Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes

Nature Microbiology

Gavriilidou, Athina; Kautsar, Satria A.; Zaburannyi, Nestor; Krug, Daniel; Müller, Rolf et al

https://doi.org/10.1038/s41564-022-01110-2

ANALYSIS
https://doi.org/10.1038/s41564-022-01110-2

nature microbiology

Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes

Athina Gavriilidou[1,7], Satria A. Kautsar[2,7], Nestor Zaburannyi[3,4], Daniel Krug[3,4], Rolf Müller[3,4], Marnix H. Medema[2,8] ✉ and Nadine Ziemert[1,5,6,8] ✉

Bacterial specialized metabolites are a proven source of antibiotics and cancer therapies, but whether we have sampled all the secondary metabolite chemical diversity of cultivated bacteria is not known. We analysed ~170,000 bacterial genomes and ~47,000 metagenome assembled genomes (MAGs) using a modified BiG-SLiCE and the new clust-o-matic algorithm. We estimate that only 3% of the natural products potentially encoded in bacterial genomes have been experimentally characterized. We show that the variation in secondary metabolite biosynthetic diversity drops significantly at the genus level, identifying it as an appropriate taxonomic rank for comparison. Equal comparison of genera based on relative evolutionary distance revealed that *Streptomyces* bacteria encode the largest biosynthetic diversity by far, with *Amycolatopsis*, *Kutzneria* and *Micromonospora* also encoding substantial diversity. Finally, we find that several less-well-studied taxa, such as Weeksellaceae (Bacteroidota), Myxococcaceae (Myxococcota), *Pleurocapsa* and Nostocaceae (Cyanobacteria), have potential to produce highly diverse sets of secondary metabolites that warrant further investigation.

Specialized metabolites (also called secondary metabolites) are biomolecules that are not essential for life but rather offer specific ecological or physiological advantages to their producers, allowing them to thrive in particular niches. These natural products (NPs) are more chemically diverse than the molecules of primary metabolism, varying in both structure and mode of action among different organisms[1]. Historically, microbial NPs and their derivatives have contributed and continue to contribute a substantial part of chemical entities brought to the clinic, especially as anticancer compounds and antibiotics[2–4]. Regrettably, the emergence of antibiotic-resistant pathogens[3] concomitant with a stagnation in antimicrobial discovery pipelines[2,4] is leading to a global public health crisis[3].

Nonetheless, genomics-based approaches to NP discovery[5,6] have revealed a largely untapped and much more diverse source of biosynthetic potential within genomes[3,7]. These findings were possible following the discovery that bacterial genes encoding the biosynthesis of secondary metabolites are usually located in close proximity to each other, forming recognizable biosynthetic gene clusters (BGCs). However, while the numbers and kinds of BGCs clearly differ across microbial genomes[7,8] and metabolomic data indicate that some biosynthetic pathways are unique to specific taxa[9], a systematic analysis of the taxonomic distribution of BGCs has not yet been performed. Similarly, while useful estimates of the chemical diversity of specific taxa have been provided[8], methodical comparisons across taxa are lacking. Because of this, the scientific community appears undecided on the best strategy for natural products discovery: should the established known NP producers be studied further or should the community be investigating underexplored taxa[7,10]? A relatively recent question is how much chemical diversity is hidden in uncultured bacteria. Metagenomic assembled genomes from uncultured bacteria have demonstrated a big potential of unknown BGCs[7]. It is unclear to what extent unexplored associated ecological niches and (micro)environments are also associated with unique and unexplored chemistry.

Here we harnessed recent advances in computational genomic analysis of BGCs to survey the enormous amount of genomic data accumulated by the scientific community so far. Using a global approach based on more than 170,000 publicly available genomes, we created a comprehensive overview of the biosynthetic diversity found across the entire bacterial kingdom. We clustered 1,185,995 BGCs into 62,449 gene cluster families (GCFs), and calibrated the granularity of the clustering to make it directly comparable to chemical classes as defined in the NPAtlas[11]. This facilitated an analysis of the variance of diversity across major taxonomic ranks, which showed the genus rank to be the most appropriate level for comparing biosynthetic diversity across homogeneous groups. This finding allowed us to conduct comparisons within the bacterial kingdom. Evident patterns emerged from our analysis, revealing popular taxa as prominent sources of both actual and potential biosynthetic diversity, and multiple yet uncommon taxa as promising producers.

Biosynthetic diversity of the bacterial kingdom

To assess the global number of GCFs found in sequenced bacterial strains, we ran AntiSMASH[12] on ~170,000 genomes from the NCBI RefSeq database[13] (Supplementary Table 1), spanning 48 bacterial phyla containing 464 families (according to the Genome Taxonomy DataBase classification, GTDB[14]). We also included almost 50,000 bacterial metagenome assembled genomes (MAGs) from 6 metagenomic projects of various origins[15–20] (Table 1 and Supplementary

[1]Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany. [2]Bioinformatics Group, Wageningen University, Wageningen, the Netherlands. [3]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS)—Helmholtz Centre for Infection Research (HZI), Saarbrücken, Germany. [4]German Center for Infection Research (DZIF), Partner site Hannover-Braunschweig, Braunschweig, Germany. [5]Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen, Germany. [6]German Centre for Infection Research (DZIF), Partnersite Tübingen, Tübingen, Germany. [7]These authors contributed equally: Athina Gavriilidou, Satria A. Kautsar. [8]These authors jointly supervised this work: Marnix H. Medema, Nadine Ziemert. ✉e-mail: marnix.medema@wur.nl; nadine.ziemert@uni-tuebingen.de

Table 1 | Input datasets and biosynthetic diversity with different BiG-SLiCE cut-offs

Dataset		Genomes	BGCs	Gene cluster families			
				T = 0.4	T = 0.5	T = 0.6	T = 0.7
Complete dataset	All RefSeq bacteria	170,549	1,060,592	**51,052**	37,785	28,057	19,152
	Bacterial MAGs[a]	47,098	125,403	**21,354**	–	–	–
	Total	217,647	1,185,995	**62,449**	–	–	–
RefSeq bacteria with known species taxonomy	Complete genomes	16,004	94,904	**16,984**	13,546	10,399	7,151
	Draft genomes	147,265	913,642	**37,123**	27,748	20,638	14,016
	Total	163,269	1,008,546	**41,870**	31,237	23,227	15,766

The 'complete dataset' was used for the computation of the actual and potential biosynthetic diversity found in all cultured (and some uncultured) bacteria. The dataset 'RefSeq bacteria with known species taxonomy' was used for pinpointing the emergence of biosynthetic diversity, for which accurate taxonomic information was needed, and for identifying groups of promising producers. The 'T's under gene cluster families represent different BiG-SLiCE l2-normalized euclidean thresholds; the values under T = 0.4 stand out due to it being considered the most suitable cut-off. BGC to GCF assignment for each threshold can be found in Supplementary Tables 2–5. [a]MAG sources: bovine rumen[15], chicken caecum[16], human gut[17], ocean[18], uncultivated bacteria[19], various sources[20].

Table 1). To accurately group similar BGCs – which probably encode pathways towards the production of similar compounds – into GCFs across such a large dataset, we used a slightly modified version of the BiG-SLiCE tool[21], which has been calibrated to output GCFs that match the grouping of known compounds in the NPAtlas database[11] (see Methods, Quantification of biosynthetic diversity with BiG-SLiCE). The resulting GCFs were then used to measure biosynthetic diversity across taxa.

The number of GCFs in RefSeq ranged from 19,152 to 51,052 depending on the cut-off used by BiG-SLiCE (Table 1). While, as expected, the pure numbers of the analysis changed on the basis of the l2-normalized euclidean threshold, the overall tendencies observed remained the same (Fig. 1a and Supplementary Fig. 1). The effect that the chosen threshold has on these results presented a challenge to our investigation, as previous estimations have also shown great heterogeneity when different thresholds were used[7,8], precluding direct comparisons of their predictions. As each BGC can be considered a proxy for its encoded pathways and their products, differing thresholds will result in different degrees of granularity in the grouping of compound structures (Extended Data Fig. 1). Nevertheless, linear relationships are not always applicable, as shown previously[22], and a specific threshold will need to be set anyway to make comparisons possible. For this, we sought to directly relate the choice of our BGC clustering threshold to the clustering of their compound structures. NPAtlas, a database of known microbial small molecules, provides hierarchical clustering of the compound structures via Morgan fingerprinting and Dice similarity scoring[11]. As many as 947 compounds in the NPAtlas are mapped to a known BGC in the MIBiG repository[23], giving us the opportunity to use them as an anchor for choosing our clustering threshold. After mapping the BiG-SLiCE groupings of known BGCs from the MIBiG to the compound clusters in the NPAtlas (Supplementary Fig. 2), we chose a threshold of 0.4, as it provided the most congruent agreements between the two groupings, with a v-score = 0.94 (out of 1.00) and ΔGCF = −17.

This calibration of thresholds of GCFs to families of chemical structures allowed us to perform a rarefaction analysis to assess how genomically encoded biochemical diversity (expressed as the number of distinct GCFs) increases with the number of sequenced and screened genomes (Fig. 1b). The curve appears far from saturated, while the slope is even steeper if the bacterial MAGs are included in the analysis. When compared to the number of chemical classes documented in the NPAtlas[11] database (Fig. 1b), it appears that, to date, only ~3% of the kingdom's biosynthetic diversity has been experimentally assessed.

In an attempt to evaluate the potential contribution of metagenomic data to NP discovery, we studied how many of the GCFs found in the MAGs datasets were unique to this dataset (Fig. 1c). Around 53.4% of GCFs in the MAGs were not found in the RefSeq strains or in the Minimum Information about a Biosynthetic Gene cluster database (MIBiG[23]). Paradoxically, in Fig. 1b, the contribution of MAGs does not reflect this finding, but this is most probably because the metagenomic dataset is of limited size and does not cover the full microbial diversity of the biosphere. An analysis of the uniqueness of GCFs found in different environments, although limited to only one[20] of the MAGs datasets, suggests that a connection exists between the biogeography of microbiomes and the uniqueness of their biosynthetic diversity, as the majority of GCFs (74.43 %) are biome-specific (Extended Data Fig. 2 and Supplementary Table 7). The latter finding is concordant with recent proof that most genes have a strong biogeography signal[24].

Variation in biosynthetic diversity drops at the genus level

To identify the most promising bacterial producers, it is important to compare them at a specific taxonomic level. Several studies indicate that there is substantial discontinuity in how BGCs are distributed across taxa: 'lower' taxonomic ranks such as species within a genus carry more similar biosynthetic diversity, than 'higher' taxonomic ranks such as phyla within a kingdom. To assess which taxonomic rank is the most appropriate to evaluate biosynthetic potential, we aimed to determine up to which taxonomic level the biosynthetic diversity remains homogeneous within that taxon. For this analysis, from our initial dataset, we left out the MAGs and only used the RefSeq bacterial strains as taxonomic assignment up to species rank (based on GTDB[14]) was available only for the latter dataset(Table 1).

