Biosynthetic potential of the global ocean microbiome

Lucas Paoli1, Hans-Joachim Ruscheweyh1,18, Clarissa C. Forneris2,16, Florian Hubrich2,18, Satria Kautsar3, Agneya Bhushan2, Alessandro Lotti2, Quentin Clayssen3, Guillem Salazar4, Alessio Milanese1, Charlotte I. Carlström1, Chrysia Papadopoulou4, Daniel Gehrig2, Mikhail Karasikov4,5,6, Harun Mustafa4,5,6, Martin Larralde2, Laura M. Carroll7, Pablo Sánchez8, Ahmed A. Zayed9, Dylan R. Cronin9, Silvia G. Acinas9, Peer Bork10,11,12, Chris Bowler2,13, Tom O. Delmont13,14, Josep M. Gasol15, Alvar D. Gossert15, André Kahles4,5,6, Matthew B. Sullivan8,16, Patrick Wincker13,14, Georg Zeller8, Serina L. Robinson2,17,18, Jörn Piel2,18 & Shinichi Sunagawa1

Natural microbial communities are phylogenetically and metabolically diverse. In addition to underexplored organismal groups, this diversity encompasses a rich discovery potential for ecologically and biotechnologically relevant enzymes and biochemical compounds. However, studying this diversity to identify genomic pathways for the synthesis of such compounds and assigning them to their respective hosts remains challenging. The biosynthetic potential of microorganisms in the open ocean remains largely uncharted owing to limitations in the analysis of genome-resolved data at the global scale. Here we investigated the diversity and novelty of biosynthetic gene clusters in the ocean by integrating around 10,000 microbial genomes from cultivated and single cells with more than 25,000 newly reconstructed draft genomes from more than 1,000 seawater samples. These efforts revealed approximately 40,000 putative mostly new biosynthetic gene clusters, several of which were found in previously unsuspected phylogenetic groups. Among these groups, we identified a lineage rich in biosynthetic gene clusters (Candidatus Eudoremicrobiaceae) that belongs to an uncultivated bacterial phylum and includes some of the most biosynthetically diverse microorganisms in this environment. From these, we characterized the phosphopentin and pythonamide pathways, revealing cases of unusual bioactive compound structure and enzymology, respectively. Together, this research demonstrates how microbiomics-driven strategies can enable the investigation of previously undescribed enzymes and natural products in underexplored microbial groups and environments.

Microorganisms drive global biogeochemical cycles, support food webs, and underpin the health of animals and plants. Their immense phylogenetic, metabolic and functional diversity represents a rich discovery potential for new taxa, enzymes and biochemical compounds, including natural products. In environmental communities, such molecules confer microorganisms with diverse physiological and ecological functions ranging from communication to competition. Addition to their original functions, these natural products and their genetically encoded production pathways include examples used for biotechnological and therapeutic applications. The identification of such pathways and compounds has largely been facilitated by studying cultivable microorganisms. However, taxonomic surveys of natural environments have revealed that the vast majority of microbial life has not yet been cultivated. This cultivation bias has limited our ability to tap into much of the microbially encoded functional diversity.
To overcome these limitations, technological advances over the past decade have enabled researchers to directly (that is, without previous cultivation) sequence pieces of microbial DNA from whole communities (metagenomes) or single cells. The possibility to assemble such pieces into larger genomic fragments and to reconstruct several metagenome-assembled genomes (MAGs) or single amplified genomes (SAGs), respectively, has opened new paths to the previously taxon-centric investigation of microorganisms (that is, microbial communities and their genetic material in a given environment)\(^ {10-12}\). Indeed, recent surveys have vastly extended the phylogenomic representation of microbial diversity on Earth\(^ {13-15} \) and revealed that most of the functional diversity in different microorganisms had previously not been captured by reference genome sequences (REFs) from cultured microorganisms\(^ {16} \). The ability to place uncovered functional diversity into the host genomic (that is, genome-resolved) context has been critical to predict yet uncharacterized microbial lineages that putatively encode new natural products\(^ {15,16} \) or to trace such compounds to their original producers\(^ {17} \).

A combinatorial approach of metagenomic and single-cell genomic analyses, for example, led to the recognition of *Candidatus Entothioneola*, a group of metabolically rich, sponge-associated bacteria, as producers of multiple classes of candidate drugs\(^ {18} \). However, despite recent attempts to establish genome-resolved explorations of various microorganisms\(^ {19-21} \), for the ocean—the largest ecosystem on Earth—over two-thirds of global metagenomic data still remain unaccounted for\(^ {22-24} \).

Thus, the biosynthetic potential of the ocean microbiome in general and its potential as a reservoir of new enzymology and natural products specifically remain largely underexplored.

To examine the biosynthetic potential of the ocean microbiome at the global scale, we first integrated ocean microbial genomes obtained using cultivation-dependent and cultivation-independent methods to establish an extensive phylogenomic and gene functional database. By mining this database, we uncovered a diverse array of biosynthetic gene clusters (BGCs), the majority of them from yet uncharacterized gene cluster families (GCFs). We further identified an uncharted family of bacteria that display the highest known diversity of BGCs in the open oceans to date. We selected two ribosomally synthesized and post-translationally modified peptide (RiPP) pathways on the basis of their genetic dissimilarity to currently known ones for experimental validation. Functional characterization of these pathways revealed examples of unexpected enzymology as well as a structurally unusual compound with protease inhibitory activity.

**Phylogenomic representation of the ocean microbiome**

We first sought to establish a global genome-resolved data resource focusing on its bacterial and archaeal constituents. To this end, we aggregated metagenomic data along with contextual information from 1,038 ocean water samples from 215 globally distributed sampling sites (latitudinal range = 141.6\(^ {\circ} \)) and several depth layers (from 1 to 5,600 m deep, covering epipelagic, mesopelagic and bathypelagic zones)\(^ {25-27} \) (Fig. 1a, Extended Data Fig. 1a and Supplementary Table 1). In addition to providing broad geographical coverage, these size-selectively filtered samples enabled us to compare different components of the ocean microbiome, including virus-enriched (<0.2 μm), prokaryote-enriched (0.2–3 μm), particle-enriched (0.8–20 μm) and virus-depleted (>0.2 μm) communities.

Using this dataset, we reconstructed a total of 26,293 predominantly bacterial and archaeal MAGs (Fig. 1b and Extended Data Fig. 1b). We generated these MAGs on the basis of assemblies from individual, rather than pooled, metagenomic samples to prevent the collapsing of natural sequence variations across samples from different locations or time points (Methods). Furthermore, we grouped genomic fragments on the basis of their abundance correlation across large numbers of samples (between 58 and 610 samples, depending on the survey; Methods). We found this to be a computationally intensive, yet important step\(^ {28} \), that was omitted in several large-scale MAG reconstruction efforts\(^ {10,19,25} \) and substantially improved both the number (mean, 2.7 times) and quality score (mean, +20%) of genomes reconstructed from the ocean metagenomes studied here (Extended Data Fig. 2a and Supplementary Information). Overall, these efforts have increased the number of ocean water microbial MAGs by a factor of 4.5 (when counting high-quality MAGs only) compared with the most comprehensive MAG resource available to date\(^ {26} \) (Methods). This set of newly created MAGs was then combined with 830 manually curated MAGs\(^ {27} \), 5,969 SAGs\(^ {28} \) and 1,707 REFs\(^ {29} \) of marine bacteria and archaea into a combined collection of 34,799 genomes (Fig. 1b).

We next evaluated the newly established resource for its improved ability to represent ocean microbial communities and to assess the impact of integrating different genome types. On average, we found that it captured about 40–60% of ocean metagenomic data (Fig. 1c), corresponding to a two- to threefold increase in coverage with a more consistent representation across depths and latitudes compared with previous reports based solely on MAGs\(^ {30} \) or SAGs\(^ {30} \). Furthermore, to obtain a systematic measure of the taxonomic diversity within the established collection, we annotated all genomes using the Genome Taxonomy Database (GTDB) Toolkit (Methods) and clustered them using a 95% whole-genome average nucleotide identity cut-off\(^ {31} \) to define 8,304 species-level clusters (species). Two thirds of these species (including new clades) were previously not represented in the GTDB and 2,790 of them were uncovered by MAGs reconstructed in this study (Fig. 1d). Moreover, we found that the different genome types were highly complementary, with 55%, 26% and 11% of the species being exclusively composed of MAGs, SAGs and REFs, respectively (Fig. 1e).

Furthermore, MAGs covered all 49 phyla detected in the water column, whereas SAGs and REFs represented only 18 and 11 of them, respectively. However, SAGs better represented the diversity of the most abundant clades (Extended Data Fig. 3a), such as the order Pelagibacterales (SAR11), with nearly 1,300 species covered by SAGs as opposed to only 390 by MAGs. Notably, REFs rarely overlapped with either MAGs or SAGs at the species level and represented >95% of the approximately 1,000 genomes that were not detected in the set of open ocean metagenomes studied here (Methods), mostly owing to representatives that were isolated from other types of marine samples (such as sediments or host-associated). To enable its broad use by the scientific community, this ocean genomic resource—which also includes unbinned fragments (for example, from predicted phages, genomic islands and fragments of genomes with insufficient data for MAG reconstruction)—can be accessed alongside taxonomic and gene functional annotations as well as contextual environmental parameters at the Ocean Microbiomics Database (OMD; https://microbiomics.io/ocean/).

**Biosynthetic potential of the global ocean microbiome**

Next, we set out to investigate the richness and the degree of novelty of the biosynthetic potential in the open ocean microbiome. To this end, we first used antiSMASH on all of the MAGs, SAGs and REFs detected in the set of 1,038 ocean metagenomes (Methods) to predict a total of 39,055 BGCs. We then clustered them into 6,907 non-redundant gene cluster families (GCCs) with protease inhibitory activity.
metagenomes, possibly indicative of widespread microbial adaptations to the ocean environment, including resistance to reactive oxygen species, oxidative and osmotic stress or uptake of iron (Supplementary Information). This functional diversity contrasted with recent analyses of -1.2 million BGCs from any of the -190,000 genomes deposited in the NCBI RefSeq database (BiG-FAM/RefSeq, hereafter RefSeq)29 that showed a dominance of non-ribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) BGCs (Supplementary Information). We also found that 44 (29%) GCCs were only remotely related to any RefSeq BGCs (\(d_{\text{RefSeq}} > 0.4\); Fig. 2a and Methods), and 53 (35%) GCCs were encoded only in MAGs, highlighting the potential for discovery of previously undescribed chemistry within the OMD. The OMD improves the genomic representation (mapping rates of metagenomic reads; Methods) of ocean microbial communities by a factor of two to three compared with previous reports based solely on SAGs (GORG)20 or MAGs (GEM)16, with a more consistent representation across depth and latitudes. \(<0.2, n = 151; 0.2–0.8, n = 67; 0.2–3, n = 180; 0.8–20, n = 30; >0.2, n = 610; <30°, n = 132; 30–60°, n = 73; >60°, n = 42; EPI, n = 174; MES, n = 45; BAT, n = 28. d. Grouping the OMD into species-level (95% average nucleotide identity) clusters identified a total of around 8,300 species, over half of which were previously uncharacterized based on taxonomic annotations using the GTDB (release 89)30. e. A breakdown of the species by genome type reveals a high complementarity of MAGs, SAGS and REFS in capturing the phylogenomic diversity of the ocean microbiome. Specifically, 55%, 26% and 11% of the species were specific to MAGs, SAGS and REFS, respectively. BATS, Bermuda Atlantic Time-series; GEM, Genomes from Earth’s Microbiomes; GORG, Global Ocean Reference Genomes; HOT, Hawaiian Ocean Time-series.
Identification of undescribed BGC-rich lineages

To complement the survey of the biosynthetic potential of the ocean microbiome, we sought to map its phylogenomic distribution and identify new BGC-rich clades. To this end, we placed the ocean microbiomes in the standardized bacterial and archaeal phylogeny backbone trees to reveal the extent of the phylogenomic coverage of the OMD. Clades without any genome in the OMD are coloured grey. The number of BGCs corresponds to the highest number of predicted BGCs per genome in a given clade. For visualization, the last 15% of the nodes were collapsed. The arrows denote BGC-rich clades (>15 BGCs) with the exception of Myxobacteroides, Gordonia (next to Rhodococcus) and Crocosphaera (next to Synechococcus).

d, An unknown species of 'Ca. Eremiobacterota' displayed the highest biosynthetic diversity (Shannon index based on natural product types). Each bar represents the genome with the highest number of BGCs within a species. TIPKS, type I PKS; T2/3PKS, type II and III PKS.

RiPPs, terpenes and other putative natural products. c, All genomes in the OMD detected across 1,038 ocean metagenomes were placed onto the GTDB backbone trees to reveal the extent of the phylogenomic coverage of the OMD. Clades without any genome in the OMD are coloured grey. The number of BGCs corresponds to the highest number of predicted BGCs per genome in a given clade. For visualization, the last 15% of the nodes were collapsed. The arrows denote BGC-rich clades (>15 BGCs) with the exception of Myxobacteroides, Gordonia (next to Rhodococcus) and Crocosphaera (next to Synechococcus).

Identification of undescribed BGC-rich lineages

To complement the survey of the biosynthetic potential of the ocean microbiome, we sought to map its phylogenomic distribution and identify new BGC-rich clades. To this end, we placed the ocean microbiomes in the standardized bacterial and archaeal phylogenomic trees of the GTDB\(^5\), and overlayed the putative biosynthetic pathways that they encode (Fig. 2c). We readily detected in ocean water samples (Methods) several BGC-rich clades (representatives with >15 BGCs) that are either well known for their biosynthetic potential, such as Cyanobacteria (Synechococcus) and Proteobacteria (such as Tistrella)\(^2,3\), or that have recently garnered attention for their natural products, such as Myxococca (Sardaracinaceae), Rhodococcus and Planctomycetota\(^34,36\). Interestingly, we found within
these clades several previously unexplored lineages. For example, those species with the richest biosynthetic potential within the phyla Planctomycetota and Myxococcota belonged to an uncharacterized candidate order and genus, respectively (Supplementary Table 3). Overall, this shows that the OMD provides access to previously uncharted phylogenomic information, including for microorganisms that may represent new targets for the discovery of enzymes and natural products.

