurred taxonomic profiles of the paraphyletic GFCs.

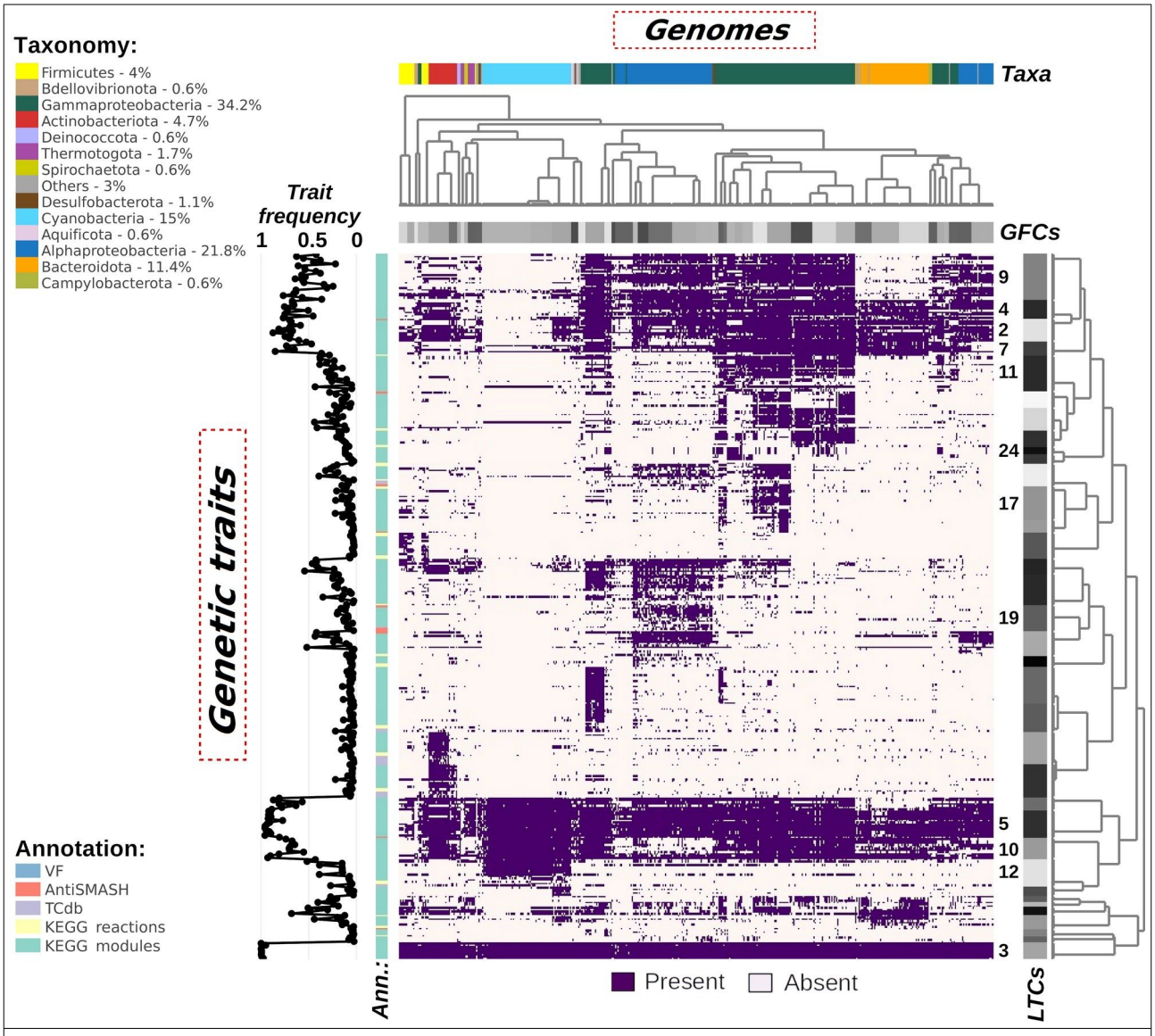

Fig. 1: **Atlas of Marine Microbial Functional Traits** showing presence/absence of genetic traits across all analysed genomes. Columns represent genomes grouped into genome functional clusters (GFCs) as shown by the horizontal grey bar. The horizontal colour bar represents the taxonomic affiliation of genomes (mainly phyla, with the exception of Proteobacteria that are represented at the class level). The number next to each taxon in the legend represents the percent from the total genomes analysed. Rows represent the genetic traits grouped into linked trait clusters (LTCs) as shown by the vertical grey bar. LTCs discussed

in the text are labelled with the related number alongside the grey bar. Left-side row annotations show the frequency of each genetic trait across all tested genomes and the annotation tool (colour coded bar) with which they have been identified. Both dendrograms are computed using the *aggExCluster* function (r package apcluster) that generates hierarchal clustering from an affinity propagation result. An interactive version of this figure is available as Supplementary file 1.

GFCs are ecologically-relevant entities in natural communities

To what extent do the GFCs represent the diversity of natural communities, or display dynamics expected from ecological units in the oceans? To test this, we mapped the 16S rDNA reads from a natural coastal community that was sampled at high temporal frequency [59] to the GFCs (Supplementary text and Supplementary Fig 4). Firstly, a significant fraction of the natural community was represented with high fidelity in the GFCs (mean 22.9% of the 16S rDNA reads, range 12.7-44.3%). Thus, despite the inherent bias derived from using only high-quality, closed genomes available (mainly) from cultured bacteria (legend of Figure 1), the GFCs represented a significant fraction of bacterial diversity. Similar results were obtained with an open-ocean community from the Eastern Mediterranean sampled each season for 3 years [60] (mean 13.9% of the 16S rDNA reads, range 0.5-60.0%), where the GFCs represented a significantly higher fraction of the microbial community on particles >11 µm compared to free-living bacteria (5-0.22 µm; Supplementary Fig. 5). Secondly, using a temporal deconvolution analysis of the coastal site [59] we found that individual 16S phylotypes belonging to the same GFC displayed significantly more synchronous temporal trends (p-value < 0.001) than phylotypes belonging to different GFCs (Supplementary text and Supplementary Fig. 6). Assuming that similar temporal trends suggest similar ecological niches, these results suggest that (at least some of) the GFCs represent bona-fide ecological units.

Specific GFCs are enriched in “interaction traits”

We next focused on selected traits potentially involved in microbial interactions – vitamin exchange, siderophore and phytohormone production and antibiosis – asking whether we could observe patterns in their distribution across the genome dataset. As shown in Figure 2, these traits are not equally distributed among the GFCs – rather, some GFCs were significantly enriched in “interaction traits” (Supplementary Fig. 7a-b). While the number of interaction traits depended to some extent on genome size, as expected from previous studies [61,62], Gammaproteobacteria and several Alphaproteobacteria encoded more interaction traits than expected by their genome size, while Bacteroidota encoded fewer (Supplementary Fig. 7c-d). Overall, GFCs grouping genomes of typical host- or particle-associated bacteria, such as of Alpha- and Gammaproteobacteria [56], are predicted to bear almost the full combination of these traits to sense (chemotaxis, quorum sensing), reach (motility and adhesion) and fight (production/resistance towards antimicrobial compounds, secretion systems) for a targeted hotspot. Contrary, some ubiquitous copiotrophs (e.g. Bacteroidota) [56] and known free-living taxa (e.g. pico-Cyanobacteria and Pelagibacterales) [29,61] possess only a scarce and scattered equipment of such traits. Below, we describe in more detail some of the main observations on the distribution of interaction traits across diversity.