We first decorated the GTDB[14] bacterial tree with GCF values from the BiG-SLiCE analysis (Fig. 2a), revealing the biosynthetic diversity found within currently sequenced genomes at the phylum rank. It immediately stood out that biosynthetic diversity was differently dispersed among the bacterial phyla, in accordance with published data[7,25]. As expected for known NP producers, the phyla Proteobacteria and Actinobacteria appeared particularly diverse[8,26,27]. However, these phyla are among the most studied and therefore the most sequenced[8,26,27]—a bias that is addressed later in the study.

Next, we examined whether the diversity of each phylum contributed to the domain's total diversity, or if there was overlap among them. For this reason, we depicted the number of unique GCFs within each phylum, as well as the pairwise overlaps (Fig. 2b). In most phyla, the vast majority (on average 73.81 ± 20.35%) of their GCFs appeared to be unique to them and not found anywhere else. This is coherent with the fact that horizontal gene transfer events, although relatively frequent for BGCs[28], are much more common among closely related taxa[29].

Once we obtained information on the diversity of different phyla, as well as the rest of the major taxonomic ranks (classes, orders, families, genera, species), we proceeded to determine at which taxonomic rank biosynthetic diversity levels no longer show high variability. Therefore, we conducted a variance analysis that included each taxonomic rank, from phylum to species. For each rank, the variance value was computed on the basis of the number of GCF values of immediately lower-ranked taxa (see Methods, Variance analysis). The distribution of these variance values for each rank is visualized in Fig. 3a.

There is a noticeable drop in the range of variance values for each rank, while diversity becomes highly homogeneous at the species level (Fig. 3a,b). The plunge is most striking from the family to the genus level (Fig. 3a), with even the outliers all falling under the 10^3 line in the genus rank. Different species within a genus are likely to display uniform biosynthetic diversity, while much dissimilarity is observed between different genera belonging to the same family (Fig. 3b). Additional statistical analysis confirmed the significance of this observation (Supplementary Fig. 3), thus pinpointing, probably for the first time, the genus rank as the most appropriate for comparative analyses.

Taxa that are sources of substantial biosynthetic diversity

The identification of the genus level as the most informative rank to measure biosynthetic diversity across taxonomy paved the way for a comprehensive comparative analysis of biosynthetic potential across the bacterial tree of life. However, to be able to systematically compare diversity values among groups, said groups need to be uniform. In this case, a common phylogenetic metric was necessary. We chose relative evolutionary divergence (RED) and a specific threshold that was based on the GTDB's range of RED values for the genus rank[14] to define REDgroups: groups of bacteria analogous to genera but characterized by equal evolutionary distance (see Methods, Definition of REDgroups). Our classification revealed the inequalities in within-taxon phylogenetic similarities among the genera, with some being divided into multiple REDgroups (for example the *Streptomyces* genus was split into 21 REDgroups: Streptomyces_RG1, Streptomyces_RG2 etc.) and some being joined together with other genera to form mixed REDgroups (for example Burkholderiaceae_mixed_RG1 includes the genera *Paraburkholderia*, *Paraburkholderia_A*, *Paraburkholderia_B*, *Burkholderia*, *Paraburkholderia_E* and *Caballeronia*). This disparity among the genera reaffirmed the importance of defining the REDgroups as a technique that allowed for fair comparisons among bacterial NP producers.

The resulting 3,779 REDgroups showed huge differences in biosynthetic diversity as measured by the numbers of GCFs found in genomes sequenced from these groups so far, with the maximum diversity at 3,339 GCFs, average at 17 GCFs and minimum at 1 GCF. Nevertheless, the variance of diversity within the REDgroups was even more uniform than in the genera (Supplementary Fig. 4). Some of the top groups (Supplementary Table 8) included known rich NP producers, such as *Streptomyces*, *Pseudomonas_E* and *Nocardia*[23,26,27,30].

Although very informative, this analysis is biased because of large differences in the number of sequenced strains among the groups, with economically or medically important strains having been sequenced more systematically than others. To overcome

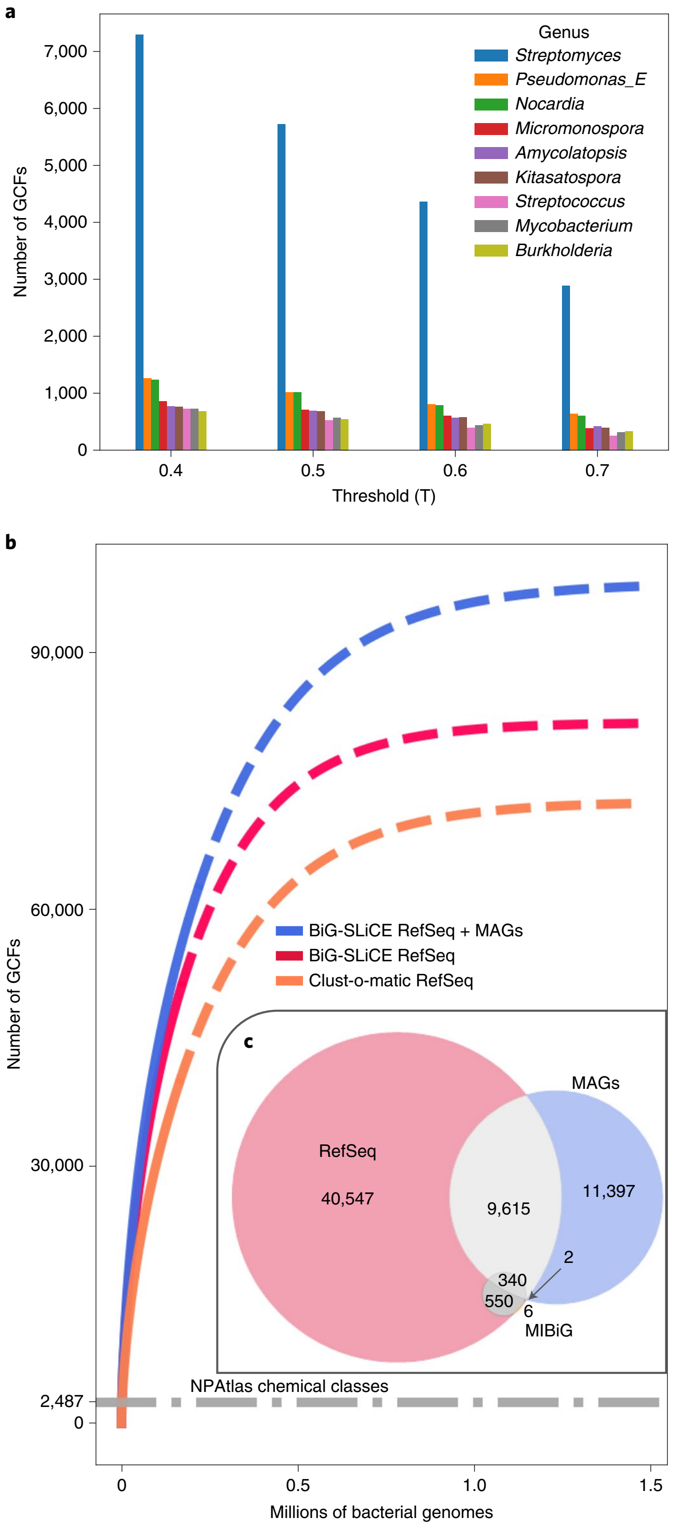


Fig. 1 | Biosynthetic diversity of the sequenced bacterial kingdom. **a**, Barplots of GCFs (as defined by BiG-SLiCE) of the nine most biosynthetically diverse genera using different thresholds (T). The absolute number of GCFs changes from threshold to threshold, but the general tendencies (highest to lowest GCF count) are consistent between them. **b**, Rarefaction curves of all RefSeq bacteria based on BiG-SLiCE (red) and clust-o-matic (orange), and rarefaction curve of the complete dataset, which includes bacterial MAGs (blue), based on BiG-SLiCE. BiG-SLiCE GCFs were calculated with T = 0.4. Clust-o-matic GCFs were calculated with T = 0.5. The solid lines represent interpolated and actual data, while the dashed lines represent extrapolated data. The number of chemical classes documented in the NPAtlas[11], which come from bacterial producers (grey dashed line; 2,487), corresponds to 2.5–3.3% of the predicted potential of the bacterial kingdom (number of GCFs at 1.6 million genomes). The *Y* values (number of extrapolated GCFs) at the right end of the graph are 97,760.12 (blue), 81,748.32 (red) and 72,411.11 (orange). **c**, Venn diagram of GCFs (as defined by BiG-SLiCE, T = 0.4) of the bacterial RefSeq, MIBiG[23] and bacterial MAGs datasets. More information on the MIBiG dataset can be found in Supplementary Table 6. About 53.4% of the GCFs of MAGs are unique (blue shape) to this dataset.