We further characterized BGC-rich clades not only by counting the maximum number of BGCs encoded by their members, but also by assessing the diversity of these BGCs, which accounts for the frequency of different candidate natural product types (Fig. 2c and Methods). We found that the most biosynthetically diverse species were represented by bacterial MAGs exclusively reconstructed in this study. These bacteria belong to the uncultivated phylum 'Candidatus Eremiobacterota', which has remained largely unexplored except in a few genomic studies. Notably, 'Ca. Eremiobacterota' spp. have been analysed only from terrestrial environments and have not been known to include any BGC-rich representatives. Here we initially reconstructed eight MAGs from the same species (with a nucleotide identity of >99%) from deep (between 2,000 m and 4,000 m) and particle-enriched (0.8–20 μm) ocean metagenomes collected by the Malapais expedition. Accordingly, we propose that this species is named 'Candidatus Eremiobacterota malapaisi', after the nereid (sea nymph) of fine gifts in Greek mythology and the expedition. 'Ca. E. malapaisi' had no previously known relatives below the order level based on phylogenomic annotation and therefore belongs to a new bacterial family for which we propose 'Ca. E. malapaisi' as the type species and 'Ca. E. Eremiobacterota' as its official name (Supplementary Information). The short-read metagenomic reconstruction...
of ‘Ca. E. malaspinii’ draft genomes was corroborated by ultra-low input, long-read metagenomic sequencing of one sample and targeted assembly (Methods) into a single 9.63 Mb linear chromosome with a 75 kb repeat as the only remaining ambiguity.

To establish a phylogenetic context for this species, we searched for closely related species through targeted genome reconstructions in additional eukaryote-enriched metagenomic samples from the Tara Oceans expedition. In brief, we aligned metagenomic reads to ‘Ca. E. malaspinii’-related genome fragments and assumed increased recruitment rates in a given sample to be indicative of the presence of additional relatives (Methods). As a result, we recovered 10 MAGs, and the combined set of 19 MAGs represents five species across three genera within the newly defined family (that is, ‘Ca. Eudoremicrobiaceae’). After manual inspection and quality control (Extended Data Fig. 6 and Supplementary Information), we found that ‘Ca. Eudoremicrobiaceae’ spp. representatives have larger genomes (8 Mb) and a richer biosynthetic potential (ranging from 14 to 22 BGCs per species) compared with members of other ‘Ca. Eremiobacterota’ clades (up to 7 BGCs) (Fig. 3a–c).

Exploring the abundance and distribution of ‘Ca. Eudoremicrobiaceae’, we found that its members are prevalent in most oceanic basins as well as throughout the water column (Fig. 3d). Locally, they account for up to 6% of ocean microbial communities, making them a numerically substantial component of the global ocean microbiome. Furthermore, we found the relative abundances of ‘Ca. Eudoremicrobiaceae’ spp. and their BGCs expression levels to be the highest in eukaryote-enriched fractions (Fig. 3c and Extended Data Fig. 7), suggesting possible interactions with particulate matter, including planktonic organisms. This observation, and the homology of some ‘Ca. Eudoremicrobiaceae’ BGCs to known pathways producing cytotoxic natural products could suggest a predatory behaviour (Supplementary Information and Extended Data...
New enzymes and natural products

We finally sought to experimentally validate the promising prospects of our microbiomics-driven work for the discovery of new pathways, enzymes and natural products. Among the different BGC classes, RfIP pathways are known to encode a wealth of chemical and functional diversity owing to the various modifications installed post-transcriptionally on a core peptide by maturase enzymes. We therefore selected two ‘Ca. Eudoremicrobiaceae’ RfIP BGCs (Fig. 3b and 4a–e) that were predicted to produce novel metabolites on the basis of their dissimilarity to any known BGC (ΔI_MIBiG and ΔI_RefSeq above 0.2).

The first RfIP pathway (ΔI_MIBiG = 0.41, ΔI_RefSeq = 0.29) was found only in the deep-ocean species ‘Ca. E. malapinii’ and encodes a precursor peptide modified by a sole maturase (Fig. 4a,b). In this maturase, we found a single functional domain homologous to the dehydration domain of lantipeptide synthetases, which normally catalyses phosphorylation and subsequent elimination [Supplementary Information]. We therefore predicted the modifications of the precursor peptide to include such a two-step dehydration. However, using tandem mass spectrometry (MS/MS) and nuclear magnetic resonance spectroscopy (NMR), we identified a poly-phosphorylated linear peptide (Fig. 4c). Although this was unexpected, we found several lines of evidence supporting that it is the final product: the absence of dehydration in two different heterologous hosts as well as in vitro assays, the identification of mutated key residues in the dehydration catalytic site of the maturase, which were consistently found in all reconstructed ‘Ca. E. malapinii’ genomes (Extended Data Fig. 9 and Supplementary Information), and finally the bioactivity of the phosphorylated product rather than the chemically synthesized, dehydrated form (Fig. 4d). Indeed, we found that it displayed low-micromolar protease inhibitory activity against neutrophil elastase within a concentration range (IC50 = 14.3 μM) comparable to other relevant natural products, although the ecological role of this unusual natural product remains to be elucidated. On the basis of these results, we propose naming this pathway ‘phospeptin’.

The second case represents a complex RfIP pathway specific to ‘Ca. Eudoremicrobiaceae’ spp. (ΔI_MIBiG = 0.46, ΔI_RefSeq = 0.33) that is predicted to encode a proteins natural product (Fig. 4e). These pathways are of particular biotechnological interest owing to the expected density and diversity of unusual chemical modifications installed by enzymes encoded in relatively short BGCs. We found that this protein differs from previously characterized ones as it lacks both the NX̂N core motif of polytheonamides and the lanthionine rings of landonamides. To overcome the limitations of common heterologous expression models, we used them along a non-standard Microvirga aerodenitrificans system to characterize the four maturase enzymes of the pathway (Methods). Using a combination of MS/MS, isotope labelling and NMR, we found that these maturases install up to 21 modifications, including 1- to 3-amino acid epimerization, hydroxylation as well as -methyltransferase family members47, 48. Thus, together with the report of a non-canonical bioactive poly-phosphorylated RfIP, our results demonstrate the resource-intensive, yet critical value of synthetic biology efforts to fully uncover the functional richness, diversity and unusual architectures of biochemical compounds.

Conclusions

Here we have demonstrated the extent of microbially encoded biosynthetic potential and its genomic context in the global ocean microbiome, facilitating future research by making the generated resources available to the scientific community (https://microbiomics.io/ocean/). We found that the majority of both its phylogenomic and functional novelty is accessible only through the reconstruction of MAGs and SAGs, particularly in underexplored microbial communities, which could direct future bioprospecting efforts. Although we focused here on ‘Ca. Eudoremicrobiaceae’ as a particularly biosynthetically ‘talented’ lineage, many of the predicted BGCs within other uncovered microbial groups are likely to encode previously undescribed enzymology that produces compounds with ecologically and/or biotechnologically relevant activities.

Online content

Any methods, additional references. Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04862-3.

1. Hug, L. A. et al. A new view of the tree of life. Nat. Microbiol. 1, 16048 (2016).
2. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J. Nat. Prod. 83, 770–803 (2020).
3. Adrio, J. L. & Demain, A. L. Microbial enzymes: tools for biotechnological processes. Biocatalysis 4, 117–139 (2019).
4. Medema, M. H., de Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. Nat. Rev. Genet. 22, 553–571 (2021).
5. Cavicchioli, R. et al. Scientists’ warning to humanity: microorganisms and climate change. Nat. Rev. Microbiol. 17, 569–586 (2019).
6. Pye, C. R., Berin, M. J., Loxley, R. S., Gerwick, W. H. & Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. Proc. Natl Acad. Sci. USA 114, 5601–5606 (2017).
7. Davies, J. Specialized microbial metabolites: functions and origins. J. Antibiot. 66, 361–364 (2013).
8. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically novel uncultured microbial cells dominate earth microbiomes. mSystems 3, e00055-18 (2018).
9. Robinson, S. L., Piet, J. & Sunagawa, S. A roadmap for metagenomic enzyme discovery. Nat. Prod. Rep. 38, 1984–2003 (2021).
10. Venter, J. C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74 (2004).
11. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65 (2010).
12. Sunagawa, S. et al. Structure and function of the global ocean microbiome. Science 348, 1293359 (2015).
13. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36, 996–1004 (2018).
14. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics and metatranscriptomics approach for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 26, 1612–1625 (2016).
15. Crites-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature 558, 440–444 (2018).
16. Nayfach, S. et al. A genomic catalog of Earth’s microbiomes. Nat. Biotechnol. 39, 498–509 (2021).
17. Zan, J. et al. A microbial factory for defensive kahalalides in a tripartite marine symbiosis. Science 364, eaww6732 (2019).
18. Wilson, M. C. et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58–62 (2014).
19. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2018).
20. Pachiaudi, M. G. et al. Charting the complexity of the marine microbiome through single-cell genomics. *Cell* **179**, 1623–1635 (2019).
21. Biller, S. J. et al. Marine microbial metagenomes sampled across space and time. *Sci. Data* **5**, 180776 (2018).
22. Salazar, G. et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* **179**, 1068–1083 (2019).
23. Acinas, S. G. et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun. Biol.* **4**, 604 (2021).
24. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
25. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 533–542 (2017).
26. Delmont, T. O. et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* **3**, 804–813 (2018).
27. Klemetsen, T. et al. The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.* **46**, D692–D699 (2018).
28. Jain, C., Rodriguez-R. L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
29. Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. Big-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.* **49**, D490–D497 (2021).
30. Kautsar, S. A. et al. MiBIG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
31. Klassen, J. L. & Currie, C. R. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMCC Genom.* **13**, 14 (2012).
32. Timmermans, M. L., Paudel, Y. P. & Ross, A. C. Investigating the biosynthesis of natural products from marine proteobacteria: a survey of molecules and strategies. *Mar. Drugs* **15**, 235 (2017).
33. Shah, S. A. A. et al. Structural diversity, biological properties and applications of natural products from cyanobacteria. A review. *Mar. Drugs* **15**, 354 (2017).
34. Wiegard, S. et al. Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. *Nat. Microbiol.* **5**, 106–140 (2020).
35. Ceniceros, A., Dijkhuizen, L., Petrusma, M. & Medema, M. H. Genome-based exploration of the specialized metabolic capacities of the genus Rhodococcus. *BMC Genom.* **18**, 593 (2017).
36. Gregory, K., Salvador, L. A., Akbar, S., Adakiloph, B. I. & Stevens, D. C. Survey of biosynthetic gene clusters from sequenced myxobacteria reveals unexplored biosynthetic potential. *Microorganisms* **7**, 181 (2019).
37. Woodcroft, B. J. et al. Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
38. Ward, L. M., Cardona, T. & Holland-Moritz, H. Evolutionary implications of anoxicogenic phototrophy in the bacterial phylum Eremiobacterota (WPS-2). *Front. Microbiol.* **10**, 1658 (2019).
39. Ji, M. et al. Candidatus Eremiobacterota, a metabolically and phylogenetically diverse terrestrial phylum with acid-tolerant adaptations. *ISME J.* **15**, 2692–2707 (2021).
40. Sunagawa, S. et al. Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
41. Pérez, J., Moraleda-Muñoz, A., Marcos-Torres, F. J. & Muñoz-Dorado, J. Bacterial predation: 75 years and counting! *Environ. Microbiol.* **18**, 766–779 (2016).
42. Arison, P. G. et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30**, 108–160 (2013).
43. Dong, S. H. et al. The enteroococcal cytolsin synthetase has an unanticipated lipid kinase fold. *Elife* **4**, e07607 (2015).
44. Ahmad, S. et al. The natural polypeptides as significant elastase inhibitors. *Front. Pharmacol.* **11**, 688 (2020).
45. Freeman, M. F. et al. Metagenome mining reveals polytheonamines as posttranslationally modified ribosomal peptides. *Science* **338**, 387–390 (2012).
46. Bösch, N. M. et al. Landomimides: antiviral ornithine-containing ribosomal peptides discovered through genome mining. *Angew. Chem. Int. Ed.* **59**, 1763–1768 (2020).
47. Motamedi, H. et al. Characterization of methyltransferase and hydroxylase genes involved in the biosynthesis of the immunosuppressants FK506 and FK520. *J. Bacteriol.* **179**, 5243–5248 (1998).
48. Labby, K. J., Watts, S. G. & Garneau-Tildesley, S. Interrupted adenylation domains: unique bifunctional enzymes involved in nonribosomal peptide biosynthesis. *Nat. Prod. Rep.* **32**, 641–653 (2015).
49. Song, H. & Naismith, J. H. Enzymatic methylation of the amide bond. *Curr. Opin. Struct. Biol.* **65**, 79–88 (2020).
50. van der Velden, N. S. et al. Autocatalytic backbone N-methylation in a family of ribosomal peptide natural products. *Nat. Chem. Biol.* **13**, 833–835 (2017).
51. Miller, F. S. et al. Conformational rearrangements enable iterative backbone N-methylation in RIP biosynthesis. *Nat. Commun.* **12**, 5355 (2021).
52. Chatterjee, J., Rechenmacher, F. & Kessler, H. N-methylation of peptides and proteins: an important element for modulating biological functions. *Angew. Chem. Int. Ed.* **52**, 254–268 (2013).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

Metagenomic data selection, assembly and binning

Metagenomic datasets from major oceanographical surveys and time-series studies with sufficient sequencing depth were included to maximize the coverage of global ocean microbial communities across ocean basins, depth layers and time. These datasets (Supplementary Table 1 and Fig. 1) included metagenomes from samples collected by Tara Oceans (virus-enriched, $n = 190$) and prokaryote-enriched, $n = 180$) BioGEOTRACES expeditions ($n = 480$), the Hawaiian Ocean Time-series (HOT, $n = 68$), the Bermuda-Atlantic Time-series Study (BATS, $n = 62$) and the Malaspina expedition ($n = 58$). Sequencing reads from all metagenomes were quality filtered using BBMap (v.38.71) by removing sequencing adapters from the reads, removing reads that mapped to quality control sequences (PhI genome) and discarding low quality reads using the parameters trimq = 14, maq = 20, maxns = 0 and minlength = 45. Downstream analyses were performed on quality-controlled reads or if specified, merged quality-controlled reads (bbmerge.sh target = 40, mindepth = 0) before they were assembled with metaSPades (v.3.11.0 or v.3.12.1 if required)\(^{35}\). The resulting scaffolded contigs (hereafter scaffolds) were finally filtered by length ($\geq 1$ kb).

The 1,038 metagenomic samples were grouped into several sets and, for each sample set, the quality-controlled metagenomic reads from all samples were individually mapped against the scaffolds of each sample, resulting in the following numbers of pairwise readset to scaffold mappings: Tara Oceans virus-enriched ($190 \times 190$), prokaryote-enriched ($180 \times 180$), BioGEOTRACES, HOT and BATS ($610 \times 610$) and Malaspina ($58 \times 58$). Mapping was performed using the Burrows–Wheeler–Aligner (BWA) (v.0.7.17-r1188)\(^{34}\), allowing the reads to map to secondary sites (with the -a flag). Alignments were filtered to be at least 45 bases in length, with an identity of $\geq 97\%$ and covering $\geq 80\%$ of the read sequence. The resulting BAM files were processed using the jgi_summarize_bam_contig_depths script of MetaBAT2 (v.2.12.1)\(^{36}\) to provide within- and between-sample coverages for each scaffold. The scaffolds were finally binned by running MetaBAT2 on all samples individually with parameters --minContig 2000 and --maxEdges 500 for increased sensitivity. We used MetaBAT2 in lieu of ensemble binning approaches as it was shown to be the best performing single bin in independent benchmarks\(^{39,40}\) and was found to be 10 to 50 times faster than other usual binners\(^{31}\). To test for the effect of abundance correlation, a randomly selected subsample of the metagenomes (10 from each of the two Tara Ocean datasets, 10 from BioGEOTRACES, five from each time-series and five from Malaspina) was additionally binned using only within-sample coverage information (Supplementary Information).

