Comparative whole-genome approach to identify traits underlying 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

Classification
BIOLOGICAL SCIENCES - Microbiology

Keywords
marine bacteria, functional clusters, functional traits, B vitamins, siderophore, phytohormones, motility and chemotaxis, secretion systems, antimicrobial compounds

Author Contributions
L.Z., D.S. and H.-P.G. designed research; L.Z. and D.S. analyzed data; T.M. contributed analytic tools; L.Z., D.S., D.Segrè and H.-P.G. wrote the paper.

ABSTRACT

Interactions among microorganisms affect the structure and function of microbial communities, potentially affecting ecosystem health and biogeochemical cycles. The functional traits mediating microbial interactions are known for several model organisms, but the prevalence of these traits across microbial diversity is unknown. We developed a new genomic approach to systematically explore the occurrence of metabolic functions and specific interaction traits, and applied it to 473 sequenced genomes from marine bacteria. The bacteria could be divided into coherent genome functional clusters (GFCs), some of which are consistent with known bacterial ecotypes (e.g. within pico-Cyanobacteria and *Vibrio* taxa) while others suggest new ecological units (e.g. *Marinobacter*, *Alteromonas* and *Pseudoalteromonas*). Some traits important for microbial interactions, such as the production of and resistance towards antimicrobial compounds and the production of phytohormones, are widely distributed among the GFCs. Other traits, such as the production of siderophores and secretion systems, as well as the production and export of specific B vitamins, are less common. Linked Trait Clusters (LTCs) include traits that may have evolved together, for example chemotaxis, motility and adhesion are linked with regulatory systems involved in virulence and biofilm formation. Our results highlight specific GFCs, such as those comprising Alpha- and Gammaproteobacteria, as particularly poised to interact both synergistically and antagonistically with co-occurring bacteria and phytoplankton. Similar efficient processing of multidimensional microbial information will be increasingly essential for translating genomes into ecosystem understanding across biomes, and identifying the fundamental rules that govern community dynamics and assembly.

Introduction

Interactions between microorganisms, such as symbiosis, competition and allelopathy, are a central feature of microbial communities (1). In aquatic environments, heterotrophic bacteria interact with microbial primary producers (phytoplankton) in many ways, potentially affecting the growth of both organisms (2, 3) with consequences for ecosystem functioning and biogeochemical cycles (4, 5). 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 (6). Thus, if and how a bacterium can interact with other bacteria and eukaryotes may have important consequences for the biological carbon pump in the current and future oceans (7, 8).

Recent studies, using specific model organisms in binary co-cultures, have started to elucidate mechanisms underlying marine microbial interactions (mostly between bacteria and phytoplankton). Many of these interactions are mediated by the exchange of metabolites used for growth or respiration. For example, bacteria associated with phytoplankton (*i.e.* within the phycosphere, (3, 9)), gain access to labile organic carbon released by the primary producers, e.g. amino acids and small sulfur-containing compounds (10–15). In return, phytoplankton benefit from an increased accessibility to nutrients via bacteria-mediated processes, e.g. nitrogen and phosphorus remineralization (16), vitamin supply (11, 17) and iron scavenging via formation of siderophores (18, 19). In addition to such metabolic interactions, direct signaling may also occur between bacteria and phytoplankton, with heterotrophic bacteria directly controlling the phytoplankton cell cycle through phytohormones (10, 20) or harming it using toxins (15, 21). Through such specific infochemical-mediated interactions, bacteria may also directly affect the rate of release of organic carbon from phytoplankton, as well as rates of mortality and aggregation (15, 20, 22).

While much is known about microbial interactions involving model organisms such as specific strains of *Roseobacter* (10, 15–17, 21), *Alteromonas* (23–25), Cyanobacteria (26) or *Vibrio* (27, 28), little is known about the potential for such interactions to occur in other species or microbial lineages. The few experimental studies that measure microbial interactions across diversity (e.g. (29–31)) are usually limited in their phylogenetic scope and are performed under conditions which are very different from those occurring in the natural marine environment. 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-analyze the previously published genomes of 421 diverse marine bacteria, providing an “atlas” of their functional metabolic capacity. The atlas includes also 52 bacteria isolated from extreme marine habitats, human and plant roots that are meant to serve as functional out-groups and represent well known symbiotic plant bacteria (i.e. Rhizobacteria). In particular, we focus on genomic traits likely to be involved in mediating interactions between heterotrophic bacteria and other organisms. These traits are estimated based on the presence of KEGG modules or of genes encoding for transporters, phytohormones and secondary metabolite production. Trait-based approaches offer a new perspective to investigating microbial diversity with a more mechanistic understanding (32) and have been used in some specific cases to highlight putative bacterial interactions (e.g., (33)). As shown below, our results identify clusters of organisms whose genomes encode similar functional capacity (defined as genome functional clusters: GFC). We propose that organisms belonging to the same GFC are likely to interact in similar ways with other microorganisms. We also identify clusters of traits that are statistically linked, and propose that these linked trait clusters (LTCs) may have evolved to function together, potentially during microbial interactions. GFCs and LTCs provide a framework to extend the knowledge on microbial interactions gained from specific model systems, leading to testable hypotheses as to the prevalence of microbial interactions across bacterial diversity.

Results & Discussion

Genome functional clusters, a framework to capture potential new ecotypes

To obtain an overview on the functional capabilities of marine bacteria, we re-annotated a set of 473 high-quality genomes, and analyzed 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, Supplementary information). Our analysis was based on metabolic KEGG modules, with the addition of specific traits which encode for mechanisms known to mediate interactions (Supplementary Table 1). These “interaction traits” include the production of certain secondary metabolites, secretion systems, vitamins and vitamin transporters, siderophores and phytohormones. As shown in Fig. 1, some traits were found across almost all genomes (“core” traits). These include basic cellular metabolisms (nucleotides, amino acids and carbohydrates), a few transport systems and cofactor biosynthesis. Other traits were

found only in specific groups of organisms. Based on the patterns observed in Fig. 1, we could cluster the genomes into 48 genome functional clusters (GFCs; see Supplementary Information for a more detailed description of the clustering method). Within each GFC, all genomes are inferred to encode similar traits, and thus these GFCs are expected to be coherent in terms of their functional and metabolic capacity, including the ways in which the related organisms interact with other microbes.

We next sought to understand the extent to which the GFCs correlate with phylogeny. If all bacteria in a GFC belong to the same phylogenetic clade, and all bacteria in this phylogenetic clade are found together in the same GFC, this would imply that the phylogenetic affiliation of these bacteria can predict the traits encoded in their genomes. We defined such GFCs as *phylogenetically coherent* and measured the coherence at multiple taxonomic ranks (i.e. genus, family, order, class or phylum; see Supplementary information and Supplementary Fig. 2a for more information). According to this metric, 22 out of the 49 GFCs were coherent (monophyletic), most of them at the genus level (Supplementary Fig. 2b,d,e). In these coherent GFCs, which include all Firmicutes and half of the Alpha- and Gammaproteobacteria GFCs (Supplementary Fig. 2b,c), the functional capacity (genome-encoded traits) seems to follow phylogeny closely. Of the remaining 29 GFCs, 22 were paraphyletic, i.e. GFCs contained genomes from multiple phylogenetic groups, and some phylogenetic groups were partitioned among multiple GFCs (Supplementary Fig. 2b-e). The paraphyletic GFCs scored their highest phylogenetic coherence at the class level (the remaining half of the Alpha- and Gammaproteobacteria GFCs) and Genus level (all Cyanobacteria GFCs). In these GFCs a significant functional similarity (e.g. functional redundancy) existed between distantly related genomes. Finally, five GFCs were polyphyletic, i.e. they include organisms from multiple phyla. Some of these GFCs group organisms isolated from extreme environments (e.g. thermal vents or hyper saline environments; GFCs 12 and 44) and marine sediment (GFC 35). Such genomes were added in the analysis as outer groups and, therefore, their phylogenetic diversity was not adequately covered (Supplementary Table 2 and Supplementary information). Overall, the observation that functionality does not strictly follow phylogeny makes it difficult to infer the function of a community from its taxonomic structure, as previously shown across the global oceans (34, 35).

Importantly, some of the GFCs correspond to previously defined ecotypes or to ecologically-defined species. For example, GFC 3 comprised all genomes of the order SAR11 (Pelagibacterales) (Supplementary Table 2), defining a group of highly abundant taxa with streamlined genomes adapted to thrive under oligotrophic

conditions (36, 37). Similarly, pico-Cyanobacterial genomes clustered separately from all other Cyanobacteria, and, within the pico-Cyanobacterial group, GFC 29 consisted of exclusively high-light *Prochlorococcus* strains. GFC 28 comprised mostly low-light *Prochlorococcus* and GFC 30 comprised some low-light type IV *Prochlorococcus* and *Synechococcus* strains. Thus, while none of the Cyanobacterial GFCs were strictly coherent, they were consistent to a large extent with previous ecological and genomic studies (reviewed by (26)). Finally, genomes belonging to the family Vibrionaceae were clustered in three different GFCs (47, 48 and 49). GFC 48 mostly grouped *Vibrio alginolyticus*, which have been shown to prefer zooplankton hosts over other organic matter particles (38), whereas GFC 47 harbored free-leaving and non-pathogenic strains of *V. furnissii* and *V. natriegens* (39, 40). GFC 49 included several pathogenic strains of more generalist *Vibrio* species characterized by a wide range of aquatic hosts (e.g. *V. splendidus*; (41)), as well as a few human pathogens (*V. cholerae* and *V. vulnificus;* (39)). Along with *Vibrio* genomes, GFC 49 contained also genomes from additional taxa (e.g., *Photobacterium* (3 strains) and *Psychromonas* (2 strains)) which are also potential pathogens or gut endobionts of crustacean and marine snails (42, 43).

The correspondence between the aforementioned GFCs and known bacterial ecotypes highlights that our analytical framework is able to categorize bacterial diversity into ecologically relevant units. Therefore, it is interesting that three groups of Gammaproteobacteria, i.e. *Alteromonas*, *Marinobacter* and *Pseudoalteromonas*, each formed their own GFC (4, 21 and 31, respectively). These organisms are all known as copiotrophs often associated with organic particles or phytoplankton (44–47). However, these GFCs can be distinguished by the presence of different metabolic and interaction traits (see following chapters for the discussion on related biological implications; Supplementary Fig. 3). For example, *Pseudoalteromonas* and *Alteromonas* bear more traits involved in the resistance against antimicrobial compounds (especially *Pseudoalteromonas*), as well as regulation for osmotic and redox stresses. They also have similar vitamin B1 and siderophore transporters, which are different from those encoded by *Marinobacter*. *Marinobacter* possess several more phosphonate and amino acid transporters, as well as specific regulatory systems for adhesion (e.g. alginate and type 4 fimbriae production) and chemotaxis. These differences suggest that these taxa form functionally coherent and separate ecological units that differ in multiple ways, including by their interaction capabilities with other microorganisms. We note, however, that the GFCs do not resolve all bacterial diversity,

e.g. that found within the *Alteromonas* (48) or within the high-light *Prochlorococcus (26)* clades.

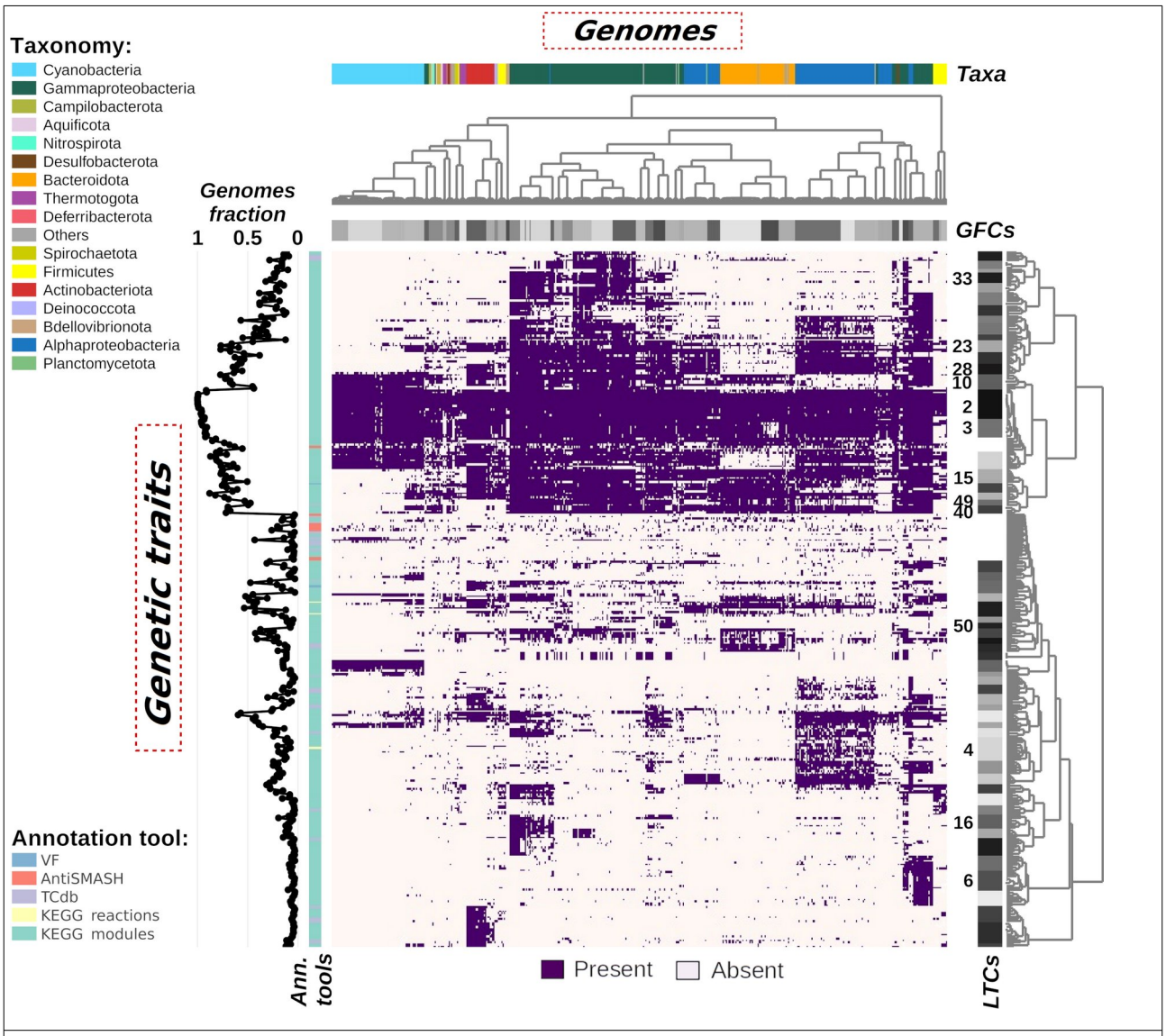


Fig. 1: ***Atlas of Marine Microbial Functional Traits*** *showing presence/absence of genetic traits across all analyzed genomes. Each column represents a genome and these are hierarchically clustered. The horizontal color bar represents the taxonomic affiliations of the genomes (mainly phyla, with the exception of Proteobacteria that are represented at the class level) and the horizontal grey bar delineates specific genome functional clusters (GFCs). Rows are the genetic traits clustered using the coefficient of disequilibrium into linked trait clusters (LTCs) (vertical grey bar); relevant LTCs are marked. Left-side row annotations show the average abundance for the genetic traits and the annotation tool (color coded bar) with which they have been annotated. An interactive version of this figure is available as Supplementary*

dataset 1.