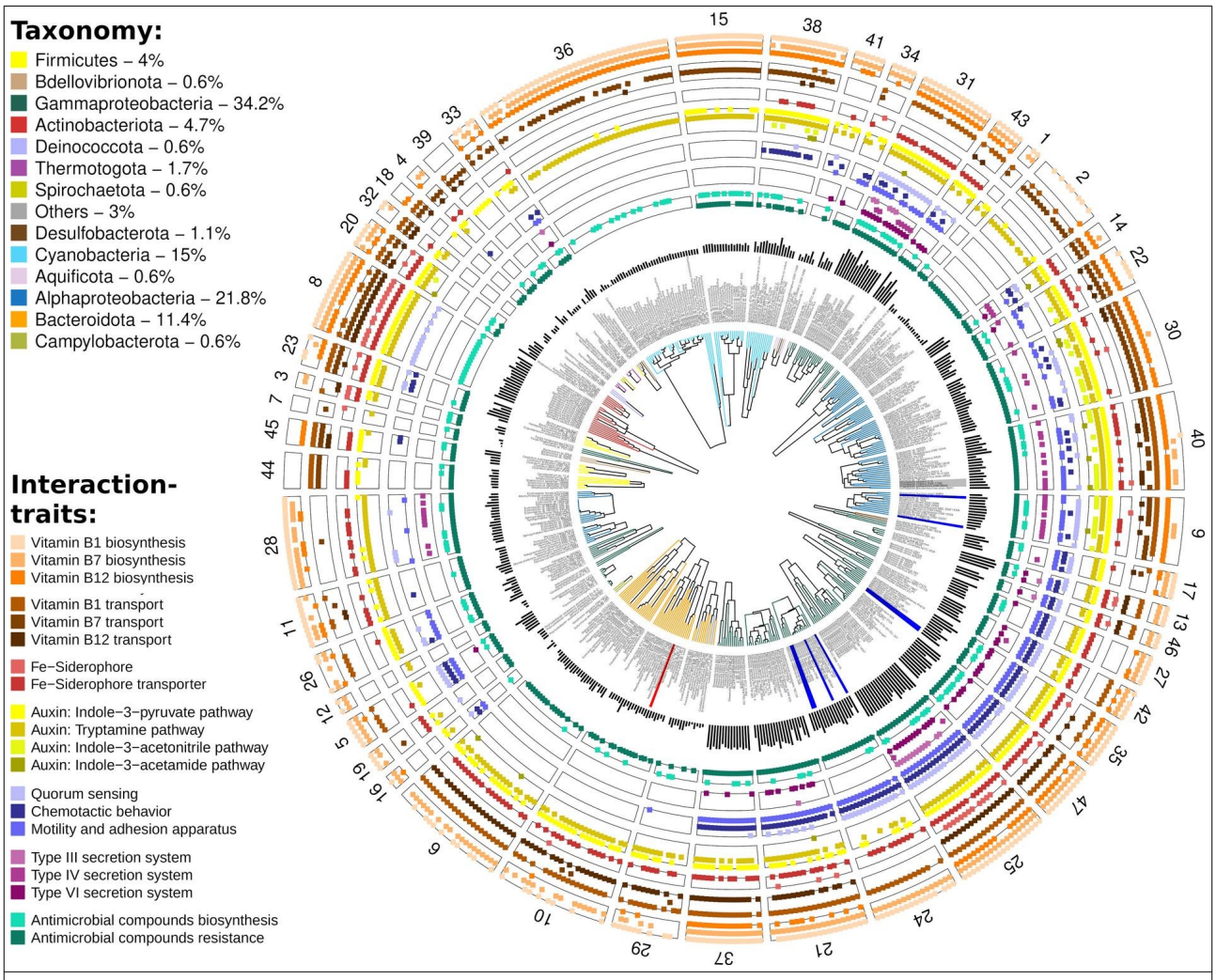

Fig. 2: Distribution across genome functional clusters (GFCs) of interaction traits known to mediate cell-cell interactions in bacterial model systems. Each slice shows the interaction traits present in a GFC and, as dendrogram, the functional similarity of genetic traits between the grouped genomes (hierarchical clustering of r-correlation matrix with complete agglomeration algorithm). The dark bars show the number of interaction traits annotated in each genome. Genomes belonging to model bacteria used to discover some of these traits are highlighted in blue if the interaction was positive (e.g. enhancing phytoplankton growth), in red if it was negative (e.g. kill the host) or in grey if the interaction shifted from positive to negative.

Many bacteria need "to shop" for their vitamins

Vitamins B_1, B_7 and B_{12} are essential cofactors for microbes. Some microorganisms (including abundant phytoplankton) are auxotrophic for these vitamins and need to obtain them from co-occurring bacteria [63,64].

Vitamins are found at low concentrations in aquatic ecosystems [65,66] and their supply can limit biogeochemical cycles, e.g. through limiting primary productivity in the Southern Ocean [67]. Less than half of all the genomes in our dataset were predicted to produce all three vitamins (~39%, including all pico-Cyanobacteria, Actinobacteriota and many Gammaproteobacteria; Fig. 3a). Of the rest, ~29% synthesized at least two B vitamins (e.g. some Alphaproteobacteria, Bacteroidota and the rest of Gammaproteobacteria, which could produce vitamin B_1 and B_7) and ~23% could produce only one type of B vitamin (or ~9% none at all). This suggests that there is a major "market" for B vitamins, and indeed almost all genomes (~83%) encoded transporters for at least one vitamin.

A more detailed analysis of the genomes suggested that marine bacteria could be divided into three main groups based on their predicted strategy for B vitamins acquisition: (1) "consumers", which lacked the biosynthetic genes but harboured the vitamin transporters (we assumed transporters were for uptake, see Supplementary text for details on transporters' directionality); (2) "independents", which encoded the biosynthetic pathways but not the relevant transporters; (3) "flexibles", which encoded both the biosynthetic pathways and transporters for a specific vitamin. Bacteria possessing the latter strategy can potentially switch from being consumers to independent or vice-versa, according to what is more efficient given the surrounding conditions (e.g. availability of extracellular vitamins). The proportion of these three groups changed with the B vitamin studied and the taxonomy of the genomes (Fig. 3b). Very few genomes were flexibles for all three vitamins (~2%), and these were mostly Actinobacteriota (Fig. 3c). There were almost equal proportions of flexibles and independents for vitamin B_1 (Gammaproteobacteria and Cyanobacteria, respectively), whereas the most common bacterial strategies for vitamin B_7 and B_{12} were independent (Fig. 3b). Many different combinations of synthesis and uptake of the three vitamins were represented in the genomes (51 out of 64 possible combinations, Fig. 3c and Supplementary Fig 8a), several being enriched in specific GFCs (Supplementary Fig 9 and Supplementary text) and, notably, most strategies required the exogenous uptake of at least one vitamin. While the perceived lack of biosynthetic capacity could be due to the utilization of precursors or to gaps in the pathway annotations (see Supplementary Fig 10 and Supplementary text), we speculate that this could be a potential manifestation of the "Black Queen" hypothesis, which

stipulates that bacteria “outsource” critical functions to the surrounding community, enabling a reduction of their metabolic cost [68]. In our dataset, the highest fraction of B-vitamin consumers, and hence putative auxotrophs, was observed for vitamin B_1, followed by B_{12} and B_7. This order, however, does not reflect the metabolic costs of producing such vitamins, as B_{12} would be the most “expensive” with about 20 genes involved [69], whereas only 4 genes are required to synthesize B_7 [63] and 5 genes for B_1 [70,71]. Therefore, we hypothesized that vitamin B_1 supplies might be more stable or frequent (e.g. as a result of higher export or lysis of producing bacteria) than that of vitamin B_{12}. Nevertheless, very few organisms were predicted to be auxotrophic for all three vitamins, suggesting that completely relying on exogenous sources for vitamins represents a risky strategy in marine pelagic environments. Taken together, these data provide a comprehensive overview of the market for B vitamins in marine environments by defining specific roles (e.g. consumer, independent, flexible/source) and identifying which bacteria (taxon and GFC) fulfil each role.

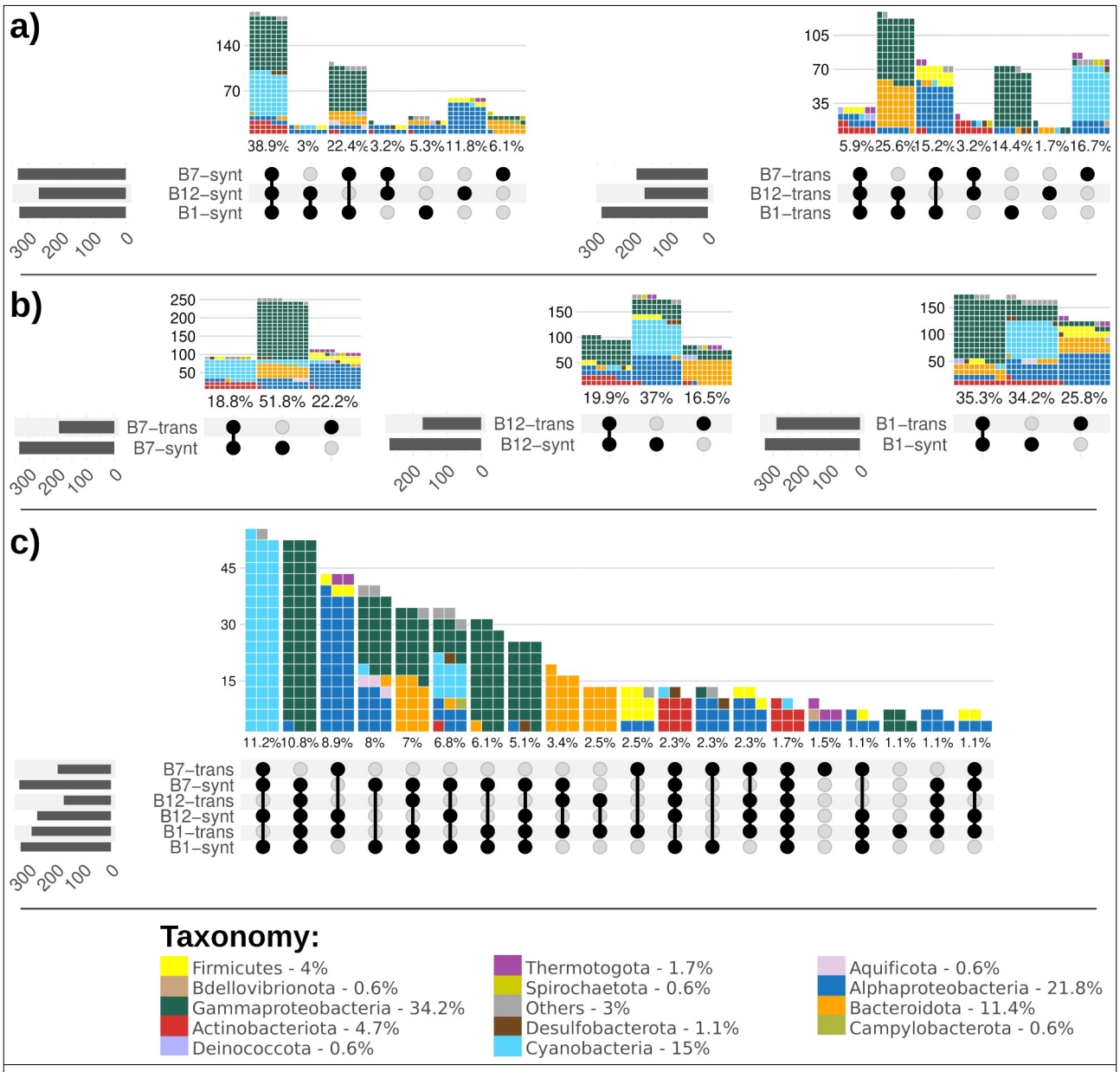

Fig. 3: Plots of intersecting sets showing different genomic configurations of traits for the biosynthesis and transport of vitamins B_1, B_{12}, and B_7 : (a) capabilities to produce or transport these vitamins, (b) different configurations to either produce and/or transport each of these vitamins, and (c) the most abundant configurations to produce and/or transport these vitamins across genomes; the remaining combinations are shown in Supplementary Fig. 8a. Overall, the left bar chart indicates the total number of genomes for each trait, the dark connected dots indicate the different configurations of traits and the waffle bar chart indicates the number (and percentage) of genomes provided with such a configuration; each piece of a waffle bar represents a genome and it is coloured according to the taxon.