Selection of additional genomes

Additional (external) genomes were included in downstream analyses, namely 830 manually curated MAGs from a subset of the Tara Oceans dataset\(^{32,41}\), 5,287 SAGs from the GORG dataset\(^{42}\), as well as 1,707 isolate REFs and 682 SAGs from the MAR databases (MarDB v.4)\(^{32,43}\). For the MarDB dataset, genomes were selected on the basis of the available metadata if the sample type matched the following regular expression: ['(Ss)ingle.?', '[Cc]ell[Cc]ulture'[Il]l]olate'.

Quality evaluation of metagenomic bins and external genomes

The quality of each metagenomic bin and external genome was evaluated using both the ‘lineage workflow’ of CheckM (v.1.0.13) and Anvi’o (v.5.5.0)\(^{36,44}\). Metagenomic bins and external genomes were retained for downstream analyses if either CheckM or Anvi’o reported a completeness/contamination of $\geq 50\%$ and a contamination/redundancy of $\leq 10\%$. These metrics were then aggregated into a mean completeness (mcpl) and a mean contamination (mctn) value to classify genome quality according to community standards\(^{48}\) as follows: high quality: mcpl $\geq 90\%$ and mctn $\leq 5\%$; good quality: mcpl $\geq 70\%$ and mctn $\leq 10\%$; medium quality: mcpl $\geq 50\%$ and mctn $\leq 10\%$; fair quality: mcpl $\leq 90\%$ or mctn $\geq 10\%$. Filtered genomes were further attributed quality scores ($Q$ and $Q'$) as follows: $Q = mcpl - 5 \times mctn$ and $Q' = mcpl - 5 \times mctn + mctn \times (strain\ heterogeneity)/100 + 0.5 \times \log(N_{\text{dis}})$ (as implemented in dRep\(^{49}\)).

Species-level clustering of the genome collection and comparison to other resources

To allow for comparative analyses between different data resources and genome types (MAGs, SAGs and REFs), the full set of 34,799 genomes was dereplicated on the basis of both whole-genome average nucleotide identity (ANI) using dRep (v.2.5.4)\(^{41}\) with a 95% ANI threshold\(^{28,62}\) (comp 0 -con 1000 -sa 0.95 -nc 0.2) and single-copy marker genes using SpecI\(^{63}\), providing species-level clustering of the genomes. A representative genome was selected on the maximum quality score defined above ($Q'$) for each of the dRep clusters, which were considered to be a proxy for species membership.

To estimate mapping rates, all 1,038 metagenomic readsets were mapped against the 34,799 genomes included in the OMD using BWA (v.0.7.17-r1188, -a). Quality-controlled reads were mapped in single-end mode and the resulting alignments were filtered to keep only alignments of $\geq 45$ bp in length and with an of identity $\geq 95\%$. The per-sample mapping rate is the percentage of reads that remained after filtering divided by the total number of quality-controlled reads. Using the same method, each of the 1,038 metagenomes was downsamples to 5 million inserts (Extended Data Fig. 1c) and mapped to the GORG SAGs within the OMD and all of the GEM\(^{46}\). The number of MAGs in the GEM catalogue\(^{46}\) that were recovered from ocean waters was determined on the basis of a keyword query on the source of metagenomes, selecting for ocean water samples (as opposed to marine sediments, for example). Specifically, we selected ‘aquatic’ as the ‘ecosystem_category’, ‘marine’ as the ‘ecosystem_type’ and filtered ‘habitat’ for ‘deep ocean’, ‘marine’, ‘marine oceanic’, ‘pelagic marine’, ‘seawater’, ‘marine’, ‘seawater’, ‘surface seawater’, ‘surface seawater’. This resulted in 5,903 MAGs (734 high quality), distributed across 1,823 OTUs (here, species).

Taxonomic and functional genome annotation

Prokaryotic genomes were taxonomically annotated using GTDB-Tk (v.1.0.2)\(^{44}\) with the default parameters against the GTDB r89 release\(^{33}\). Anvi’o was used to identify eukaryotic genomes on the basis of domain prediction and completion of $\geq 50\%$ and redundancy of $\leq 10\%$. The taxonomic annotation of a species is defined as the one of its representative genome. Excluding eukaryotes (148 MAGs), each genome was functionally annotated by first calling complete genes using prokka (v.1.14.5)\(^{65}\) with the ‘Archaea’ or ‘Bacteria’ parameter specified as appropriate, which also reported non-coding genes and CRISPR regions, among other genomic features. The predicted genes were annotated by identifying universal single-copy marker genes (usMGs) with fetchMGs (v.1.2)\(^{44}\), assigning orthologous groups with emapper (v.2.0.1)\(^{32}\) based on eggNOG (v.5.0)\(^{46}\) and performing queries against the KEGG database (release 2020-02-10)\(^{49}\). This last step was performed by aligning the proteins to the KEGG database using DIAMOND (v.0.9.30)\(^{50}\) with a query and subject coverage of $\geq 70\%$. The results were further filtered on the basis of the bitscore being $\geq 50\%$ of the maximum expected bitscore (reference against itself) according to the NCBI Prokaryotic Genome Annotation Pipeline\(^{51}\). The gene sequences were additionally used as input to identify BGGS in the genomes using antiSMASH (v.5.1.0)\(^{52}\) with the default parameters and the different cluster blasts turned on. All genomes and annotations have been compiled along with contextual metadata into the OMD, which is available online (https://microbiomes.io/ocean/).

Gene-level profiling

Similar to the methods described previously\(^{21,22}\), we clustered the >56.6 million protein-coding genes from the bacterial and archaeal
Article

genes of the OMD at 95% identity and 90% coverage of the shorter gene using CD-HIT (v.4.8.1) to filter against the database. The longest sequence was selected as the representative gene of each gene cluster. The 1,038 metagenomes were then mapped to the database with BWA (-a) and the resulting BAM files were filtered to retain only alignments with a minimum identity of ≥95% and ≥45 bases aligned. Length-normalized gene abundance was calculated by first counting insertions from best unique alignments and then, for ambiguously mapped inserts, adding fractional counts to the respective target genes in proportion to their unique insert abundances.

Species-level profiling with mOTUs

The genomes in the extended OMD (augmented with additional MAGs from 'Ca. Eudoremiobacteria', see below) were added to the database (v.2.5.1) of the metagenomic profiling tool mOTUs to generate an extended mOTUs reference database. Only genomes with at least six out of the ten uscMGs in single copy were kept (23,528 genomes). The extension of the database resulted in 4,494 additional species-level clusters.

Clustering and selection of BGCs

All BGCs from MAGs, SAGs and REFs in the OMD (see above) were combined with the ones identified across all the metagenomic scaffolds (antiSMASH v.5.0, default parameters) and processed with BiG-SLICE (v.1.1) for feature (PFAM domains) extraction. On the basis of these features, we computed all-versus-all cosine distances between BGCs and clustered them (average linkage) into GCFs and GCCs, using a 0.2 and a 0.8 distance threshold, respectively. These thresholds are an adaptation of those previously used with Euclidean distances and a 0.8 distance threshold, respectively. These thresholds are an adaptation of those previously used with Euclidean distances (989 genomes; 95% REFs, 5% SAGs and 99.9% belonging to MarDB) contained within 644 mOTU clusters were not detected. This reflects the various additional marine isolation sources (most of the genomes not detected are associated with organisms isolated from, for example, sediments, marine hosts) of the MarDB genomes. To remain focused on the open ocean environment in this study, we excluded them from downstream analyses if they were not detected or not included in the extended mOTU database established in this study.

Abundance and prevalence of GCFs and GCCs

The metagenomic abundance of a BGC was estimated on the basis of distances to the respective database. These minimum distances were then averaged (mean) per GCF or GCC as appropriate. A GCF was considered to be novel if the distance to the database was above 0.2, which corresponds to (on average) the complete separation between the GCF and the reference. For GCCs, we selected 0.4, twice the GCF-defining threshold, to capture remote relationships with the reference.

Phylogenetic distribution of BGCs

The filtered set of BGCs encoded by genomes in the OMD (in scaffolds ≥5 kb and excluding MarDB REFs and SAGs that were not detected in the 1,038 metagenomes, see above) along with their predicted product classes were displayed on the GTDB bacterial and archaeal trees on the basis of the GTDBtk phylogenomic placement of the genomes (see above). We first reduced the data on a per-species basis, using the genome with most BGCs in that species as a representative. For visualization, the representatives were further binned along the tree and, similarly, for each binned clade, the genome containing the most BGCs was selected as representative. BGC-rich species (at least one genome with >15 BGCs) were further analysed by computing the Shannon diversity index of the product types encoded in these BGCs. Chemical hybrids and other complex BGCs (as predicted by antiSMASH) were considered to be from the same product type if all of the predicted product types were identical, irrespective of their order within the cluster (for example, a proteusin–bacteriocin hybrid is identical to bacteriocin–proteusin hybrid).

Long-read sequencing of 'Ca. Eudoremiobacterium'

Leftover DNA (an estimated 6 ng) from the sample Malaspina MP1648, corresponding to the biosample SAMN05421553 and matching the short-read Illumina metagenomic readset SRR3962772, was processed for an ultralow input PacBio sequencing protocol to produce a 20 Gb Hifi Pacbio metagenome using the PacBio kits SMRTbell gDNA Sample amplification kit (100–980–000) and the SMRTbell Express Template Prep kit 2.0 (100–938–900). In brief, the remaining DNA was sheared using a Covaris (g-TUBE, 52104), repaired and purified (ProNex beads). The purified DNA was then library prepared, amplified, purified (ProNex
Targeted binning of ‘Ca. Eremiobacterota’

After the reconstruction of the first two ‘Ca. Eremiobacterota’ MAGs, we identified six additional ones with ANI > 99% (these are included in Fig. 3) that were initially filtered out on the basis of contamination estimates (later identified as gene duplications, see below). We additionally recovered bins identified as ‘Ca. Eremiobacterota’ from a different study and used them along with the eight MAGs from our study as a reference for subsampled mapping (5 million reads) of metagenomic reads from 633 eukarote-enriched (>0.8 μm) samples using BWA (v.0.7.17-r1188, -a flag). On the basis of enriched specific mappings (after 95% alignment identity and 80% read coverage filtering), 10 metagenomes (expected coverage, ±5%) were selected for assembly and 49 additional metagenomes (expected coverage, ±1%) for abundance correlation. Using the same parameters as described above, these samples were binned and 10 additional ‘Ca. Eremiobacterota’ MAGs were recovered. These 16 MAGs (which excludes the two that were already in the database) bring the total number of genomes in the extended OMD to 34,515. The MAGs were assigned to taxonomic ranks on the basis of their genomic similarity and GTDB placement. The 18 MAGs were dereplicated using dRep into 5 species (within-species ANIs were >99%) and 3 genera (within-genus ANIs ranged between 85% and 94%) within the same family. Species representatives were manually selected on the basis of completeness, contamination and N50. Proposed naming is available in Supplementary Information.

Manual evaluation of ‘Ca. Eremiobacterota’ MAGs

To evaluate the completeness and contamination of ‘Ca. Eremiobacterota’ MAGs, we assessed the presence of uscMGs, in addition to lineage- and domain-specific single-copy marker gene sets used by CheckM and Anvi’o. The identification of duplications among 2 out of the 40 uscMGs was confirmed by phylogenetic reconstruction (see below) to rule out any potential contamination (which would have corresponded to 5% on the basis of these 40 marker genes). Additional inspection of the representative MAGs of the five ‘Ca. Eremiobacterota’ species confirmed low rates of contaminants in these reconstructed genomes on the basis of abundance correlation and sequence composition (Supplementary Information) using the Anvi’o interactive interface.

Phylogenomics of ‘Ca. Eudoremicrobiaceae’

For phylogenomic analyses, we selected the representative MAGs of the five ‘Ca. Eudoremicrobiaceae’ species, all ‘Ca. Eremiobacterota’ genomes available in GTDB (r89) and representatives of additional phyla (including UBP13, Armaminodota, Patescibacteria, Dormibacterota, Chloroflexota, Cyanobacteria, Actinobacteria and Planctomycetota). All of these genomes were annotated as previously described to extract single-copy marker genes and to annotate BGCs. GTDB genomes were retained on the basis of the completeness and contamination criteria mentioned above. The phylogenomic analysis was performed using the Anvi’o phylogenomics workflow. The tree was constructed with IQTREE (v.2.0.3) (default parameters and -bb 1000) and an alignment (MUSCLE, v.3.8.155) of 39 concatenated ribosomal proteins identified by Anvi’o, with positions trimmed for coverage in at least 50% of the genomes and using Planctomycetota as the outgroup based on the GTDB tree topology. Individual trees for the 40 uscMGs were built using the same tools and parameters.

Trait and lifestyle prediction of ‘Ca. Eudoremicrobiaceae’

We used Traitar (v.1.1.2) with the default parameters (phenotype, from nucleotides) to predict general microbial traits. We investigated the potential predatory lifestyle on the basis of a previously developed predatory index, which relies on the protein-coding gene content of a genome. Specifically, we used DIAMOND to compare the proteins from a genome to the OrthoMCL database (v.4) using the parameters –more-sensitive –id 25 –query-cover 70 –subject-cover 70 –top 20 and counted genes that matched predatory and non-predatory marker genes. The index is the difference between the number of predatory and non-predatory markers. As an additional control, we also analysed the genome of ‘Ca. Entotheneillia’ factor TSY1 on the basis of its similar characteristics to ‘Ca. Eudoremicrobiota’ (large genome size and bio- synthetic potential). We further tested a potential link between predatory and non-predatory marker genes with the biosynthetic potential of ‘Ca. Eudoremicrobiaceae’ and found at most one gene (from either type, that is, predatory/non-predatory, of marker genes) overlapping with BGCs, suggesting that BGCs do not confound the predatory signal. Additional annotations of the genomes to specifically investigate secretion systems, pilus and flagella were performed using TXSSCAN (v.1.0.2) for unordered replications.