Missing common trait clusters highlight basic metabolic differences

Just as the genomes (represented by the columns in Fig. 1) could be grouped into GFCs, the genetic traits that drive this clustering could themselves be grouped into linked trait clusters (LTCs). The traits within each LTC were found together in the genomes more often than expected by chance, and thus may be linked functionally (Supplementary Fig. 4). For example, LTC 10 includes pathways for assimilatory sulfate reduction, siroheme and heme biosynthesis, and biotin biosynthesis. The sulfate reduction and siroheme pathways are functionally linked, as siroheme is a prosthetic group for assimilatory sulfite reductases (49, 50). While heme and siroheme are different molecules and participate in different pathways, siroheme can be "hijacked" for the production of heme in sulfate-reducing bacteria (51). Finally, once reduced, sulfur can be incorporated into essential molecules such as amino acids (methionine and cysteine), membrane lipids and the B vitamins thiamine and biotin (52). The pathways for the biosynthesis of biotin and its precursor (pimeloyl-ACP) are indeed among the traits encoded in LTC 10. Moreover, every pair of traits in this LTC is much more likely to co-occur within a genome than random pairs of traits (the mean r^2 within this LTC is 0.48, compared with 0.08 among all traits and 0.03 among traits not clustered into any LTC; Supplementary Fig. 4b). Taken together, these results support an evolutionary relationship among the traits included in LTC 10. Therefore, the LTC concept may be useful to identify traits that are likely to be functionally connected and may have evolved together.

Based on their occurrence across genomes (Supplementary Fig. 4c), we divided the LTCs into three subgroups: “core” (present in >90% of genomes), “common” (<90% and ≥30%) and “ancillary” (≤30%). LTCs 2 and 3 (mean r^2 of 0.90 and 0.88) were the most commonly found across genomes (>93%; Fig. 1) and represent the core LTCs. They included KEGG modules involved in the core metabolism (glycolysis, pentose phosphate pathway and the first three reactions of the TCA cycle), as well as pathways responsible for nucleotide, amino acid and cofactor metabolism. In addition, LTC 2 also included an F-type ATPase, cofactor biosynthetic pathways (Coenzyme A and FAD) and a few transporters, e.g. ABC type II (for a variety of small ions and macromolecules)

and Tat protein exporters (see Supplementary Table 3 for complete list of all LTCs and their respective genetic traits).

Other common LTCs (15, 23, 28 and 40), which were found in 61-70% of the genomes, also included traits involved in core metabolisms, e.g. parts of the TCA cycle in the LTC 15 (mean r^2 = 0.64). The lack of the full TCA cycle in some organisms, such as pico-Cyanobacteria, is consistent with previous studies (Supplementary text and Supplementary Fig. 5) (53). Moreover, our results show that some heterotrophic bacteria also lack part of the TCA cycle (e.g. Spirochaetota, Thermotogota and Firmicutes, corresponding to GFCs 40, 42, 43, 8, 9 and 41), confirming and broadening previous studies (54–57). However, other important cellular functions were also found in LTCs not present in all genomes, for example genes involved in cell wall assembly (LTC 40 and 49, mean r^2 of 0.71 and 0.60) and variants of RNA degradosome and polymerase (LTCs 23 and 28, mean r^2 of 0.45 and 0.70; more details in the supplementary information). Additionally, as the linkage of the traits inside each LTC is not complete, the presence or absence of a LTC is not a fully robust indication for the presence or absence of each trait. Finally, a large fraction of the genes in each genome was annotated as hypothetical or not annotated at all (range 22-66%), a strong reminder of the limitations of current genomic and physiological knowledge. We therefore interpreted the patterns of LTCs associated with microbial interactions, keeping in mind that these represent bioinformatics predictions requiring experimental validation.

Many bacteria need “to shop” for their vitamins

Vitamins B1, B7 and B12 are among the most essential cofactors for microbes (58, 59). They are metabolically expensive to produce as they often need several enzymes (e.g. about 20 for vitamin B12) (60), and their availability in the dissolved extracellular pools is extremely limited across all aquatic ecosystems (61, 62). Some phytoplankton are known to require exogenous vitamins from co-occurring heterotrophic bacteria, and indeed vitamin B12 may be a co-limiting micro-nutrient for primary productivity, e.g. in the Southern Ocean (63). We thus analyzed the presence and absence of the pathways for the production of these vitamins, and of their transport across membranes, among our set of marine genomes. Less than half of all genomes are predicted to produce all these vitamins (~45%, including all pico-Cyanobacteria - GFC 27, 28 and 40 - and many Gammaproteobacteria; Fig. 2b). Of the rest, ~33%

synthesize at least two B vitamins (e.g. Alphaproteobacteria, which produce mainly vitamin B1 and B12, and the rest of Gammaproteobacteria, which produce vitamin B1 and B7) and ~15% can produce only one type of B vitamin (or ~7% none at all). This suggests that there is a major “market” for B vitamins, and indeed almost all genomes encoded transporters for at least one vitamin.

A more detailed analysis of the genomes suggests that marine bacteria can be divided into three main groups based on their predicted strategy for B vitamins acquisition: (1) “consumers”, which lack the biosynthetic genes but harbor the vitamin transporters; (2) “independents”, which encode the biosynthetic pathways but not the relevant transporters; (3) “flexibles”, which encode both the biosynthetic pathways and transporters for a specific vitamin (Fig. 2a). 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 changes with the B vitamin studied and the taxonomy of the genomes. Very few genomes were “flexibles” for all three vitamins (~4%), and these were mostly Actinobacteriota (Fig. 3a). In contrast, the most common strategy for vitamin B1 was the “flexible” (just over half of the genomes), and almost half of these were Gammaproteobacteria (Fig. 2a). There were almost equal proportions of flexible, consumers and independents for vitamin B12, whereas the most common bacterial strategy for vitamin B7 was independent (Fig. 2a). Genomes with flexible strategy for vitamin B1 and B12 were quite common (52% and 32% of total genomes) and for several of them it was possible to speculate on their role as ‘providers’ for such vitamins. Indeed, ~20% of the genomes bearing both biosynthetic pathways and transporters for these two vitamins encoded a bidirectional transporter (Supplementary Table 4) that could allow to export and “share” the vitamin.

Notably, there were 64 possible combinations of traits for the synthesis and uptake of the three vitamins, and we could identify genomes corresponding to 45 of these possible combinations (Fig. 2c and Supplementary Fig. 6a). This plasticity is also reflected in the clustering of the related vitamin traits together with different metabolic traits in several different LTCs (Supplementary Table 3). Taken together, these results suggest a wide diversity of strategies to obtain vitamins, most of which require an exogenous uptake for a number of vitamins. We speculate that this represents a manifestation of the “Black Queen” hypothesis, where bacteria can “outsource” critical functions to the surrounding community, enabling a reduction of the metabolic cost (64). Several of these strategies were specifically associated with

one taxon and related GFCs (Supplementary text and Supplementary Fig. 6b). For example, all Cyanobacteria could produce all three vitamins, whereas their genomes encoded only vitamin B7 transporters, suggesting that if this group is a source (i.e. provider) of vitamin B1 and B12, these vitamins become available to the rest of the community through cell death rather than metabolic exchange between living cells.

In our dataset, the highest fraction of B vitamins consumers, and hence putative auxotrophs, was observed for vitamin B12, followed by B7 and B1. This order may be related to the metabolic costs of producing each vitamin. About 20 genes are required for *de-novo* aerobic production of B12 (60), whereas only 4 genes are required to synthesize B7 (58) and 5 genes for B1 (65, 66). Notably, very few organisms were predicted to be auxotrophic for all three vitamins, suggesting that completely relying on exogenous sources for vitamins represent a risky strategy of difficult implementation in marine pelagic environments. Some of the putatively auxotrophic organisms actually encode parts of their vitamin biosynthesis pathways, and therefore may depend on the uptake of a precursor rather than of the vitamin itself (Supplementary text, Supplementary Fig.s 7a-d). Our results, together with experimental observations of vitamin limitation in laboratory cultures and in nature (58, 61, 67), and of shifts in the capacity of marine communities to produce vitamins (68), argue for an important role of vitamins or their precursors and their exchange between organisms in shaping marine microbial communities.

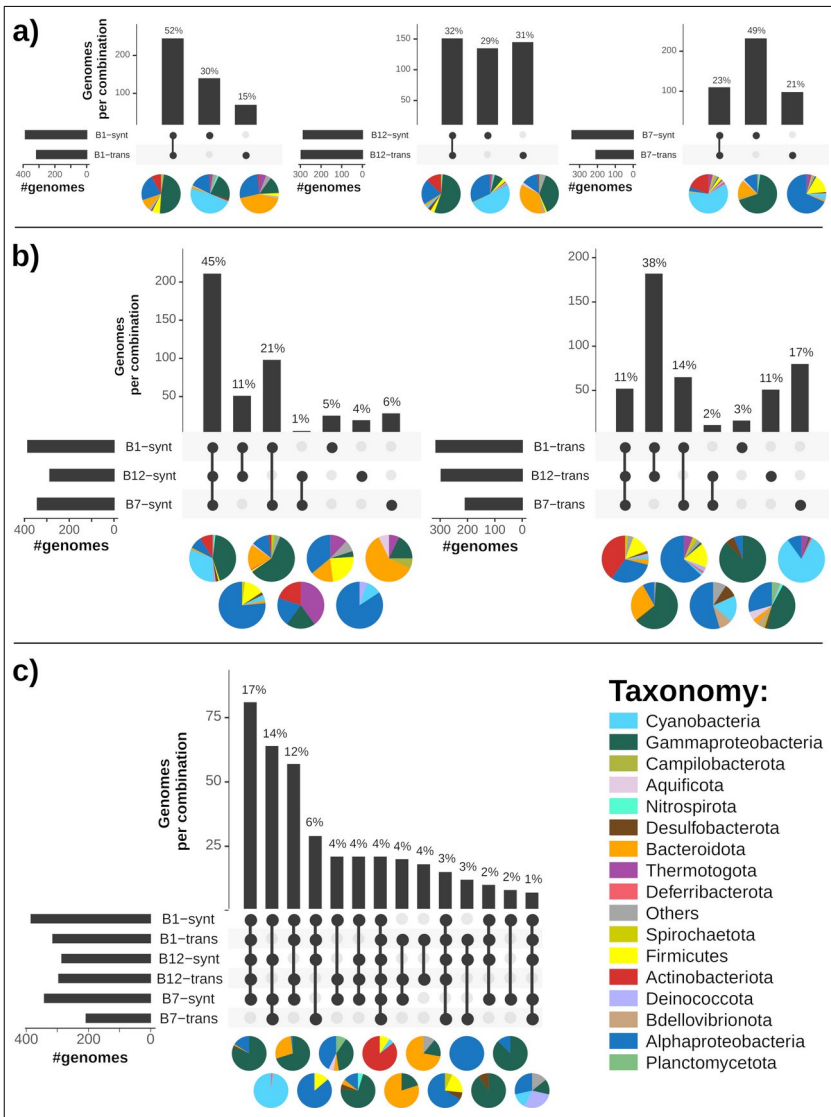


Fig. 2: UpSet plots exploring the different genomic configurations of traits involved in the biosynthesis and transport of vitamins B1, B12, and B7. (a) Different strategy of acquisition of B vitamins among genomes in relation to their ability to produce or/and transport a certain vitamin. (b) Genome partitioning in relation to either production or transport of all selected B vitamins. (c) Most abundant combinations of all these traits across genomes; the remaining combinations are shown in Supplementary Fig. 6a. 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 upper bar chart indicates the number of genomes provided with such configuration. Pie charts show the relative abundance of the different taxa represented in each configuration.

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

We next focused on the traits that may mediate synergistic microbial interactions through the production and exchange of “common goods” such as siderophores (69), as well as of specific phytohormones like auxin (70). Siderophores are organic molecules that bind iron, increasing its solubility and bioavailability (69). Siderophore production by heterotrophic bacteria can stimulate the growth of phytoplankton, providing a potential trading good for synergistic interactions (18). However, high affinity uptake of siderophores can also serve as a mechanism for competition for iron (71, 72). As shown in Fig. 3, approximately 32% of the genomes have the capacity to produce siderophores, and these genomes were primarily grouped in GFCs 5 (Bdellovibrionota), 9, 18 (Firmicutes), 34 (Actinobacteriota), as well as in GFCs 1, 13, 15 and 48 (Gammaproteobacteria). By contrast, many more genomes, from multiple phyla, encoded siderophore transporters (76% of the genomes, Fig. 3). Furthermore, microorganisms can utilize siderophore-bound iron also without the need for siderophore transporters, e.g. using ferric reductases located on the plasma membrane (73) or via direct endocytosis (74). For example, about half of the genomes that produce siderophores can produces vibrioferrin (~15% of total genomes), yet the vibrioferrin-bound iron is likely accessible to many more organisms, including phytoplankton, upon photolysis (18). Thus, in agreement with recent considerations (75), we highlight the role of siderophores as “keystone molecules” (76) and take this as an indication that the organisms producing them have an important role in the functionality of microbial communities (e.g., references (77, 78)).

Several recent studies have shown that bacteria can influence the growth of phytoplankton through the production of phytohormones (10, 15, 21), and indeed one auxin hormone, indole-3-acetic acid (IAA), has been identified in natural marine samples (10). Almost half of the genomes (~49%) in our dataset are predicted to produce IAA, including some Cyanobacteria. Four pathways for the production of IAA were identified, with some organisms encoding more than one pathway. The indole-3-acetamide pathway is the most common one and is present in nearly all GFCs comprising genomes of Alphaproteobacteria and Actinobacteriota, as well as in some other taxa (Fig. 3). The second most common is the tryptamide pathway, whereas the last two pathways are rarer and limited to Alphaproteobacteria (indole-3-acetonitrile) and some genomes of Cyanobacteria and Actinobacteriota (indole-3-pyruvate). 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 signaling (e.g. (10, 15)). 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 (79) and references therein). Therefore, it is currently unclear whether IAA is simply a byproduct of cellular metabolism or whether it is a key molecule utilized by diverse bacteria to influence the growth of phytoplankton.