Production of siderophores and phytohormones – key mechanisms of synergistic microbial interactions

The production and exchange of "common goods" such as siderophores [72], as well as of specific phytohormones like auxin [73] represent traits that may mediate synergistic microbial interactions. As shown in Fig. 2, approximately 10% of the genomes have the capacity to produce siderophores (mainly Actinobacteriota and Gammaproteobacteria), while almost half of the genomes, from multiple taxa, encoded siderophore transporters (45% of the genomes). The occurrence of siderophores biosynthetic traits was partially following the GFC clustering (e.g. GFC 8, Actinobacteriota and GFC 25, Gammaproteobacteria) and partially scattered across single genomes in different taxa. In contrast, the distribution of the transporters mainly followed the GFC grouping (Fig. 2). Furthermore, microorganisms can utilize siderophore-bound iron also without the need for siderophore transporters, e.g. using ferric reductases located on the plasma membrane [74] or via direct endocytosis [75]. In this regard, 5% of the genomes encoded the capacity to produce vibrioferrin (Supplementary Fig. 11), which is available to a wide range of organisms upon photolysis [21]. Field studies revealed that siderophore biosynthesis is widespread in the ocean [76], and that bacteria producing e.g. vibrioferrin can represent a relevant percentage of the total bacterial communities [77]. Thus, siderophores can be considered as "keystone molecules" (*sensu* [78]), produced by a limited subset of organisms but utilizable by a wide range of bacteria [79–81].

Several recent studies have shown that bacteria can influence the growth of phytoplankton through the production of phytohormones [13,18,24], and indeed the auxin hormone indole-3-acetic acid (IAA) has been identified in natural marine samples [13]. Nearly all genomes (~92%) in our dataset are predicted to produce IAA. Four pathways for the production of IAA were identified, with some organisms encoding more than one pathway. The tryptamine pathway was the most common one and was present in nearly all GFCs comprising genomes of Alphaproteobacteria, Gammaproteobacteria, Cyanobacteria and Actinobacteriota (Fig. 2). The indole-3-pyruvate pathway was the second most common with an almost identical distribution to the tryptamine pathway (missing in GFCs 36 and 15, pico-Cyanobacteria, and 28, Alphaproteobacteria), whereas the last two pathways were rarer (<10% of genomes) and limited to Alphaproteobacteria (indole-3-

acetonitrile) and some genomes of Cyanobacteria and Actinobacteriota (indole-3-acetamide). It is tempting to speculate that the widespread distribution of the capacity to produce IAA, and the diversity of biosynthetic pathways, suggest that many heterotrophic bacteria can directly increase phytoplankton growth through specific signalling (e.g. [13,18]). However, all pathways for IAA production are tightly intertwined with the metabolism of tryptophan, either involved in tryptophan catabolism (to cleave the amino group for nitrogen metabolism) or as a "release valve" to avoid the accumulation of toxic intermediates (e.g. α-keto acid indolepyruvate and indoleacetaldehyde). Additionally, IAA can be catabolized as a carbon source for growth (see [82] and references therein). Given the wide distribution of the tryptamine and indole-3-pyruvate pathways, we hypothesize that they might be more linked to the metabolism/catabolism of tryptophan, whereas the indole-3-acetonitrile and indole-3-acetamide pathways could be responsible for the production of IAA involved in phytoplankton-bacteria interactions. This hypothesis is supported by the presence of the latter traits in GFCs 9 and 40 that group model organisms known to interact through auxin with phytoplankton (Fig. 2) [13,18,24], and by the fact that genes specifically related to these two pathways were found to be up-regulated in one of those studies [13].

Traits underlying potential antagonistic interactions

Experimental measurements of interactions among marine bacteria suggest that antagonism is common (>50% of the tested isolates) [32,34], but in most cases the mechanisms behind antagonistic interactions are unclear. Antimicrobial compounds may underlie many antagonistic interactions in marine environments (e.g. [83–85]), and indeed genes encoding for the production of such compounds were found in several bacteria in our dataset (Supplementary Tables 5 and 6). Interestingly, antimicrobial compounds were predicted to be produced in ~30% of genomes (inner ring in Fig. 2), including also GFCs poor in other interaction traits. The most abundant traits across GFCs were bacteriocin and betalactone production [86,87] (Supplementary Fig. 12). Genes involved in the resistance to antimicrobial compounds were also relatively common (78% of genomes; Fig. 2), however, along with "specific" resistance traits (e.g., efflux pumps specific for antibiotics), we

noticed many KEGG modules that were annotated as “antimicrobial compounds resistance” but were also involved in other cellular functions (e.g. cell division, protein quality control and transport of other compounds; Supplementary Table 5) [88,89]. All GFCs which grouped genomes of Cyanobacteria, Actinobacteriota, and Bacteroidota possessed only these “generic” resistance traits (Supplementary Fig. 12), suggesting that such clades are less efficient in resisting microbial “chemical warfare”. In support of this hypothesis, some Cyanobacteria strains are indeed used as markers for antibiotic contamination because of their sensitivity (e.g., references [90,91]), and Bacteroidota are often inhibited when co-cultured with other bacteria that express antagonistic behaviour [32,34]. Overall, these genome-based predictions are in agreement with previous experimental results [32,34], which suggested that Alpha- and Gammaproteobacteria commonly inhibited other bacteria, whereas Bacteroidota had a low inhibitory capacity and were the most sensitive to inhibition by other bacteria.

Antimicrobial compounds or toxins often need to be delivered into the target organism, e.g. using type IV or type VI secretion systems (T4SS and T6SS, respectively). Approximately 24% of the strains encoded T4SS or T6SS, and these were found primarily in GFCs containing Alpha- and Gammaproteobacteria (Fig. 2). The two secretion systems had different distributions among the GFCs, with only GFC 25 and 31 (comprising *Vibrio* and *Burkholderia*, genera of Gammaproteobacteria) bearing both systems. The T4SS system can perform multiple roles, including conjugation, DNA exchange and toxin delivery in bacteria-bacteria or bacteria-eukaryote interactions [92]. T4SSs were detected more frequently in Alphaproteobacteria (5 out of 8 GFCs). To date, T6SSs are known to be involved only in antagonistic interactions, including among marine bacteria [93], suggesting that the presence of this trait is a high-confidence predictor of the ability to directly inhibit other cells ([94] and references therein). In our dataset, T6SSs occurred almost exclusively in GFCs comprising Gammaproteobacteria, specifically in *Marinobacter* and *Vibrio*, suggesting a strong capacity for contact-mediated antagonistic interactions in these taxa. Type III secretion systems (T3SS), which deliver effector molecules that maintain the bacterial association with the host [95], were found only in a few genomes as the *Vibrio* clustered in GFC 25. This GFC grouped known zooplanktonic hosts [55], suggesting a more specific role for T3SS in metazoan host-microbe interactions.

Linked trait clusters (LTCs) delineate functional connectivity between individual interaction traits

While individual traits may be important in determining the outcome of microbial interactions, such interactions are often highly complex, requiring multiple traits such as motility, signalling and metabolic interactions to operate together (e.g. [13,18,20,23,24]). If these interaction mechanisms are evolutionarily conserved, traits that are functioning together to mediate such interactions should co-evolve, meaning that selection would favour maintaining all relevant traits in the same genome [96,97]. To identify cases of co-evolving traits, we used linkage disequilibrium analysis and clustered traits which were found together more often than expected by chance (adjusted p-value < 0.05) into Linked Trait Clusters (LTCs; Figure 1, Supplementary Table 7 and Supplementary Fig 13). For example, LTC 10 includes pathways for assimilatory sulfate reduction, siroheme and heme biosynthesis, as well as vitamins B_1 and B_7 biosynthesis. These traits appeared together more often than random pairs of traits (mean r within this LTC is 0.38, compared to 0.09 among all traits; Supplementary Fig. 13b) and they are also functionally linked. In fact, siroheme is a prosthetic group for assimilatory sulfite reductases [98,99] and, in sulfate-reducing bacteria, siroheme can be "hijacked" for the biosynthesis of heme [100]. Finally, once reduced, sulphur can be incorporated into essential molecules such as amino acids (methionine and cysteine) and membrane lipids, as well as into vitamins B_1 and B_7 [101].

We focus first on LTCs grouping traits not directly involved in interactions. Two LTCs, 3 and 5 (mean r of 0.30 and 0.35), occurred in nearly all genomes (>93%; Fig. 1) and, as expected, they linked traits which mediate for core metabolic functions, common to any cell. These include biosynthesis of nucleotide (DNA and RNA) and amino acids, as well as core metabolic pathways (glycolysis, pentose phosphate pathway and the first three reactions of the TCA cycle). In contrast, other LTCs (i.e. 2, 4 and 7) highlighted cases in which major metabolic pathways such as the TCA cycle or pathways for the cell wall assembly were missing in specific bacterial clades. These absence-patterns were consistent with previous studies (see Supplementary Text for

detailed description). The LTC concept can, therefore, be used to identify traits that may function together, providing hypotheses as to new modes of interaction that can be tested experimentally.