Transcriptomic profiling of ‘Ca. E. taraoceanii’

Transcriptomic profiling was performed by mapping 623 metatranscriptomes from Tara Oceans prokaryote- and eukaryote-enriched fractions (v.0.7.17-r1188, -a flag) to the five representative ‘Ca. Eudoremicrobiaceae’ genomes. After 80% read coverage and 95% identity filtering, the BAM files were processed using FeatureCounts (v.2.0.1) (using the parameters featureCounts --primary -O --fraction -t CDS,tRNA -F GTF -g ID -p) to compute the number of inserts per gene. The resulting profiles were normalized to gene length and mOTU marker gene abundances (median length-normalized insert count of genes with insert count of >0) and log-transformed to obtain relative per-cell expression levels of each gene, which also accounts for between-samples differences in sequencing effort. Such ratios allow for comparative analyses by mitigating the issues of compositionality when working with relative abundance data. Only samples with >5 out of the 10 mOTU marker genes were considered for further analyses, ensuring that a large enough fraction of the genome was detected.

The normalized transcriptomic profiles of ‘Ca. E. taraoceanii’ were dimension-reduced using UMAP and the resulting representation was used for unsupervised clustering using HDBSCAN (see above) to identify expression states. The significance of differences between the identified clusters was tested by PERMANOVA in the original (non-reduced) distance space. Differential expression between these states was tested across 201 KEGG pathways identified in the genome (see above) and functional groups, namely, BGCs, secretion systems and flagellar genes from TXSSCAN, degradative enzymes (proteases and peptidases) from prokka and predatory and non-predatory markers from the predatory index. For each sample, we computed the median normalized expression for each category (note that BGC expression was itself computed as the median expression of the biosynthetic genes of that BGC) and tested for significance (FDR-corrected Kruskal–Wallis test) across the different states.

Experimental validation of a novel phosphorylated RiPP pathway (‘Ca. E. malaspinii’, HLJLDLBE BGC 75.1)

Materials for heterologous expression. Synthetic genes were purchased from GenScript and PCR primers were ordered from Microsynth. Phusion polymerase from Thermo Fisher Scientific was used for DNA amplification. NucleoSpin plasmid and NucleoSpin Gel and PCR Clean-up kits from Macherey–Nagel were used to purify DNA. Restriction enzymes and T4 DNA ligase were purchased from New England Biolabs. Chemicals were purchased from Sigma-Aldrich, with the exception of isopropyl-β-D-1-thiogalactopyranoside (IPTG) (Biosynth) and 1,4-dithiothreitol (DTT, AppliChem) and used without further purification. The antibiotics chloramphenicol (CM), spectinomycin dihydrochloride (Sm), ampicillin (Amp), gentamicin (Gt) and carbenicillin (Cbn) were purchased from AppliChem. The medium components Bacto Tryptone and Bacto Yeast Extract were purchased from BD Biosciences. Sequencing-grade trypsin was purchased from Promega.
Cloning of embA, embM and orf3 (embI) for protein expression. Gene sequences were extracted from BGC 75.1 predicted by antiSMASH on the type material of 'Ca. E. malaspinii' (Supplementary Information).

The genes embA (locus, MALA_SAMN05422137_META-scaffold_127-gene_5), embM (locus, MALA_SAMN05422137META-scaffold_127-gene_4) and embAM (including intragenic region) were ordered as synthetic constructs in pUC57(AmpR), with and without codon-optimization for expression in Escherichia coli. The gene embA was subcloned into the first multiple cloning site (MCSI) of pACYCDuet-1(IncW) and pCDFDuet-1(SmR) with BamHI and HindIII cut sites. The genes embM and embMopt (codonoptimized) were subcloned into MCSII of pCDFDuet-1(IncW) with BamHI and HindIII and in the second multiple cloning site (MCSII) of pCDFDuet-1(IncW) and of pRSFDuet-1(IncW) with NdeI/Xhol. The embM cassette was subcloned into pCDFDuet1(IncW) with the BamHI and HindIII cut sites. The gene orf3/embI (locus, MALA_SAMN05422137 META-scaffold_127-gene_3) was constructed by overlap extension PCR with primers EmbI_F_plmb, EmbI_R_plmb, Plmb_F_EmbA and Plmb_R_EmbA is provided in Supplementary Table 6.

A list of the Gibson assembly primers EmbI_F_plmb, EmbI_R_plmb, Plmb_F_EmbA and Plmb_R_EmbA is provided in Supplementary Table 6. Transformation of E. coli DH5α for conjugation.

All constructs generated above were introduced into chemically competent E. coli DH5α and plated onto LB agar with appropriate antibiotic selection. Plasmids were purified from single colonies and sequenced using sequencing primers to verify proper insertion of genes (Supplementary Table 6). The genes embA and embAM were additionally subcloned in a modified plMB509m(Getty) vector for M. aerodenitrificans expression through Gibson assembly, with the inclusion of an N-terminal His, purification tag in the final EmbA protein product.

A list of the Gibson assembly primers EmbI_F_plmb, EmbI_R_plmb, Plmb_F_EmbA and Plmb_R_EmbA is provided in Supplementary Table 6. Transformation of E. coli DH5α, isolation of plasmids and validation of the correct clones by sequencing was performed by introduction of NHis6-EmbA plMB509m and NHis6-EmbAM plMB509m into E. coli SM10 for conjugation.

Heterologous expression and purification for protein isolation for in vitro assays. Chemically competent E. coli BL21(DE3) was introduced into pCDFDuet-1-EmbAMCSI1 and pCDFDuet-1-EmbAMCSI1. The same conditions were used for expression of both N-terminally His6-tagged proteins. Overnight cultures were prepared from single colonies and used to inoculate (1%, v/v) TB medium (2 × 200 ml) in 11-baffled Erlenmeyer flasks supplemented with spectinomycin (50 mg ml−1).

Cells were grown at 37 °C, 200 rpm, until an optical density at 600 nm (OD600) of around 1.0, cooled on an ice bath for 20 min and induced with a final concentration of 0.5 mM IPTG. The cultures were further incubated at 16 °C, 180 rpm for 18–20 h, and subsequently collected by centrifugation (8,000 g for 20 min) and frozen.

Purifications of NHis6-EmbA, NHi6-EmbM and NHis6-EmbM were carried out at 4 °C, using the same procedure for both proteins. Cell pellets were resuspended in 5 ml of lysis buffer (50 mM Tris, 300 mM NaCl, 5 mM imidazole, 10% glycerol, pH 7.8). The suspension was supplemented with lysozyme (1 mg ml−1), DNase I (10 U ml−1) and protease inhibitor cocktail (0.2%, v/v) and stirred at 37 °C for 30 min. After cooling the suspension for 15 min on an ice bath, cells were lysed by sonication (30% amplitude, 10s on/off cycles, for a total of 3 min), and the clarified lysate was obtained by centrifugation (27,000 g for 30 min). The supernatant was loaded onto 4 ml of Ni-NTA agarose resin in the presence of lysis buffer in a fritted purification column. The lysate was washed with 10 column volumes (CV) of lysis buffer, 3 CV of wash buffer (50 mM Tris, 300 mM NaCl, 40 mM imidazole, 10% glycerol, pH 7.8) and finally eluted with 3 CV of elution buffer (50 mM Tris, 300 mM NaCl, 250 mM imidazole, 10% glycerol, pH 7.8) in 1.5 ml fractions. Elution fractions were analysed by SDS–PAGE, pooled and concentrated in spin concentrators with the appropriate molecular weight cut-off. NHis6-EmbA and NHis6-EmbM were buffer-exchanged using a PD MiniTrap G25 column pre-equilibrated with G25 buffer (50 mM Tris, 300 mM NaCl, 10% glycerol, pH 8.0). The concentration of buffer-exchanged proteins was determined by measuring the absorbance of purified proteins at 280 nm and using the calculated values for molecular mass and extinction coefficient for each protein.

In vitro enzymatic activity assays with EmbA and EmbM. Extensive screening of enzymatic reaction parameters, including temperature, time, enzyme and substrate concentration, buffer pH and salinity resulted in the following best condition set for EmbM turnover: EmbA was added to a glass vial to a final concentration of 200 mM. Final concentrations of 2 mM of MgCl2 and 2 mM of adenosine 5′-triphosphate (ATP) were added to the reactions. EmbM was not added to control reactions and added to a final concentration of 10 mM in turnover experiments. The enzymatic reaction was stirred at 37 °C for 72 h, and the reaction mixture was supplemented with 2 mM of ATP every 24 h. Reaction scales ranged from 100 μl for analytical purposes, to 3 ml for product isolation.

Modified EmbA was proteolyzed with trypsin for MS analysis and with La T150 for MS analysis and product isolation90. Trypsin cleavage was performed by diluting the reaction mixture with 2× trypsin buffer (50 mM Tris, 5 mM CaCl2, pH 8.0), adding 1:20 trypsin:EmbA and incubating overnight at 37 °C. La T150 cleavage was performed by adding La T150 at a 1:10 ratio to EmbA, and incubating at room temperature overnight for small-scale reactions, or for 24 h, for reactions larger than 1 ml.

Large-scale enzymatic reactions were purified using solid-phase extraction (SPE) with a Phenomenex Strata C18-E reverse phase column (2 g sorbent). The sorbent was first washed with 24 ml MeOH and equilibrated with 24 ml H2O (+0.1% formic acid). The proteolysis reactions were loaded on the sorbent, which was then washed with 24 ml H2O (+0.1% formic acid). The peptide products were eluted with 24 ml of 1:1 MeCN:H2O (+0.1% formic acid) and 12 ml MeCN (+0.1% formic acid). Elution fractions were pooled, dried on a Genevac concentrator and stored at −20 °C.

Co-expression of EmbA, EmbM and Orf3 (EmbI) in heterologous hosts and purification. A wide variety of E. coli co-expression conditions of EmbA, EmbM and Orf3 were screened (Supplementary Table 6). In general, chemically competent BL21(DE3) or Tuner(DE3) were transformed with different combinations of the constructs described above, and selected with appropriate antibiotics on LB plates. Overnight cultures were prepared from single cultures and 1% (v/v) of culture was used to inoculate 200 ml of medium (TB, LB, XPPM) supplemented with appropriate antibiotic selection in 11 baffled Erlenmeyer flasks. Cells were grown at 37 °C, 200 rpm until an OD600 of around 1.2. For low-temperature growths, cultures were subsequently cooled in ice baths for 20 min, induced with 0.5 mM IPTG and incubated at 16 °C at 180 rpm. Incubation times varied from 24 h to 7 days. High temperature growths were induced with a final concentration of 0.5 mM IPTG without cooling and incubated at 37 °C. 200 rpm from 6 h to 16 h. Cultures were collected by centrifugation (8,000 g for 20 min), frozen and purified as described above. Proteolysis reactions with trypsin or La T150 for MS analysis were performed as described in the previous section.

Transformation of M. aerodenitrificans DSMZ 15089 with NHis6-EmbA plMB509m and NHis6-EmbM plMB509m was achieved according to a published procedure90. Culturing conditions followed an adaptation of a previously described method90. Cultures were collected, stored and purified as described above. Trypsin digestion was used for MS analysis.

β-Elimination of phosphorylated peptides. Phosphorylated peptide intermediates obtained by co-expression of EmbA and EmbM
were submitted to β-elimination conditions at the 100 μl scale. EmbA (200 μM) in G25 buffer was used either as an intact protein or as a trypsin-digest product. The pH of the solution was adjusted to pH 13 with 1 M NaOH, and the elimination reaction proceeded at 37 °C for 4 h to afford dehydrobutyryl (Dhb)-containing products. Derivatization of Dhb was performed by adding a final concentration of 50 mM DTT to the reaction. The pH of the solution was adjusted to 7 with HCl (aq.) before MS analysis.

β-Elimination of phosphorylated peptide intermediates obtained through in vitro enzymatic reactions was performed using EmbA that had been proteolyzed with LahtT150 and purified using SPE as described above. EmbA (0.6 mmol) was resuspended in 3 ml of H2O, and the pH of the solution was adjusted to 14 with 1 M NaOH. The reaction stirred at 37 °C for 4 h, and was subsequently neutralized to pH 7 with HCl (aq.). SPE purification was performed as described above.

**HPLC–HR-MS and MS/MS analyses.** HPLC–HR-MS and MS/MS analyses were performed on a Thermo Scientific Dionex UltiMate 3000 UHPLC coupled to a Thermo Scientific Q Exactive mass spectrometer using heated electrospray ionization in positive ion mode with a Phenomenex Kinex 2.6 μm XB-C18 100 Å (150 × 4.6 mm) column. The column temperature was set to 50 °C and the flow rate to 0.5 ml min⁻¹. Samples were centrifuged before injections and target peptides were eluted with a gradient of 15–55 % MeCN (0.1% formic acid) over 15 min. Full MS was performed at a resolution of 70,000 (AGC target, 1 × 10⁶; maximum IT, 100 ms) and parallel reaction monitoring was performed at a resolution of 17,500 (AGC target, 2 × 10⁵; maximum IT, 100 ms); isolation windows were 100 ms and parallel reaction monitoring was performed at a resolution of 17,500 (AGC target, 2 × 10⁵; maximum IT, 100 ms); isolation windows were 100 ms.

**NMR analysis.** 2D [¹³C, H] HSQC spectra with multiplicity editing were recorded at natural [¹³C abundance on -4 mM solutions of full length EmbA in unmodified and modified form. Spectra were recorded at 25 °C on a Bruker AVANCE 600 MHz spectrometer equipped with a TCI CryoProbeTM. The following spectral parameters were used: 2.048 complex points at a spectral width of 16 ppm, centred at 4.7 ppm in the direct 1H dimension and 512 complex points at a spectral width of 80 ppm centred at 42 ppm in the indirect ¹³C dimension. A number of scans of 40 was used, which resulted in a measurement time of 24 h per spectrum.

**Antibiotic activity assays.** E. coli DSM1103, Staphylococcus aureus ssp. aureus ATCC 29213, Pseudomonas aeruginosa DSM 1117, Acinetobacter baumanii DSM 30007, Enterococcus faecalis DSM 2570, Rhodococcus sp. L233, Aquimarina sp. Aq135, Rheinheimera aquimarina B26, Vibrio sp. (salt marsh isolate), Pseudoalteromonas rubra DSM 6842, Saccharomyces cerevisiae W301-1A and Pichia pastoris (Komagataella phaffii) NRR Y-11430 were tested for antimicrobial activity with the dehydrated peptide from BGC 75.1. Bioactivity assays were carried out in accordance with the 2003 guidelines of the Clinical and Laboratory Standards Institute (CLSI) using the microtiter method.