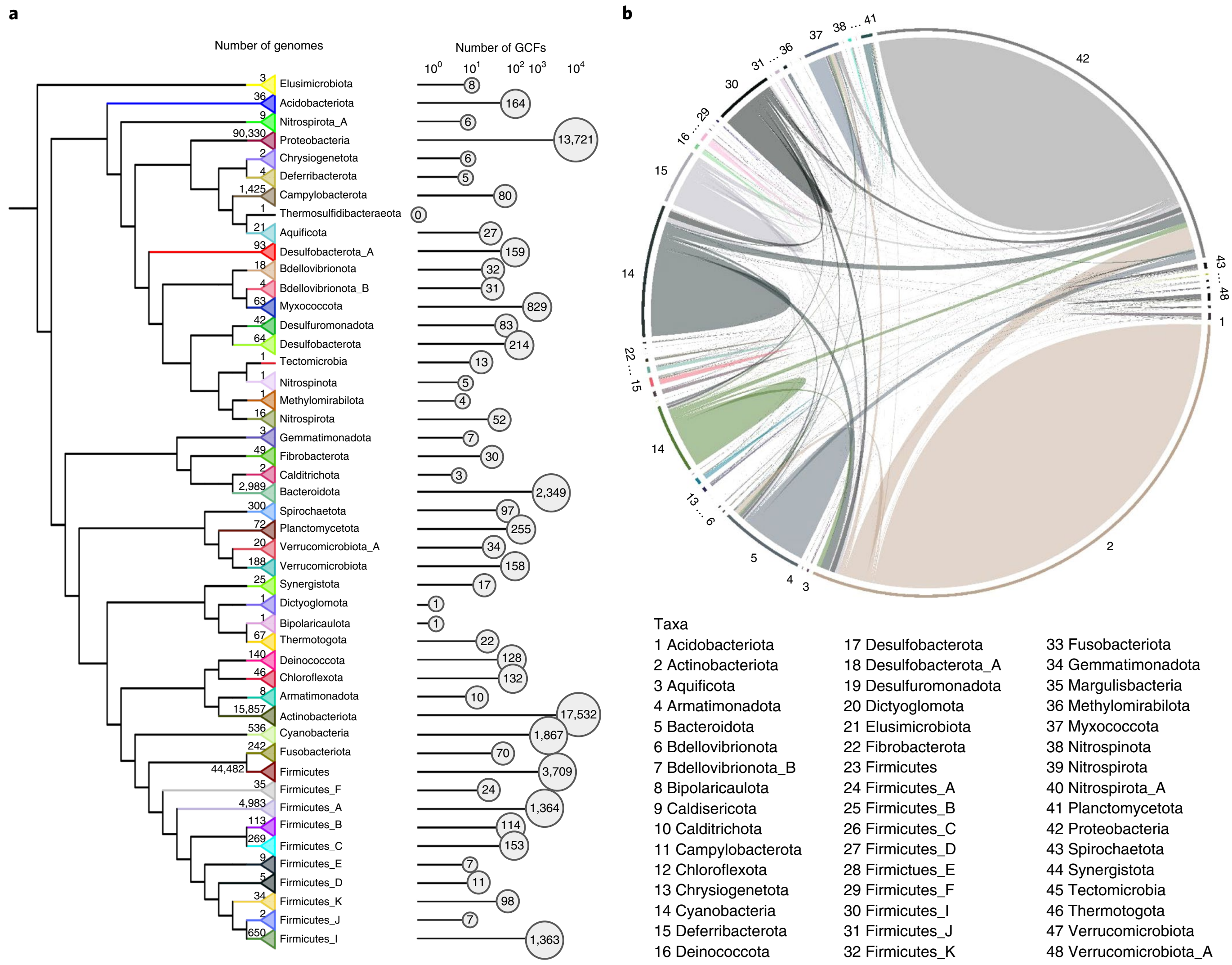


Fig. 2 | Comparison of biosynthetic diversity among phyla. **a**, The GTDB[14] bacterial tree was visualized with iTOL[63] v6.5.2, decorated with GCF values (as defined by BiG-SLiCE at T = 0.4), collapsed at the phylum rank and accompanied by barplot of GCFs in logarithmic scale (10^0 to 10^4). The number of genomes belonging to each phylum is displayed next to the tree's leaf nodes. **b**, GCFs (as defined by BiG-SLiCE, T = 0.4) unique to phyla (solid shapes) and with pairwise overlaps between phyla (ribbons), visualized with circlize[64]. Each phylum has a distinct colour. Actinobacteriota (2) and Proteobacteria (42) seem particularly rich in unique GCFs.

this bias, rarefaction analyses were conducted for each REDgroup (Fig. 4b and Supplementary Table 8), as performed in previous studies[31,32]. Additionally, to examine how effectively this method overcomes the sequencing bias, a random sampling approach was taken (see Methods, Random sampling), which showed comparable results to the original analysis (Supplementary Table 9). With all the information on REDgroups, and to provide a global overview of the actual biosynthetic diversity and the potential number of GCFs, we modified and complemented the bacterial tree from Parks et al.[14], as shown in Fig. 4a (Extended Data Fig. 3). The dispersion of these values across the various phyla can also be seen, with the exceptional outliers standing out: Streptomyces_RG1, Streptomyces_RG2, Amycolatopsis_RG1, Kutzneria and Micromonospora. All these are groups known for their NP producers[8,26,27,33] and they remain at the top (Supplementary Table 8), seemingly having much unexplored biosynthetic potential.

To ensure that our conclusions are not the product of algorithmic artefacts, we re-ran the analysis using an alternative method of quantifying biosynthetic diversity, which was developed independently, yet for the same purpose. This alternative approach, called clust-o-matic, is based on a sequence similarity all-versus-all distance matrix of BGCs and subsequent agglomerative hierarchical clustering to form GCFs (see Methods, Quantification of biosynthetic diversity with clust-o-matic). Similar to BiG-SLiCE, we calibrated the threshold for clust-o-matic on the basis of the NPAtlas clusters. When comparing the results (Fig. 4c,d and Supplementary Table 8), the two algorithms appeared to identify very similar trends despite slight differences in absolute numbers.

Streptomyces, even when split into multiple REDgroups, is in the top groups both based on the known biosynthetic diversity and on the estimated potential values. A total of 5,908 (+103 Streptomyces_B, +39 Streptomyces_C, +16 Streptomyces_D) GCFs appear to be unique to the group, even among other phyla (Fig. 5a). This is in agreement with previous studies investigating how much overlap there is among the main groups of producers[34]. Of note, streptomycetes appear to be the source of a good percentage of the biosynthetic diversity attributed to the Actinobacteria phylum, as seen in Fig. 5b.

However, taxa less popular for NP discovery also show promise, as was evidenced by a comparison of our results with data from the NPASS database of Natural Products[35] (Fig. 5c). Among the

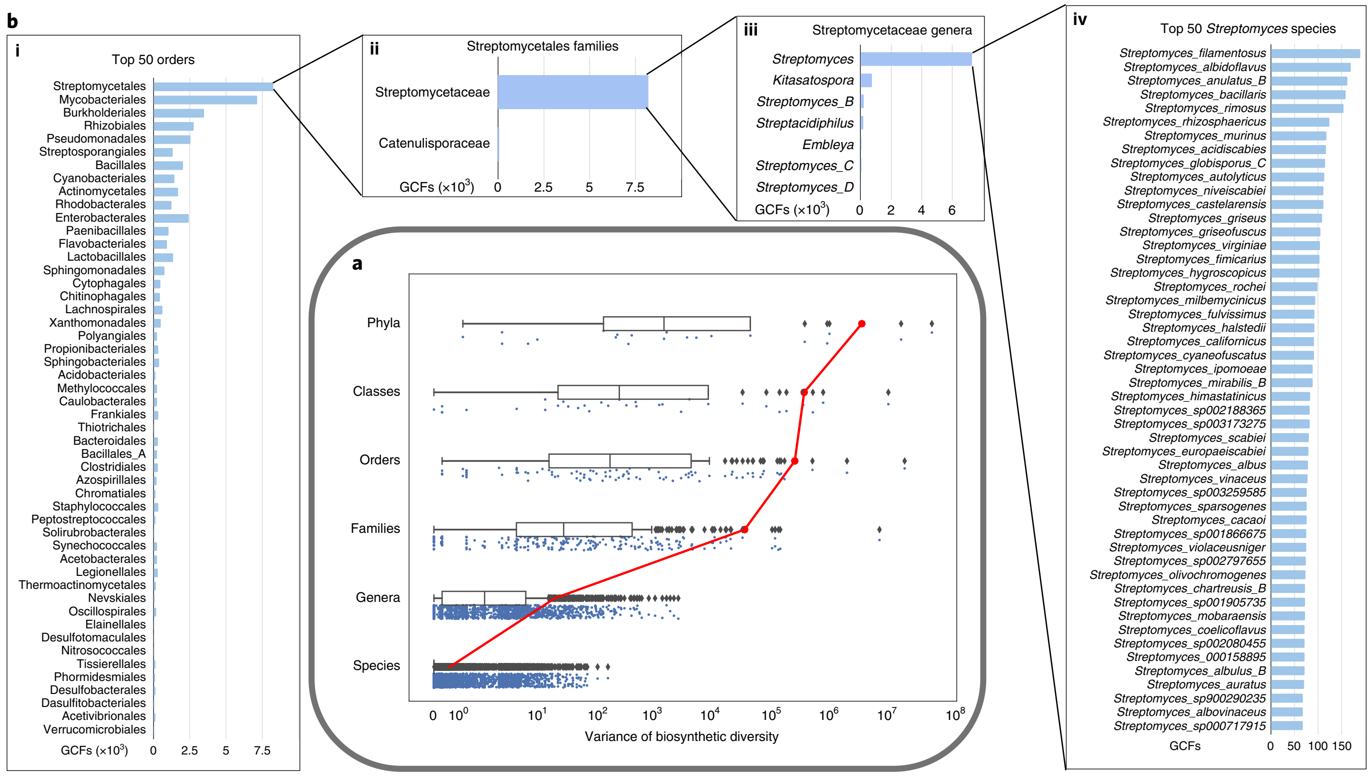


Fig. 3 | Relations of taxonomic levels to variability in biosynthetic diversity. a, Modified 'raincloud plots'[65] of major taxonomic ranks (*X* axis in logarithmic scale). Each boxplot represents the dispersion of variance values of a certain taxonomic rank, computed from the number of GCFs (defined by BiG-SLiCE at T = 0.4) of the immediately lower rank. The boxplots' center line represents the median value; the box limits represent the upper and lower quartiles. Whiskers represent 1.5× interquartile range. Points outside of the whiskers are outliers. Sample sizes: Phyla *n* = 21, Classes *n* = 33, Orders *n* = 89, Families *n* = 224, Genera *n* = 1,607, Species *n* = 13,065. Jittered raw data points are plotted under the boxplots for better visualization of the values' distribution. The red line connects the mean variance values of each rank. There is a noticeable drop in dispersion of variance values from the family rank to the genus rank (see also Supplementary Fig. 3), indicating that genera are suitable taxonomic groups to be characterized as diverse and be compared to each other. **b**, Biosynthetic diversity of various taxa, measured in absolute numbers of distinct GCFs (as defined by BiG-SLiCE, T = 0.4) from currently sequenced genomes. Top 50 most diverse orders (**i**), Streptomycetales families (**ii**), Streptomycetaceae genera (**iii**), top 50 most diverse *Streptomyces* species (**iv**). The difference in variance is visible in **i**–**iii**, but becomes homogeneous at the species level as shown in **iv**.

20 overall most promising REDgroups, we found at least 6 groups that show promise but whose members are either not catalogued in the database as NP sources or are connected to few (<15) known compounds: Amycolatopsis_RG1, Kutzneria, Xanthobacteriaceae_mixed_RG1, Mycolicibacterium_RG1, Nonomuraea and Kitasatospora_RG1. The Amycolatopsis_RG1 group only includes three rare species: *Amycolatopsis antarctica*, *marina* and *nigrescens*. Other promising REDgroups with very few known producers include Cupriavidus (from Proteobacteria phylum), Weeksellaceae_mixed_RG1 (from Bacteroidota phylum) and Pleurocapsa (from Cyanobacteria phylum). More information about the promising underexplored taxa can be found in Supplementary Table 8.

Discussion

Using two different algorithms, we mined deposited bacterial sequencing data to identify BGCs and grouped them into GCFs according to chemical families of encoded compounds. We identified maximal emergence of the highest biosynthetic diversity close to the genus rank and chose to further investigate analogous taxonomic groups (REDgroups). Rarefaction analysis identified the highest biosynthetic potential and the most promising bacterial taxa among many known diverse groups, as well as multiple promising understudied producers. To the best of our knowledge, this is the largest survey of secondary metabolite production to date, and our study provides a reproducible pipeline to underpin drug discovery efforts.