Analysis of LTCs which include interaction traits such as secretion systems suggested that these are linked also to traits encoding for chemotaxis, motility and adhesion, a typical set of traits a bacterium would need to locate, reach and settle on an organic matter particle or eukaryotic hosts (phytoplankton, zooplankton, fish). Moreover, other genetic traits (e.g. metabolic or regulatory) were linked within these LTCs and they may also be involved in microbial interactions (Fig. 4). For example, LTC 11 (mean r = 0.40) included, in addition to T6SS, traits for adhesion, a flagellar regulatory system, quorum sensing (controlling for swarming and biofilm formation), chemotaxis and a nitrogen transporter with the regulation system. The same LTC also encoded for the biosynthesis of ubiquinone, vitamins B_6 (pyridoxal) and B_7, and for two regulatory systems (*BarA-UvrY*, *RstB-RstA*), which are known to regulate virulence, metabolism, biofilm formation, stress resistance, quorum sensing and secretion systems (see [102,103]). This LTC was common in the GFCs grouping Gammaproteobacteria such as *Pseudoalteromonas* (GFC 21) *Alteromonas* (GFC 24), *Marinobacter* (GFC 35), *Shewanella* (GFC 37) and *Vibrio* (GFCs 25 and 47; Supplementary Tables 3 and 7). All of these organisms are known as particle and phytoplankton associated bacteria (e.g. [44–47,56]), and in such micro-environments they can engage in microbiome interactions using these linked traits (i.e. biosynthesis of B vitamins, quorum sensing and T6SS; [93,104]). Interestingly, the other two secretion systems, T4SS and T3SS, were also linked with the regulation of nitrogen metabolism and with vitamin B_7 or B_{12} transport as part of LTC 25 and LTC 17, respectively. We propose that this provides further support for a putative functional link between microbial interactions (injection of effector molecules), nitrogen and vitamin metabolism.

Notably, LTC 17 (which encodes the T3SS; mean r = 0.47) included also amino acid and sugar transporters, and two regulation systems (*UhpB-UhpA*, *CitA-CitB*). The *UhpB-UhpA* genes control the motility and colonization of fish pathogens [105] and, not surprisingly, the LTC was found in GFCs 46 and 47 which included known pathogenic bacteria (grouping *Aeromonas* and *V. natriegensis* genomes; Supplementary Tables 3 and 7) [106,107]. LTC 17 was also found in GFCs that grouped non-pathogenic but still host associated taxa (GFC 25, which includes *V. alginolyticus* and GFC 13 grouping Enterobacteriaceae) [108], supporting the role of this LTC

(as well as of the T3SS it encodes) in interactions with eukaryotic hosts. Moreover, *CitA-CitB* regulation system controls the citrate metabolism in response to changes of amino acid concentration or pH [109]. We speculate that the link between microbial interaction traits, citrate metabolism and protein and amino acid transporters within LTC 17 may be relevant for a host-associated bacteria during the dispersal stage (lower amino acid concentration and change in pH changes).

Finally, the analysis of LTC 19 (mean r = 0.34) lent further support for the hypothesis that some IAA production pathways are involved in phytoplankton-bacteria interactions and not just in the tryptophan catabolism (see above). This LTC included the indole-3-acetonitrile and indole-3-acetamide pathways, along with other key microbial-phytoplankton interaction traits such as quorum sensing [24,110] and taurine degradation [13,16]. Other linked genetic traits hint to additional molecular mechanisms: the manganese/iron transporter suggests a micronutrient-dependent response, the transport of capsular polysaccharide may be involved in resistance to host defence and pathogenicity [111], and the biosynthesis of putrescine can stimulate phytoplankton growth, productivity and stress tolerance, as shown previously in plants [112]. Moreover, the LTC was found complete in GFCs 9, 30 and 40 (Supplementary Table 7) which grouped genomes of bacteria known to interact via IAA with phytoplankton (e.g. *Dinoroseobacter*, *Sulfitobacter* and *Phaeobacter*) [13,18,24,113]. We propose that possessing this LTC indicates that the relevant bacterium is capable of affecting the growth of phytoplankton through a combination of specific hormone signalling pathways and metabolic interactions.

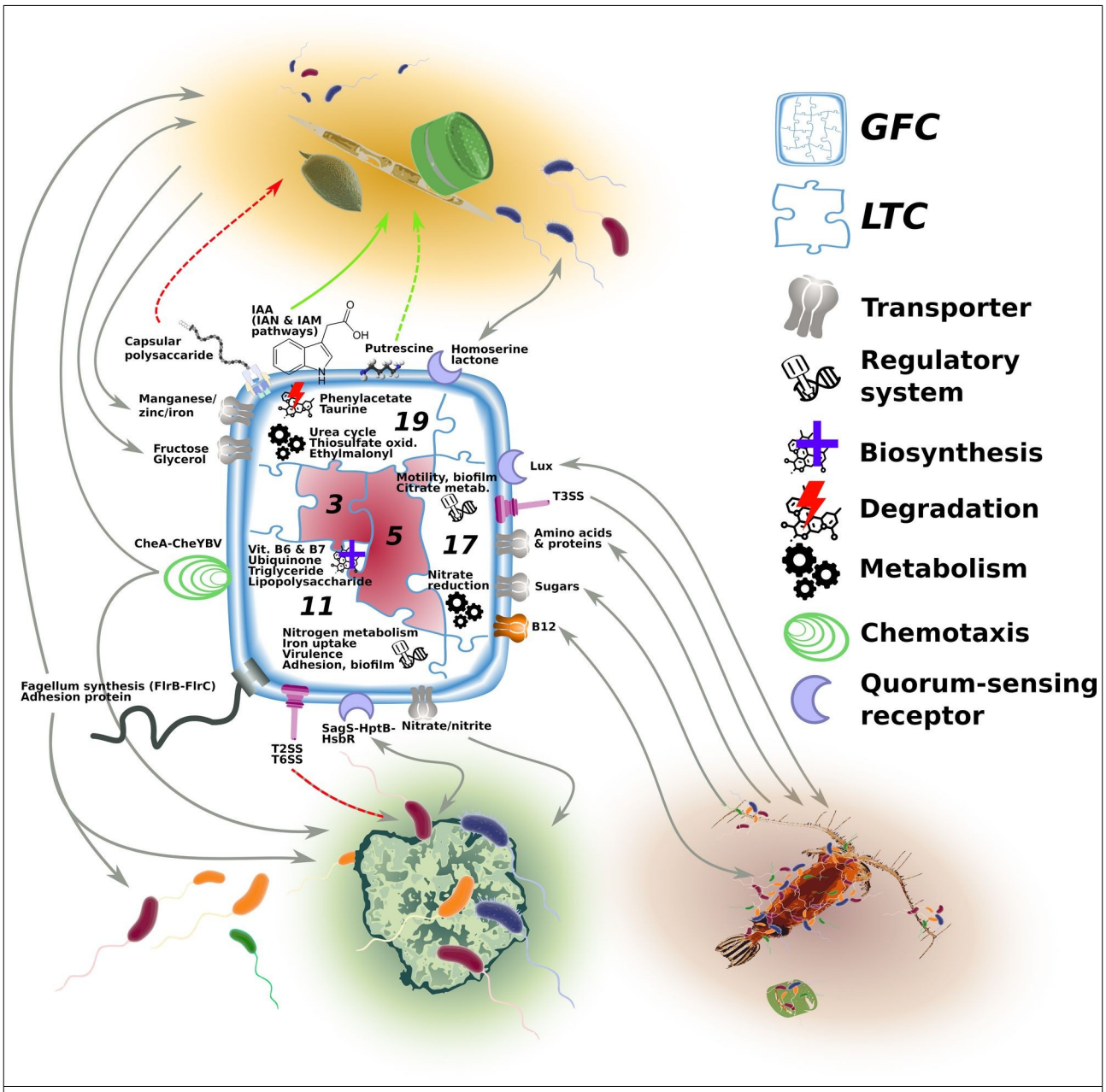

Fig. 4: Conceptual representation of the predicted interaction dynamics for an hypothetical bacterium analysed with our trait-based approach. The bacterium is assigned to a GFC, visualized here as a jigsaw puzzle. Every puzzle piece represents one of the linked trait clusters (LTCs) possessed in that GFC. LTCs 3 and 5 (marked in red) are the core LTCs and they are present in any GFC. In this example, the GFC also possess the ancillary LTCs 11, 17 and 19. Each of these ancillary LTC predicts specific interaction types with bacteria, organic matter particles, phytoplankton or other eukaryotic hosts (e.g. fish, zooplankton). Green arrows indicate traits with positive effects (e.g. enhancing growth), grey arrows traits mediating metabolites/chemical exchange, movement or attachment, and red arrows traits with negative effects (e.g.

pathogenicity). A solid arrow is used when the mediated mechanism has already been described in literature, while a dashed arrow indicates a potential new mechanism.

Summary and outlook

We present a framework that extrapolates from studies of specific model organisms to predict the interaction potential of other bacteria based on the traits encoded in their genomes. By focusing on biologically relevant traits (including specific interaction traits), we reduced a highly complex genomic dataset to a tractable matrix of organisms by functions. By organizing the ensuing genomic information into GFCs, we further simplified the interpretation of complex genomic datasets, while at the same time highlighting the non-trivial grouping of organisms by phenotypic traits, sometimes irrespective of taxonomic boundaries. The LTCs provided evidence for the functional and evolutionary linkage between traits, raising hypotheses as to how these traits act together in the context of complex processes such as microbial interactions. This approach can be easily scaled to different ecosystems (e.g. freshwater or terrestrial) and expanded to include information from additional data sources (e.g. metabolomics or high-throughput functional assays).