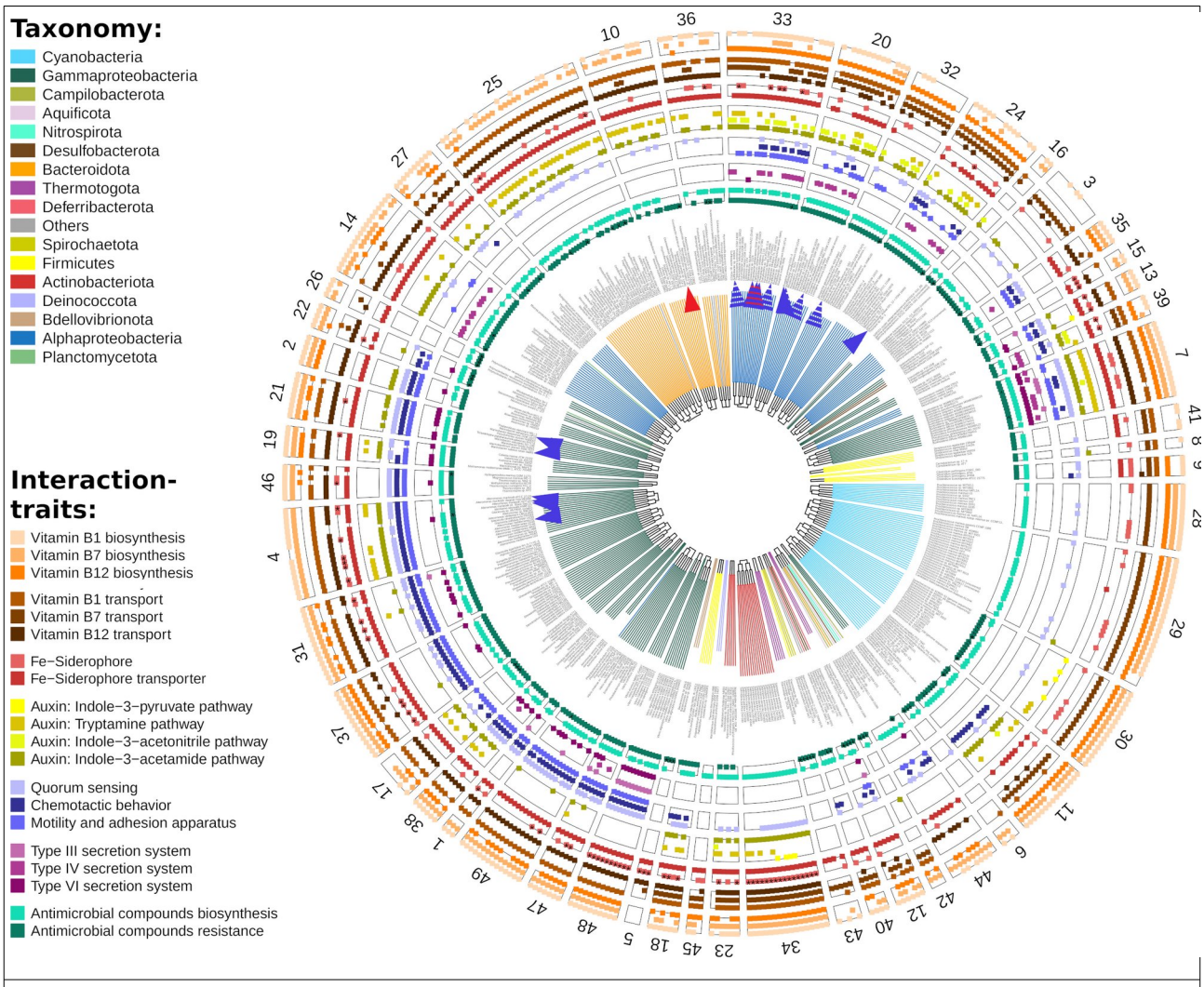


Fig. 3: Overview on interaction-traits across genome functional clusters (GFCs). Each slice shows the interaction traits present in a GFC and, as dendrogram, the functional similarity of genetic traits of the grouped genomes. Known interactions (described in the referenced literature) between bacteria and phytoplankton are marked with arrows; a blue arrow indicates a positive interaction (an enhancing effect on phytoplankton growth), a red arrow a negative interaction, and a blue-red arrow a positive interaction that eventually becomes negative. A dashed arrow indicates a

known interaction involving a bacterial genotype with a high similarity to one of the genomes included in the analysis. Asterisks in siderophores annotation indicate the presence of the specific vibrioferrin synthetic pathway along with the secondary metabolite pathway, while asterisks in the antimicrobial resistance annotation indicate that only generic resistance traits were annotated for that genome.

Traits underlying potential antagonistic interactions

Experimental measurements of interactions among marine bacteria suggest that antagonism is common (>50% of the tested isolates) (29, 31), but in most cases the mechanisms behind antagonistic interactions are unclear. Perhaps surprisingly, genes encoding for the production of putative antimicrobial compounds (as detected by antiSMASH) (80) and KEGG modules; Supplementary Table 5 and 6) were found in almost all bacteria in our dataset, including GFCs poor in other interaction traits (77% of genomes, inner ring in Fig. 3). The most abundant traits across GFCs were bacteriocin and betalactone production (Supplementary Fig. 8). These two classes of compounds group several known molecules with potent bioactivity against bacteria and fungi (81, 82), however their implication and role in natural environments are still understudied. Genes involved in the resistance to antimicrobial compounds were also relatively common (78% of genomes). We noted, however, that many pathways annotated in KEGG as resistance mechanisms against antimicrobial peptides have also other cellular functions (e.g. cell division, protein quality control and transport of other compounds; Supplementary Table 5) (83, 84). Cyanobacteria and Bacteroidota are notable examples of this as they don’t possess any specific resistance traits other than the generic ones (Fig. 3 and Supplementary Fig. 8), suggesting they might be less efficient in resisting a chemical warfare. Some Cyanobacteria strains are indeed used as markers for antibiotic contamination because of their sensitivity (e.g., references (85, 86)) and Bacteroidota often succumb when co-cultured with other bacteria that express antagonistic behaviour (29, 31). Overall, these genome-based predictions are in agreement with the experimental results of (29, 31), which suggested that Alpha- and Gammaproteobacteria commonly inhibited other bacteria, whereas Bacteroidota showed the lowest inhibitory capacity and were the most sensitive to inhibition by other bacteria.

In contrast to the predicted widespread potential for allelopathy, the capacities to sense and move towards target organisms (quorum sensing, chemotaxis, motility and adhesion), and to directly inject effector molecules (secretion systems), were more limited (respectively 61% and 24% of all bacteria, Fig. 3). Chemotaxis, motility, adhesion and quorum sensing occurred together, primarily in the GFCs that are rich in other interaction traits. These GFCs comprise Alpha- and Gammaproteobacteria (Fig. 3). In a few cases these traits were grouped within the same LTC (e.g. LTC 50, which contains traits for chemotaxis, flagellar motility and/or adhesion, or the LTC 6 which groups quorum sensing and potential resistance to antimicrobial compounds). Often, GFCs that had these “antagonism LTCs” also encoded type IV or type VI secretion systems (T4SS and T6SS, respectively). Notably, the T4SS and T6SS have different distributions among the GFCs, with only GFC 7 and 47 (comprising *Burkholderia* and *Vibrio*, 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 (87). T4SSs were detected more frequently in Alphaproteobacteria (5 out of 8 GFCs). Similar to T4SSs, the T6SS system can also deliver effector molecules into other bacterial or eukaryotic cells by using a contractible sheath-like structure (88). To date, T6SSs are known to be involved only in antagonistic interactions, suggesting that the presence of this trait is a high-confidence predictor of the ability to directly antagonize other cells ((89) 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) were found only in a few genomes (e.g. *Burkholderia* and *Aeromonas*), which are often considered metazoan-associated and sometimes pathogenic bacteria (90). T3SS allows to delivering effector molecules that maintain the bacterial association with the host (91).

Genome functional clusters differ in their overall potential for interactions

When considering antagonistic and synergistic interaction traits together, it is clear that some GFCs encode significantly more interaction traits than others (Fig. 3), both in terms of diversity and richness (Supplementary Fig. 9a). One potential driver for the difference in richness and diversity of interaction traits could be the reduction in genome size associated with oligotrophic lineages such as pico-Cyanobacteria and SAR11, and indeed the number of interaction traits can partly be explained by genome

size (Supplementary Fig. 9b-c) (92). However, genome size could not explain by itself the number of different interaction traits, with Gammaproteobacteria and several Alphaproteobacteria encoding more interaction traits than predicted by genome size, and Bacteroidota encoding fewer than expected.

Conceptual Fig. 4 demonstrates four of the main auxiliary LTCs that we predict to link traits involved in microbial interactions. LTCs 33 and 50 show the linkage of traits for chemotaxis, motility and adhesion, a typical set of traits a bacterium would need to locate, reach and settle on an organic matter particle. LTC 33 (mean r^2 = 0.63) includes, in addition, a two-component regulatory system (BarA-UvrY), the biosyntesis of pyridoxal (one form of vitamin B6) and ubiquinone. BarA-UvrY genes are known to regulate virulence, metabolism, biofilm formation, stress resistance, quorum sensing and secretion systems (see (93) and references therein). This LTC is common in the GFCs grouping Gammaproteobacteria (10 out of 19) such as *Alteromonas* (GFC 4), *Marinobacter* (GFC 21)*, Pseudoalteromonas* (GFC 31), *Shewanella* (GFC 37) and *Vibrio* (GFCs 47 and 48; Supplementary Table 3). All of these organisms are known as particle and phytoplankton associated bacteria (e.g. (23, 38, 44, 45, 94, 95)). LTC 50 (mean r2 = 0.56) includes, in addition to traits for chemotaxis, motility and adhesion, also a two-component regulatory system (AlgZ-AlgR), which is a key element in the regulation of twitching motility, alginate production and biofilm formation during *Pseudomonas aeruginosa* infections (96). This LTC is found complete, for example, in GFC 21 representing mainly *Marinobacter* (Supplementary Table 3). A model bacterium which is a member of this GFC, *M. adhaerens* HP15, has been shown to indeed interact with phytoplankton through attachment (up-regulation of type-4 fimbriae synthesis which is controlled by the PilS-PilR regulatory system included in LTC 50), subsequently increasing host aggregation (97).

LTC 16 (mean r2 = 0.33) includes interaction traits for quorum sensing and vitamin B12 transport along with amino acid and sugar transporters, and the RstB-RstA regulation system. Similar to the other regulatory mechanisms described above, the RstB-RstA system is involved in adhesion, biofilm formation, motility and hemolysis (98). These effects have been shown in in *Vibirio alginolyticus* and, not surprisingly, these linked traits (LTC 16) are all found in GFC 48 (which includes *V. alginolyticus*) and in GFCs 1 and 47 (grouping *Aeromonas* and *V. natriegensis* genomes; Supplementary Table 3). *V. alginolyticus* strains are found frequently associated with macro-algae (99) supporting the role of this LTC in interactions with phytoplankton.

Finally, our results allow us to propose that the production of phytohormones may be linked with other interaction traits. Specifically, LTC 4 (mean r^2 = 0.16), which contains the indole-3-acetonitrile pathway, is found complete in GFCs 20, 33 and 39 (Supplementary Table 3). These GFCs group genomes of bacteria known to interact via IAA-mediated mechanisms with phytoplankton (10, 15, 21) and plants (GFC 39) (100). LTC 4 also contains T4SS, which have been shown to play a pivotal role in the delivery of another phytohormone, i.e. cytokinin, into host cells (101). While the mean linkage within this LTC is relatively low (0.16), we speculate that, for some bacteria, the T4SS may be used for injecting IAA into its target cells, as the phytohormone is negatively charged under physiological conditions (pKa= 4.8) and thus not likely to cross membranes passively (79). Injection of IAA through T4SS into target cells may also explain why addition of IAA to phytoplankton growth media did not produce the expected phenotype in a phytoplankton model organism (10), and why physical connection between bacterial and host cells is often observed in related model systems (15, 21). Moreover, other genetic traits linked within LTC 4 might hint at additional molecular mechanisms underlying the interaction potential of GFCs containing this LTC. For example, the manganese/iron transporter suggests a micronutrient-dependent response, the transport of capsular polysaccharide may be involved in resistance to host defense and pathogenicity (102) , and the biosynthesis of putrescine may hint that this molecule functions as a signal stimulating growth, productivity, and stress tolerance, as shown previously in plants (103). These genome-based hypotheses can now be tested experimentally.

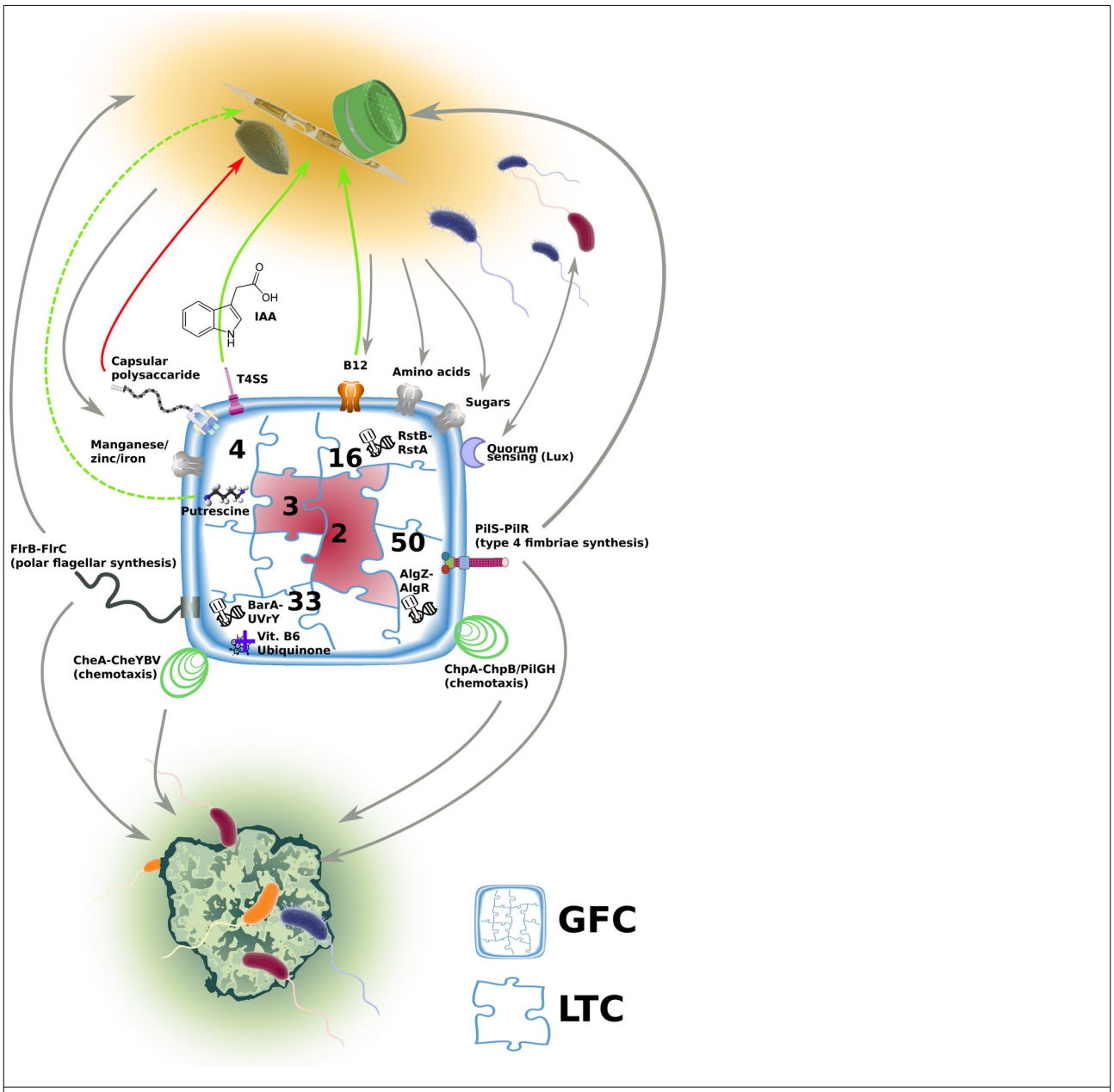


Fig. 4: Conceptual representation of a hypothetical bacterial cell characterized by the presence of the core linked trait clusters (LTCs) 2 and 3 (marked in red) and different common/ancillary LTCs mediating interactions with phytoplankton, other bacteria and particles. Green arrows indicate positive effects (e.g. enhancing growth), grey arrows are for metabolites/chemical exchange, movement or attachment, and red arrows for negative effects (e.g. pathogenicity). LTCs 4 and 16 could be mainly involved in interactions with phytoplankton, while LTCs 33 and 50 in interactions with organic matter particles.