The biosynthetic capacity of the bacterial kingdom was previously assessed by Cimermancic et al.[7], but the dataset analysed was only 33,000 BGCs compared with the 1,185,995 BGCs we analysed. Additionally, they used ClusterFinder, which is known as a more exploratory identification tool[7,36]. Projects that exploit publicly available genomic data are reliant on the quality of genomes sequenced as well as the efficiency of available genome mining methods, which have some limitations[37]. For instance, the study of GCF uniqueness among taxa may be affected by antiSMASH's imperfect BGC boundary prediction[12]. Although BiG-SLiCE converts BGCs into features only on the basis of domains related to biosynthesis[21], genomic context unrelated to the biosynthetic pathway of a BGC could still have a role in the GCF assignment; this issue cannot be fully addressed with currently available tools. However, antiSMASH's ability to discern cluster limits and detect BGCs from cultured strains and MAGs is comparable to alternative tools, while its ability to predict different BGC types is unparalleled[38], as is apparent from its common use in NP research[7,9,25,30,32,39]. Of note, the fact that it is rule-based[12] implies the possibility of undetected types of clusters and increases the likelihood that our calculations have underestimated the true biosynthetic potential of bacterial organisms.

Furthermore, our pipeline appears to be the first to use the GTDB[14] taxonomy for studying global bacterial biosynthetic diversity. This enabled us to avoid misclassifications of NCBI taxonomic placement[40–43]. The use of rarefaction curves allowed us to infer

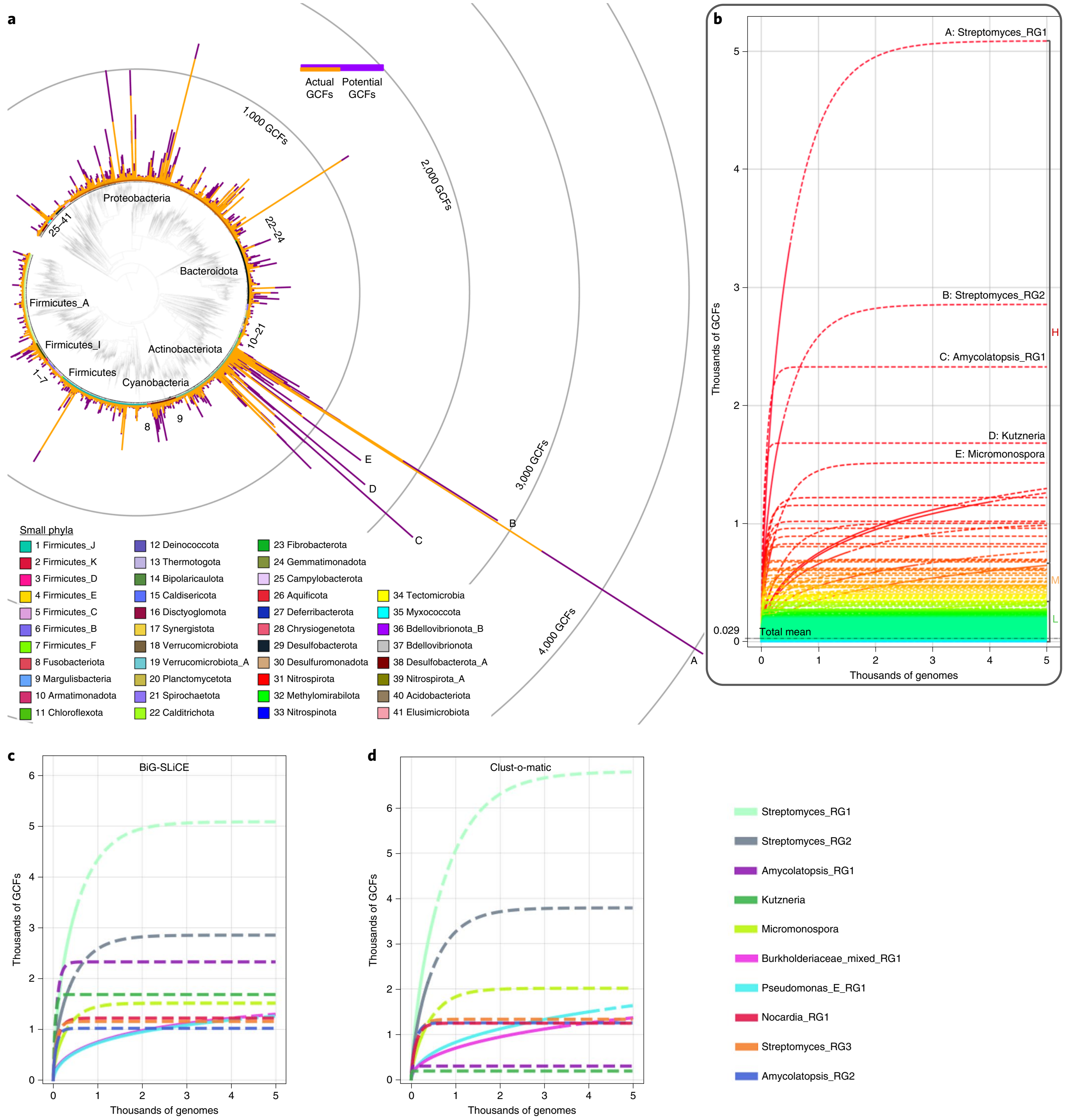


Fig. 4 | Overview of actual and potential biosynthetic diversity of the bacterial kingdom, compared at the REDgroup level. a, GTDB[14] bacterial tree up to the REDgroup level, visualized with iTOL[63] v6.5.2, colour coded by phylum, decorated with barplots of actual (orange) and potential (purple) GCFs as defined by BiG-SLiCE (T = 0.4). Top REDgroups with the most potential GCFs include the following: A, Streptomyces_RG1; B, Streptomyces_RG2; C, Amycolatopsis_RG1; D, Kutzneria; E, Micromonospora. Phyla known to be enriched in NP producers are immediately visible (Actinobacteriota, Protobacteriota), with the most promising groups coming from the Actinobacteriota phylum (the highest peak belongs to a REDgroup containing *Streptomyces* strains). Simultaneously, within the underexplored phyla, there seems to be notable biosynthetic diversity and potential. An interactive version of Fig. 4a can be accessed online (Extended Data Fig. 3). **b–d**, Rarefaction curves of REDgroups (BiG-SLiCE T = 0.4) (**b**), of the most promising REDgroups (BiG-SLiCE T = 0.4) (**c**) and of the most promising REDgroups (clust-o-matic T = 0.5) (**d**). In **b–d**, the solid lines represent interpolated and actual data, while the dashed lines represent extrapolated data. In **b**, the letters 'L', 'M' and 'H' correspond to low- (0–389 pGCFs), medium- (390–649 pGCFs) and high-diversity (>650 pGCFs) producers. The 'L' range includes 3,737 REDgroups (shades of green), the 'M' range includes 22 (shades of yellow/orange), while the 'H' range includes 20 REDgroups (shades of red). The vast majority of REDgroups belong to the low-diversity producers (the mean of all REDgroups' pGCFs is 29). The labels of the most promising REDgroups are indicated (the letters A–E correspond to the peaks in **a**). *Streptomyces* strains are included in several of them. Although the exact numbers differ, the similarities between the two methods are apparent.

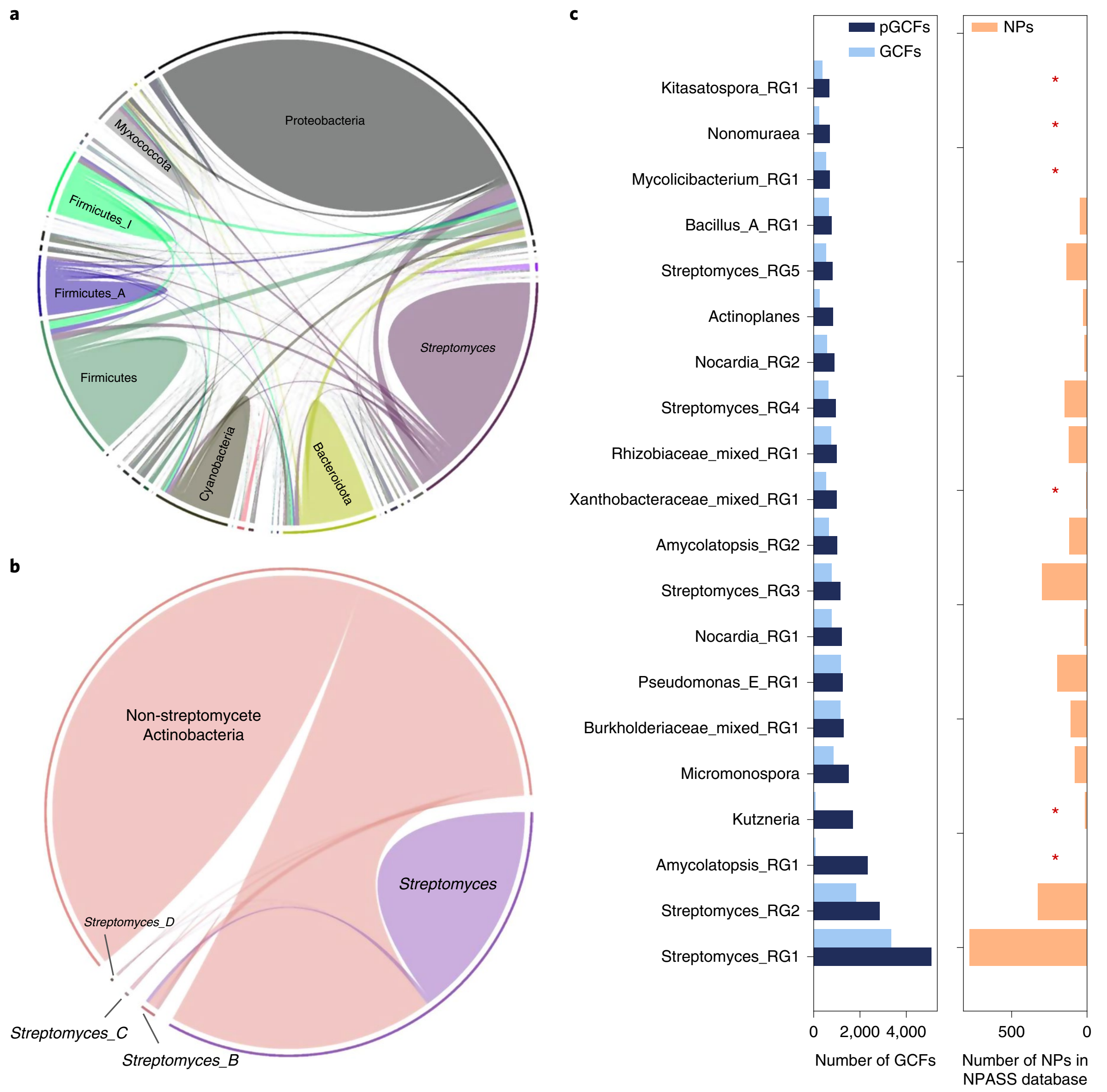


Fig. 5 | Unique diversity in the known producer *Streptomyces* and promising potential of less popular taxa. a, Unique GCFs (as defined by BiG-SLiCE, T = 0.4) of phyla and *Streptomyces* (solid shapes) and pairwise overlaps of phyla-phyla and phyla-*Streptomyces* (ribbons), visualized with circlize[64]. Each taxon has a distinct colour. The smaller shapes and ribbons represent smaller phyla that can be seen in Extended Data Fig. 4. The genus *Streptomyces* appears to have a very high amount of unique GCFs comparable to entire phyla, such as Proteobacteria. **b**, Unique GCFs (as defined by BiG-SLiCE, T = 0.4) of non-streptomycete Actinobacteriota and all *Streptomyces* genera (solid shapes) and pairwise overlaps between Actinobacteriota and *Streptomyces* (ribbons), visualized with circlize[64]. The *Streptomyces* genus, only one of many belonging to the Actinobacteriota phylum, appears to be responsible for a large percentage of the phylum's unique diversity (see big pink ribbon). **c**, Left: potential (pGCFs) and actual number of GCFs (as defined by BiG-SLiCE, T = 0.4) of the top 20 most promising REDgroups. Right: number of NPs found in the NPASS database[35] that originate from species included in each REDgroup. The REDgroups with few (<15) to no known NPs associated with them are marked with red stars on the right side of the graph. Several of the displayed groups are in the latter category: Amycolatopsis_RG1, Kutzneria, Xanthobacteraceae_mixed_RG1 (containing the genera *Bradyrhizobium, Rhodopseudomonas, Tardiphaga* and *Nitrobacter*), Mycolicibacterium_RG1, Nonomuraea and Kitasatospora_RG1.