Applying this approach to a wide diversity of bacterial taxa, we showed that marine bacteria encode different configurations of interaction traits and can, therefore, engage in different interaction types. Known particle associated taxa of Alpha- and Gammaproteobacteria possessed the full set of traits to interact with particles and living hosts, while Bacteroidota, a known ubiquitous copiotroph taxon, did not have this capacity. Actinobacteriota and Cyanobacteria represented potential sources for B_7 in the B vitamin market, while most Alphaproteobacteria appeared as obligate customers. We suggest that siderophores, and vibrioferrin in particular, are keystone molecules being produced by only a few bacteria (Actinobacteriota and Gammaproteobacteria) but affecting a much larger diversity of potential users, in agreement with the Black Queen Hypothesis [68]. Finally, the production of IAA might be more common than expected, and in some cases (e.g. GFCs 9, 30 and 40 encoding LTC 19) this may be linked with other traits involved in phytoplankton-bacteria interactions.

As with any bioinformatic approach, also our workflow aimed to identifying functional traits on genome sequences, has inherent limitations. Firstly, on average, only 63% of genes were functionally annotated across genomes (range ~40-80%), with the remaining genes either annotated as hypothetical or not annotated at all. This is a strong reminder of the limitations of current genomic and metabolic knowledge. Secondly, only a fraction of the natural microbial diversity was represented by high-quality sequenced genomes, leading to a bias in the representation of different clades in the analysed dataset (Figure 1). While the analysed genomes did represent a significant fraction of natural diversity (Supplementary Fig.s 4 and 5), a future work, both culture dependent and independent, is needed to close these gaps and obtain an unbiased view of the traits encoded by marine microorganisms. Finally, the GFC and LTC concepts are both statistical in nature, representing the probability of bacteria having similar functional capacity, and of traits being functionally and/or evolutionarily linked. In support of the GFC concept, a study of marine Vibrionaceae suggested that ecologically-cohesive unit (similar to GFCs) are likely to interact in a similar way [30]. However, some interaction phenotypes, such as the ability to inhibit multiple target bacteria ("super-killers"), are not phylogenetically conserved [30,114]. Experimental studies using both established and new model systems across multiple scales of diversity (e.g. between and within GFCs) are needed to test the GFC framework, whereas genetic manipulation of linked traits to test their effect on microbial interactions will be required to determine to what extent traits within LTCs are functionally linked. Nevertheless, GFCs and LTCs describe how putative interaction traits vary across microbial diversity, and thus can be used to quantify how the fraction of natural communities potentially "ready to interact" changes over space and time. In turn, this should help to elucidate fundamental rules that govern community dynamics and assembly in the oceans, and the roles played by microbial interactions will in global ecosystem-level processes and biogeochemistry.

Materials and Methods

Genome selection

A dataset of complete and high-quality draft genomes of marine bacteria was compiled performing an extensive research on metadata available from NCBI (http://www.ncbi.nlm.nih.gov/genome), JGI (https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=FindGenomes&page=genomeSearch), and MegX (https://mb3is.megx.net/browse/genomes) websites. Although the focus of the analysis was on bacteria inhabiting the marine pelagic environment, some genomes from organisms isolated in extreme marine environments (i.e. thermal vents, saline and hypersaline environments, estuaries) and sediment, as well as from human and plant symbionts (*Sinorhizobium* and *Mesorhizobium*) were kept for comparison. The final list of 473 genomes includes all of the genomes that, using CheckM [115], were defined "closed" (i.e. each DNA molecule, such as chromosome and plasmids, was represented as a single sequence in the fasta file) or high-quality draft genomes (>90% of completeness, <10% of contaminations, >18 tRNA genes and all 3 rDNAs present) [116]. The final dataset included 473 complete genomes with 117 closed genomes and with >81% genomes that were >99% complete. Of these 473 genomes, 421 were isolated in marine pelagic and coastal zones, 34 in extreme environments (e.g. salt marsh or hydrothermal vent), 6 in marine sediment and, of the remaining, 8 were human associated and 4 plant roots associated (Supplementary Table 3).

Genome annotation

Genome taxonomic classification was obtained using the protein phylogeny workflow implemented in GTDB-tk v1.4.0 [50] with the command *classify_wf* (standard settings), and all retrieved genomes were functionally re-annotated using a standardized pipeline. In brief, gene calling and first raw annotation steps were performed with Prokka v1.14.5 (standard settings and '--rnammer' for rRNA prediction) [117]. The amino acid sequences translated from the identified coding DNA sequences of each genome were annotated against pre-computed hidden Markov model profiles of KEGG Ortholog (KEGG database v94.0) using kofamscan v1.2.0 (standard settings) [118]. Additional targeted analyses were performed to annotate secondary metabolites, phytohormones, specific transporters and utilization of sulfur-metabolites. The genbank files generated

by Prokka were submitted to a local version of Anti-SMASH v5.1.2 (--clusterblast --subclusterblast --knownclusterblast --smcogs --inclusive --borderpredict --full-hmmer --asf –tta), which generated a list of predicted secondary metabolite biosynthesis gene clusters [119]. Different pathways for the biosynthesis of the phytohormone indole-3-acetic acid (IAA) were manually identified by annotating (blastp v2.9.0 best hit, e-value <1e-5 and bit score > 60 [120]) the translated amino acid sequences against the KEGG orthologies required to generate IAA from trypthopan in the KEGG map01070 (Supplementary Table 8; Supplementary file 3). Translated amino acid sequences were also used as input for a GBlast search (BioV suite; standard settings) [121] to identify trans-membrane proteins mediating the transport of B vitamins and siderophores(Supplementary Table 4). For downstream analysis, only annotations with a trans-membrane alpha-helical overlap score >1 and a blast e-value <1e-6 were retained. Production and transport of photoactive siderophores (i.e. vibrioferrin) [21] were identified with functional annotations from the most similar (BLAST best hit) sequences. The predicted protein sequences were blasted (blastp v2.9.0; e-value <1e-5 and amino acid similarity >30%) [122] against a reference dataset assembled using all available sequences of related genes (pvsABCDE and pvuBCDE operons) [123] available in UniProt (Supplementary Table 9 and Supplementary file 4). The same approach was carried out to annotate the dimethylsulfoniopropionate (DMSP) degradation pathways of demetilation and cleavage blasting (e-value <1e-70) against a reference dataset that contained all Uniprot sequences of the genes listed in [124,125] (Supplementary Table 9 and Supplementary file 5). Catabolic pathways of other two sulfur-metabolites, 2,3-dihydroxypropane-1-sulfonate (DHPS) and taurine, were identified checking for the presence of key reactions. The known routes to degrade DHPS share all a sulfopropanediol 3-dehydrogenase (hpsN gene; K15509) [14], while taurine can enter the TCA cycle either via taurine-pyruvate aminotransferase (tpa gene; K03851) or via taurine dehydrogenase (tauXY genes; K07255+K07256), followed by a sulfoacetaldehyde acetyltransferase (xsc gene; K03852) [126].

KEGG module reconstruction

KEGG orthologies (**KO**s) annotations generated by kofamscan were recombined in KEGG modules (**KM**s) using an in-house R script. The KMs represent minimal functional units describing either pathways, structural

complexes (e.g. transmembrane pump or ribosome), functional sets (essential sets as Aminoacyl-tRNA synthases or nucleotide sugar biosynthesis) or signaling modules (phenotypic markers as pathogenicity). Briefly, using the R implementation of KEGG REST API [127], the script fetches the diagrams of all KMs from the KEGG website. Each diagram represents a reactions' scheme of a KM listing all known KOs that can perform each of the reactions necessary to complete that scheme (Supplementary Fig. 1b). The completeness of a KM in a genome was assessed as the number of required reactions for which at least one KOs was annotated and only complete KMs were retained in downstream analyses (e.g., a KM with 7 out of 8 reactions is incomplete and would be discarded). However, to partially compensate for possible annotation issues, one missing reaction was allowed in KMs with ≥3 reactions (i.e., a KM with 7 out of 8 annotated reactions is considered complete; see Supplementary text and Supplementary Table 10 for more details).

Genetic and interaction traits identification

The annotated complete KMs, secondary metabolites, phytohormones and transporters represent the genetic traits identified in the genomes (556 genetic traits). From this list, the subset of interaction traits was manually extracted based on current knowledge about processes that likely play a role in microbial interactions (list of picked interaction traits in Supplementary Table 2). Within the KMs we identified traits related to vitamin biosynthetic pathways, quorum sensing, chemotaxis, antimicrobial resistance, motility and adhesion (Supplementary Table 5). Since the ecological role of most secondary metabolites is still unclear, a careful literature search was performed to identify and retain only the secondary metabolite clusters with a proposed function which can be linked to microbial interaction processes, such as siderophore production, quorum sensing and antimicrobial compound biosynthesis (Supplementary Table 6). The phytohormone annotations revealed the capability of producing indoleacetic acid (auxin) through four different pathways (Supplementary Table 8). Vitamin and siderophore transporters were identified in the transporter annotations looking for the related transporter families (e.g. TonB, Btu) and the substrate information (Supplementary Table 4).

Mapping to environmental datasets

The selected datasets represent amplicon time series generated with Illumina sequencing of the V4 region of the 16S rDNA of bacterial communities sampled in a coastal (between 1-5m of depth, Canoe Cove, Nahant, MA, USA) [59] and a pelagic (between 10-500m of depth, n-1200 station, Easter Mediterranean sea) [60] site. Mapping between the full 16S rDNAs extracted from genomes and amplicon sequences was performed with BLAST (blastn v2.9.0, e-value <1e-5) using different identity threshold to filter the best hits: 100%, 97%, 94.5% and 86.5%. The first two values represent a new proposed threshold and the most commonly used threshold to define operational taxonomic units (OTUs) [128], while the last two values were suggested as threshold to classify OTUs at genus and family levels [129]. In addition, for each identity threshold we inspected the top blast hits (up to 20) and calculate a mapping specificity index as the number of hits assigned to a same GFC over the total number of hits (Supplementary Fig.s 4a and 5a). To avoid spurious mapping, we only retained in downstream analyses sequences with a specificity index = 1 (i.e. all best blast hits belonged to the same GFC).