494

Summary and conclusions

Summary, caveats and conclusions

We have presented two concepts, the GFCs and the LTCs, which provide an approach to extrapolate from studies of specific model organisms to the diversity of microbes, based on the traits encoded in their genomes. It is clear that marine bacterial taxa differ in interaction traits encoded in their genomes, resulting in specific groups of bacteria that are more likely to interact among themselves, or with other organisms, e.g. key phytoplankton species. The genomic data suggest that the production of phytohormones and siderophores is likely to be more widespread among heterotrophic bacteria than previously appreciated, and that the former may be injected through specific secretion systems.

Our work contributes two significant innovations to the field of marine biogeochemistry. First, by focusing on specific biologically relevant pathways (those associated with interactions) and by looking for coherent representation of multiple genes belonging to these pathways, we reduce a highly complex genomic dataset to a tractable and intelligible matrix of organisms by functions. This is a stepping stone that could be useful as a basic reference for studying functional traits in marine microbiomes for multiple follow-up applications. Second, by organizing the ensuing genomic information into GFC and LTC, we further simplify the interpretation of this complex dataset, while at the same time highlighting the nontrivial grouping of organisms by phenotypic traits, often irrespective of phylogeny. The GFC and LTC concepts, however, come with some important caveats. First, both GFCs and LTCs are statistical in nature, representing the probability of bacteria having a similar functional capacity, or of traits being functionally and/or evolutionarily linked. Second, the sequenced genomes represent only a part of the marine bacterial diversity, and bacterial taxa are unevenly represented (Figs. 1 and 3). Finally, a large fraction of the genes in each genome is currently uncharacterized while a few can be miss-annotated, and these instances may change our inferences (primarily inferring the lack of a trait). These biases may be, in part, responsible for some of the “blurred” squares of the “checkerboard” that represent the Atlas of functional potential of marine bacteria (Fig. 1). Nevertheless, the GFC and LTC analyses enable us to identify prevalent patterns in the ability of specific bacterial groups to interact both physically (e.g. through motility and adhesion to substrates) and through the exchange of vitamins, toxins and other info-chemicals. Clearly, the GFC and LTC concepts are not sufficient to explain the mechanism of microbial interactions, which can be highly

complex (e.g. (10, 15, 17, 20, 21)). Therefore, the patterns observed through GFCs and LTCs provide valuable hypotheses that can be tested in well-defined experiments. The GFC and LTC framework 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). Thus, GFCs and LTCs provide an important tool for ecologists, in defining a quantitative probability of whether specific traits are present and thus one or the other interaction may occur. This constitutes a significant advancement in understanding microbial functions and interactions, which underlay the fundamental rules that govern community dynamics and assembly, and the role of microbes in global ecosystem-level processes (e.g. biogeochemical cycles of C, N, P and S).

Materials and Methods

Genome selection

A dataset of complete 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. All the genomes in the final list were downloaded from NCBI and JGI repositories and checked for completeness manually, and by mean of the software CheckM (104). Only genomes whose chromosome was a continuous sequence in the fasta file (plus plasmids when present) or which met the criteria proposed for high-quality draft genomes (>90% of completeness, <10% of contaminations, >18 tRNA genes and all 3 rRNA genes present) (105) were retained for downstream analyses. 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 2).

Genome annotation

All retrieved genomes were re-annotated using a standardized pipeline. In brief, gene calling and first raw annotation steps were performed with Prokka (106). 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 (107). Additional targeted analyses were performed to annotate secondary metabolites, phytohormones and specific transporters. 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 (80). Phytohormones were manually identified mapping annotated KOs to the KEGG compounds belonging to the different phytohormones pathways present in the KEGG map01070. Only phytohormones that had at least the last 3 reactions present were considered in the analysis. Translated amino acid sequences were also used as input for a GBlast search (BioV suite) (108) to identify trans-membrane proteins, specifically vitamin B and siderophore transporters; as recommended by the authors, only the annotations with a trans-membrane alpha-helical overlap score >1 and a blast e-value $<1^{-6}$ were retained. Manual annotations were performed to specifically identify production of photoactive siderophores (i.e. vibrioferrin) (18), blasting predicted protein sequences (blastp) (109) against reference dataset assembled using all available sequences of related genes (pvsABCDE operon) (110) available in UniProt.

KEGG module reconstruction

KEGG orthologies (**KO**s) annotations generated by kofamscan were recombined in KEGG modules (**KM**s) using an in-house R script (https://github.com/lucaz88/R_script/blob/master/_KM_reconstruction.R). The KMs represent minimal functional units describing either the 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 (111), the script fetches the diagrams of all KMs from the KEGG website. Each diagram represents a pathway scheme of a KM listing all known KOs that can perform each of the reactions necessary to complete the pathway (Supplementary figure 1b). The completeness of a KM in a genome is calculated as the percentage of reactions for which at least one KOs was annotated over the total number of necessary

reactions (e.g., a KM with 7 out of 8 reactions annotated is 87.5% complete). KMs were considered complete based on the following rules: 1) KMs with fewer than 3 reactions required all reactions; 2) one gap was allowed in KMs with ≥3 reactions (i.e., a KM with 7 out of 8 annotated reactions was considered 100% complete).

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 1). 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), salicylic acid and ethylene however the last two were found in < 1% of genomes so only auxin production was included in the analysis (Supplementary Table 7). 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).

Statistical analyses

The presence/absence matrix of genetic traits in genomes served as basis to cluster the former into linked trait clusters (LTCs) and the latter into genome functional clusters (GFCs). The GFCs were generated feeding a genome functional similarity matrix calculated using Phi coefficient (i.e. Pearson correlation for binary variables; *phi* function, package sjstats) into the affinity propagation algorithm implemented in the *apcluster* function (q=0.5 and lam=0.5; r package apcluster) (112). 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 (113). Briefly, the functional similarity matrix is used to construct a network where nodes and edges are known to be genomes and their pairwise Phi 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.

For the LTC delineation, a similarity matrix of genetic traits was built calculating the square of Pearson's correlation coefficient (r^2) (114). Out of a total of 556 genetic traits, 434 were retained for downstream analysis as they were found in >3% of the genomes (>14 genomes) and r^2 was computed using the *ld* function (r package snpStats) (115). The LTCs were automatically extracted from the hierarchical clustering (*hclust* function in r, method = "ward.D2") of the dissimilarity matrix ($1-r^2$) using the function *cutreeDynamic* (method = "hybrid", deepSplit = 4 and minClusterSize = 3; r package dynamicTreeCut) (116). Similar to the context of linkage disequilibrium (LD) (117), using r^2 as representative index (118), r^2 have to be carefully interpreted as two genetic traits can be non-randomly associated because these traits are interactively linked to fitness or simply because they are closely located on the chromosome (i.e. lower chances of recombination). However, as we used r^2 to compare the genetic traits which commonly involve multiple genes, the second possibility is less likely. While exploring the functional potential, a LTC was considered complete in a genome when >60% of the grouped genetic traits were present and it was considered complete in a GFC when the average completeness of the included genomes was >60%.

All analyses were performed in R 3.6.0 (119).

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 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 lead us to shape this manuscript,.

References

1. M. E. Hibbing, C. Fuqua, M. R. Parsek, S. B. Peterson, Bacterial competition: Surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8**, 15–25 (2010).

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

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

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

5. L. Zoccarato, H.-P. Grossart, “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*, C. J. Hurst, Ed. (Springer Nature Switzerland AG 2019, 2019), pp. 13–52.

6. D. L. Kirchman, *Processes in Microbial Ecology* (Oxford University Press, 2012) https:/doi.org/10.1093/acprof:oso/9780199586936.001.0001 (July 15, 2016).

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

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

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

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

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

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

13. M. A. Moran, B. P. Durham, Sulfur metabolites in the pelagic ocean. *Nat. Rev. Microbiol.* **17**, 665–678 (2019).

14. C. Paul, M. A. Mausz, G. Pohnert, 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).

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

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

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

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

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

20. H. M. van Tol, S. A. Amin, E. V. Armbrust, Ubiquitous marine bacterium inhibits diatom cell division. *ISME J.*, 1–12 (2016).

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

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

23. D. Aharonovich, D. Sher, 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).

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

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

26. S. J. Biller, P. M. Berube, D. Lindell, S. W. Chisholm, Prochlorococcus: the structure and function of collective diversity. *Nat. Rev. Microbiol.* **13**, 13–27 (2014).

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

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

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

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

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

32. S. Krause, *et al.*, Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Front. Microbiol.* **5**, 251 (2014).

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

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

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

36. S. J. Giovannoni, *et al.*, Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science (80-.).* **309**, 1242–1245 (2005).

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

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

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

40. T. M. Lux, R. Lee, J. Love, Complete genome sequence of a free-living Vibrio furnissii sp. nov. strain (NCTC 11218). *J. Bacteriol.* **193**, 1487–1488 (2011).

41. S. P. Preheim, *et al.*, Metapopulation structure of Vibrionaceae among coastal marine invertebrates. *Environ. Microbiol.* **13**, 265–275 (2011).

42. H. S. Aronson, A. J. Zellmer, S. K. Goffredi, The specific and exclusive microbiome of the deep-sea bone-eating snail, Rubyspira osteovora. *FEMS Microbiol. Ecol.* **93**, fiw250 (2017).

43. S. B. Prayitno, J. W. Latchford, Experimental infections of crustaceans with luminous bacteria related to Photobacterium and Vibrio. Effect of salinity and pH on infectiosity. *Aquaculture* **132**, 105–112 (1995).

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

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

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

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

48. M. López-Pérez, F. Rodriguez-Valera, Pangenome evolution in the marine bacterium *Alteromonas*. *Genome Biol. Evol.* **8**, evw098 (2016).

49. L. M. Siegel, M. J. Murphy, H. Kamin, 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).

50. A. I. Scott, A. J. Irwin, L. M. Siegel, J. N. Shoolery, Sirohydrochlorin. Prosthetic Group of a Sulfite Reductase Enzyme and Its Role in the Biosynthesis of Vitamin B12. *J. Am. Chem. Soc.* **100**, 316–318 (1978).

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

52. A. M. Cook, T. H. M. Smits, K. Denger, “Sulfonates and Organotrophic Sulfite Metabolism” in *Microbial Sulfur Metabolism*, (Springer Berlin Heidelberg, 2008), pp. 170–183.

53. S. Zhang, D. A. Bryant, The tricarboxylic acid cycle in cyanobacteria. *Science (80-.).* **334**, 1551–1553 (2011).

54. M. Neumann-Schaal, D. Jahn, K. Schmidt-Hohagen, Metabolism the Difficile Way: The Key to the Success of the Pathogen Clostridioides difficile. *Front. Microbiol.* **10**, 219 (2019).

55. J. Hoskins, *et al.*, Genome of the bacterium Streptococcus pneumoniae strain R6. *J. Bacteriol.* **183**, 5709–5717 (2001).

56. S. Wushke, *et al.*, A metabolic and genomic assessment of sugar fermentation profiles of the thermophilic Thermotogales, Fervidobacterium pennivorans. *Extremophiles* **22**, 965–974 (2018).

57. C. M. Fraser, *et al.*, Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. *Nature* **390**, 580–586 (1997).

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

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

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

61. C. P. Suffridge, *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).

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

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

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

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

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

67. R. W. Paerl, *et al.*, Prevalent reliance of bacterioplankton on exogenous vitamin B1 and precursor availability. *Proc. Natl. Acad. Sci.* **115**, E10447–E10456 (2018).

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

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

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

71. D. A. Hutchins, A. E. Witter, A. Butler, G. W. Luther, Competition among marine phytoplankton for different chelated iron species. *Nature* **400**, 858–861 (1999).

72. F. Joshi, G. Archana, A. Desai, Siderophore cross-utilization amongst rhizospheric bacteria and the role of their differential affinities for Fe3+ on growth stimulation under iron-limited conditions. *Curr. Microbiol.* **53**, 141–147 (2006).

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

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

75. J. Kramer, Ö. Özkaya, R. Kümmerli, Bacterial siderophores in community and host interactions. *Nat. Rev. Microbiol.* **18**, 152–163 (2020).

76. R. K. Zimmer, R. P. Ferrer, Neuroecology, chemical defense, and the keystone species concept in *Biological Bulletin*, (2007), pp. 208–225.

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

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

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

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

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

82. P. D. Cotter, R. P. Ross, C. Hill, Bacteriocins-a viable alternative to antibiotics? *Nat. Rev. Microbiol.* **11**, 95–105 (2013).

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

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

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

86. 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" (2006) (June 12, 2020).

87. E. Grohmann, P. J. Christie, G. Waksman, S. Backert, Type IV secretion in Gram-negative and Gram-positive bacteria. *Mol. Microbiol.* **107**, 455–471 (2018).

88. Y. J. Park, *et al.*, Structure of the type VI secretion system TssK–TssF–TssG baseplate subcomplex revealed by cryo-electron microscopy. *Nat. Commun.* **9**, 1–11 (2018).

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

90. A. C. Silver, *et al.*, Interaction between innate immune cells and a bacterial type III secretion system in mutualistic and pathogenic associations. *Proc. Natl. Acad. Sci.* **104**, 9481–9486 (2007).

91. A. P. Macho, C. Zipfel, Targeting of plant pattern recognition receptor-triggered immunity by bacterial type-III secretion system effectors. *Curr. Opin. Microbiol.* **23**, 14–22 (2015).

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

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

94. I. Torres-Monroy, M. S. Ullrich, Identification of Bacterial Genes Expressed During Diatom-Bacteria Interactions Using an in Vivo Expression Technology Approach. *Front. Mar. Sci.* **5**, 200 (2018).

95. J. Deng, *et al.*, “Genomic Variations Underlying Speciation and Niche Specialization of Shewanella baltica” (2019).

96. Y. Okkotsu, A. S. Little, M. J. Schurr, The pseudomonas aeruginosa AlgZR two-component system coordinates multiple phenotypes. *Front. Cell. Infect. Microbiol.* **4**, 82 (2014).

97. A. Gärdes, M. H. Iversen, H.-P. Grossart, U. Passow, M. S. Ullrich, Diatom-associated bacteria are required for aggregation of Thalassiosira weissflogii. *ISME J.* **5**, 436–445 (2011).

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

99. K. N. Murthy, R. Mohanraju, P. Karthick, C. Ramesh, 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).

100. S. Spaepen, J. Vanderleyden, Auxin and plant-microbe interactions. *Cold Spring Harb. Perspect. Biol.* **3**, 1–13 (2011).