the biosynthetic potential of bacterial groups, as done in some smaller-scale projects[7,8,31,32]. This method aims to enable fair comparisons among incomplete samples[44]. However, while overestimation is not expected to happen, for those groups that contain very few genomes, there is a tendency to underestimate their potential capacity[44]. Hence, sequencing bias of popular taxa still affects our results. We tried to minimize the bias within the pipeline as much as possible while retaining high diversity of bacterial taxa; therefore, we decided not to exclude REDgroups with very few members from the dataset. We also ran an additional random sampling analysis using the most populated REDgroups and confirmed the reproducibility of our results. Nonetheless, the remaining bias will only be eliminated with the inclusion of increased biodiversity in sequencing projects[17,20].

Our analysis identified a plethora of unexplored taxonomic groups with substantial biosynthetic potential[9,10,45–47]. At the same time, it revealed undiscovered biosynthetic diversity present in well-characterized NP producers. For example, multiple Proteobacteria taxa were identified among the top producers: *Pseudomonas*, *Pseudoalteromonas*, *Paracoccus* and *Serratia* among others. This is in accordance with the known biosynthetic potential of the Proteobacteria phylum[33]. Furthermore, we identified taxa that are less well represented in sequence databases as being potentially useful sources of secondary metabolites, including myxobacterial genera *Cystobacter*, *Melittangium*, *Archangium*, *Vitiosangium*, *Sorangium* and *Myxococcus*[9,30,48], and *Chryseobacterium* and *Chryseobacterium_A*[49] from the Bacteroidota phylum. However, the most diverse groups of metabolites are predicted to be produced by actinobacterial strains of well-known and well-studied NP producers such as *Actinoplanes*, *Amycolatopsis*, *Micromonospora*, *Mycobacterium*, *Nocardia* and *Streptomyces*[8,26,27,34]. These bacteria produce most of known natural product antibiotics[26] and our analysis confirms recent analyses of biosynthetic novelty in the genomes of rare Actinobacteria, suggesting that there is still much more natural product diversity to be discovered in this group as more diversified strains get sequenced[8,26,27,50].

Streptomyces is a genus of the Actinobacteria phylum that contains some of the most complex bacteria that we know of, albeit by far not the most sequenced in our dataset (Supplementary Fig. 5). These bacteria have been known as NP producers for a long time[34], as single strains containing a high number of BGCs have been discovered, taking up to 10% of their genome[51]. However, members of other genera contain comparable absolute numbers of BGCs. This appears to be the first time that a systematic comparison of the diversity of the encoded compounds within bacterial genera has been conducted, revealing how diverse *Streptomyces* are compared with all others[34]. The factors that cause this taxonomic group to stand out are not completely clear but are probably related to their sophisticated lifestyle. Many observations suggest that NP biosynthesis drives speciation within the *Streptomyces* genus[8]. The exploration of factors that led to the rise of biosynthetic diversity in *Streptomyces* to such an impressive degree will be the subject of further investigations in the future.

Having the genomic capacity for the biosynthesis of secondary metabolites does not always herald the discovery of a novel chemistry[52,53]. Sometimes, the bacterium in question cannot be grown or BGCs are not expressed in laboratory conditions[26,45,47,52,53]. This issue is related to the complexity of BGCs; we have only just scratched the surface of their intricate regulation and connection to primary metabolism[5,45,52,54]. However, efforts to decode biosynthetic mechanisms for the activation of silent clusters need to be tailored to specific producer groups[26,27,53], such as groups phylogenetically related to promising producers, for example members of the Pseudonocardiaceae family (REDgroups Amycolatopsis_RG1 & Kutzneria in Fig. 4; these and more REDgroups are shown in Supplementary Table 8), partly because each phylum has unique diversity (Fig. 2b).

Original approaches to the prioritization issue of NP research continue to emerge, fuelled by the advances in metagenomics and computational tools that enable the use of the biosynthetic potential of unculturable bacteria from environmental samples[55]. Furthermore, apart from the few metagenomic projects whose MAGs we incorporated in the first part of our analysis, there are multiple such projects publicly available, some of which have been the focus of NP studies[56]. Although the reconstruction of genomes from metagenomes remains a challenge[57] and the assembly will often miss BGCs[58], which has indirectly prevented their comparison to the cultured bacteria in the current project, metagenomics is proving to be a promising source of information on NPs and their producers[7,34,45,55,56], as made apparent in the present investigation. We expect the effect of this field on NP research to become more evident in subsequent years.

The collection of microbial data from a large variety of habitats points to another interesting aspect, namely the relation between the biome of origin of the producers and the uniqueness of their biosynthetic diversity. Although this connection has been investigated to some extent[24,25,32,33,47], drawing more definitive conclusions will require the use of a wider-scale dataset and the availability of more detailed and standardized metadata of producers' genomes.

Our analysis provides a global overview of diverse known and promising understudied NP-producing taxa. We expect this to greatly help overcome one of the main bottlenecks of natural products discovery: the prioritization of producers for research[55].

Methods

BGC dataset. We obtained 170,585 complete and draft bacterial genomes (Table 1) from RefSeq[13] on 27 March 2020. Furthermore, a dataset of 47,098 MAGs was included in the first part of the analysis (see Results, Biosynthetic diversity of the bacterial kingdom). For the rest of the study, we used only 161,290 RefSeq bacterial genomes whose taxonomic classification up to the species level was known (Table 1). All genomes were analysed with antiSMASH (version 5)[12], which identified their BGCs (Supplementary Table 1). The entirety of the MIBiG[23] database (accessed on 27 March 2020) was included in parts of our analysis (their IDs can be found in Supplementary Table 6).

Taxonomic classification. Due to multiple indications regarding a lack of accuracy of NCBI's taxonomic classification of bacterial genomes[40–43], we chose to use the GTDB[14] instead. The bacterial tree of 120 concatenated proteins (GTDB release 89), as well as the classifications of organisms up to the species level, were included in the analysis.

Quantification of biosynthetic diversity with BiG-SLiCE. For a bacterium to be regarded as biosynthetically diverse, we considered not the number of BGCs as important, but rather how different these BGCs are to each other. To quantify this diversity, we analysed all BGCs with the new BiG-SLiCE tool[21], which groups similar clusters into GCFs. However, the first version of this tool has an inherent bias towards multi-protein family BGCs, producing uneven coverage between BGCs of different classes (that is, due to their lack of biosynthetic domain diversity, all lanthipeptide BGCs may be grouped together using the Euclidean threshold of $T=900$, which in contrast is ideal for clustering Type-I Polyketide BGCs). To alleviate this issue and provide a fair measurement of biosynthetic diversity between the taxa, we modified the original distance measurement by normalizing the BGC features under L^2-norm, which produces a cosine-like distance when processed by the Euclidean-based BIRCH algorithm. This use of cosine-like distance virtually balances the measured distance between BGCs with 'high' and 'low' feature counts (Supplementary Fig. 6a), in the end providing an improved clustering performance when measured using the reference data of manually curated MIBiG GCFs (Supplementary Fig. 6b).

The GTDB[14] (release 89) bacterial tree was pruned so that it included only the organisms that are part of our dataset. Then, having both the taxonomic classification of all bacteria, as well as how many GCFs their BGCs group into, the pruned GTDB tree was decorated with the number of GCF values at each node. This allowed for the evaluation of the biosynthetic diversity of any clade, including the main taxonomic ranks. To pick a single threshold for subsequent taxonomic richness analysis, we compared BiG-SLiCE results on 947 MIBiG BGCs versus the compound-based clustering provided by the NPAtlas database[11] (Supplementary Fig. 2). A final threshold of $T=0.4$ was chosen on the basis of its similarity to the NPAtlas's compound clusters (v-score = 0.9X, GCF counts difference = +XX).

Quantification of biosynthetic diversity with clust-o-matic. We aimed to repeat and evaluate the reproducibility of the BGC-to-GCF quantification step of BiG-SLiCE with an alternative, independently derived algorithm. For this, instead of grouping BGCs into GCFs on the basis of biosynthetic domain diversity, we developed an algorithm that considers full core biosynthetic genes. Biosynthetic gene clusters that were detected in the input data by antiSMASH 5.1 were parsed to deliver core biosynthetic protein sequences. Those protein sequences were subjected to all-against-all multi-gene sequence similarity search with DIAMOND[59] 2.0 using default settings. Only one best hit per query core gene per BGC was allowed, divided by a total core protein length, resulting in the final pairwise BGC score always being within the range of 0 to 1. Pairwise BGC similarity scores were used to build a distance matrix that was later subjected to agglomerative hierarchical clustering in Python programming language (package scipy.cluster.hierarchy). The same process as described in the paragraph above (for BiG-SLiCE in that case) was performed for identification of the most suitable threshold for the clust-o-matic algorithm. The determined optimal threshold of 0.5 was then used to generate GCFs, which were then fed into the next steps in parallel to the original set of GCFs obtained from BiG-SLiCE.

Biogeographic analysis. One[20] of the MAGs datasets was accompanied by sufficient metadata that allowed for a study of a potential connection between biosynthetic diversity patterns and the biomes of origin of the corresponding MAGs. The GCFs for each ecosystem type were collected by combining information from Supplementary Tables 1 and 2 of this project and from the Nayfach paper[20] Supplementary Information. This led to the creation of

Supplementary Table 7. Then, the largest occurring intersections were computed and visualized in Extended Data Fig. 2 using the UpSet[60] visualization technique.