The evaluation of the GFC concept was performed by using the coastal time series, because it offers a high (daily) temporal resolution of changes in the bacterial community composition and the authors applied a time deconvolution analysis to characterize the OTU temporal dynamics [59]. The OTUs included in the deconvolution analysis represented ~97% of the total sequences in the dataset and for each pair of OTUs, the authors calculated the frequency interaction score (the higher the score, the more synchronous were the temporal dynamics of both OTUs, and vice-versa). From the list of all pairs, we only considered OTU pairs for which at least one of the two OTUs mapped to a GFC. We wanted to exclude cases of OTUs pairs that could belong to a same, but yet unknown, GFC. To test for the GFC concept, we compared the frequency interaction score between OTU pairs mapped to the same OTU (i.e. both OTUs mapped to the same GFC) versus OTU pairs mapped to different GFCs (e.g. the two OTUs mapped to different GFCs, or one of the OTU was unmapped). For each identity threshold, the normal distribution of the frequency interaction score was assessed with the Shapiro test (r's function *shapiro.test*; p.values << 0.001), and t-test (r's function t.test) was performed to test for a significant difference in the mean ranks between the two groups of OTU pairs. We

repeated the test by randomly assigning OTU pairs to the “same GFC” and “different GFC” groups by keeping the same group sizes and by creating two groups of equal size.

Statistical analyses

The presence/absence matrix of genetic traits across genomes served as basis to cluster the former into linked trait clusters (LTCs) and the latter into genome functional clusters (GFCs). Both clustering approaches implemented the Pearson coefficient *r* (also known as phi coefficient when applied to dichotmous variables) to calculate the genome and trait correlation matrices. While no pair of genomes scored a negative r value (as they all shared some “core” functional traits), negative correlations between pairs of genetic traits were threshold to zeros to ensure that the trait clustering was only driven by positive correlation. Moreover, only pairwise correlations with a FDR-corrected p-value < 0.05 (chi-square test, df = 1) were retained. The parsed correlation matrices were fed as similarity matrices into the affinity propagation algorithm implemented with the *apcluster* function (q=0.5; r package apcluster) [130]. This machine learning algorithm was chosen because it does not require the number of clusters to be determined *a priori*, allowing instead this feature to emerge from the data [131]. Briefly, a functional similarity matrix is used to construct a network where nodes and edges are known to be genomes (or genetic traits) and their pairwise *r* correlation, respectively. Starting from a random set of exemplar nodes, clusters are created by expansion towards the adjoining, most similar, nodes. Through iterations of this procedure, the algorithm tries to maximize the total similarity between nodes within each cluster, eventually converging towards the best set of clusters. Clustering robustness and accuracy of both GFCs and LTCs were tested by performing a sensitivity analysis of the ‘q’ parameter which controls the clustering sensitivity in the *apcluster* function, and by down-sampling overrepresented taxa at 80%, 60% and 40% of their genome coverage (see Supplementary information and Supplementary Fig. 15).

For the LTC delineation, genetic traits were pre-filtered to remove noisy signals and only traits found in ≥3% of the genomes (≥14 genomes) were used in the clustering (i.e. 379 genetic traits out of 578 in total). Similar to the context of linkage disequilibrium [132], *r* indicates non-random association between genetic traits

because those traits are interactively linked to fitness, or simply because they are closely located on the chromosome (i.e. lower chances of recombination). However, as the genetic traits analysed commonly involve multiple genes, the second possibility is less likely. While exploring the functional potential, a LTC was considered complete in a genome when >50% of the grouped genetic traits were present and it was considered complete in a GFC when the average completeness of the included genomes was >50%. The genetic traits belonging to LTCs that were never found to be complete in at least one GFC were considered as unclustered.

All analyses were performed in R 4.0.4 [133].

Authors' contributions

LZ, DS and H-PG designed the research. LZ and DS analyzed the data. TM contributed with analytic tools. LZ wrote the paper with substantial revisions from DS, DSegrè and H-PG. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and material

All genomes are available in the online repositories of NCBI and JGI under the accession codes listed in Supplementary Table 3. The codes for functional annotation, reconstruction of KEGG modules and statistical analysis are available on GitHub (https://github.com/lucaz88/genome_comparison_code).

Acknowledgements

This work was supported by the Human Frontier Science Program (HFSP) through the grant number RGB 0020/2016. We thank Anna Godhe, Mats Töpel and Oskar Johansson for providing early access to some of the genomes included in the analysis. We thank the entire IAMM team (Dikla Aharonovich, David Bernstein,

Falk Eigemann, Elena Forchielli, Tal Luzzatto-Knaan, Shany Ofaim, Melissa Osborne, Solvig Pinnow, Dalit Roth-Rosenberg, Angela Vogts and Maren Voss) for the constructive discussions and ideas that led us to shape this manuscript. A sincere thank you to Silvia Munari for her diligent proofreading of this manuscript.

References

1. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: Surviving and thriving in the microbial jungle. *Nature Reviews Microbiology* vol. 8 15–25 (2010).

2. Zimmerman, A. E. *et al.* Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat. Rev. Microbiol.* **18**, 21–34 (2020).

3. Bjorbækmo, M. F. M., Evenstad, A., Røsæg, L. L., Krabberød, A. K. & Logares, R. The planktonic protist interactome: where do we stand after a century of research? *ISME J.* **14**, 544–559 (2020).

4. Amin, S. A., Parker, M. S. & Armbrust, E. V. Interactions between diatoms and bacteria. *Microbiol. Mol. Biol. Rev.* **76**, 667–84 (2012).

5. Seymour, J. R., Amin, S. A., Raina, J.-B. & Stocker, R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nat. Microbiol.* **2**, 17065 (2017).

6. Farooq Azam and Francesca Malfatti. Microbial structuring of marine ecosystems. *Nat. Rev. Microbiol.* **5**, 782–791 (2007).

7. Zoccarato, L. & Grossart, H.-P. Relationship Between Lifestyle and Structure of Bacterial Communities and Their Functionality in Aquatic Systems. in *Advances in Environmental Microbiology - The Structure and Function of Aquatic Microbial Communities* (ed. Hurst, C. J.) 13–52 (Springer Nature Switzerland AG 2019, 2019). doi:10.1007/978-3-030-16775-2_2.

8. Kirchman, D. L. *Processes in Microbial Ecology*. (Oxford University Press, 2012). doi:10.1093/acprof:oso/9780199586936.001.0001.

9. Worden, A. Z. *et al.* Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science (80-.).* **347**, 1257594 (2015).

10. Gibert, J. P. Temperature directly and indirectly influences food web structure. *Sci. Rep.* **9**, 5312 (2019).

11. Levine, J. M., Bascompte, J., Adler, P. B. & Allesina, S. Beyond pairwise mechanisms of species coexistence in complex communities. *Nature* **546**, 56–64 (2017).

12. Bell, W. & Mitchell, R. Chemotactic and Growth Responses of Marine Bacteria to Algal Extracellular Products. *Biol. Bull.* **143**, 265–277 (1972).

13. Amin, S. A. *et al.* Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**, 98–101 (2015).

14. Durham, B. P. *et al.* Cryptic carbon and sulfur cycling between surface ocean plankton. *Proc. Natl. Acad. Sci.* **112**, 453–457 (2015).

15. Durham, B. P. *et al.* Sulfonate-based networks between eukaryotic phytoplankton and heterotrophic bacteria in the surface ocean. *Nat. Microbiol.* **4**, 1706–1715 (2019).

16. Moran, M. A. & Durham, B. P. Sulfur metabolites in the pelagic ocean. *Nature Reviews Microbiology* vol. 17 665–678 (2019).

17. Paul, C., Mausz, M. A. & Pohnert, G. A co-culturing/metabolomics approach to investigate chemically mediated interactions of planktonic organisms reveals influence of bacteria on diatom metabolism. *Metabolomics* **9**, 349–359 (2013).

18. Segev, E. *et al.* Dynamic metabolic exchange governs a marine algal-bacterial interaction. *Elife* **5**, e17473 (2016).

19. Christie-Oleza, J. A., Sousoni, D., Lloyd, M., Armengaud, J. & Scanlan, D. J. Nutrient recycling facilitates long-term stability of marine microbial phototroph–heterotroph interactions. *Nat. Microbiol.* **2**, 17100 (2017).

20. Wang, H., Tomasch, J., Jarek, M. & Wagner-Döbler, I. A dual-species co-cultivation system to study the interactions between Roseobacters and dinoflagellates. *Front. Microbiol.* **5**, 311 (2014).

21. Amin, S. A. *et al.* Photolysis of iron-siderophore chelates promotes bacterial-algal mutualism. *Proc. Natl. Acad. Sci.* **106**, 17071–17076 (2009).

22. Keshtacher-Liebso, E., Hadar, Y. & Chen, Y. Oligotrophic Bacteria Enhance Algal Growth under Iron-Deficient Conditions. *Appl. Environ. Microbiol.* **61**, 2439–2441 (1995).

23. van Tol, H. M., Amin, S. A. & Armbrust, E. V. Ubiquitous marine bacterium inhibits diatom cell division. *ISME J.* 1–12 (2016) doi:10.1038/ismej.2016.112.