**E. coli, S. aureus, A. baumanii, and P. aeruginosa overnight cultures were grown in LB at 37 °C. P. rubra, R. aquimarins (37 °C), V. spartinae (30 °C), P. rubra and Aquimarina sp. Aq135 (24 °C) were cultured in marine broth at their respective growth temperature optimised in parentheses. Rhodococcus sp. L233 was cultured in R2A medium at 24 °C. S. cerevisiae and P. pastoris were cultured in YPD medium at 28 °C. An overview of strains, growth conditions and taxonomy is provided in Supplementary Table 6.**

Microbial seed cultures were initiated by inoculating 5 ml of medium for each strain and by incubating overnight shaking at 200 rpm. Each culture was then diluted with their respective growth medium to an initial OD₆₀₀ of 0.02 in 80 μl volume per well in sterile 96-well plates (one per strain tested). Assays were set up in duplicates, with appropriate controls including solvent (water) controls and positive controls consisting of two different broad-spectrum antibiotics (chloramphenicol and ampicillin) with final concentrations of 50 μM. Cycloheximide and benzylpenicillin were used as positive controls for S. cerevisiae and P. pastoris.

Two final concentrations of the peptide, using water as a solvent, were tested: 50 μM and 25 μM. Then, 96-well plates were parafilmed and incubated at room temperature without shaking. OD₆₀₀ was determined after 10 s of plate agitation at the following time points: 2 h, 4 h, 6 h, 8 h, 18 h, 21 h, 24 h and 48 h.

**Protease inhibition assays.** Inhibition assays against neutrophil elastase and cathepsin B were performed using the Neutrophil Elastase Inhibitor Screening Kit (MAK213, Sigma-Aldrich) and the Cathepsin B Inhibitor Screening Kit (MAK200, Sigma-Aldrich). Assays were performed according to the manufacturer’s protocol. A microplate reader spectrofluorometer (Tecan Infinite M200 Pro) was used to measure fluorescence and data were processed in Prism 9 to calculate IC₅₀ values.

Inhibition assays with trypsin and chymotrypsin were set up in 96-well microtitre plates (black, clear bottom). To each well, 25 μl of stock solutions of different concentrations of peptides in 50 mM Tris, pH 8 buffer (assay buffer); 2 ml of enzyme stock solution (chymotrypsin, V1062, Promega, 1 mM final concentration; trypsin, V5111, Promega, 3 mM final concentration); and 48 ml of protease buffer (40 mM Tris, 10 mM CaCl₂, pH 8) were added. Phenylmethylsulfonyl fluoride was used as a positive control. The plates were incubated at room temperature for 10 min. Subsequently, 23 ml of assay buffer and 2 ml of a 500 mM solution of substrate (chymotrypsin: Suc-Ala-Ala-Pro-Phe-AMC, 314-v, Pep- tanova; trypsin: Boc-Ile-Glu-Gly-Arg-AMC, 3094-v, Peptanova) in DMSO were added to the wells. Enzyme activity was measured at 37 °C for 1 h, by measuring the fluorescence emission of the hydrolysed product (λₑₓ = 342 nm, λₑₘ = 440 nm) in a microplate reader spectrofluorometer (Tecan Infinite M200 Pro). Data were processed in Prism 9 to calculate IC₅₀ values.

**MTT assays.** Inhibition against HeLa cells was tested for the phosphorylated, chemically dehydrated and control (unmodified) forms of peptide 75.1. Stock HeLa cells were resuspended in 10 ml HEPES-buffered high-glucose Dulbecco’s modified Eagle medium (DMEM) supplemented with GlutaMAX (Gibco). The medium also contained 10% fetal calf serum and 50 mg/ml gentamicin. The cells were centrifuged for 5 min at 1,000 g and room temperature. The medium was discarded, and the cells were resuspended in 10 ml fresh medium. The cells were put in a culture dish and incubated for 3–4 days at 37 °C. The cells were checked under the microscope and treated further only when 60–80% of the surface was covered with cells. The medium was removed from the culture flask and the cells were washed with 10 ml phosphate-buffered saline (PBS). The PBS was discarded and the cells treated with 2 ml trypsin-EDTA solution. When the cells were detached, 10 ml of medium was added and centrifuged for 5 min at 1,000 g and room temperature. The supernatant was discarded and 10 ml fresh medium was added. Then, 2 ml of the cell suspension was put into a fresh culture flask containing 10 ml medium. Cells healthy enough for cytotoxicity assays were counted and diluted to 10,000 cells per ml solution. Then, 96-well plates were filled with 200 μl cell suspension per well. All of the plates were incubated overnight at 37 °C. The outer wells were not used for the assay. A starting concentration of 100 μM of phosphorylated, chemically dehydrated and control (unmodified) forms of peptide 75.1 (2 μl of 1.25 mM stock solutions in assay buffer); 2 ml of enzyme stock solution (chymotrypsin, V1062, Promega, 1 mM final concentration; trypsin, V5111, Promega, 3 mM final concentration); and 48 ml of protease buffer (40 mM Tris, 10 mM CaCl₂, pH 8) were added. Phenylmethylsulfonyl fluoride was used as a positive control. The plates were incubated at room temperature for 10 min. Subsequently, 23 ml of assay buffer and 2 ml of a 500 mM solution of substrate (chymotrypsin: Suc-Ala-Ala-Pro-Phe-AMC, 314-v, Pептанова; trypsin: Boc-Ile-Glu-Gly-Arg-AMC, 3094-v, Pептанова) in DMSO were added to the wells. Enzyme activity was measured at 37 °C for 1 h, by measuring the fluorescence emission of the hydrolysed product (λₑₓ = 342 nm, λₑₘ = 440 nm) in a microplate reader spectrofluorometer (Tecan Infinite M200 Pro). Data were processed in Prism 9 to calculate IC₅₀ values.
Experimental validation of a novel proteusin pathway (‘Ca. E. malaspinii’, HHLDDBE BGC 3.1)

Cloning (E. coli and M. aerodenitrificans). Genes of the type material of ‘Ca. E. malaspinii’ (Supplementary Information) in the antiSMASH-predicted BGC 3.1 (region I) on MALA_SAMN05422137, META-gscaffold_4, including Nhis-ereA (gene_139), ereA (gene_140), ereM (gene_141), ereD (gene_142) and ereD (gene_143) were codon-optimized and ordered as synthetic constructs as described in Supplementary Table 5 for expression in E. coli. Genes were further subcloned by Gibson assembly using the primers listed in Supplementary Table 5. All generated constructs were introduced into chemically competent E. coli DH5α and plated on LB agar with appropriate antibiotic selection. Plasmids were purified from single colonies and Sanger sequencing was used to verify proper insertion of genes (Supplementary Table 5).

*Nhis-ereAD, Nhis-ereAIMD, Nhis-ereADB* and *Nhis-ereAIMDB* with intergenic regions substituted with the *M. aerodenitrificans* ribosome-binding site (TAGGAGGATGCGGG) were subcloned by Gibson assembly in a modified pLMB509mGt(4k) vector89. Introduction into E. coli DH5α, isolation of plasmids and validation of correct clones by sequencing was followed by introduction of all constructs into E. coli SM10 for conjugation into *M. aerodenitrificans*. Conjugation was performed as described previously89.

Heterologous expression and protein purification. Chemically competent E. coli BL21(DE3) cells were transformed with all of the constructs listed in Supplementary Table 5. Overnight cultures were prepared from single colonies and used to inoculate (1%, v/v) TB medium (1 × 50 ml) in 250 ml baffled Erlenmeyer flasks supplemented with appropriate antibiotics. Cells were grown at 37°C, 200 rpm, until an OD₆₀₀ of around 0.8–1.2 and induced with a final concentration of 1 mM IPTG. The cultures were further incubated at 16°C, 180 rpm for 16 h, 24 h or 48 h as specified per experiment and subsequently collected by centrifugation (27,000 × g for 45 min). The supernatant was loaded into 1 ml of lysis buffer (50 mM Tris, 300 mM NaCl, 5 mM imidazole, 10% glycerol, pH 8.0). The dried samples were solubilized in 75 μl water, 25 μl 1 N NaHCO₃ and 125 μl N-<sup>2</sup>-Dinitrophenyl)-<sup>1</sup>β-D-ribofuranosylamine (4 mg ml<sup>−1</sup> in acetone) and the mixture was heated for 1h at 45°C. The reaction was neutralized with 25 μl 1 N HCl and diluted with 250 μl acetonitrile/water (1:1). The samples were transferred to a HPLC vial and subsequently analysed using HPLC–HR-MS. Amino acids were verified by mass and retention time when compared to authentic amino acid standards treated in the same way as described above.

Purification of *SAM* for *in vitro* EreM assays. Before the enzymatic assays, 25 mg of commercially available *SAM* (Sigma-Aldrich) was dissolved in 1 ml and injected in five portions (200 μl each) to conduct additional purification on an Agilent 1260 HPLC equipped with a SemiPrep Hydro RP (4 × 250 × 10 mm) column at ambient temperature. HPLC separation was performed using the following method (solvent A, H₂O + 0.1% FA; solvent B, ACN + 0.1% FA; flow rate: 1 ml min<sup>−1</sup>): 2 min at 30% B; 2–9 min from 30% to 100% B; 9–10 min at 100% B; 10–10 min from 100% B, 10.1–14 min at 30%.

The HPLC was coupled to a Thermo Scientific Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer using heated electrospray ionization in positive-ion mode (spray voltage, 3,500 V; capillary temperature, 280°C; probe heater temperature, 350°C; S-lens level, 70). Full MS was detected at a resolution of 70,000 (AGC target, 1 × 10<sup>6</sup>; maximum IT, 100 ms).

Digests were incubated overnight at room temperature and analysed by HPLC–HR-MS and MS/MS after approximately 16 h.

HPLC–HR-MS method for Marfey’s analysis. Analytical HPLC–HR-MS samples (5–20 μl injections) were separated on a Dionex Ultimate 3000 RSLC equipped with the Phenomenex Kinetex C18 (2.6 μm, 100 A, 150 × 4.6 mm) column heated to 50°C. HPLC separation was performed according to the following method (solvent A, H₂O + 0.1% FA; solvent B, ACN + 0.1% FA; flow rate: 1 ml min<sup>−1</sup>): 2 min at 30% B; 2–9 min from 30% to 100% B; 9–10 min at 100% B; 10–10 min from 100% B, 10.1–14 min at 30%.

The HPLC was coupled to a Thermo Scientific Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer using heated electrospray ionization in positive-ion mode (spray voltage, 3,500 V; capillary temperature, 280°C; probe heater temperature, 350°C; S-lens level, 70). Full MS was detected at a resolution of 70,000 (AGC target, 1 × 10<sup>6</sup>; maximum IT, 100 ms).

Heterologous expression and protein purification. Chemically competent E. coli BL21(DE3) cells were transformed with all of the constructs listed in Supplementary Table 5. Overnight cultures were prepared from single colonies and used to inoculate (1%, v/v) TB medium (1 × 50 ml) in 250 ml baffled Erlenmeyer flasks supplemented with appropriate antibiotics. Cells were grown at 37°C, 200 rpm, until an OD₆₀₀ of around 0.8–1.2 and induced with a final concentration of 1 mM IPTG. The cultures were further incubated at 16°C, 180 rpm for 16 h, 24 h or 48 h as specified per experiment and subsequently collected by centrifugation (6,000g for 15 min) and frozen.

For purification from *M. aerodenitrificans*, 20 ml precultures in nutrient broth with the appropriate antibiotics were inoculated from freeze stocks and grown overnight at 30°C. Then, 1% (v/v) cultures were used to inoculate 20 ml TB starter cultures with the appropriate antibiotics. Then, 1% (v/v) of starter cultures were used to inoculate 200 ml in 11 non-baffled Erlenmeyer flasks. Cells were grown at 30°C, 180 rpm, until an OD₆₀₀ of around 0.8–1.2 and induced with a final concentration of 1 mM IPTG. The cultures were further incubated at 30°C, 180 rpm for 16 h, 24 h, 48 h or 72 h as specified for each experiment and subsequently collected by centrifugation (6,000g for 15 min) and frozen.

For protein purification, cell pellets were resuspended in 5 ml g<sup>−1</sup> of lysis buffer (50 mM Tris, 300 mM NaCl, 5 mM imidazole, 10% glycerol, pH 8.0). Cells were lysed by sonication (30% amplitude, 10 s on/off cycles, for a total of 3 min), and the clarified lysate was obtained by centrifugation (27,000g for 45 min). The supernatant was loaded onto 1 ml of Ni-NTA agarose resin that had been equilibrated with a lysis buffer (50 mM Tris, 300 mM NaCl, 40 mM imidazole, 10% glycerol, pH 7.8) and finally eluted with 3 CV of elution buffer (50 mM Tris, 300 mM NaCl, 250 mM imidazole, 10% glycerol, pH 7.8). Elution fractions were analysed by SDS–PAGE, pooled and buffer-exchanged into GST buffer (50 mM Tris, 300 mM NaCl, 10% glycerol, pH 8.0) in Amicon Ultra-4 10K spin filter devices (Millipore Sigma). The concentration of spun, buffer-exchanged proteins was determined using the Roti-Quant Bradford reagent (Carl Roth).

Proteolytic cleavage for analysis of core peptides and generation of the core region. For the LahTI50 digest, 8 μl of purified LahTI50 (ref. 89) at a concentration of 20 μM was added at a 1:10 ratio to 72 μl of approximately 200 μM of spin-concentrated protein in glass MS vial inlets.
EreM were dissolved in SGT buffer (50 mM Tris, 300 mM NaCl, 10% glycerol, pH 8.0) to a total volume of 100 μl and incubated in glass MS vial inserts overnight at 37°C. Then, 5 μM LaHt was added to the in vitro assays the next day and incubated for 2 h at room temperature before analysis by HPLC–HR-MS/MS.

In vitro EreM assays with 13C-labelled 13CH3-SAM using an enzyme cascade starting with 13C-labelled 13CH3-l-methionine for NMR spectroscopy, 13CH3-l-methionine (5 mM), 10 mM adenosine-triphosphate, 100 mM KCl, 30 mM MgCl2, 1 μM MTAN, 11 μM SAM synthase, 40 μM epimerized Nhs-EreA precursor and 40 μM EreM were dissolved in 50 mM Tris at pH 8.0 to a total volume of 500 μl and incubated in an Eppendorf tube overnight at 37°C. Then, 50 μl of D2O was added to the assay mixture and transferred into an NMR tube. Proton and HSQC spectroscopy was performed on the Bruker 600 MHz NMR spectrometer equipped with a cryoprobe. To determine the location of protons attached to 13C-labelled carbons, an additional 1H NMR was recorded with parameters that enable decoupling of carbons and protons. These parameters cause a splitting of the respective protons attached to 13C-labelled carbons.

In vitro EreM assays. (NH4)2FeSO4 (800 μM) was incubated with 50 μM EreM in SGT buffer (50 mM Tris, 300 mM NaCl, 10% glycerol, pH 8.0) for 20 min on ice. Then, 1 mM 2-oxoglutarate and 1 mM dihithionietol were added and incubated for another 20 min on ice. Epimerized Nhs-EreA precursor (5 μM) was added to yield a total volume of 100 μl and the reaction mixture was incubated in Eppendorf tubes for 20 min at 30°C. The reaction was quenched by boiling the assay mixture at 95°C for 10 min. LaHt (5 μM) was added to the in vitro assays and incubated for 2 h at room temperature before analysis using HPLC–HR-MS/MS.