101. K. A. Aly, L. Krall, F. Lottspeich, C. Baron, The type IV secretion system component VirV5 binds to the trans-zeatin biosynthetic enzyme Tzs and enables its translocation to the cell surface of Agrobacterium tumefaciens. *J. Bacteriol.* **190**, 1595–1604 (2008).

102. M. Frosch, U. Edwards, K. Bousset, B. Krauße, C. Weisgerber, 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).

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

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

105. R. M. Bowers, *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).

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

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

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

109. C. Camacho, *et al.*, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

110. T. Tanabe, *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).

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

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

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

114. W. G. Hill, A. Robertson, Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).

115. D. Clayton, H. T. Leung, An R package for analysis of whole-genome association studies in *Human Heredity*, (2007), pp. 45–51.

116. P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).

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

118. J. M. VanLiere, N. A. Rosenberg, Mathematical properties of the r2 measure of linkage disequilibrium. *Theor. Popul. Biol.* **74**, 130–137 (2008).

119. R Core Team, R: A Language and Environment for Statistical Computing (2020).

SUPPLEMENTARY INFORMATION

Comparative whole-genome approach to identify traits underlying microbial interactions

Luca Zoccarato*a, Daniel Sher*b, Takeshi Mikic, Daniel Segrèd,e, Hans-Peter Grossart*a,f,g

Corresponding authors(*)

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

This PDF file includes:

Supplementary text

Figures S1 to S9

Other supplementary materials for this manuscript include the following:

Tables S1 to S7

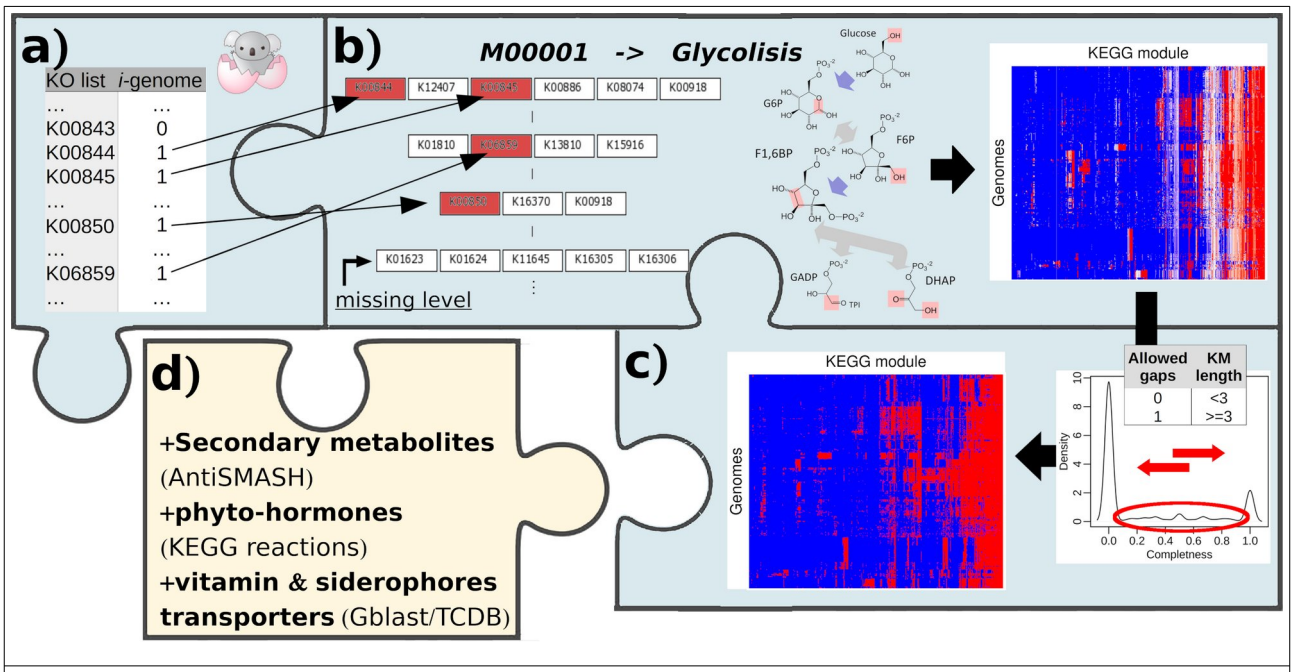
Datasets S1 to S2

Remarks on the annotation pipeline

A relevant aspect in our analysis which needs to be kept in mind, we included only closed bacterial genomes (i.e. a single, high quality sequence of each DNA molecule such as chromosome or plasmid) or high-quality draft genomes (estimated using CheckM, see method section). The rational was to provide a comprehensive description of the full functional potential of pelagic marine bacteria which requires high-quality genomes to achieve the best information (i.e. functional genes; (Fadeev et al. 2016). Gene annotation is *per-se* a challenging step, in particular, when it deals with environmental genomes for which many genes are still unknown and, therefore, cannot be properly annotated (~50% of the predicted coding sequences were annotated to KEGG orthologues (**KOs**) in our analysis).

A further step to improve cross-comparability among genomes was to re-annotate all of them with a standardized pipeline. For each genome, a list of genetic traits, termed as functional profile, was generated mainly recovering complete KEGG modules (**KMs**) from the annotated KOs (Supplementary Fig. 1a). We decided to focus our analysis on complete KMs instead of looking at the level of single annotated genes to seek for a more robust prediction of the genomic-inferred metabolic potential. To determine the presence of a KM, the detection of a single KO is not sufficient (unless the whole KEGG module is represented by a single KO), but the presence of a minimum complete set of related KOs is required. For example, for a KM with 5 required steps, and each step can be performed by several KOs, at least 1 KO per level has to be annotated in order to consider this genetic trait as complete (Supplementary Fig. 1b). This criterion was slightly relaxed by allowing for some gaps in bigger modules (i.e. 1 gap in KMs with > 3 steps) to account for possible annotation issues (i.e. missing KOs due to the absence of suitable reference genes in the database, unknown genes, miss-annotations, etc.; Supplementary Fig. 1c). The functional profiles were further enriched with the annotation of other genetic traits using specific tools, e.g. for secondary metabolites (antiSMASH), phytohormones (KEGG pathway map01070), transporters for B vitamins and siderophore (BioV suite) (Supplementary Fig. 1d).

Additionally, the presence of a complete genetic trait does not necessarily translate into an expressed phenotype. Although several genetic traits are constitutively active, many of them can be under fine regulatory expression and be manifested only under specific environmental and physiological conditions.



Supplementary Fig. 1: Annotation workflow for the identification of genetic traits in genomes (a-c). Annotated KEGG orthologies (a) were recombined in all known KEGG modules (KM; b) and labeled as present or absent using a custom r script. The script, taking into account some completions rules, generates a presence/absence matrix (c). Further annotations were performed using antiSMASH for detection of secondary metabolites, KEGG reactions of the pathways map01070 for detection of phytohormones and Gblast against the Transporter Classification Database (TCDB) to identify B vitamins and siderophore transporters (d).

16S rRNA gene classification

16S rRNA genes were identified using the RNAmmer option (Lagesen et al. 2007) during the prokka genome annotation step. The extracted sequences were classified using a local installation of SINA v1.6.1 (Pruesse et al. 2012) against the SILVA reference database SSU NR 99 v138 (Quast et al. 2013). In case of multiple 16S rRNA gene copies in a genome we ensured that classification of each copy was coherent with all other copies. Taxonomy of Cyanobacteria genomes were manually curated using NCBI taxonomy associated with the NCBI Taxonomy IDs of the genomes.

GFCs and their phylogeny

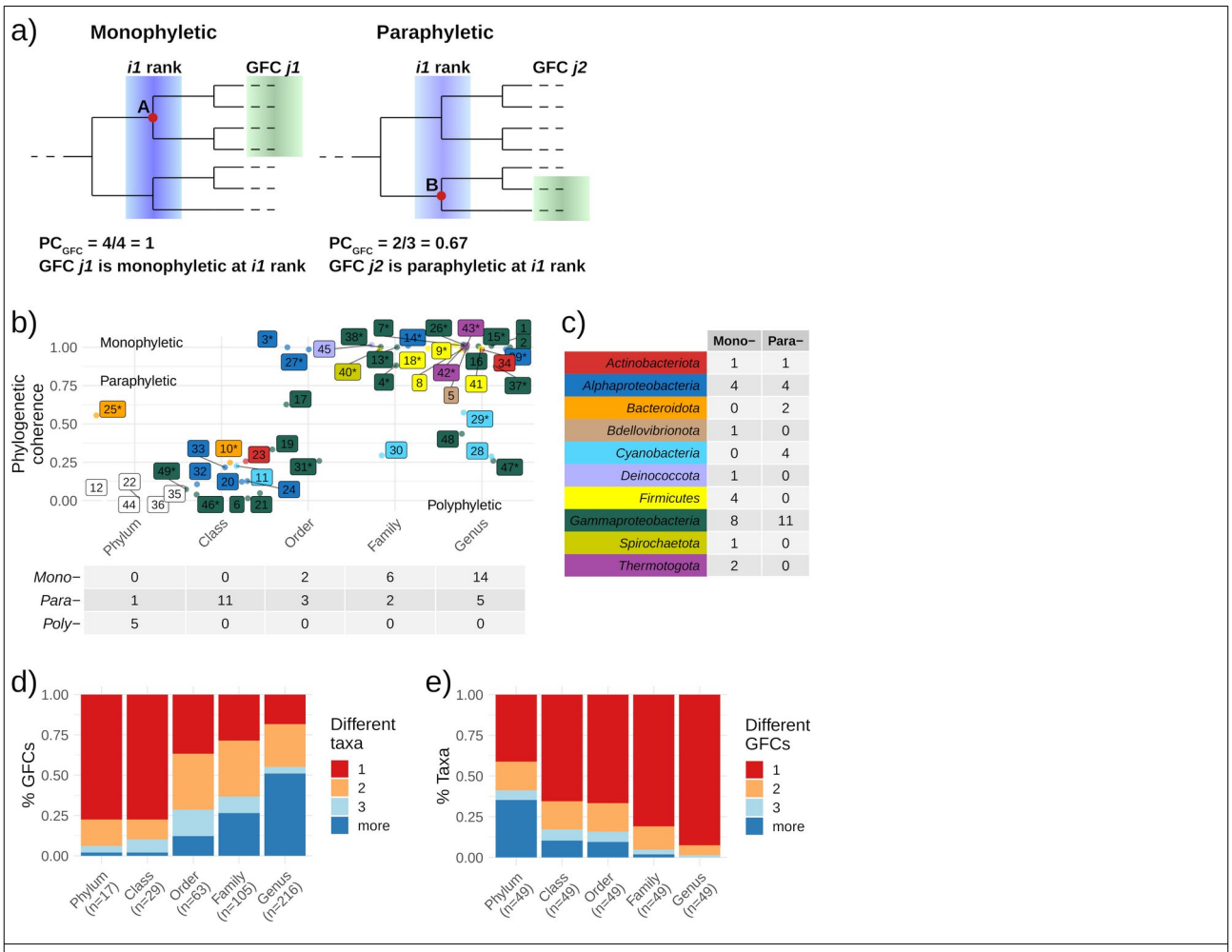
The genome clustering analysis retrieved a total of 49 genome functional clusters (GFCs). As shown in Supplementary Fig. 2d, most of these GFCs include only genomes

of the same phylum (38), and fewer than 3 different families (14 with 1 and 17 with 2). At the genus levels more than half of the GFCs group together 3 or more different taxa. From the opposite perspective, the taxa level, more than half of the phyla were represented by 2 or more GFCs while nearly all genera by a single GFC (Supplementary Fig. 2e).

Similarly to (Koeppel and Wu 2013), the phylogenetic coherence of each GFC was calculated based on the "local" phylogenetic coherence score (PC_{GFC}; Supplementary Fig. 2a):

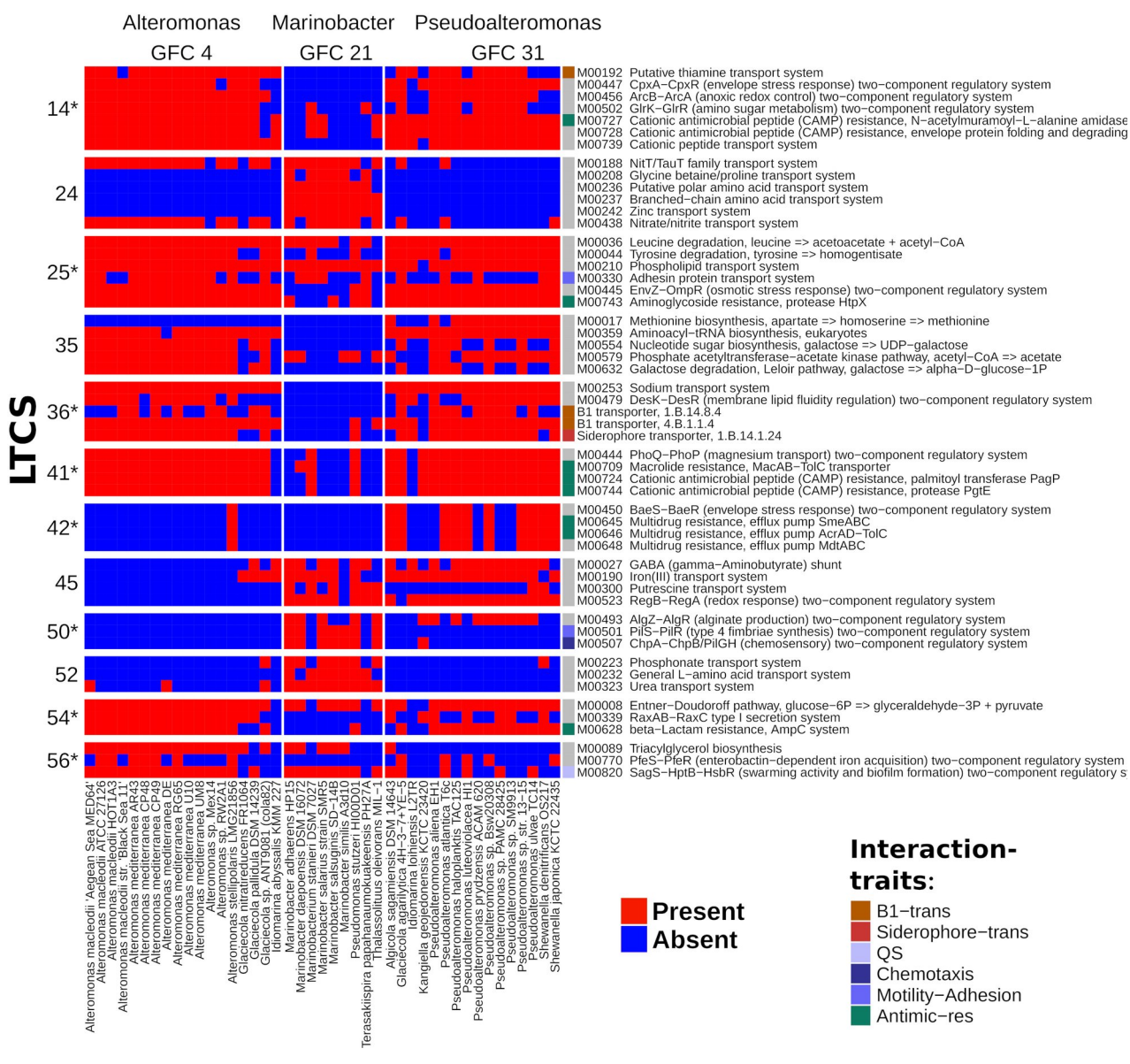
$$PC_{GFC} = N_{GFC}/N_{taxon}$$

where N_{GFC} is the number of genomes grouped in a GFC, and N_{taxon} is the total number of genomes (included in our genome atlas) belonging to the last common ancestor of the GFC. The level of phylogenetic coherence of a GFC corresponds to the taxonomic rank of the last common ancestor. A PC_{GFC} of 1 indicates that all genomes belonging to the last common ancestor are included in the respective GFC which is considered to be monophyletic. A $PC_{GFC} < 1$ indicates instead a paraphyletic GFC. As the current genome availability doesn't allow for a uniform coverage of all bacterial taxa, a possible issue leading to taxonomic promiscuity of the paraphyletic GFCs might be related to the inclusion of "singleton" taxon (i.e. a single genome that represents a different taxon, as in GFCs 2, 7 or 29; see Supplementary Table 2 and Supplementary Fig. 2b). To avoid interpretation biases due to these singletons, we excluded a maximum of one singleton genome per GFC in the computation of the phylogenetic coherence. Nonetheless, ~30% of the paraphyletic GFCs included more evenly represented taxa as GFCs 14, 30 and 33, which grouped different genera with each genus represented by 3 or more genomes.