24. Seyedsayamdost, M. R., Case, R. J., Kolter, R. & Clardy, J. The Jekyll-and-Hyde chemistry of Phaeobacter gallaeciensis. *Nat. Chem.* **3**, 331–335 (2011).

25. Grossart, H.-P. *et al.* Interactions between marine snow and heterotrophic bacteria: Aggregate formation and microbial dynamics. *Aquat. Microb. Ecol.* **42**, 19–26 (2006).

26. Aharonovich, D. & Sher, D. Transcriptional response of Prochlorococcus to co-culture with a marine Alteromonas: differences between strains and the involvement of putative infochemicals. *ISME J.* 1–15 (2016) doi:10.1038/ismej.2016.70.

27. Coe, A. *et al.* Survival of Prochlorococcus in extended darkness. *Limnol. Oceanogr.* **61**, 1375–1388 (2016).

28. Hou, S. *et al.* Benefit from decline: The primary transcriptome of *Alteromonas macleodii str. Te101* during Trichodesmium demise. *ISME J.* **12**, 981–996 (2018).

29. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. Prochlorococcus: the structure and function of collective diversity. *Nat. Rev. Microbiol.* **13**, 13–27 (2015).

30. Cordero, O. X. *et al.* Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science (80-.).* **336**, 48–51 (2012).

31. Tai, V., Paulsen, I. T., Phillippy, K., Johnson, D. A. & Palenik, B. Whole-genome microarray analyses of Synechococcus-Vibrio interactions. *Environ. Microbiol.* **11**, 2698–2709 (2009).

32. Long, R. A. & Azam, F. Antagonistic interactions among marine bacteria. *Appl. Environ. Microbiol.* **67**, 4875–4983 (2001).

33. Sher, D., Thompson, J. W., Kashtan, N., Croal, L. & Chisholm, S. W. Response of Prochlorococcus ecotypes to co-culture with diverse marine bacteria. *ISME J.* **5**, 1125–1132 (2011).

34. Grossart, H.-P., Schlingloff, A., Bernhard, M., Simon, M. & Brinkhoff, T. Antagonistic activity of bacteria isolated from organic aggregates of the German Wadden Sea. *FEMS Microbiol. Ecol.* **47**, 387–396 (2004).

35. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108–110 (1999).

36. Espariz, M., Zuljan, F. A., Esteban, L. & Magni, C. Taxonomic identity resolution of highly phylogenetically related strains and selection of phylogenetic markers by using genome-scale methods: The bacillus pumilus group case. *PLoS One* **11**, 1–17 (2016).

37. Hernández-González, I. L., Moreno-Hagelsieb, G. & Olmedo-Álvarez, G. Environmentally-driven gene content convergence and the Bacillus phylogeny. *BMC Evol. Biol.* **18**, 1–15 (2018).

38. Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**, 1623–1635 (2019).

39. Lauro, F. M. *et al.* The genomic basis of trophic strategy in marine bacteria. *Proc. Natl. Acad. Sci.* **106**, 15527–33 (2009).

40. Newton, R. J. *et al.* Genome characteristics of a generalist marine bacterial lineage. *ISME J.* **4**, 784–798 (2010).

41. Krause, S. *et al.* Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Frontiers in Microbiology* vol. 5 251 (2014).

42. Bordron, P. *et al.* Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. *Microbiologyopen* **5**, 106–117 (2016).

43. Raina, J.-B., Fernandez, V., Lambert, B., Stocker, R. & Seymour, J. R. The role of microbial motility and chemotaxis in symbiosis. *Nat. Rev. Microbiol.* (2019) doi:10.1038/s41579-019-0182-9.

44. Holmström, C. & Kjelleberg, S. Marine Pseudoalteromonas species are associated with higher organisms and produce biologically active extracellular agents. *FEMS Microbiol. Ecol.* **30**, 285–293 (1999).

45. Eilers, H., Pernthaler, J., Glöckner, F. O. & Amann, R. Culturability and in situ abundance of pelagic Bacteria from the North Sea. *Appl. Environ. Microbiol.* **66**, 3044–3051 (2000).

46. Lupette, J. *et al.* Marinobacter Dominates the Bacterial Community of the Ostreococcus tauri Phycosphere in Culture. *Front. Microbiol.* **7**, 1–14 (2016).

47. Sonnenschein, E. C., Syit, D. A., Grossart, H.-P. & Ullrich, M. S. Chemotaxis of Marinobacter adhaerens and its impact on attachmentto the diatom Thalassiosira weissflogii. *Appl. Environ. Microbiol.* **78**, 6900–6907 (2012).

48. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

49. Maistrenko, O. M. *et al.* Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.* **14**, 1247–1259 (2020).

50. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2020).

51. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

52. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science (80-.).* **353**, 1272–1277 (2016).

53. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive earth's biogeochemical cycles. *Science (80-.).* **320**, 1034–1039 (2008).

54. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: A phylogenetic perspective. *Science (80-.).* **350**, aac9323 (2015).

55. Hunt, D. E. *et al.* Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science (80-.).* **320**, (2008).

56. Teeling, H. *et al.* Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *Elife* **5**, (2016).

57. Arias-Andres, M., Klümper, U., Rojas-Jimenez, K. & Grossart, H. P. Microplastic pollution increases gene exchange in aquatic ecosystems. *Environ. Pollut.* **237**, 253–261 (2018).

58. Galand, P. E., Pereira, O., Hochart, C., Auguet, J. C. & Debroas, D. A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *ISME J.* **12**, 2470–2478 (2018).

59. Martin-Platero, A. M. *et al.* High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat. Commun.* **9**, 1–11 (2018).

60. Rosenberg, D. R. *et al.* Particle-associated and free-living bacterial communities in an oligotrophic sea are affected by different environmental and anthropogenic factors. *bioRxiv* 2020.04.20.051391 (2020) doi:10.1101/2020.04.20.051391.

61. Grote, J. *et al.* Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio* **3**, e00252-12 (2012).

62. Shih, P. M. *et al.* Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci.* **110**, 1053–1058 (2013).

63. Croft, M. T., Warren, M. J. & Smith, A. G. Algae Need Their Vitamins. *Eukaryot. Cell* **5**, 1175–1183 (2006).

64. Romine, M. F., Rodionov, D. A., Maezato, Y., Osterman, A. L. & Nelson, W. C. Underlying mechanisms for syntrophic metabolism of essential enzyme cofactors in microbial communities. *ISME J.* **11**, 1434–1446 (2017).

65. Suffridge, C. P. *et al.* B Vitamins and Their Congeners as Potential Drivers of Microbial Community Composition in an Oligotrophic Marine Ecosystem. *J. Geophys. Res. Biogeosciences* **123**, 2890–2907 (2018).

66. Sañudo-Wilhelmy, S. A. *et al.* Multiple B-vitamin depletion in large areas of the coastal ocean. *Proc. Natl. Acad. Sci.* **109**, 14041–14045 (2012).

67. Browning, T. J. *et al.* Nutrient co-limitation at the boundary of an oceanic gyre. *Nature* **551**, 242–246 (2017).

68. Morris, J. *et al.* The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *MBio* **3**, e00036 (2012).

69. Fang, H., Kang, J. & Zhang, D. Microbial production of vitamin B12: A review and future perspectives. *Microb. Cell Fact.* **16**, 1–14 (2017).

70. Jurgenson, C. T., Begley, T. P. & Ealick, S. E. The Structural and Biochemical Foundations of Thiamin Biosynthesis. *Annu. Rev. Biochem.* **78**, 569–603 (2009).

71. McRose, D. *et al.* Alternatives to vitamin B1 uptake revealed with discovery of riboswitches in multiple marine eukaryotic lineages. *ISME J.* **8**, 2517–2529 (2014).

72. Vraspir, J. M. & Butler, A. Chemistry of Marine Ligands and Siderophores. *Ann. Rev. Mar. Sci.* **1**, 43–63 (2009).

73. De Smet, I. *et al.* Unraveling the evolution of auxin signaling. *Plant Physiol.* **155**, 209–21 (2011).

74. Morrissey, J. & Bowler, C. Iron utilization in marine cyanobacteria and eukaryotic algae. *Front. Microbiol.* **3**, 43 (2012).

75. Kazamia, E. *et al.* Endocytosis-mediated siderophore uptake as a strategy for Fe acquisition in diatoms. *Sci. Adv.* **4**, eaar4536 (2018).

76. Gärdes, A. *et al.* Detection of photoactive siderophore biosynthetic genes in the marine environment. *BioMetals* **26**, 507–516 (2013).

77. Yarimizu, K. *et al.* Distribution of dissolved iron and bacteria producing the photoactive siderophore, vibrioferrin, in waters off Southern California and Northern Baja. *BioMetals* **32**, 139–154 (2019).

78. Madhusoodanan, J. Can single molecules bind together entire ecosystems? *Proc. Natl. Acad. Sci. U. S. A.* **113**, 10451–10452 (2016).

79. Coale, T. H. *et al.* Reduction-dependent siderophore assimilation in a model pennate diatom. *Proc. Natl. Acad. Sci.* **116**, 23609–23617 (2019).

80. Basu, S., Gledhill, M., de Beer, D., Prabhu Matondkar, S. G. & Shaked, Y. Colonies of marine cyanobacteria Trichodesmium interact with associated bacteria to acquire iron from dust. *Commun. Biol.* **2**, 1–8 (2019).

81. Kramer, J., Özkaya, Ö. & Kümmerli, R. Bacterial siderophores in community and host interactions. *Nature Reviews Microbiology* vol. 18 152–163 (2020).

82. Patten, C. L., Blakney, A. J. C. & Coulson, T. J. D. Activity, distribution and function of indole-3-acetic acid biosynthetic pathways in bacteria. *Crit. Rev. Microbiol.* **39**, 395–415 (2012).