Variance analysis. To pinpoint the emergence of biosynthetic diversity, the within-taxon homogeneity was compared among the main taxonomic ranks. For each rank, the variance value was computed (with NumPy[61]) on the basis of the number of GCF values of immediately lower-ranked taxa, provided that there were at least two such taxa. For example, a phylum that includes only one class in our dataset was omitted from this computation. However, a phylum with two or more classes would be assigned a variance value computed from its classes' number of GCF values. The distribution of these variance values was plotted for each rank in Fig. 3a. We noticed a significant reduction in variance from the family to the genus rank, which was confirmed with an additional statistical test (Supplementary Fig. 3 and Supplementary Methods). A similar variance analysis was performed to compare genera and REDgroups (Supplementary Fig. 4), but in this case variance was calculated on the basis of the strains' biosynthetic diversity.

Definition of REDgroups. To study the biosynthetic diversity of genera, we attempted to achieve uniform taxa. For taxonomic rank normalization, the creators of GTDB used RED[14], which is a metric that relies heavily on the branch length of a phylogenetic tree and is consequently dependent on the rooting. The GTDB developers provided us with a bacterial tree decorated with the average RED values of all plausible rootings at each node. Since GTDB accepts a range of RED values for each taxonomic rank placement[14], we chose the median of GTDB genus RED values, namely 0.934, as a cut-off threshold. Any clade in the GTDB bacterial tree with an assigned RED value higher than the threshold was considered one group (Supplementary Fig. 7) that we named 'REDgroup'. For REDgroup naming conventions, see Supplementary Fig. 7.

Rarefaction analysis. The extrapolation of potential number of GCF values was achieved by conducting rarefaction analyses using the iNEXT R package[62]. A GCF presence/absence table (GCF-by-strain matrix) was constructed for each group considered and was then used as 'incidence-raw' data in the iNEXT main function, where 500 points were inter- or extrapolated with an endpoint of 5,000 for the REDgroups, and an endpoint of 1.6 million (about 8 times the number of strains in the 'complete dataset') in each group for the RefSeq analyses (where 2,000 points were inter- or extrapolated). By default, the number of bootstrap replications is 50.

Random sampling. To test whether the above methods (creation of REDgroups and the subsequent rarefaction analyses) surmounted the inherent sequencing bias in our dataset, a random sampling technique was used. A reduced dataset that included only those REDgroups containing at least 20 members was tested. For each REDgroup, a sample of 20 genomes was randomly chosen (using the Python 'random' module), while preserving the species diversity of the group. The latter was achieved by ensuring that genomes belonging to as many species as possible are included in each sample; if all species of a REDgroup were included but the genomes were fewer than 20, the remaining 'spots' were distributed evenly among a random sample of the REDgroup's species. One hundred iterations of this process were calculated for all REDgroups in this reduced dataset and rarefaction analyses were conducted for the random samples in each iteration. Finally, the average pGCF value for each REDgroup from all iterations was calculated and values are reported in Supplementary Table 9.

Identification of unknown producers. We investigated the genera included in the most promising REDgroups to find out whether they include species that are producers of known compounds. Hence, the species names were cross-referenced with the species named as producers in the NPASS depository[35] (accessed on 15 October 2020), taking care to match the GTDB-given names to the NCBI-given names that the database uses.

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

Data availability

The datasets generated and analysed during the study are available in the Zenodo repository: https://doi.org/10.5281/zenodo.6365726. Source data are provided with this paper.

Code availability

The clust-o-matic code is available at https://github.com/Helmholtz-HIPS. The modified BiG-SLiCE script (that accepts as input a regular BiG-SLiCE output folder, then outputs the GCF membership in a tsv file) is available both in our Zenodo repository (file name: perform_l2norm_clustering.py) and at the following link: https://github.com/medema-group/bigslice/blob/master/misc/useful_scripts/perform_l2norm_clustering.py.

Received: 11 August 2021; Accepted: 24 March 2022;
Published online: 2 May 2022

References

1. O'Connor, S. E. Engineering of secondary metabolism. *Annu. Rev. Genet.* **49**, 71–94 (2015).
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. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).
4. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., International Natural Product Sciences Taskforce & Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
5. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).
6. 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).
7. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
8. Doroghazi, J. R. et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
9. Hoffmann, T. et al. Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nat. Commun.* **9**, 803 (2018).
10. Lewis, K. The science of antibiotic discovery. *Cell* **181**, 29–45 (2020).
11. van Santen, J. A. et al. The natural products atlas: an open access knowledge base for microbial natural products discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).
12. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
13. Haft, D. H. et al. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).
14. 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).
15. Stewart, R. D. et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
16. Glendinning, L., Stewart, R. D., Pallen, M. J., Watson, K. A. & Watson, M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. *Genome Biol.* **21**, 34 (2020).
17. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39,** 105–114 (2021). https://doi.org/10.1101/762682
18. 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).
19. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
20. Nayfach, S. et al. Author correction: a genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 521 (2021).
21. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medemac, M. H. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *GigaScience* **10**, giaa154 (2021).
22. Navarro-Muñoz, J. C. et al. A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
23. Kautsar, S. A. et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
24. Coelho, L. P. et al. Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2021).
25. Sharrar, A. M. et al. Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type. *mBio* **11**, e00416–e00420 (2020).
26. Barka, E. A. et al. Taxonomy, physiology, and natural products of actinobacteria. *Microbiol. Mol. Biol. Rev.* **80**, 1–43 (2016).
27. Genilloud, O. Actinomycetes: still a source of novel antibiotics. *Nat. Prod. Rep.* **34**, 1203–1232 (2017).
28. Chevrette, M. G. et al. The confluence of big data and evolutionary genome mining for the discovery of natural products. *Nat. Prod. Rep.* **38**, 2024–2040 (2021).
29. Chase, A. B., Sweeney, D., Muskat, M. N., Guillén-Matus, D. & Jensen, P. R. Vertical inheritance governs biosynthetic gene cluster evolution and chemical diversification. *mBio* **12**, e02700-21 (2021).
30. Männle, D. et al. Comparative genomics and metabolomics in the genus *Nocardia*. *mSystems* **5**, e00125-20 (2020).
31. Ziemert, N. et al. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc. Natl Acad. Sci. USA* **111**, E1130–E1139 (2014).
32. Adamek, M. et al. Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in *Amycolatopsis* species. *BMC Genomics* **19**, 426 (2018).

33. Buijs, Y. et al. Marine Proteobacteria as a source of natural products: advances in molecular tools and strategies. *Nat. Prod. Rep.* **36**, 1333–1350 (2019).
34. Bérdy, J. Bioactive microbial metabolites: a personal view. *J. Antibiotics* 58, 1–26 (2005).
35. Zeng, X. et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **46**, D1217–D1222 (2018).
36. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
37. Miller, M. E. et al. Increased virulence of *Puccinia coronata* f. sp. *avenae* populations through allele frequency changes at multiple putative *Avr* loci. *PLoS Genet.* **16**, e1009291 (2020).
38. Chavali, A. K. & Rhee, S. Y. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Brief. Bioinform.* **19**, 1022–1034 (2018).
39. Adamek, M., Alanjary, M. & Ziemert, N. Applied evolution: phylogeny-based approaches in natural products research. *Nat. Prod. Rep.* **36**, 1295–1312 (2019).
40. Ciufo, S. et al. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* **68**, 2386–2392 (2018).
41. Martínez-Romero, E. et al. Genome misclassification of *Klebsiella variicola* and *Klebsiella quasipneumoniae* isolated from plants, animals and humans. *Salud Publica Mex.* **60**, 56–62 (2018).
42. Mateo-Estrada, V., Graña-Miraglia, L., López-Leal, G. & Castillo-Ramírez, S. Phylogenomics reveals clear cases of misclassification and genus-wide phylogenetic markers for *Acinetobacter*. *Genome Biol. Evol.* **11**, 2531–2541 (2019).
43. Rekadwad, B. N. & Gonzalez, J. M. Correcting names of bacteria deposited in National Microbial Repositories: an analysed sequence data necessary for taxonomic re-categorization of misclassified bacteria - ONE example, genus *Lysinibacillus*. *Data Brief* **13**, 761–778 (2017).
44. Chao, A. et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **84**, 45–67 (2014).
45. Hug, J. J., Bader, C. D., Remškar, M., Cirnski, K. & Müller, R. Concepts and methods to access novel antibiotics from Actinomycetes. *Antibiotics* **7**, 44 (2018).
46. Ling, L. L. et al. A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
47. Subramani, R. & Sipkema, D. Marine rare Actinomycetes: a promising source of structurally diverse and unique novel natural products. *Mar. Drugs* **17**, 249 (2019).
48. Weissman, K. J. & Müller, R. Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.* **27**, 1276–1295 (2010).
49. Dahal, R. H., Chaudhary, D. K., Kim, D.-U., Pandey, R. P. & Kim, J. *Chryseobacterium antibioticum* sp. nov. with antimicrobial activity against Gram-negative bacteria, isolated from Arctic soil. *J. Antibiotics* **74**, 115–123 (2021).
50. Schorn, M. A. et al. Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology* **162**, 2075–2086 (2016).
51. Baltz, R. H. Gifted microbes for genome mining and natural product discovery. *J. Ind. Microbiol. Biotechnol.* **44**, 573–588 (2017).
52. Seyedsayamdost, M. R. Toward a global picture of bacterial secondary metabolism. *J. Ind. Microbiol. Biotechnol.* **46**, 301–311 (2019).
53. Wohlleben, W., Mast, Y., Stegmann, E. & Ziemert, N. Antibiotic drug discovery. *Microb. Biotechnol.* **9**, 541–548 (2016).
54. van Bergeijk, D. A., Terlouw, B. R., Medema, M. H. & van Wezel, G. P. Ecology and genomics of Actinobacteria: new concepts for natural product discovery. *Nat. Rev. Microbiol.* **18**, 546–558 (2020).
55. Tracanna, V., de Jong, A., Medema, M. H. & Kuipers, O. P. Mining prokaryotes for antimicrobial compounds: from diversity to function. *FEMS Microbiol. Rev.* **41**, 417–429 (2017).
56. Chen, R. et al. Discovery of an abundance of biosynthetic gene clusters in shark bay microbial mats. *Front. Microbiol.* **11**, 1950 (2020).
57. Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic assembly: overview, challenges and applications. *Yale J. Biol. Med.* **89**, 353–362 (2016).
58. Mantri, S. S. et al. Metagenomic sequencing of multiple soil horizons and sites in close vicinity revealed novel secondary metabolite diversity. *mSystems* **6**, e0101821 (2021).
59. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
60. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
61. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
62. Hsieh, T. C., Ma, K. H. & Chao, A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol. Evol.* **7**, 1451–1456 (2016).
63. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
64. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
65. Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R. & Kievit, R. A. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res.* **4**, 63 (2019).

Acknowledgements

A.G. is grateful for the support of the Deutsche Forschungsgemeinschaft (DFG; Project ID No. 398967434-TRR 261). N. Ziemert was supported by the German Center for Infection Research (DZIF) (TTU 09.716). M.H.M. was supported by a European Research Council Starting Grant 948770-DECIPHER. S.K. was supported by the Graduate School for Experimental Plant Sciences (EPS) of Wageningen University. Work in the lab of R.M. was supported by BMBF (16GW0243), DFG and DZIF (807-5-8-0982600). A.G. and N. Ziemert thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2124 – 390838134 for the infrastructural support. A.G. thanks M. Direnc Mungan for valued discussions on optimizing the analysis, as well as C. Bagci for his imaginative suggestion on dealing with large data. We also thank L. do Presti for invaluable comments on the manuscript.