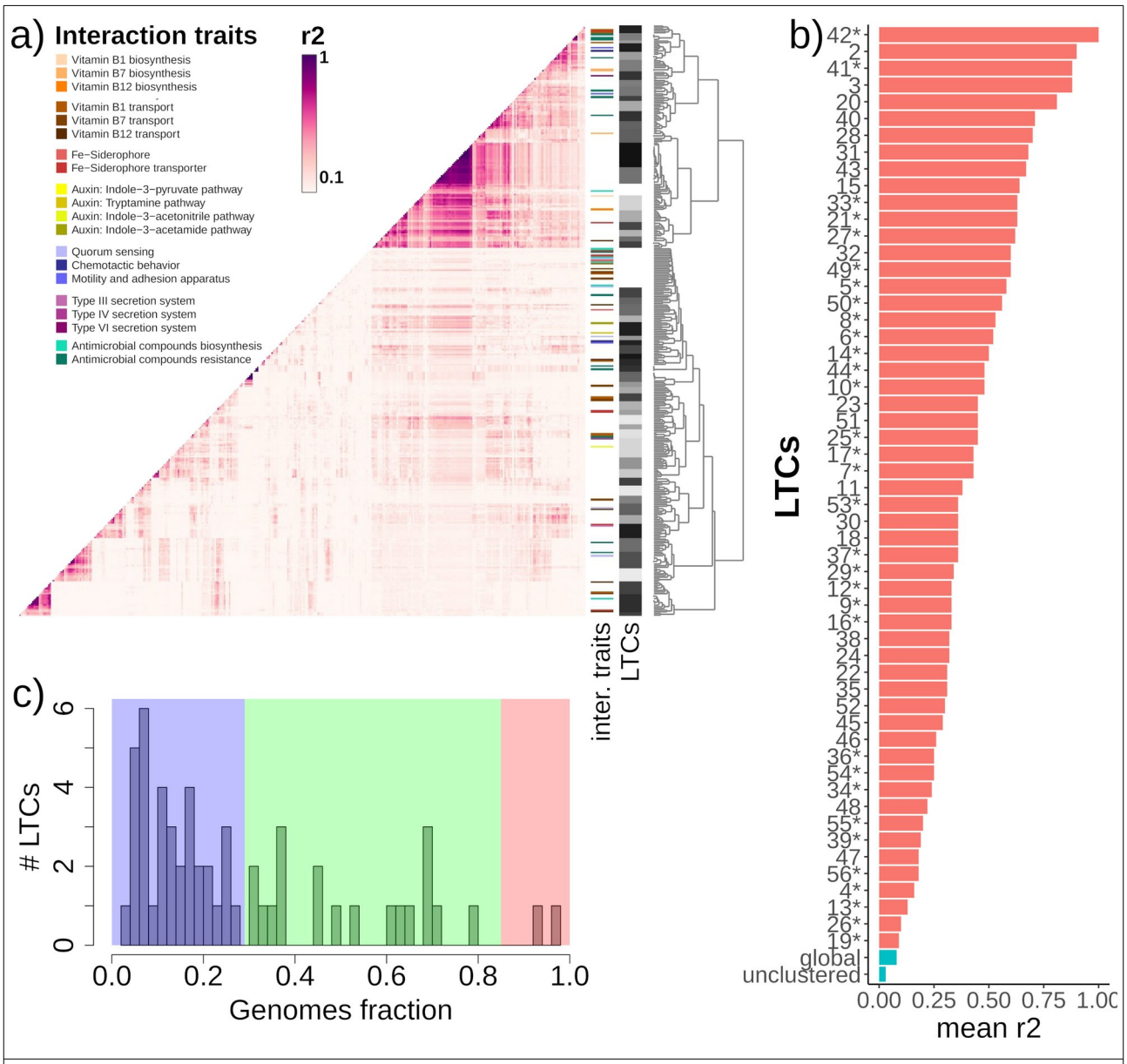
Supplementary Fig. 2: (a) Schematic example of calculating the phylogenetic coherence score in the genome functional clusters (GFCs) 'j1' and 'j2' carried out at the taxonomic rank 'i1'. GFC 'j1' is monophyletic at that rank as it includes all genomes belonging to the related taxon 'A' (4/4) while GFC 'j2' is paraphyletic as it included only some genomes of the related taxon 'B' (2/3). (b) Highest phylogenetic coherence scores of each GFC separated by taxonomic ranks. Table underneath x-axis shows the amount of mono-, para- or polyphyletic GFCs for each rank; GFCs containing 1 "singleton" genome belonging to a different clade are marked with an asterisk; dots are color coded according to their taxonomy as shown in panel c. (c) Table summarizing the number of mono- and paraphyletic GFCs across different phyla. (d) Percentage of GFCs including 1, 2, 3 or more different taxa for each taxonomic rank. (e) Percentage of taxa included in 1, 2, 3 or more different GFCs for each taxonomic rank.

1044



Supplementary Fig. 3: Insight analysis on the different linked trait clusters (LTCs) possessed by Alteromonas, Marinobacter and Pseudoalteromonas genome functional clusters (GFCs). The asterisk on the LTC names indicate that a LTC contains also one or more interaction traits. All tree GFCs share a set of complete LTCs, namely:2, 3, 8*, 10*, 15, 17*, 21*, 23, 27*, 28, 33*, 40, 43 and 49*; check Supplementary Table 3 for more details abut these LTCs.*

1045



Supplementary Fig. 4: (a) Pairwise r^2 correlations among all genetic trait pairs; vertical color bar indicates the presence of different interaction traits while the vertical gray bar delineates specific linked trait clusters (LTCs). An interactive version of the same figure is available as Supplementary dataset 2. (b) Mean r^2 value of all pair of genetic traits included in each LTC; mean r^2 between all detected genetic traits ('global') and between all non-clustered genetic traits are shown as indication for random correlation within the dataset. (c) Histogram of LTC occurrence across genomes showing the division in "core" (present in ≥ 90% of genomes; red), "common" (< 90% and ≥ 30%; green) and "ancillary" (≤ 30%; blue) LTC.

1046

Absence-pattern among common LTCs

Some of the genetic traits included in the common LTCs 15, 23, 28 and 40 (found in 61-70% of the genomes) were consistently missing in some GFCs and/or taxonomic groups (Supplementary Fig. 5).

Broken TCA cycle

LTC 15 was absent in all Cyanobacteria (GFCs 11, 28, 29 and 30), Thermotogota (GFCs 42 and 43), Spirochaetota (GFC 40), Deferribacterota (GFC 12) and most GFCs of Firmicutes (8, 9 and 41). It included KEGG modules involved in the Citrate cycle (M00009, M00011, M00149), the biosynthesis of the amino acid serine and of phosphatidylethanolamine, the degradation oh histidine to glutamate and beta-oxidation (breakdown of fatty acids). Of note, LTC 15 is missing in ~30% of genomes and it encodes the complete TCA cycle (M00009) as well as the second carboxylation part (from 2-oxoglutarate to oxaloacetate; M00011). These genomes, however, still have the first carboxylation part of this cycle (from oxaloacetate to 2-oxoglutarate; M00010) which is included in the core LTC 2.

For almost four decades it was thought that Cyanobacteria indeed lack a full TCA cycle, until two new enzymes were discovered that, together catalyze the conversion of 2-oxoglutarate to succinate and thus functionally replace 2-oxoglutarate dehydrogenase and succinyl-CoA synthetase (Zhang and Bryant 2011). In our annotation, Cyanobacterial genomes lack these two ‘classical’ reactions of the TCA cycle (M00009), and therefore this pathway is flagged as incomplete, based on our rule of one gap for traits with up to 10 reactions. A manual search revealed that the 2-OGDC gene (K01652), which catalyzes the conversion of 2-oxoglutarate to succinic semialdehyde, is present in all Cyanobacterial genomes, whereas the SSADH gene (K00135) that catalyzes the conversion of succinic semialdehyde to succinate is present only in the genomes of Cyanobacteria belonging to GFC 11. All of the pico-Cyanobacteria genomes lack the SSADH gene, consistent with the current view that these organisms lack a full TCA cycle (Zhang and Bryant 2011).

In addition, our results show that other heterotrophic bacteria lack several or almost all of the reaction of this pathway. To the former case belongs genomes of Spirochaetota and Thermotogota, (grouped in GFCs 37 and 42, respectively) while to the latter case belong several Firmicutes (GFCs 9 and 38) and other Thermotogota (GFC 41) genomes. These findings confirm and broaden what has been previously

reported for some genomes of these taxa (Fraser et al. 1997; Hoskins et al. 2001; Wushke et al. 2018; Neumann-Schaal et al. 2019).

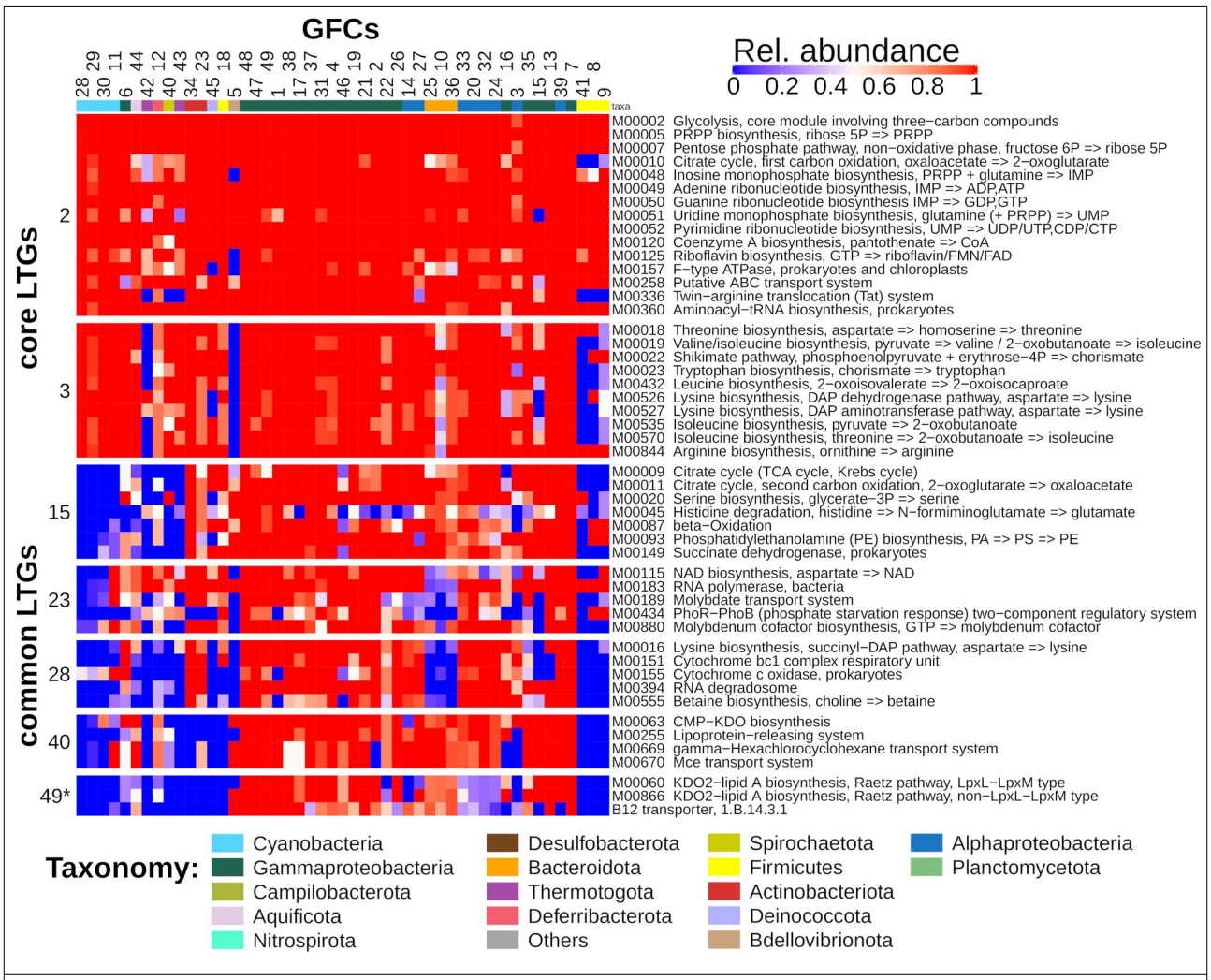
Differences in cell wall composition

In the LTC 40 and 49 (present in 70% and 65% of genomes), the absence of any pathway for the production of keto-deoxyoctulosonate (KDO, core constituent of lipopolysaccharide; M00060, M00063 and M00866) and of transporters for lipoprotein (M00255; Lol machinery), phospholipids (M00670; *mla* machinery) and gamma−Hexachlorocyclohexane (M00669; similar to *mla* machinery; Supplementary Fig. 5) provided another direct way to compare and validate our findings. The emerging incompleteness pattern of this LTC in GFCs 8, 9, 18, 41 (Firmicutes), 23, 34 (Actinobacteriota), and 45 (Deinococcota) could be largely explained by the bacterial cell wall type. In fact, all these GFCs grouped gram positive bacteria which are missing an outer membrane and, therefore, they don’t require any KDO (lipopolysaccharide transporter (M00320; Lpt machinery) was an unclustered trait), nor export of any lipoprotein out of the cell membrane (Silhavy et al. 2010). They also don’t need to exchange any phospholipids between the two membranes (Malinverni and Silhavy 2009). Despite being gram-negative, Thermotogota (GFCs 42 and 43), Spirochaetota (GFC 40) and most Cyanobacteria (GFCs 28, 29 and 30) were consistently missing some of the genetic traits of the LTC 40. Experimental evidence supports the lack of lipopolysaccharide in Thermotogota (Plötz et al. 2000) and of KDO in the cell wall of Cyanobacteria (Durai et al. 2015) and Spirochaetota (Vinogradov et al. 2004). However the lack of any lipoprotein transporter in Cyanobacteria likely reflects a consistent gap in the annotation as these organisms are known to possess lipoproteins in the outer cell membrane (Durai et al. 2015).