83. Long, R. A. *et al.* Antagonistic interactions among marine bacteria impede the proliferation of Vibrio cholerae. *Appl. Environ. Microbiol.* **71**, 8531–8536 (2005).

84. Long, R. A., Qureshi, A., Faulkner, D. J. & Azam, F. 2-n-pentyl-4-quinolinol produced by a marine Alteromonas sp. and its potential ecological and biogeochemical roles. *Appl. Environ. Microbiol.* **69**, 568–576 (2003).

85. Paz-Yepes, J., Brahamsha, B. & Palenik, B. Role of a Microcin-C-like biosynthetic gene cluster in allelopathic interactions in marine Synechococcus. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12030–12035 (2013).

86. Robinson, S. L., Christenson, J. K. & Wackett, L. P. Biosynthesis and chemical diversity of β-lactone natural products. *Nat. Prod. Rep.* **36**, 458–475 (2019).

87. Cotter, P. D., Ross, R. P. & Hill, C. Bacteriocins-a viable alternative to antibiotics? *Nature Reviews Microbiology* vol. 11 95–105 (2013).

88. Guo, L. *et al.* Lipid A Acylation and Bacterial Resistance against Vertebrate Antimicrobial Peptides systemic illnesses termed enteric fevers, which are char-acterized by microorganism colonization of the intes-tine, followed by systemic spread to tissues rich in. *Cell* **95**, 189–198 (1998).

89. Hinz, A., Lee, S., Jacoby, K. & Manoil, C. Membrane proteases and aminoglycoside antibiotic resistance. *J. Bacteriol.* **193**, 4790–4797 (2011).

90. Yasser, E.-N. & Adli, A. Toxicity of Single and Mixtures of Antibiotics to Cyanobacteria. *J. Environ. Anal. Toxicol.* **05**, 274 (2014).

91. EMEA. *European Medicines Agency. Doc ref. EMEA/CHMP/SWP/4447/00. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/10/WC500003978.pdf*. http://www.emea.eu.int (2006).

92. Grohmann, E., Christie, P. J., Waksman, G. & Backert, S. Type IV secretion in Gram-negative and Gram-positive bacteria. *Molecular Microbiology* vol. 107 455–471 (2018).

93. Guillemette, R., Ushijima, B., Jalan, M., Häse, C. C. & Azam, F. Insight into the resilience and susceptibility of marine bacteria to T6SS attack by Vibrio cholerae and Vibrio coralliilyticus. *PLoS One* **15**, 1–19 (2020).

94. Cianfanelli, F. R., Monlezun, L. & Coulthurst, S. J. Aim, Load, Fire: The Type VI Secretion System, a Bacterial Nanoweapon. *Trends in Microbiology* vol. 24 51–62 (2016).

95. Macho, A. P. & Zipfel, C. Targeting of plant pattern recognition receptor-triggered immunity by bacterial type-III secretion system effectors. *Current Opinion in Microbiology* vol. 23 14–22 (2015).

96. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 4285–4288 (1999).

97. Liu, C., Wright, B., Allen-Vercoe, E., Gu, H. & Beiko, R. Phylogenetic Clustering of Genes Reveals Shared Evolutionary Trajectories and Putative Gene Functions. *Genome Biol. Evol.* **10**, 2255–2265 (2018).

98. Siegel, L. M., Murphy, M. J. & Kamin, H. Reduced nicotinamide adenine dinucleotide phosphate-sulfite reductase of enterobacteria. I. The Escherichia coli hemoflavoprotein: molecular parameters and prosthetic groups. *J. Biol. Chem.* **248**, 251–264 (1973).

99. Scott, A. I., Irwin, A. J., Siegel, L. M. & Shoolery, J. N. Sirohydrochlorin. Prosthetic Group of a Sulfite Reductase Enzyme and Its Role in the Biosynthesis of Vitamin B12. *Journal of the American Chemical Society* vol. 100 316–318 (1978).

100. Bali, S. *et al.* Molecular hijacking of siroheme for the synthesis of heme and d1 heme. *Proc. Natl. Acad. Sci.* **108**, 18260–18265 (2011).

101. Cook, A. M., Smits, T. H. M. & Denger, K. Sulfonates and Organotrophic Sulfite Metabolism. in *Microbial Sulfur Metabolism* 170–183 (Springer Berlin Heidelberg, 2008). doi:10.1007/978-3-540-72682-1_14.

102. Zere, T. R. *et al.* Genomic targets and features of BarA-UvrY (-SirA) signal transduction systems. *PLoS One* **10**, e0145035 (2015).

103. Huang, L., Xu, W., Su, Y., Zhao, L. & Yan, Q. Regulatory role of the RstB-RstA system in adhesion, biofilm production, motility, and hemolysis. *Microbiologyopen* **7**, e00599 (2018).

104. Rolland, J. L., Stien, D., Sanchez-Ferandin, S. & Lami, R. Quorum Sensing and Quorum Quenching in the Phycosphere of Phytoplankton: a Case of Chemical Interactions in Ecology. *J. Chem. Ecol.* 1–11 (2016) doi:10.1007/s10886-016-0791-y.

105. Chen, J. *et al.* The UhpA mutant of Edwardsiella piscicida enhanced its motility and the colonization in the intestine of tilapia. *Fish Shellfish Immunol.* **104**, 587–591 (2020).

106. Janda, J. M. & Abbott, S. L. The genus Aeromonas: Taxonomy, pathogenicity, and infection. *Clin. Microbiol. Rev.* **23**, 35–73 (2010).

107. Darshanee Ruwandeepika, H. A. *et al.* Pathogenesis, virulence factors and virulence regulation of vibrios belonging to the Harveyi clade. *Rev. Aquac.* **4**, 59–74 (2012).

108. Murthy, K. N., Mohanraju, R., Karthick, P. & Ramesh, C. Phenotypic and molecular characterization of epiphytic vibrios from the marine macro algae of Andaman Islands, India. *Indian J. Geo-Marine Sci.* **45**, 304–309 (2016).

109. van Mastrigt, O., Mager, E. E., Jamin, C., Abee, T. & Smid, E. J. Citrate, low pH and amino acid limitation induce citrate utilization in Lactococcus lactis biovar diacetylactis. *Microb. Biotechnol.* **11**, 369–380 (2018).

110. Wang, H. *et al.* Identification of Genetic Modules Mediating the Jekyll and Hyde Interaction of Dinoroseobacter shibae with the Dinoflagellate Prorocentrum minimum. *Front. Microbiol.* **6**, 1262 (2015).

111. Frosch, M., Edwards, U., Bousset, K., Krauße, B. & Weisgerber, C. Evidence for a common molecular origin of the capsule gene loci in Gram-negative bacteria expressing group II capsular polysaccharides. *Mol. Microbiol.* **5**, 1251–1263 (1991).

112. Chen, D., Shao, Q., Yin, L., Younis, A. & Zheng, B. Polyamine function in plants: Metabolism, regulation on development, and roles in abiotic stress responses. *Frontiers in Plant Science* vol. 9 1945 (2019).

113. Wienhausen, G., Noriega-Ortega, B. E., Niggemann, J., Dittmar, T. & Simon, M. The exometabolome of two model strains of the Roseobacter group: A marketplace of microbial metabolites. *Front. Microbiol.* **8**, 1–15 (2017).

114. Russel, J., Røder, H. L., Madsen, J. S., Burmølle, M. & Sørensen, S. J. Antagonism correlates with metabolic similarity in diverse bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10684–10688 (2017).

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

116. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

117. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

118. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

119. Blin, K. *et al.* AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).

120. Zhang, P. *et al.* The distribution of tryptophan-dependent indole-3-acetic acid synthesis pathways in bacteria unraveled by large-scale genomic analysis. *Molecules* **24**, 1–14 (2019).

121. Reddy, V. S. & Saier, M. H. BioV Suite - A collection of programs for the study of transport protein evolution. *FEBS J.* **279**, 2036–2046 (2012).

122. Amin, S. A., Green, D. H., Waheeb, D. Al, Gärdes, A. & Carrano, C. J. Iron transport in the genus Marinobacter. *BioMetals* **25**, 135–147 (2012).

123. Tanabe, T. *et al.* Identification and Characterization of Genes Required for Biosynthesis and Transport of the Siderophore Vibrioferrin in Vibrio parahaemolyticus. *J. Bacteriol.* **185**, 6938–6949 (2003).

124. Moran, M. A., Reisch, C. R., Kiene, R. P. & Whitman, W. B. Genomic Insights into Bacterial DMSP Transformations. *Ann. Rev. Mar. Sci.* **4**, 523–542 (2012).

125. Bullock, H. A., Luo, H. & Whitman, W. B. Evolution of dimethylsulfoniopropionate metabolism in marine phytoplankton and bacteria. *Front. Microbiol.* **8**, 637 (2017).

126. Cook, A. M. & Denger, K. Metabolism of taurine in microorganisms: a primer in molecular biodiversity? in *Advances in Experimental Medicine and Biology* (ed. Oja, S. S.) 3–13 (Springer New York, 2006). doi:10.1007/978-0-387-33504-9_1.

127. Tenenbaum, D. KEGGREST: Client-side REST access to KEGG. R package version 1.29.0. (2020).

128. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).

129. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).

130. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. Apcluster: an R package for affinity propagation clustering. *Bioinformatics* **27**, 2463–2464 (2011).

131. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science (80-.).* **315**, 972–976 (2007).

132. Kimura, M. Theoretical foundation of population genetics at the molecular level. *Theor. Popul. Biol.* **2**, 174–208 (1971).

133. R Core Team. R: A Language and Environment for Statistical Computing. (2021).