Author contributions

A.G., S.A.K., N. Zaburannyi and D.K. performed the analysis. S.A.K. and N. Zaburannyi contributed analysis tools. A.G., D.K., R.M., M.H.M. and N. Ziemert wrote the paper. All authors contributed to the conception and design of the analysis, read and agreed to the published version of the manuscript.

Competing interests

M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-022-01110-2.

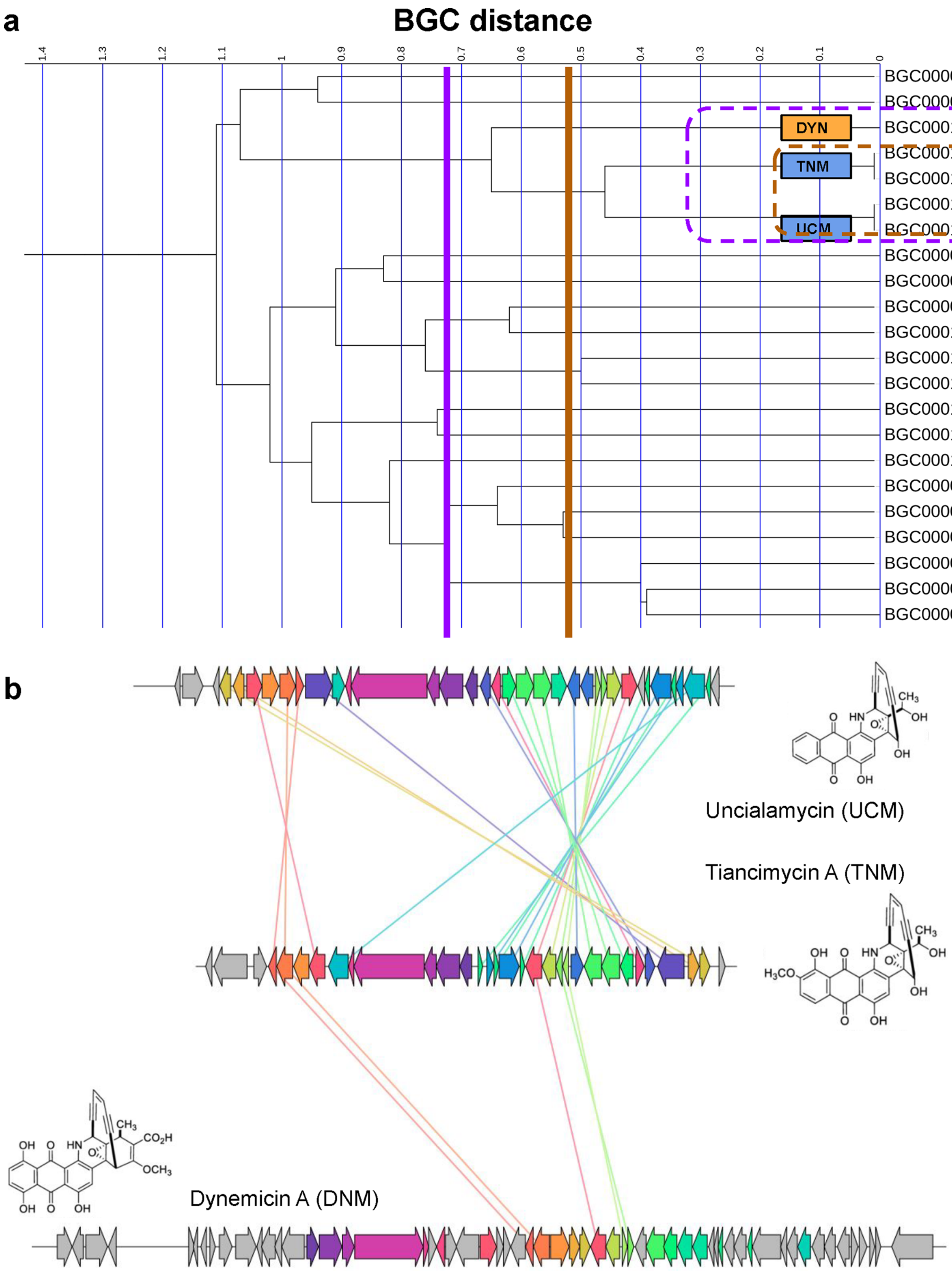
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-022-01110-2.

Correspondence and requests for materials should be addressed to Marnix H. Medema or Nadine Ziemert.

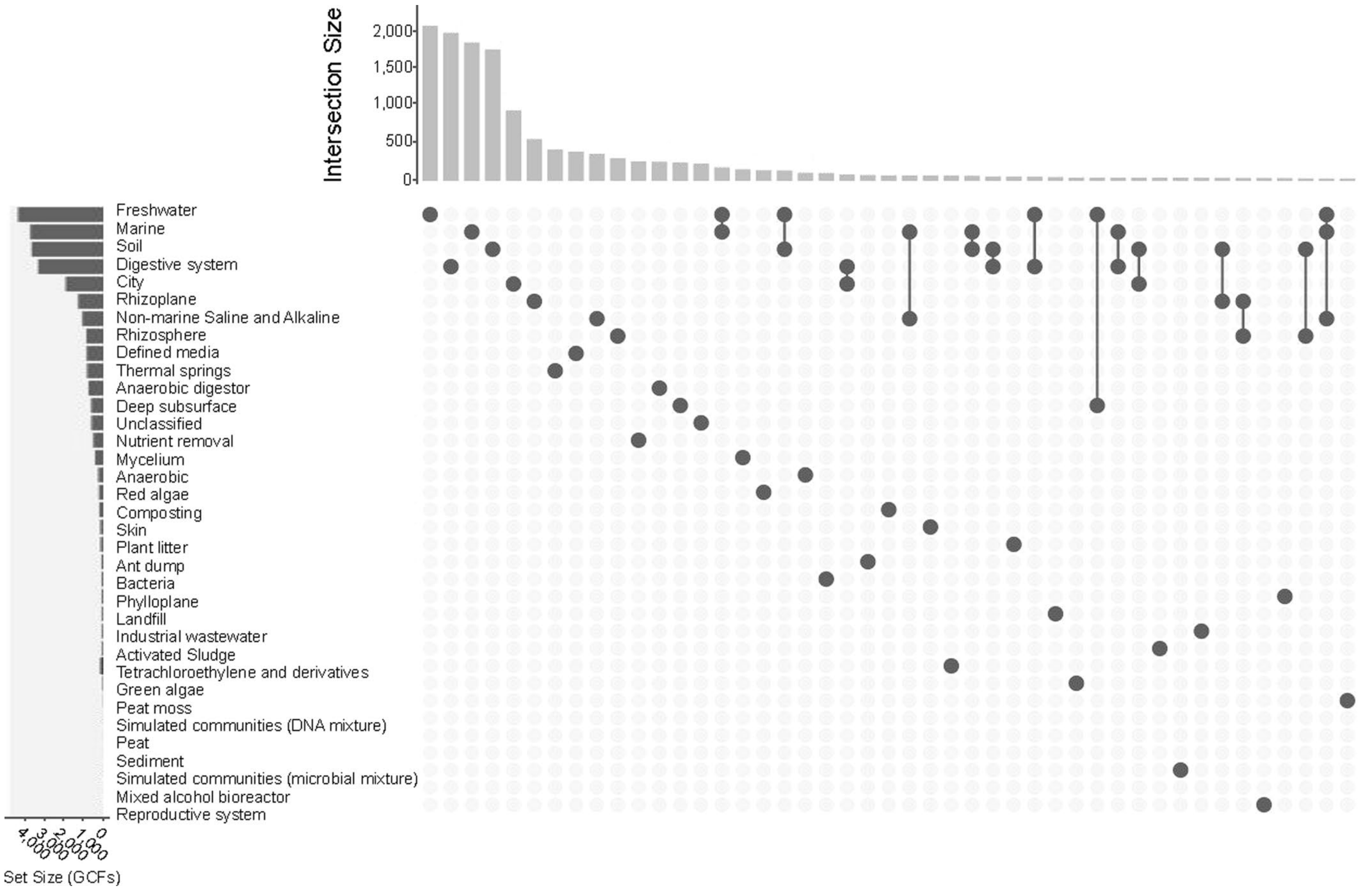
Peer review information *Nature Microbiology* thanks Nigel Mouncey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

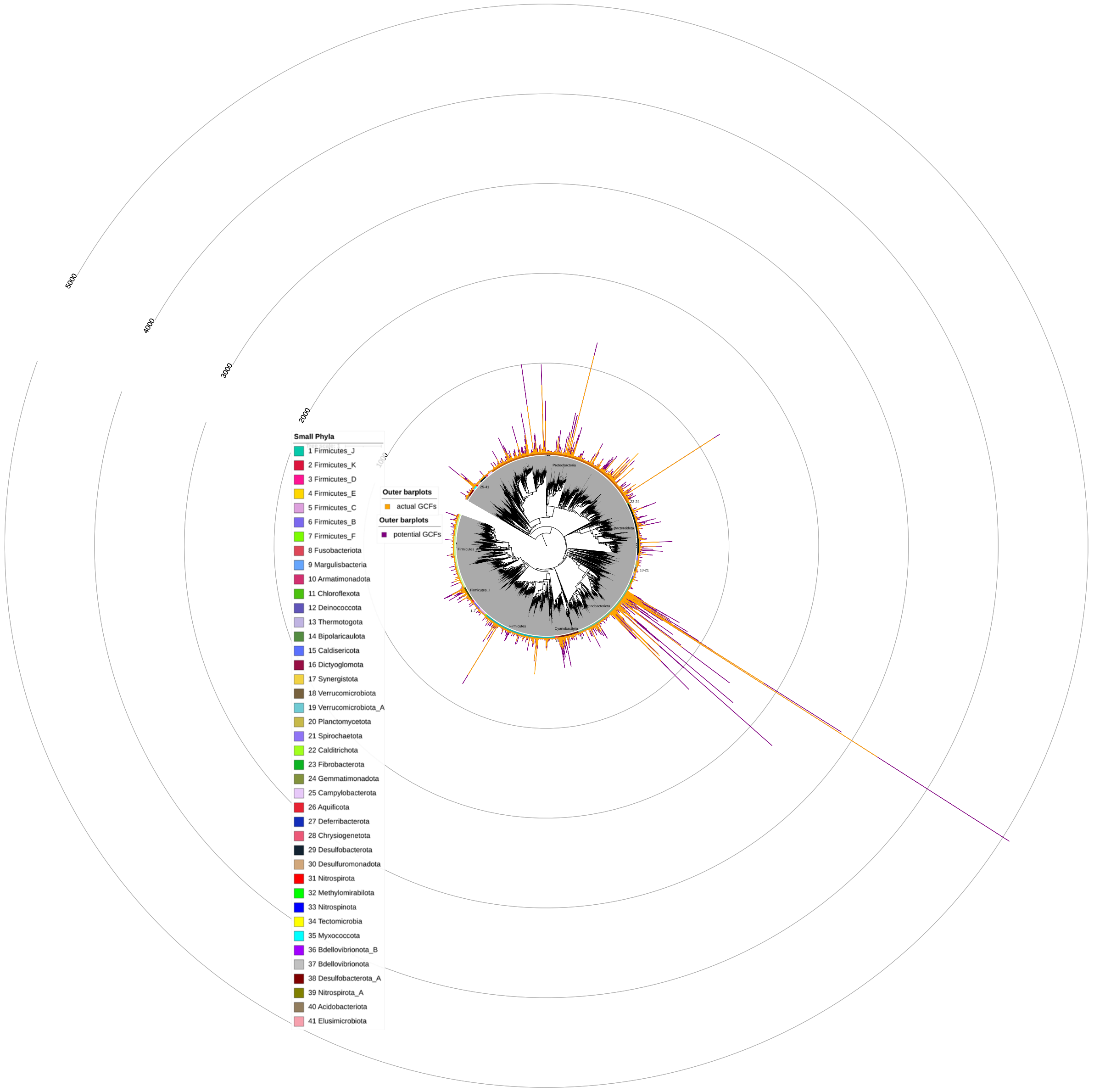
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