Variants of RNA degradosome and polymerase

All Bacteroidota (GFCs 10, 25 and 36) and pico-Cyanobacteria (GFCs 10, 25 and 36) consistently lacked LTCs 23 and 28 grouping, among others, the biosynthesis of the RNA degradosome and polymerase. Absence of the two RNA related traits offered a good chance for validation. Indeed, Bacteroidota and pico-Cyanobacteria are known to possess different and unique types of σ factors and other subunits of the RNA polymerase (Vingadassalom et al. 2005; MacGregor 2015) and they do not possess the typical RNA degradosome complex (Kaberdin et al. 2011; Zhang et al. 2014; Aït-Bara and Carpousis 2015). The RNA degradosome was missing also in Thermotogota (GFCs 42 and 43), Spirochaetota (GFC 40), Deferribacterota (GFC 12) and in Firmicutes (GFCs

8, 9 and 41) for which similar conclusion could be drawn from the literature (Kaberdin et al. 2011; Aït-Bara and Carpousis 2015).



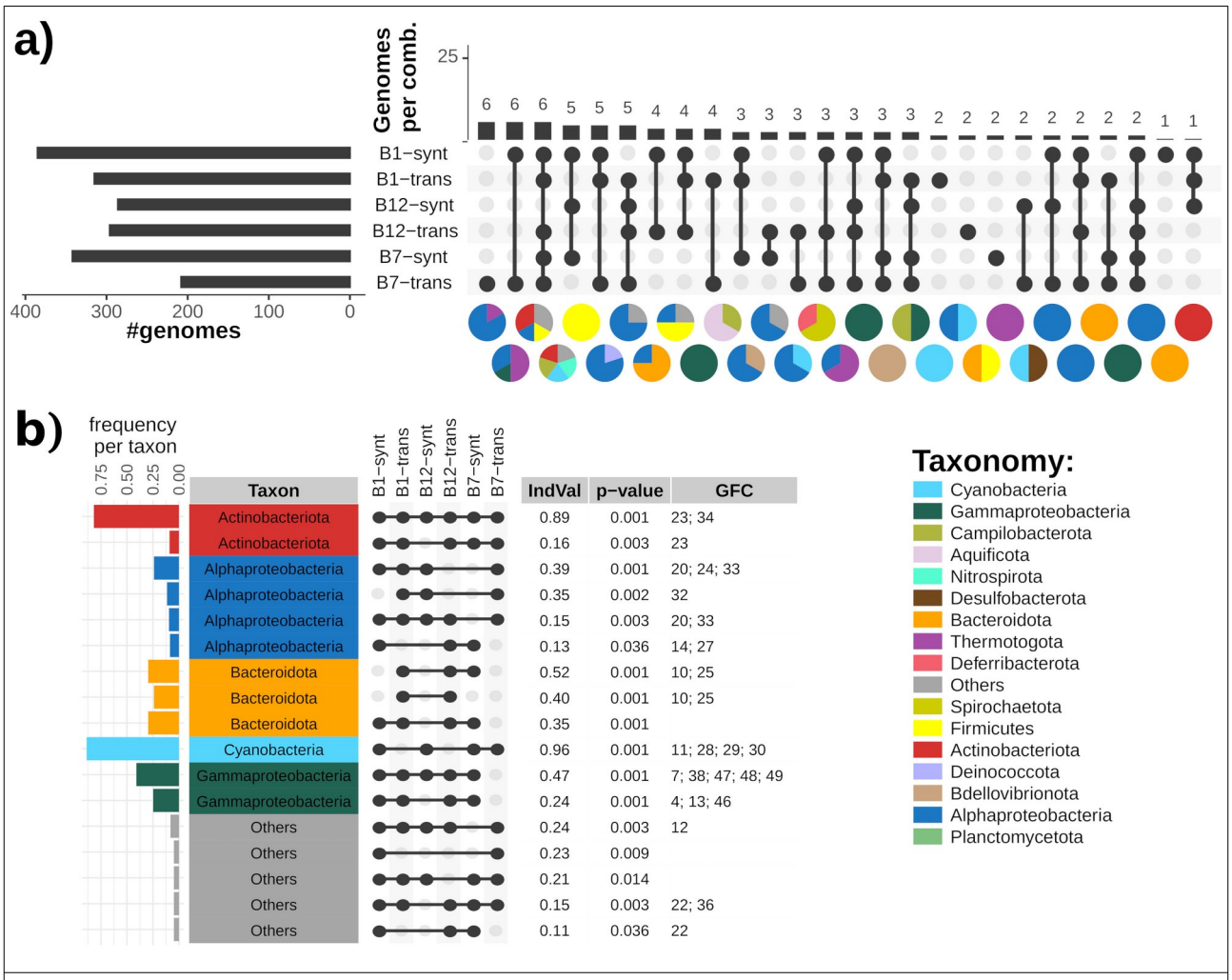
Supplementary Fig. 5: Relative abundance and absence-patterns of genetic traits grouped into core linked trait clusters (LTCs) 2 and 3, and in the common LTGs 15 (TCA cycle insight), 23, 28 (RNA polymerase and degradosome insight), 40 and 49 (cell wall insight) across all genome functional clusters (GFCs). The horizontal color bar indicate the most represented taxon in each GFC.

Combinations of vitamin traits'

As shown in Fig. 2c, some of the most frequent combinations of synthesis and uptake of vitamins B1, B12, and B7 appear more often in certain taxa than others. These patterns suggest the existence of taxa-specific evolutionary strategies for handling

these B vitamins. To test for this notion, we carried out an indicator species analysis as implemented in the function *multipatt* (func = "r.g", duleg = F, max.order = 2; package indicspecies) (De Cáceres and Legendre 2009; Cáceres et al. 2010). The function is design to identify one or multiple species that can be used as indicators for certain habitats because of their strong species- habitat association. We performed the analysis using the most abundant B vitamins strategies (Fig. 2c) instead of species, while taxonomic groups and GFCs were used instead of habitats. We carried out the analysis mainly at the phyla level (except for Proteobacteria which we analyzed at the class level). Thereby, taxa represented by only 1-2 genomes were grouped together as 'Others'. We also tested for specific associations of B vitamin strategies with combinations of site groups accounting for any possible taxa pair for combinations up to 5 different GFCs.

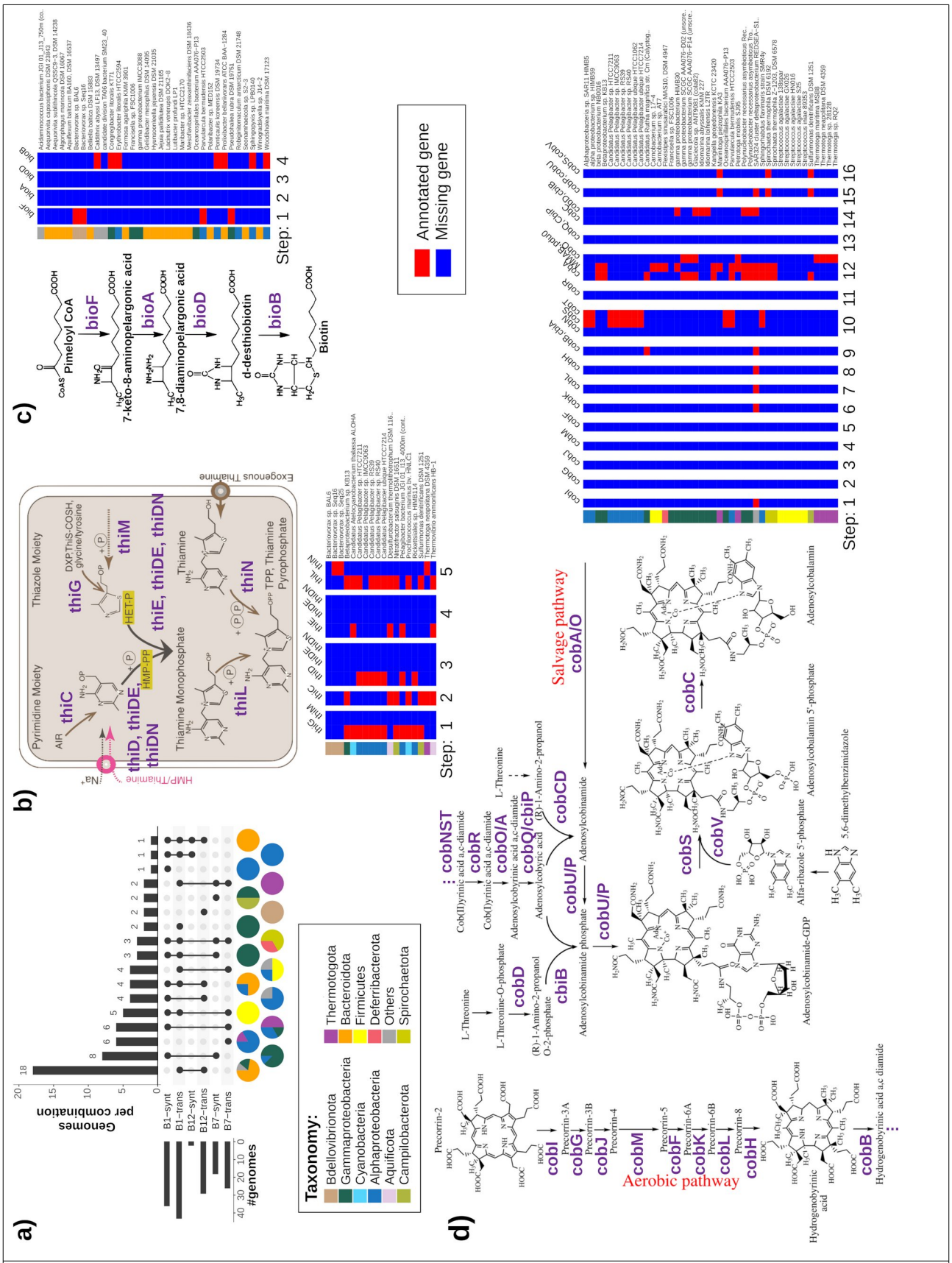
We identified a specific B vitamin strategy for 5 taxa (as 'Others' represent an artificial group containing Phyla with only a few representative genomes) and for 36 GFCs (Supplementary Fig. 6b). Except for Cyanobacteria and Actinobacteriota, all other taxa had more than one associated strategy which in some cases partitioned singularly into different GFCs, sometimes different GFCs of the same taxon possessed the same strategy, and in other cases the same strategy was present in GFCs of different taxa.



Supplementary Fig. 6: (a) UpSet plot showing the less common combinations of traits involved in the biosynthesis and transport of vitamins B1, B12, and B7. Horizontal bars indicate the total number of genomes for each trait, the dark connected dots indicate the different configurations of traits and vertical bars indicate the number of genomes possessing each configuration. Pie charts show the relative abundance of the taxonomic groups represented in each configuration. (b) Table with strategies for B vitamins uptake associated to specific taxa and genome functional clusters (GFCs). Horizontal bars show the frequency of strategies in each taxon and 'IndVal' expresses the strength of the strategy-taxon associations.

Problematic vitamin cases

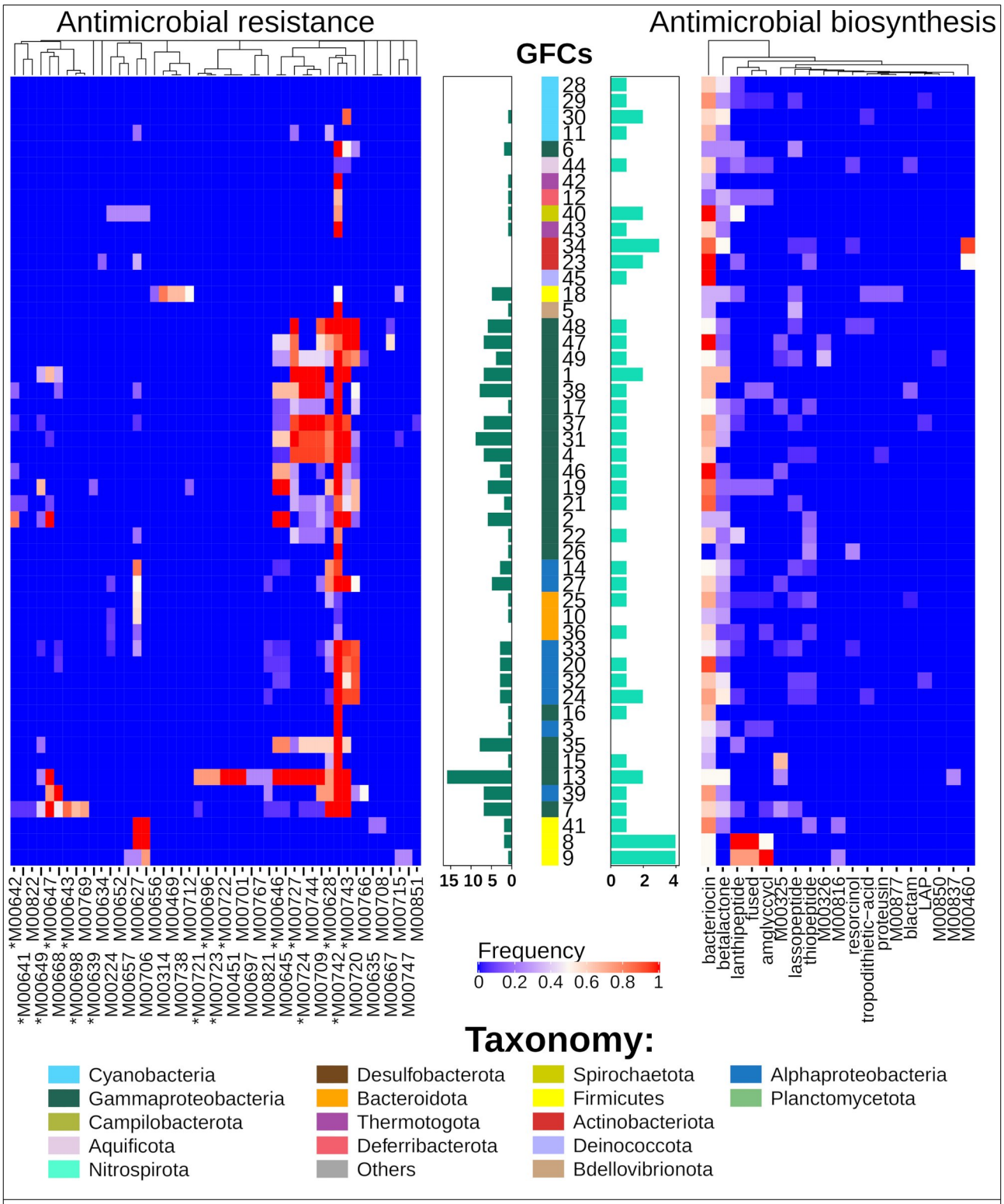
A small portion of genomes showed a fourth type of genetic configuration where neither the biosynthetic pathway nor the transporter for one (13%) or more (3%) vitamins have been detected (Supplementary Fig. 7a). These problematic cases could reflect limitations of our annotation process (e.g. unknown transporters or alternative genes/ pathways), however, there is growing evidence of organisms that even without complete biosynthetic pathways, but are able to grow on vitamin B intermediates (Wienhausen et al. 2017; Paerl et al. 2018). Therefore, we looked for the presence of possible fragmented biosynthetic pathways that could suggest forms of auxotrophy towards specific B-vitamin intermediates. We found that nearly all genomes with problematic vitamin B1 annotations possessed truncated biosynthetic pathways and need exogenous precursor supplies (Supplementary Fig. 7b). Contrary, for the problematic cases of vitamin B7 or B12, the related genomes owned only a few annotated genes or any at all (Supplementary Fig. 7c,d). This could point to an annotation issue of the related transporter, to yet unknown alternative biosynthetic pathways and reactions, or that these vitamins are not essential. Vitamin B12 is involved in amino acid and nucleotide synthesis, as well as in fatty- and amino acid breakdown; while vitamin B7 is involved in a “side” path of the TCA cycle and in the urea cycle. Most of these processes, however, have vitamin-independent routings (see in (Romine et al. 2017)) and genomes without biosynthetic capability nor vitamin B12 dependent genes have been reported (Shelton et al. 2019) pointing that these vitamins, in some cases, might not be essential.