Site-directed mutagenesis to generate core variants. Primers for site-directed mutagenesis (Supplementary Table S) were synthesized and used to amplify template DNA from EreAD-pET Duet (Amp®). Mutagenesis was accomplished using PCR amplification, KLD treatment and enrichment, and transformation into E. coli DH5α for isolation of plasmids and validation of the correct clones by sequencing. Owing to the highly repetitive nature of the core sequence, truncated core variants were also generated during site-directed mutagenesis and also tested for EreM modification (Supplementary Table S).

Orthogonal D2O-based induction system for labelling epimerized residues. E. coli BL21 (DE3) cells were co-transformed with Nhs-ereA in pACYCDuet-1 and ereE in pCDFBAD/Myc-His A and plated on LB agar containing chloramphenicol (25 μg ml−1) and ampicillin (100 μg ml−1) and grown for 20 h at 37°C or until colonies appeared. These colonies were used to inoculate 20 ml LB with chloramphenicol (25 μg ml−1) and ampicillin (100 μg ml−1), and the cultures were grown overnight. The next day, four separate 50 ml Falcon tubes containing TB medium, (15 ml), chloramphenicol (25 μg ml−1) and ampicillin (100 μg ml−1) were inoculated with 150 μl and shaken at 37°C, 250 rpm to an OD600 of 1.6–2. Cultures were cooled on ice for 30 min, induced with IPTG (0.1 mM final concentration) and shaken (180 rpm at 16°C) for 18 h. The cultures were centrifuged (10 min at 10,000g) and the supernatant was removed. The cell pellets were then washed with TB medium (1 x 15 ml) to remove any residual IPTG. This process was followed by washes with 1 ml TB in D2O. The washed cell pellets were resuspended in 15 ml TB medium in D2O containing ampicillin (100 μg ml−1 in D2O) and L-arabinose (0.2% w/v in D2O) and shaken (180 rpm at 16°C) for 18 h. The cultures were combined and centrifuged (30 min at 15,000g) and the pellets were resuspended in 4 ml lysis buffer per gram of cell pellet and purified as described above. For the control, the same procedure was followed with normal TB medium.

Generation of the M. aerodenitrificans Δaer mutant. The suicide vector pSW8197 (ref. 25) was used as a basis to create a stable and markerless deletion of the aeronemamide (aer) cluster in M. aerodenitrificans. The primer pairs Aeri1f/r and Aeri2f/r were used to amplify 500 bp homologous regions up- and downstream of the aer cluster. The resulting DNA products were assembled into PCR-amplified pSW8197 (pSWAerKO-f/r) using Gibson assembly and transformed into E. coli SY327 electrocompetent cells and sequence-verified after plasmid extraction from the resulting colonies.

The resulting plasmid (pSW8197_akero) was then transformed into chemically competent E. coli ST18 donor cells56 and selected with 50 μg ml−1 kanamycin and 5-aminolevulinic acid (required by E. coli st18). Conjugation with the M. aerodenitrificans wild type was performed as previously described, plated on LB agar plates with 50 μg ml−1 5-aminolevulinic acid and incubated at 37°C for 24 h. Integrants were selected by plating on nutrient agar plates containing 25 μg ml−1 kanamycin and 400 μg ml−1 ampicillin at 30°C and confirmed by PCR (akero seqf/pswAraC-R). Positive integrants were grown non-selectively in 5 ml nutrient broth overnight at 30°C and plated on nutrient agar with 0.5% (w/v) L-arabinose (to induce counter-selectable ccdB toxin) and 100 μg ml−1 ampicillin, resulting in colonies with successful deletions or wild-type revertants. Successful deletion mutants of the aer cluster were identified using PCR with the primers aero seqf/R, which anneal to regions on the genome outside the flanking regions used to construct the vector. The resulting PCR product was verified by sequencing.

Phylogenetic analysis of FkbM-family proteins. PfamScan classified EreM from ‘Ca. E. malapinii’ as belonging solely to the FkbM methyltransferase (PF05050) family. To identify other FkbM-family proteins involved in natural product biosynthesis, the FkbM-family methyltransferase HMM (Methyltransf_21.HMM in Pfam_A) was used to query all protein-coding sequences in MiBiG (v.2.0)53 using hmmsearch in HMmer v.3.1b2 (http://hmmer.org/) with the default parameters and the--cut nc PFAM noise cut-off. Hits within 37 characterized BGCs in MiBiG were identified (Supplementary Table S) and associated literature was manually assessed for experimental evidence of FkbM-family methyltransferase activity. Eight proteins were excluded on the basis of the FkbM hit falling outside the defined BGC cluster boundaries and having no apparent role in biosynthesis based on the final natural product structure (Supplementary Table S). Four FkbM family members had experimental evidence for O-methyltransferase activity in the form of heterologous expression or genetic studies. A total of 25 FkbM-family proteins were documented in publications by authors to have likely O-methyltransferase activity on the basis of the final natural product structure, biosynthetic logic and bioinformatic evidence. The summed 29 FkbM-family proteins were aligned using MUSCLE (v.3.8.1551)81 with two outgroups involved in proteusin biosynthesis, PoyE (AFS60641.1) and AerE (AFS60641.1) from a different methyltransferase protein family (PF0175). The protein alignment was assessed and all columns containing 50% or more gaps were removed using Trimal v.1.2. The trimmed alignment was used for phylogenetic model selection using IQ-TREE (v.2.0.3)82 with the V5+F+R5 model was selected based on trimming IQ-TREE (v.2.0.3)96 and the V5+F+R5 model was selected based on trimming. The V5+F+R5 model was selected based on trimming IQ-TREE (v.2.0.3)96 and the V5+F+R5 model was selected based on trimming IQ-TREE (v.2.0.3)96 and the V5+F+R5 model was selected based on trimming IQ-TREE (v.2.0.3)96 and the V5+F+R5 model was selected based on trimming.
Method A: solvent A: H2O + 0.5% FA; solvent B: 1-propanol + 0.5% FA; flow rate: 0.5 ml min⁻¹; 2 min at 20% B; 2–14 min from 20% to 80% B; 14–17 min at 80% B; 17–17 min from 80% to 20% B; 17–20 min at 20% B.
Method B: solvent A: H2O + 0.1% FA; solvent B: acetonitrile + 0.1% FA; flow rate: 1.0 ml min⁻¹; 2 min at 2% B; 2–12 min from 2% to 100% B; 12–16 min at 100% B; 16–17 min from 100% to 2% B; 17–20 min at 2%.

The HPLC was coupled to a Thermo Scientific Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer using heated electrospray ionization in positive-ion mode (spray voltage, 3,500 V; capillary temperature, 320 °C; ion capture lens, 70 V). Full MS was detected at a resolution of 70,000 (AGC target: 1 × 10⁶; maximum IT, 100 ms). MS² fragmentation was performed at a resolution of 35,000 (AGC target, 2 × 10⁴; maximum IT, 100 ms, isolation window, 4.0 m/z). Normalized collision energy was 20, 25 and 30 for +3 charge states. Parallel reaction monitoring was performed at a resolution of 17,500 (AGC target 2 × 10⁵; maximum IT, 100 ms; isolation windows, 1.4 m/z) and a normalized collision energy of 18, 20 and 24 for +2 and +3 charge states.

Statistics and reproducibility
Wherever appropriate, correction for multiple testing was performed using false-discovery rate correction. Wherever appropriate and if not specified otherwise, statistical tests performed were two-sided. The box plots were plotted in R (v.4.0.2–v.4.1.2) using ggplot2 (v.3.3.0–v.3.3.5) and defined as follows: the bottom and top hinges correspond to the first and third quartiles (the 25th and 75th percentiles), the top whisker extends from the hinge to the largest value no further than 1.5 × IQR from the hinge (where the IQR is the interquartile range, or distance between the first and third quartiles). The bottom whisker extends from the bottom hinge to the smallest value at most 1.5 × IQR. Data points beyond the end of the whiskers are outliers are plotted individually, except for in Fig. 1c, owing to the large number of points and space constraints.

Fig. 1e was plotted using the R package UpSetR (v.1.4.0)⁹⁸. The trees in Fig. 2 were plotted using the R package ggtree (v.3.3.0.901)⁹⁹.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability
The metagenomic and metatranscriptomic data used in this study were downloaded from the European Nucleotide Archive (ENA) and a summary of their accessions is provided in Supplementary Table I. Publicly available genomes were downloaded from FigsTree (https://doi.org/10.6084/m9.figshare.4902923) for manually curated MAGs from Tara Oceans, from ENA using the project accession PRJEB33281 for GORG and from https://mpp2.sfb.uit.no/databases/ for MarDB. The GEM MAGs were downloaded from https://portal.nersc.gov/GEM/. MAGs contained in the GTDB r89 were downloaded from https://www.genome.gov/gtdb/eggnog. The MIBIG and Big-FAM databases can be accessed at https://mibig.secondarymetabolites.org/ and https://bigfam.bioinformatics.nl/, respectively. The data produced in this study, including metagenomic assemblies, bins and MAGs have been deposited at the ENA under the accession PRJEB45951 and a list of individual accession identifiers is provided in Supplementary Table I. Other supporting data have been deposited at Zenodo (https://doi.org/10.5281/zenodo.4474310), and the OMD can be interactively accessed online (https://microbiomes.io/zenodo/). Additional material generated in this study is available on request.

Code availability
The code used for the analyses performed in this study is accessible at GitHub (https://github.com/SushiLab/magpipe/) and archived at Zenodo (https://doi.org/10.5281/zenodo.6393817).
Morinaka, B. I. et al. Radical S-adenosylmethionine epimerases: regioselective introduction of diverse β-amino acid patterns into peptide natural products. Angew. Chem. Int. Ed. 53, 8503–8507 (2014).

Morinaka, B. I. et al. Natural noncanonical protein splicing yields products with diverse β-amino acid residues. Science 359, 779–782 (2018).

Le Roux, F., Binesse, J., Saulnier, D. & Mazel, D. Conservation of a Vibrio splendidus mutant lacking the maltoprotease gene vsm by use of a novel counterselectable suicide vector. Appl. Environ. Microbiol. 73, 777–784 (2007).

Thoma, S. & Schobert, M. An improved Escherichia coli donor strain for diparental mating. FEMS Microbiol. Lett. 294, 127–132 (2009).

Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1544 (2020).

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522 (2018).

Conway, J. R., Lex, A. & Gehlenborg, N. UpsiRe: an R package for the visualization of intersecting sets and their properties. Bioinformatics 33, 2938–2940 (2017).

Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36 (2017).

Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci. Data 5, 170203 (2018).

Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for large alignments. PLoS ONE 5, e9490 (2010).

Conibeer, A. C., Rosengren, K. J., Becker, C. F. W. & Kaehlig, H. Random coil shifts of posttranslationally modified amino acids. J. Biol. Chem. 283, 587–599 (2018).

Galís, D. E. & Crist, D. R. Use of NOE difference spectra to determine configurations and conformations of imidate esters. Magn. Reson. Chem. 23, 327–328 (1985).

Meese, C. O. & Walter, W. Unusual 13C/177se couplings in the 13C NMR spectra of (2H, 13C) analogues. Arch. Pharm. 323, 957–965 (1990).

Mordhorst, S., Siegrist, J., Müller, M., Richter, M. & Andexer, J. N. Catalytic alkylation using a cyclic S-adenosylmethionine regeneration system. Angew. Chem. Int. Ed. 56, 4037–4041 (2017).

Sterner, O., Etzel, W., Mayer, A. & Anke, H. Omphalotin, a new cyclic peptide with potent nematicidal activity from Omphalotus olearius II. Isolation and structure determination. Nat. Prod. Lett. 10, 33–38 (1997).

Acknowledgements We thank the members of Tara Oceans, which has been driven by a collaborative engagement of the Tara Ocean Foundation, countless scientific project members and associated institutional partners. We are grateful for the availability of the data generated in the context of the BioGEOTrACES, Malapsina, HOT, BATS and GORG projects as well as the data compiled as part of the Mar Databases project, which formed an essential basis of this work. We thank the staff at the ETH Zürich IT Services, HPC facilities and Biomolecular NMR Spectroscopy Platform as well as the Functional Genomics Center Zurich, in particular A. Patrignani, for their support, S. Booker for the gift of the pBAD42-BtuCEDFB plasmid, J. Vorholt, E. Rocha and R. Neher for their feedback throughout this project, F. Hügenholtz and T. J. Williams for their comments on the preprint version of this Article, and A. Jermy for his advice during the writing of this manuscript. This Article is contribution number 130 of Tara Oceans. The work was supported by funding from the ETH and the Helmut Horten Foundation; the Swiss National Science Foundation (SNSF) through project grants 205321_184855/1 to S.S., 205320_185077 to J.P. and the NCCR-Microbiomes (SIN4F-180575) to S.S.; by the Gordon and Betty Moore Foundation (https://doi.org/10.37807/GSBMF5204) and the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 10000392 (MARBLES) to J.P.; by an ETH research grant ETH-21 16-2 to J.P.; and by the Peter and Traudi Engelform Foundation and by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 897571 to C.C.F. S.L.R. was supported by an ETH Zurich postdoctoral fellowship 20-1 FE-07. M.L., L.M.C. and G.Z. were supported by EMBL Core Funding and the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft, project no. 395317507; SFB 1371 to G.Z.). M.B.S. was supported by the NSF grant 1OCE1829831. C.B. was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement Diatomic, no. 835067). S.G.A. was supported by the Spanish Ministry of Economy and Competitiveness (PID2020-116499GB-I00). M.K. and H.M. were funded by the SNSF grant 407540, 167331 as part of the Swiss National Research Programme 75 ‘Big Data’. M.K., H.M. and A.K. are also partially funded by ETH core funding (to G. Rätsch).

Author contributions S.S., S.L.R. and J.P. conceived and co-supervised the study. L.P., S.S. and S.L.R. designed the research. L.P., H.-J.R., Q.C., S.K., G.S. and A.M. performed bioinformatic and statistical analyses. C.C.F., F.H., S.L.R. and J.P. conceived and co-supervised the study. L.P., H.-J.R., Q.C., S.K., G.S. and A.M. performed bioinformatic and statistical analyses. C.C.F., F.H., S.L.R. and J.P. performed the experimental work. L.P., C.P., H.-J.R., D.G., M.K., H.M., M.L., L.M.C., G.Z. and A.K. designed and implemented web and data resources. A.A.Z., D.R.C., M.B.S., T.Q.D., P.S., S.G.A., J.M.G., P.W., P.B. and C.B. contributed samples, data and scientific input. L.P. and S.S. drafted the manuscript with S.L.R., C.C.F., F.H. and J.P. All of the authors reviewed the submitted manuscript and approved the final version.

Competing interests The authors declare no competing interests.