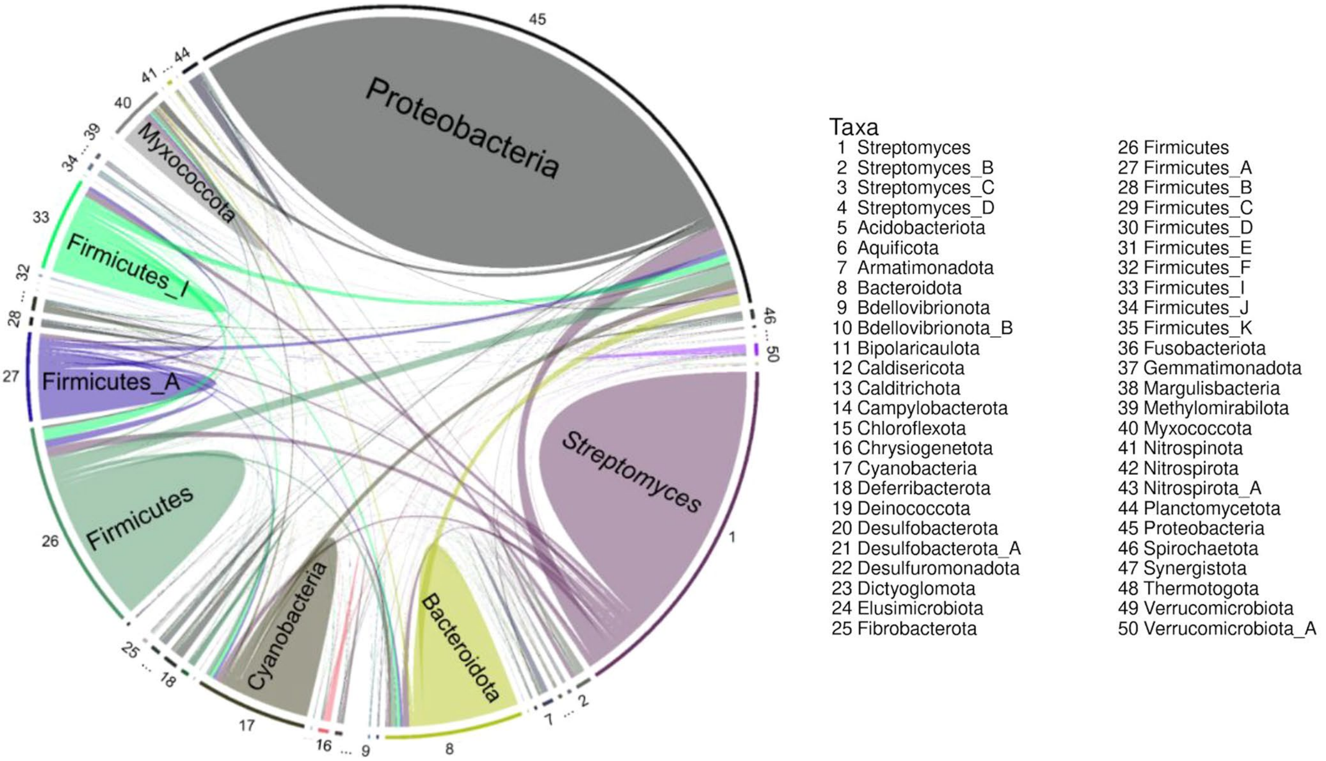
Extended Data Fig. 1 | Illustrating the correlation between BGC clustering thresholds and the grouping of their pathway products. a) a snippet of a complete-linkage hierarchical dendrogram constructed by doing a pairwise distance comparison of L2-normalized BGC features within the MIBiG dataset, highlighting the grouping of BGCs for the enediynes Uncialamycin (UCM) and Tiancimycin (TNM) under the threshold T = 0.5, and further grouping with another related enediyne BGC, Dynemicin (DNM) under the looser threshold of T = 0.7. **b**) Comparative genes analysis generated using the clinker tool[92] v0.0.23 shows how UCM and TNM BGCs are much more similar to each other than to DNM (same-colored genes indicate <70% amino acid similarity, while colored edges indicate <50% amino acid similarity), which is consistent with the structural diversity of their compounds (pictured).



Extended Data Fig. 2 | Intersections and distribution of biosynthetic diversity values among different ecosystem types. The bar plot on the left depicts the number of Gene Cluster Families (GCFs as defined by BiG-SLiCE with T = 0.4) found in each biome type. The bar plot on top shows the size (number of GCFs) of each intersection. Which sets (biome types) are included in each intersection can be seen in the matrix below the bar plot, where the dark dots pinpoint included sets. If more than one set is part of an intersection, connecting lines are drawn for better visibility. The data presented in this graph come only from the MAGs in the GEMS dataset (see Supplementary Table 1), which was the only one with sufficient metadata. Only the top 63 most sizable intersections are depicted here, and only the 35 ecosystem types (with the most GCFs out of the 63) that were part of them are shown on the left. The data indicate that there is barely any overlap between the ecosystem types; most GCFs (74.43 %) are specific to a single biome (a complete overview of unique GCFs per ecosystem type can be found in Supplementary Table 7), while the largest intersection (the one including most habitats - not visible in this Figure) includes 50 of the 63 ecosystem types.



Extended Data Fig. 3 | Overview of actual and potential biosynthetic diversity of bacterial kingdom, compared at REDgroup level. Extended Data Fig. 3 is interactive and can be accessed online on iTOL: https://itol.embl.de/shared/1B6W5n9MixSdJ. GTDB bacterial tree up to REDgroup level (for more details see Methods - REDgroup definition), colour-coded by phylum, decorated with barplots of actual (orange) and potential (purple) Gene Cluster Families (GCFs) as defined by BiG-SLiCE (T = 0.4). Potential GCFs were computed by rarefaction analyses (for more details see Results - Well known and less popular taxa as sources of biosynthetic diversity). REDgroups names are displayed around the tree as leaf node labels; hovering over them provides further taxonomic information (for full REDgroup metadata see Supplementary Table 1). Phyla known to be enriched in NP producers are immediately visible (Actinobacteriota, Protobacteriota), with the most promising groups coming from the Actinobacteriota phylum (the highest peak belongs to a REDgroup containing Streptomyces strains). Simultaneously, within the underexplored phyla, there seems to be notable biosynthetic diversity and potential. This Figure is meant to be explored by zooming in and out, searching for keywords and visualizing different kinds of information by switching between Tree Views. Any other attempt at modification (for example turning datasets on and off) may result in an unreadable graph.



Extended Data Fig. 4 | Unique diversity in the known producer *Streptomyces*. Unique GCFs, as defined by BiG-SLICE (T = 0.4), of bacterial phyla and Streptomyces (solid shapes) and pairwise overlaps of phyla - phyla and phyla - Streptomyces (ribbons). Each taxon has a distinct colour. The genus Streptomyces (1) appears to have a very high amount of unique GCFs comparable to entire phyla, such as Proteobacteria (43).

nature portfolio

Corresponding author(s): Nadine Ziemert, Marnix H. Medema

Last updated by author(s): Mar 20, 2022

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 estimates 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

All genomes used were downloaded from the publicly available RefSeq database and from specific publications that are specified in our Results and their Accession numbers are included in Supplementary Table 1. Some BGCs were obtained from the MiBIG database (IDs included in Supplementary Table 6). Information on compound producers was obtained from NPASS, as specified in the Methods.

Data analysis

Our analysis was conducted using the following software: modified BiG-SLiCE algorithm (https://github.com/medema-group/bigslice/blob/master/misc/useful_scripts/perform_l2norm_clustering.py), clust-o-matic (https://github.com/Helmholtz-HIPS), R package iNEXT (v2.0.20), python library NumPy (v1.19.1), webserver iTOL (v6.5.2), R package circlize (v0.4.13), UpSet visualisation (local version downloaded on 17.11.2021), clinker (v0.0.23).

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 datasets generated and analyzed during the current study are available in the following zenodo repository: https://doi.org/10.5281/zenodo.5159210.

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 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	Our dataset included all publicly available bacterial genomes from RefSeq, and several thousand published bacterial Metagenome Assembled Genomes.
Data exclusions	No data was exlcuded from the analysis under Results section "Biosynthetic diversity of the bacterial kingdom". All genomes with missing species taxonomic information were excluded from the rest of the analyses, as it was important to associate them with a specific species in order to compare taxa.
Replication	The analysis was completed with two independently developed algorithms and our results were confirmed with both. The rarefaction analyses of the REDgroups were replicated 100 times using random sampling, which confirmed our initial results.
Randomization	This is not relevant to our analysis because we did not conduct experiments in the lab. The genomes we used were separated into taxonomic groups for a part of the analysis as described in the Methods section.
Blinding	This is not relevant to our analysis because everything was done in silico.

Behavioural & social sciences study design

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

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

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

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National*

Research sample

Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data collection

Describe the data collection procedure, including who recorded the data and how.

Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? □ Yes □ No

Field work, collection and transport

Field conditions

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

Location

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

Access & import/export

Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

Disturbance

Describe any disturbance caused by the study and how it was minimized.

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

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

State the source of each cell line used.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

□ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

Identify the organization(s) that approved the study protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
□	□	Public health
□	□	National security
□	□	Crops and/or livestock
□	□	Ecosystems
□	□	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
□	□	Demonstrate how to render a vaccine ineffective
□	□	Confer resistance to therapeutically useful antibiotics or antiviral agents
□	□	Enhance the virulence of a pathogen or render a nonpathogen virulent
□	□	Increase transmissibility of a pathogen
□	□	Alter the host range of a pathogen
□	□	Enable evasion of diagnostic/detection modalities
□	□	Enable the weaponization of a biological agent or toxin
□	□	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

□ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*