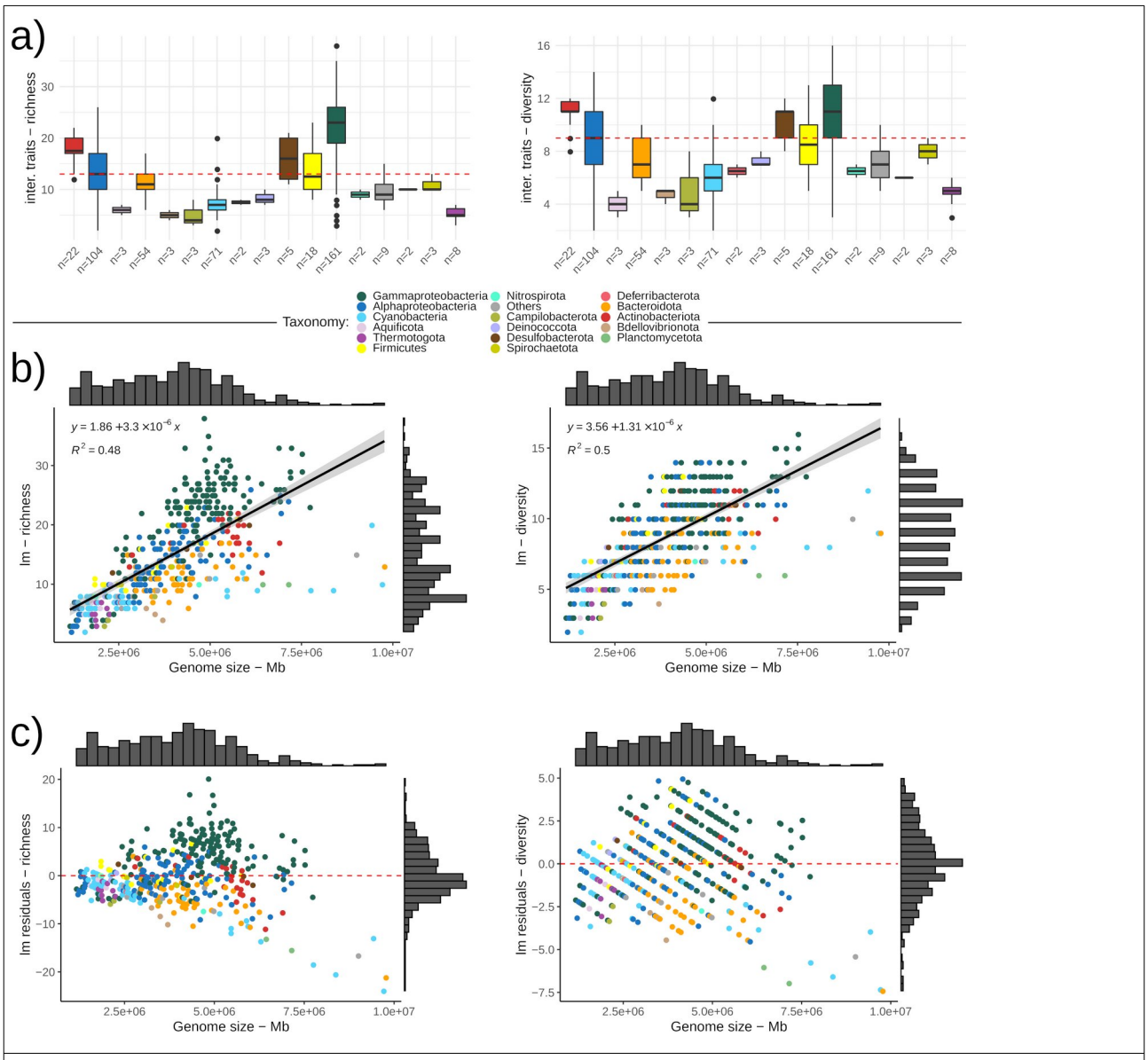
Supplementary Fig. 7: (a) UpSet plot showing incomplete combinations where both the biosynthetic pathway and the transporter for one (or more, 1% of genomes) of

the B vitamin were missing. Horizontal bars indicate the total number of genomes for each trait, the dark connected dots indicate the different configurations of traits and vertical bars indicate the number of genomes possessing each configuration. Pie charts show the relative abundance of the taxonomic groups represented in each configuration. (b-d) Schematic overview of the biosynthetic pathways of (b) vitamin B1 (modified figure from (McRose et al. 2014), (c) B7 (modified figure from (Croft et al. 2006) and (d) B12 (modified figure from (Fang et al. 2017). The related presence-absence maps show the annotated and missing genes involved in each pathway in genomes that were missing the genetic traits for both biosynthesis and transport of that vitamin. The genes are arranged as in the schematic view of the related pathway.

1165



Supplementary Fig. 8: Distribution of antimicrobial resistance and biosynthetic traits in genome functional clusters (GFCs). Resistance traits marked with an asterisk are considered generic as they might be involved in other cellular metabolisms (Supplementary Table 4). Bar plots at the center shown the counts of traits with a frequency > 0.5 in a GFC.



Supplementary Fig. 9: (a) Genome enrichment of interaction traits in terms of total number of interaction traits and diversity of interaction traits (as of different types e.g. vitamin B1 biosynthesis, type 4 secretion system,..). Box-plots show the distribution in different taxa and the dashed red lines represent the median value of trait richness and diversity across all genomes. (b) Linear regression between genomes size and interaction trait richness and diversity. (c) Plot of the residuals of the linear regression models. Genomes (i.e. dots) above the dashed red lines encode a higher number or a higher diversity of interaction traits that expected according to their genome size, while genomes underneath the lines bear less interaction traits than expected.

References

Aït-Bara, S., Carpousis, A.J., 2015. RNA degradosomes in bacteria and chloroplasts: Classification, distribution and evolution of RNase E homologs. Mol. Microbiol. 97, 1021–1035. https://doi.org/10.1111/mmi.13095

Cáceres, D.M., Legendre, P., Moretti, M., 2010. Improving indicator species analysis by combining groups of sites. Oikos 119, 1674–1684. https://doi.org/10.1111/j.1600-0706.2010.18334.x

Croft, M.T., Warren, M.J., Smith, A.G., 2006. Algae Need Their Vitamins. Eukaryot. Cell 5, 1175–1183. https://doi.org/10.1128/EC.00097-06

De Cáceres, M., Legendre, P., 2009. Associations between species and groups of sites: Indices and statistical inference. Ecology 90, 3566–3574. https://doi.org/10.1890/08-1823.1

Durai, P., Batool, M., Choi, S., 2015. Structure and effects of cyanobacterial lipopolysaccharides. Mar. Drugs. https://doi.org/10.3390/md13074217

Fadeev, E., De Pascale, F., Vezzi, A., Hübner, S., Aharonovich, D., Sher, D., 2016. Why close a bacterial genome? The plasmid of Alteromonas macleodii HOT1A3 is a vector for inter-specific transfer of a flexible genomic Island. Front. Microbiol. 7, 1–13. https://doi.org/10.3389/fmicb.2016.00248

Fang, H., Kang, J., Zhang, D., 2017. Microbial production of vitamin B12: A review and future perspectives. Microb. Cell Fact. 16, 1–14. https://doi.org/10.1186/s12934-017-0631-y

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., Van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M.D., Horst, K., Roberts, K., Hatch, B., Smith, H.O., Venter, J.C., 1997. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390, 580–586. https://doi.org/10.1038/37551

Hoskins, J., Alborn, J., Arnold, J., Blaszczak, L.C., Burgett, S., Dehoff, B.S., Estrem, S.T., Fritz, L., Fu, D.J., Fuller, W., Geringer, C., Gilmour, R., Glass, J.S., Khoja, H., Kraft, A.R., Lagace, R.E., LeBlanc, D.J., Lee, L.N., Lefkowitz, E.J., Lu, J., Matsushima, P., McAhren, S.M., McHenney, M., McLeaster, K., Mundy, C.W., Nicas, T.I., Norris, F.H.,

O'Gara, M., Peery, R.B., Robertson, G.T., Rockey, P., Sun, P.M., Winkler, M.E., Yang, Y., Young-Bellido, M., Zhao, G., Zook, C.A., Baltz, R.H., Jaskunas, S.R., Rosteck, J., Skatrud, P.L., Glass, J.I., 2001. Genome of the bacterium Streptococcus pneumoniae strain R6. J. Bacteriol. 183, 5709–5717. https://doi.org/10.1128/JB.183.19.5709-5717.2001

Kaberdin, V.R., Singh, D., Lin-Chao, S., 2011. Composition and conservation of the mRNA-degrading machinery in bacteria. J. Biomed. Sci. https://doi.org/10.1186/1423-0127-18-23

Koeppel, A.F., Wu, M., 2013. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. Nucleic Acids Res. 41, 5175–5188. https://doi.org/10.1093/nar/gkt241

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T., Ussery, D.W., 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108. https://doi.org/10.1093/nar/gkm160

MacGregor, B.J., 2015. Abundant intergenic TAACTGA direct repeats and putative alternate RNA polymerase β' subunits in marine beggiatoaceae genomes: Possible regulatory roles and origins. Front. Microbiol. 6, 1397. https://doi.org/10.3389/fmicb.2015.01397

Malinverni, J.C., Silhavy, T.J., 2009. An ABC transport system that maintains lipid asymmetry in the Gram-negative outer membrane. Proc. Natl. Acad. Sci. 106, 8009–8014. https://doi.org/10.1073/pnas.0903229106

McRose, D., Guo, J., Monier, A., Sudek, S., Wilken, S., Yan, S., Mock, T., Archibald, J.M., Begley, T.P., Reyes-Prieto, A., Worden, A.Z., 2014. Alternatives to vitamin B1 uptake revealed with discovery of riboswitches in multiple marine eukaryotic lineages. ISME J. 8, 2517–2529. https://doi.org/10.1038/ismej.2014.146

Neumann-Schaal, M., Jahn, D., Schmidt-Hohagen, K., 2019. Metabolism the Difficile Way: The Key to the Success of the Pathogen Clostridioides difficile. Front. Microbiol. 10, 219. https://doi.org/10.3389/fmicb.2019.00219

Paerl, R.W., Sundh, J., Tan, D., Svenningsen, S.L., Hylander, S., Pinhassi, J., Andersson, A.F., Riemann, L., 2018. Prevalent reliance of bacterioplankton on exogenous vitamin B1 and precursor availability. Proc. Natl. Acad. Sci. U. S. A. 115, E10447–E10456. https://doi.org/10.1073/pnas.1806425115

Plötz, B.M., Lindner, B., Stetter, K.O., Holst, O., 2000. Characterization of a novel lipid A containing D-galacturonic acid that replaces phosphate residues. The structure of the lipid A of the lipopolysaccharide from the hyperthermophilic bacterium Aquifex pyrophilus. J. Biol. Chem. 275, 11222–11228. https://doi.org/10.1074/jbc.275.15.11222

Pruesse, E., Peplies, J., Glöckner, F.O., 2012. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28, 1823–1829. https://doi.org/10.1093/bioinformatics/bts252

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590-6. https://doi.org/10.1093/nar/gks1219

Romine, M.F., Rodionov, D.A., Maezato, Y., Osterman, A.L., Nelson, W.C., 2017. Underlying mechanisms for syntrophic metabolism of essential enzyme cofactors in microbial communities. ISME J. 11, 1434–1446. https://doi.org/10.1038/ismej.2017.2

Shelton, A.N., Seth, E.C., Mok, K.C., Han, A.W., Jackson, S.N., Haft, D.R., Taga, M.E., 2019. Uneven distribution of cobamide biosynthesis and dependence in bacteria predicted by comparative genomics. ISME J. 13, 789–804. https://doi.org/10.1038/s41396-018-0304-9

Silhavy, T.J., Kahne, D., Walker, S., 2010. The bacterial cell envelope. Cold Spring Harb. Perspect. Biol. https://doi.org/10.1101/cshperspect.a000414

Vingadassalom, D., Kolb, A., Mayer, C., Rybkine, T., Collatz, E., Podglajen, I., 2005. An unusual primary sigma factor in the Bacteroidetes phylum. Mol. Microbiol. 56, 888–902. https://doi.org/10.1111/j.1365-2958.2005.04590.x

Vinogradov, E., Paul, C.J., Li, J., Zhou, Y., Lyle, E.A., Tapping, R.I., Kropinski, A.M., Perry, M.B., 2004. The structure and biological characteristics of the Spirochaeta aurantia outer membrane glycolipid LGLB. Eur. J. Biochem. 271, 4685–4695. https://doi.org/10.1111/j.1432-1033.2004.04433.x

Wienhausen, G., Noriega-Ortega, B.E., Niggemann, J., Dittmar, T., Simon, M., 2017. The exometabolome of two model strains of the Roseobacter group: A marketplace of microbial metabolites. Front. Microbiol. 8, 1–15. https://doi.org/10.3389/fmicb.2017.01985

Wushke, S., Fristensky, B., Zhang, X.L., Spicer, V., Krokhin, O. V, Levin, D.B., Stott, M.B., Sparling, R., 2018. A metabolic and genomic assessment of sugar fermentation profiles of the thermophilic Thermotogales, Fervidobacterium pennivorans. Extremophiles 22, 965–974. https://doi.org/10.1007/s00792-018-1053-4

Zhang, J.Y., Deng, X.M., Li, F.P., Wang, L., Huang, Q.Y., Zhang, C.C., Chen, W.L., 2014. RNase E forms a complex with polynucleotide phosphorylase in cyanobacteria via a cyanobacterialspecific nonapeptide in the noncatalytic region. RNA 20, 568–579. https://doi.org/10.1261/rna.043513.113

Zhang, S., Bryant, D.A., 2011. The tricarboxylic acid cycle in cyanobacteria. Science (80-.). 334, 1551–1553. https://doi.org/10.1126/science.1210858