Additional information Supplementary information The online version contains supplementary material available at https://doi.org/10.10138/415686-022-04862-3.

Correspondence and requests for materials should be addressed to Serina L. Robinson, Jörn Piel or Shinichi Sunagawa.

Peer review information Nature thanks Bradley Moore and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Depth distribution of the metagenomes used in this study; overview of the bioinformatic pipeline and proxies for sequencing depth. (a) 1,038 publicly-available ocean microbial community genomes (metagenomes) were collected across all major depth layers (1 - 5,601 m) in the context of different ocean expeditions and time series programmes; EPI - epipelagic layer; MES - mesopelagic layer; BAT - bathypelagic layer; ABY - abyssopelagic layer. (b) Quality-controlled, high-throughput DNA sequencing reads from ocean microbial community samples were individually assembled into metagenomic scaffolded contigs (scaffolds). Sequencing reads from large subsets (n ranging from 58 to 610) of all samples were aligned to scaffolds of each individual sample to compute relative copy-number abundances for each scaffold in each sample. Based on a combination of tetranucleotide frequency, within-sample co-abundance and between-sample abundance correlations, scaffolds were grouped into a total of 62,874 metagenomic bins, each with total nucleotide sequence lengths of > 200 kb. These metagenomic bins were filtered for genome completeness and contamination, resulting in 26,293 metagenome assembled genomes (MAGs). These MAGs were complemented with external sets of MAGs, single amplified genomes (SAGs) and genomes from cultured isolates (REFs). The combined set of 34,799 genomes was clustered at the species level using a 95% average nucleotide identity (ANI) and, along with taxonomic and functional annotations, abundance profiles and contextual information, compiled into the Ocean Microbiomics Database (OMD); see methods for details (Methods). (c) Comparing mapping rates obtained from mapping subsampled readsets compared to those obtained from mapping the total number of reads shows that this procedure yields almost identical results at considerably less computational costs. (d) mOTUs counts as a good proxy for sequencing depth. We find a strong correlation in prokaryote-, particle-enriched and virus-depleted communities, while this correlation is more variable in virus-enriched communities. This observation is actually in support of using the mOTUs count rather than sequencing depth when focusing on the bacteria and archaeal component of microbial communities, as we do here.
Extended Data Fig. 2 | Impact of abundance correlation on MAGs recovery and quality, quality improvement over other ocean MAGs datasets, recovery of mobile genetic elements and evaluation of genome chimerism. (a) In this study, MAGs were reconstructed using abundance correlation information (Extended Data Fig. 1b) (Methods), which resulted in both higher cumulative quality scores per sample and individual quality scores per MAG. The ratio of cumulative quality scores (Supplementary Information) of MAGs binned with and without differential coverage information was on average (median) 2.3 across the different datasets. Per individual MAGs, a mean quality score increase of 20% was achieved. The number of samples used for differential coverage profiling are indicated above the boxplots. The colours of the boxplots reflect the different datasets as indicated in Fig. 1b. (b) We investigated the bin membership of >80 M scaffolds across size and fragment type. These scaffolds were annotated to identify chromosomes, plasmids and phages (Supplementary Information). The difference between chromosomes and plasmids binning rates provides an evaluation of the bias of the MAG reconstruction against hypervariable regions within the genomes. Annotations were integrated to classify scaffolds as follows, chromosomes (‘eukrep = Prokarya & plasmfinder prediction = chromosome & cbar prediction = Chromosome & plasmfinder plasmid != NaN & deepvirfinder p-value > 0.05 & virsorter score not in [1,2] & deepvirfinder p-value > 0.05’), viruses (‘virsorter score = 1 & deepvirfinder p-value < 0.01 & eukrep = Prokarya & plasmfinder prediction = ’plasmid & cbar prediction = ’Plasmid’) or unannotated. By benchmarking the quality of the MAGs reconstructed in this study (Supplementary Information), we found that combining single-sample assemblies with large-scale abundance correlations achieved on average significantly higher community-defined quality scores than and two datasets of automatically generated MAGs, dataset #1 and dataset #2, and (d) even manually curated MAGs. ‘n’ denotes the number of possible comparisons (i.e. number of shared species) with the different MAGs sets. All genomes in the extended OMD were evaluated for chimerism using the taxonomic annotation of 10 universal single copy marker genes (Supplementary Information). (f) For each taxonomic level, the genomes were classified as: “No annotation” if a maximum of one gene out of 10 was annotated; “Agreeing” if all genes had the same annotation; “Majority agreeing” if more than half agreed and “Not agreeing” otherwise. The evaluation was split for the genomes origin (y-axis). (g) Percentage of “Not agreeing” annotations over all the annotated clades (i.e. the sum of “Agreeing”, Majority agreeing” and “Not agreeing”) otherwise. The evaluation was split for the genomes origin (y-axis).
Extended Data Fig. 3 | Different genome reconstruction strategies capture complementary phylogenomic diversity; trends in community genome sizes across the global ocean microbiome. (a) Reconstructed MAGs, external MAGs, SAGs as well as REFs detected across the set of 1,038 ocean metagenomes were placed on the GTDB backbone trees revealing that the different genome types (MAGs, SAGs and REFs) capture complementary phylogenomic diversity. Similar to Fig. 3, the green-to-blue colours of the branches indicate the number of genomes in that part of the tree. The inner layer denotes the taxonomy of specific clades (some indicated by arrows due to limited space). The outer layer represents the percentage of genomes across the binned tree for each genome type. Clades without any genome from the OMD were left in grey. For visualization purposes, the last 15% of the nodes are collapsed. (b, c) The average genome size per sample was significantly larger in deeper waters (Kruskal Wallis test, p-value < 2*10^{-16}, n = 1,038) and was inversely correlated with temperature (linear model). (d) Comparing genome sizes from MAG-based predictions and reference genomes for 85 mOTUs (species-level) clusters with at least one reference genome. Genome sizes are estimated using MAGs of good quality and above only (completeness above 70%), a criteria that is met for > 80% of the mOTUs clusters.
Extended Data Fig. 4 | Structure and drivers of the ocean microbiome biosynthetic potential; evaluation of BGC completeness using length and number of genes between predicted and characterized BGCs. (a) The abundances of GCFs (Methods) were used to compute distances between the 1,038 metagenomic samples. Using dimension reduction and density based clustering (Methods), we identified three sample clusters. (b) A prediction strength analysis strongly supports clustering the data into 3 groups (largest number of clusters above the 0.9 threshold). This is also confirmed by the Silhouette Index (data not shown). (c) These clusters were broken down by community origin, including size fractions, depth layers and ocean basins. We found significant differences in BGC class abundances (FDR-corrected pairwise Wilcoxon tests, p-value < 2*10^{-16}, n = 1,038) and average genome sizes (FDR-corrected pairwise Wilcoxon tests, p-value < 2*10^{-16}, n = 1,038) (Methods) between the clusters (Supplementary Table 2). (d) We found temperature and depth to be significantly different between the sample clusters identified based on biosynthetic potential composition (Kruskal Wallis test, p-value < 2*10^{-16}, n = 1,038). RiPP - Ribosomally synthesized and Post-translationally modified Peptide; NRPS - Non-Ribosomal Peptide Synthetase; T1PKS - Type I Polyketide Synthase; T2/3PKS - Type II and III Polyketide Synthases. BGC length distributions across BGC classes are not significantly different (Wilcoxon test, significance denoted by * with p-value < 10^{-5}, n >> 30) between the set of BGCs studied in this work (antiSMASH) and the characterized BGCs in MIBiG with the exception of the polyketides and non-ribosomal peptide synthetases, which may be expected based on the particularly large clusters they can encompass (e) and the BGCs studied in this work (antiSMASH) to have a similar or higher number of genes than the characterized BGCs in MIBiG (f).
Extended Data Fig. 5 | GCF novelty across latitude, depth layers and size fractions for each BGC class and distribution of nucleoside BGCs across genomic and metagenomic fragments. (a) We estimated the discovery potential of different microbial communities by counting the number of new GCFs (Methods) detected in a sample after rarefaction of per-cell GCFs abundance profile to 2,000 cells. Although well studied communities (non-polar epipelagic prokaryote-enriched (0.2–3 μm) and virus-depleted (>0.2 μm)) displayed the highest discovery potential for terpenes, least explored communities (polar, deep, virus- and particle-enriched) were found to have the highest potential for NRPS, PKS, RiPPs or other natural products discovery. Polar is defined as absolute latitude > 60º. NRPS: Non-Ribosomal Peptide Synthetases; PKS: Polyketide Synthase; RiPP: Ribosomally Synthesized and Post-translationally modified Peptide. (b) An overview of the putative terpenoid diversity. A phylogenetic tree of all terpene biosynthetic core genes (as defined by antiSMASH) identified in the OMD, in the context of the 195 MIBiG terpene biosynthetic core genes, provides an overview of the terpenoid diversity and novelty. Briefly, the 31,398 terpene biosynthetic core genes identified across all predicted BGCs were filtered (length ≥ 120aa, removing < 2% of the sequences), dereplicated (using MMSSEQS13.45111 clustering, 60% identity) into 2,904 protein sequences and aligned with the 195 MIBiG proteins using MAFFT v7.310. The resulting alignment was trimmed with trimal to remove positions with more than 50% gaps and used to build the tree using FastTree v2.1.10. The inner annotation layers indicate whether a gene is coming from a MIBiG cluster and if this one was annotated as a carotenoid or hopene cluster. The outer layers correspond to the biosynthetic core gene domain according to antiSMASH categories. Plants were used to root the tree. (c) Investigation of the proportion of BGCs binned within a MAG by product type showed that nucleosides were most rarely encoded in MAGs. (d) Breakdown by fragment type of the BGCs in the remaining metagenomic fragments. Strikingly, nucleoside BGCs were rarely encoded on predicted chromosome fragments and most often in predicted phage fragments (Supplementary Information). For this analysis, we refined the prediction described in Extended Data Fig. 2b with prophages (\texttt{virsorter category=prophage & virsorter score >} = 1 \& \texttt{eukrep=Prokarya & plasflow prediction!=plasmid & cbar prediction!=Plasmid}), \texttt{phages (\texttt{virsorter category=phage & virsorter score >} = 1 \& \texttt{deepvirfinder p-value<0.01 & eukrep=Prokarya & plasflow prediction!=plasmid & cbar prediction!=Plasmid})} and \texttt{putative phages (not in \texttt{phages & (\texttt{virsorter category=phage & virsorter score >} = D\| \texttt{deepvirfinder p-value<0.05}) & eukrep= Eukarya & plasflow prediction!=plasmid & cbar prediction!=Plasmid})}. 
Extended Data Fig. 6 | Manual inspection of Ca. Eudoremicrobiaceae MAGs and phylogeny of the duplicated marker gene COG0124. (a–f) Anvi’o interface of representatives of the five Ca. Eudoremicrobiaceae species reveals stable abundance correlation patterns across the vast majority of the genomes, indicative of low contamination rates (Supplementary Information). (b) Inspection of the assembly graph for Ca. E. malaspinii (Supplementary Information) showed that all scaffolds from the representative genomes were connected with the exception of a single 20 kb one. (g) Investigating the evolutionary history of duplicated single-copy marker genes (here COG0124), we found consistent duplication across Ca. Eudoremicrobiaceae and the parent order UBP9, thus ruling out the duplication as a signal of contamination in the binning process. The different evolutionary history of the second copy of COG0124 (right-hand side of the tree), with closer relationship to Actinobacteria suggests that introgression events (including before the UBP9 and Ca. Eudoremicrobiaceae split) could be the origin of the increased genome size and biosynthetic potential observed in Ca. Eudoremicrobiaceae. (h) Similar patterns can be found in the second duplicated marker gene (COG0522), although duplication was not detected across all Ca. Eudoremicrobiaceae spp. representatives.
Extended Data Fig. 7 | BGCs are the most differentially expressed genes in Ca. E. taraoceanii natural populations and are expressed in nature across the Ca. Eudoremicrobiaceae family. The 25 metatranscriptomic samples used for the Ca. E. taraoceanii expression analyses were selected based on the detection of at least 6 out of 10 universal single-copy marker genes. (a) Four discrete expression states explained 29.4% of the overall transcriptomic variance (PERMANOVA, p-value < 0.001, n = 28) across Ca. E. taraoceanii populations. One state (cluster 1) was exclusive to larger organismal size fractions. Leaves represent transcriptomic profiles and the dendrogram represents dimensionality-reduced distances (Methods). Genes associated with BGCs, secretion systems, degradative enzymes and predatory markers were differentially expressed across the states and represented the most discriminatory categories compared to 200 KEGG pathways (Supplementary Table 4). (b) We investigated the metagenomic detection of the 8,500 genes encoded by the Ca. E. taraoceanii representative, using methodology identical to the transcriptomic analyses (Methods). In samples where the 10 marker genes were detected, we counted the number of genes with one or more insert(s). We found that the 8,500 genes were detected in several ocean basins and different size fractions, with variation in detection rates likely due to variable sequencing depths across samples and datasets. This indicates, at least for the gene set covered by the reconstructed genome, that niche partitioning may be driven by gene expression changes rather than gene content variation. (c) Distribution of the number of genes depending on the number of samples they were detected in. (d) Number of genes detected across the different metatranscriptomic samples. All BGCs encoded by (e) Ca. Autonomicrobium septentrionale, (f) Ca. Amphithomicrobiurn indiannii and (g) Ca. Amphithomicrobiurn mesopelagicum representatives were found to be expressed in the natural environment (in the 623 Tara Oceans metatranscriptomic samples). Some displayed near constitutive expression while others appear to be tightly regulated across the metatranscriptomes studied here. Filled circles indicate samples where active transcription was detected. Orange data points indicate values below or above a log2 fold change from the constitutive expression rate of housekeeping genes. All the BGCs encoded by Ca. E. malaspinii were also found to be expressed (Fig. 3c). The expression of Ca. E. malaspinii BGCs could not be investigated since that species was not sufficiently abundant in the epipelagic and mesopelagic ocean, the only layers for which metatranscriptomes were available.
Extended Data Fig. 8 | Visual representations of BGCs encoded by Ca. E. malaspinii. Visual representations and manual annotations of some Ca. Eudoremicorbium specific BGCs discussed in Supplementary Information, i.e. BGC 2.2(a), BGC 54.1(b) and BGC 34.1(c). Colour-coding corresponds to predicted enzyme domains and modifications. These can be interactively explored here: https://sunagawalab.ethz.ch/share/microbiomics/ocean/db/1.0/marine_eremios/annotations/MALA_SAMN05422137_METAG_HLLJDLBE/antismash/MALA_SAMN05422137_METAG_HLLJDLBE-antismash/.
Extended Data Fig. 9 | EmbM structural prediction and comparison to CylM (PDB: 5DZT); NMR and Mass spectrometry data for modified EmbA peptides. (a) CylM crystal structure. Coloured domains are involved in phosphorylation/dehydration and the domain in grey is responsible for cyclization. (b) EmbM structure prediction, highlighting similarities to CylM. (c) CylM active site. Residues in pink are proposed to be involved in phosphorylation and residues in purple are necessary for elimination. (d) Modelled active site of EmbM. (e) Multiple sequence alignment showing that mutated residues in the catalytic site are conserved across the independent Ca. E. malaspinii reconstructions. (f) Overlay of 2D [13C,1H] HSQC spectra of EmbA and modified EmbA (EmbAM). Multiplicity editing leads to positive signals for CH and CH3 groups (EmbA: blue, EmbAM: red) and negative signals for CH2 groups (EmbA: cyan, EmbAM: magenta). Regions of interest are identified with boxes and major buffer signals are labelled. (g) Serine Cβ region. Serine Cβ moieties are identified by the negative sign of the signal (CH2-group), and the average chemical shift of 63.8 ppm. A change of the Cβ chemical shift of typically +3 ppm is expected upon a phosphorylation event, but there are no negative signals visible in the expected region in the EmbAM spectrum (grey box). (h) and (i): threonine Cγ and Cβ regions, respectively, as identified by chemical shift and sign of signals. In the EmbAM spectra, additional signals are visible at expected chemical shifts for phosphorylated threonine residues, i.e. at a 13C chemical shift of 20.5 ppm for Cγ (grey arrows in h) and 70 ppm for Cβ (grey arrows in i). (j) HR-MS/MS fragmentation of EmbA core at different modification stages (cleaved with LaHT150). (k) Mass spectrum of dehydrated EmbA species: unmodified, single- and double dehydrated EmbA core (top); unmodified, single- and double dehydrated EmbA cleaved with trypsin (middle); and unmodified, single- and double dehydrated, DTT adduct of EmbA cleaved with trypsin (bottom).
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | In vitro EreM 13C-labelling experiments. NMR and MS2-fragmentation data; EreM phylogenetic tree; EreM synthetic core mass shifts and MS2-fragmentation data. (a) Mass spectra of the LahT-digested, single methylated Nhis-EreA from in vitro EreM assays with 13C-labelled SAM (top, red) and non-labelled SAM (bottom, grey). Top: Mass spectrum of the LahT-released 48 aa long EreA core with an N-terminal extension of two glycine residues (m/z = 1471.1263 Da) and the corresponding 13C-labelled methylated (m/z = 1476.1310 Da) core with an N-terminal extension of two leader-derived glycine residues. The mass shift of 5.00 Da (z = 3) is highlighted by a red arrow. Bottom: Mass spectrum of the LahT-released 48 aa long EreA core with an N-terminal extension of two glycine residues (m/z = 1471.1272 Da) and the corresponding methylated (m/z = 1475.7971 Da) core with an N-terminal extension of two glycine residues. The mass shift of 4.67 Da (z = 3) is highlighted by a grey arrow. (b) MS2-fragmentation detected for the 13C-labelled core with an N-terminal extension of two glycine residues (m/z = 1476.1310 Da). All y-ions show masses corresponding to fragments with the addition of a 13C-labelled methyl group (red). All b-ions show masses corresponding to a fragment with no modification (black). The resulting fragmentation pattern suggests 13C-labelled methylation at the C-terminal cysteine residue (red box). MS2-fragmentation data are available in Supplementary Table 5. (c) Overlay of a C-H decoupled (red) and standard (blue) proton NMR of an in vitro EreM assay with 13C-labelled SAM. The peak splitting of the singlets at 2.03 ppm and 2.88 ppm indicates the 13C-H bonds for these protons. (d) HSQC NMR of an in vitro EreM assay with 13C-labelled SAM. The spectrum shows two single signals at 2.03/17.3 ppm (yellow box) and 2.88/25.9 ppm (red box). Another four signals are detected downfield: 3.46/70.0 ppm, 3.35/70.0 ppm, 3.64/62.2 ppm and 3.69/74.6 ppm (grey box). Comparison with the literature suggest the presence of a 13C-S bond at 2.03/17.3 ppm (yellow box) from residual 13CH3-l-methionine and of a 13C-N bond at 2.88/25.9 ppm (red box) from a methylated amide105–109. The remaining four signals are suggested to originate from the Tris-buffer of the reaction mixture (grey box). (e) Maximum-likelihood tree of FkbM-family methyltransferase (PF05050) Hidden Markov Model (HMM) hits within BGCs for natural products in the MIBiG 2.0 database (Supplementary Table 5). Outgroups involved in proteusin biosynthesis from a different methyltransferase protein family (PF05175) are shown in grey text. Branch support values are estimated using the 5,000 ultrafast bootstrap approximation17 in IQ-TREE 28. Letter in ‘MT-type’ column indicates documented N- or O-methyltransferase activity from publications based on genetic knockout or heterologous expression studies (coloured) or bioinformatic evidence, biosynthetic logic, and final natural product structure (grey). To date, EreM from this study is the only FkbM-family enzyme with reported N-methyltransferase activity in a characterized biosynthetic pathway. Coloured points in BGC type columns indicate the majority of FkbM-family enzymes are contained with PKS, NRPS, or Other (e.g., nucleoside antibiotic) biosynthetic pathways. Thus EreM is also the only FkbM-family methyltransferase characterized in a RiPP cluster to date. (f) EreA core variants generated in this study. Mutation or truncation sites are highlighted in yellow. (g) Mass shifts of +14.01 Da corresponding to methylation of the EreA core and variants were observed expressed with EreM as compared to controls without EreM (data not shown for core variants, but results are in accordance with natural core controls). All EreA variant + EreM co-productions were tested with and without EreD, but EreD co-productions are pictured. since epimerized (EreA + EreD) cores have better solubility and higher concentrations. (h) Proteinase K-generated fragments of the wild-type EreA core following co-productions with EreIMD reveal a mixture of variable methylation patterns. (i) MS2-fragmentation of the wild-type EreA core after co-production with EreIMD. Mass shifts corresponding to up to 6 non-radical methylations (+84.09 Da) were observed and were localized to valine residues (highlighted in light blue, N-Me). Dashed lines around boxes indicate uncertainty regarding the position. MS2-fragmentation data are in Supplementary Table 5.
Extended Data Fig. 11 | EreI mass shift and high resolution tandem mass spectroscopy (MS$^2$); EreD retention time shift and ODIS and advanced Marfey’s analysis for d-Val/d-Ala.

(a) Mass spectra and MS$^2$-fragmentation of the LahT-digested Nhis-EreA modified by EreI, EreM, and EreD. Bottom: A mass shift of +15.99 Da corresponding to incorporation of one oxygen into the mono-methylated core (EreA + EreIMD, [M+3H]$^3+$ = 1443.1060 Da) was observed after co-expressions of ereAIMD. Top: No oxygen incorporation was observed in Nhis-EreA modified by EreM and EreD (EreA + EreMD) controls lacking the aspartinyl-asparaginyl β-hydroxylase protein family protein, EreI. Notably, the +15.99 Da modification was only observed on methylated, LahT-released EreA cores ([M+3H]$^3+$ = 1437.7754 Da) and not on the non-methylated core ([M+3H]$^3+$ = 1433.1124 Da) as observed by in vivo ereAI or ereAID co-expressions and in vitro assays with purified NHis-EreI and NHis-EreA or NHis-EreA modified by EreD. (b) MS$^2$-fragmentation of [M+3H]$^3+$ = 1443.1060 Da. The data localize oxygen (Ox) incorporation to the C-terminus of the peptide but cannot distinguish between terminal cysteine (C46) or valine (V45). MS$^2$-fragmentation data including calculated and observed masses for all b- and y-ions are available in Supplementary Table 5. (c) Extracted ion chromatograms (EICs) at 1433.1123 Da of LahT-digested precursors Nhis-EreA (black trace, top) and epimerized Nhis-EreA from co-productions with radical SAM epimerase EreD (orange trace, bottom), which show a retention time shift of 0.47 min. No mass shift was observed, since l- to d- amino acid epimerization is a mass-neutral modification, requiring the use of the orthogonal D$_2$O induction systems (ODIS) to localize epimerization sites. (d) MS$^2$-fragmentation of [m/z = 1447.47308 Da] patterns enabled localization of epimerized residues to V10, A12, A14, V16, V18, V29 (orange asterisks) and either V44 or V45 (grey asterisk). MS$^2$-fragmentation data are available in Supplementary Table 5. (e) Expressions using ODIS results in a shift of +7.04 Da corresponding to the incorporation of 7 deuterium atoms. A mixture of products is observed due to slowdown of epimerization in the presence of deuterium. (f) EICs from advanced Marfey’s analysis of epimerized and modified cores from two heterologous hosts: E. coli (pink) and M. aerodenitrificans (light and dark blue) consistently yielded d-Val and d-Ala (grey shading) as the only d-amino acids detected in EreA cores, as compared to d-Thr, d-Asp, and d-Ser standards (not shown), for which no corresponding peaks were detected. Both d-Val and d-Ala were measured in ratios of 1:3 to their l-amino acid counterparts. Based on EreA core amino acid composition, these ratios correspond to approximately 2 d-Ala and 5 d-Val per core consistent with ODIS results.
Extended Data Fig. 12 | EreB mass shift, MS²-fragmentation data and advanced Marfey’s analysis for tert-Leu. (a) Co-production of Nhis-EreA with epimerase EreD and the B₁₂-dependent radical SAM C-methyltransferase EreB in M. aerodenitrificans Δaer with a knocked-out aeronamide BGC yielded a mixture of C-methylated products with mass shifts corresponding to up to 7 methylations (+98.13 Da). (b) Alignment of a mixture of differently-modified fragments detected by MaxQuant analysis of proteinase K digested Nhis-EreA following co-productions with EreDB. (c) Representative MS²-fragmentation of EreA core following co-production with EreDB at m/z = 1466.4819 Da. Observed and calculated masses for b- and y-ions are in Supplementary Table 5. Modification sites (dark blue, C-Me) were localized to V9, V13, V15, V35, V37, V44, and V45. (d) Total ion chromatogram (TIC, black) and EICs from advanced Marfey’s analysis of C-methylated core from co-productions of EreA + EreDB in M. aerodenitrificans. The unspiked sample (dark grey) is compared to identical samples that were spiked with synthetic standards: tert-Leu (orange), l-Leu (blue), allo-Ile (brown), and dl-Ile (purple). The grey box is an inset on a narrower retention time from the left panel highlighting a peak shoulder from the M. aerodenitrificans EIC corresponding to tert-Leu (yellow shading).
Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a  Confirmed

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ The statistical test(s) used AND whether they are one- or two-sided

Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ A description of all covariates tested

☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated measures of uncertainty (e.g. confidence intervals)

☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

Give P values as exact values whenever suitable.

☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about: availability of computer code

| Data collection | Data collected from ENA was downloaded from ENA using their API. All other publicly available data was downloaded directly from the sources specified in the data availability statement. |
|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data analysis   | The data was analyzed using a pipeline developed as part of this study as well as with additional ad hoc python and R scripts and with the following softwares: BBMap v.38.71, metaSPAdes v3.11.1 and v3.12, BWA v0.7.17-r1188, MetaBAT2 v2.12.1, CheckM v1.0.13, Anvi’o v5.5.0, dRep v2.5.4, Speci, GTDB-Tk v1.2.2, Prokka v1.14.5, fetchMGA v1.2, emapper v2.0.1, DIAMOND v0.9.30, antiSMASH v5.1.0 and v5.0.0, CD-HIT v4.8.1, mOTUs v2.5.1, Big-Slice v3.1, Iqtree v2.0.3, MUSCLE v3.8.1551, trimal v1.4.1, Trailar v1.1.2, TXSSCAN v1.0.2, FeatureCounts v2.0.1, PlasmidFinder v2.1, PlasFlow v1.1.0, S Barker v2.1, VirSorter v1.0.5, DeepVirFinder v1.0.0, EukReps v0.6.5, contigs v1.0.0, STAG v0.7.0, GECCO v0.4.4, MMSEQS v13.45111, MAGFFT v7.310, Python >= 3.6 with the packages pandas [v1.0.0-1.3.4], biopython [v1.73], umap-learn [v0.5.2], hdbscan [v0.8.28], scikit-learn [v1.0.2] and R [v4.0.0-0.4.1.2] with the packages ggplot2 [v3.3.0-3.3.5], tidyverse [v1.3.1], vegan [v2.5.7], ggrepel [v3.3.0.901], and tidytree [v0.3.6], treeio [v1.19.1] and UpSetR [v1.4.0]. The code used in this study is accessible at https://github.com/SushiLab/mappipe and archived at Zenodo [https://doi.org/10.5281/zenodo.6393817]. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.
Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The metagenomic and metatranscriptomic data used in this study was downloaded from the European Nucleotide Archive (ENA) and their accessions are summarized in Supplementary Table 1. Publicly available genomes were downloaded from https://dsi.org/10.6084/m9.flgshare.4902923 for manually curated MAGs from Tara Oceans, from ENA using the project accession PRJEB33281 for GORC and from https://mmp2.sfb.uit.no/databases/ for MarDB. The GEM MAGs were downloaded from https://portal.nersc.gov/GEM/. MAGs contained in the GTDB r89 were downloaded from https://data.gtdb.ecogenomic.org/releases/release89/. The MiBi6 and Big-FAM databases can be accessed at https://mibig.secondarymetabolites.org/ and https://bigfam.bioinformatics.nl/, respectively. The data produced in this study, including metagenomic assemblies, bins and MAGs have been deposited at the European Nucleotide Archive under the accession PRJEB69561 and individual accessions are summarized in Supplementary Table 1. Other supporting data has been deposited on Zenodo (https://doi.org/10.5281/zenodo.4474310), and the OMD can be interactively accessed at https://microbiomes.io/ocean/. Additional material generated in this study is available upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see https://nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size | Samples sizes were defined by the availability of published data that were used to perform the analyses.
Data exclusions | No data were excluded from the analyses.
Replication | All post-translational modification of the peptides reported in this study were supported by replicated experiments, bio-activity assays included replicates (n >= 3) and all replicates were successful.
Randomization | For the different analyses conducted in this study, all samples were processed similarly and thus randomization was not necessary.
Blinding | For the different analyses conducted in this study, all samples were processed similarly and thus blinding was not necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a | Involved in the study
☒ ☐ Antibodies
☒ ☐ Eukaryotic cell lines
☒ ☐ Palaeontology and archaeology
☒ ☐ Animals and other organisms
☐ ☐ Human research participants
☒ ☐ Clinical data
☑ ☐ Dual use research of concern

Methods

n/a | Involved in the study
☒ ☐ ChIP-seq
☒ ☐ Flow cytometry
☒ ☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s) | HeLa cells.

March 2021
| Authentication       | Not authenticated. |
|----------------------|---------------------|
| Mycoplasma contamination | Not tested.       |
| Commonly misidentified lines (See [IGAC register](#)) | Not applicable. |