A global survey of specialized metabolic diversity encoded in bacterial genomes

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

*These authors contributed equally: Athina Gavriilidou, Satria A. Kautsar.

**These authors jointly supervised this work: Marnix H. Medema, Nadine Ziemert.

[+]e-mail: marnix.medema@wur.nl; nadine.ziemert@uni-tuebingen.de

[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, Auf der Morgenstelle 28, 72076 Tübingen, Germany.

[2]Bioinformatics Group, Wageningen University, the Netherlands

[3]Chemistry Department, Scripps Research Florida, United States

[4]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS)—Helmholtz Centre for Infection Research (HZI), Campus E8 1, 66123 Saarbrücken, Germany.

[5]German Center for Infection Research (DZIF), Partner site Hannover-Braunschweig, Inhoffenstr. 7, 38124 Braunschweig, Germany.

[6]Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen, Germany.

[7]German Centre for Infection Research (DZIF), Partnersite Tübingen, Tübingen, Germany.

Abstract

Bacterial secondary metabolites have been studied for decades for their usefulness as drugs, such as antibiotics. However, the identification of new structures has been decelerating, while multi-resistant pathogens continue to emerge. It is unclear how much chemical diversity exists in Nature and whether discovery efforts should be focused on established antibiotic producers or rather on understudied taxa. Here, we surveyed around 170,000 bacterial genomes and several thousand of Metagenome Assembled Genomes (MAGs). Our results indicate that only 3% of the genomic potential for natural products has been experimentally discovered. Our analysis connects the emergence of most biosynthetic diversity in evolutionary history close to the taxonomic rank of genus, identifies *Streptomyces* as by far the most biosynthetically diverse, followed by producers from genera such as *Amycolatopsis*, *Kutzneria* and *Micromonospora*, and highlights multiple promising high-producing taxa that have thus far been less investigated, such as Weeksellaceae (Bacteroidota), Myxococcaceae (Myxococcota), *Pleurocapsa*, Nostocaceae (Cyanobacteria).

Background

Secondary metabolites (also called specialized 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 new chemical entities brought to the clinic, especially as anticancer compounds and antibiotics[2–6]. Regrettably, the emergence of antibiotic-resistant pathogens[5,7–10] concomitant to a stagnation of antimicrobial discovery pipelines[3,6,10] is leading to a global public health crisis[5,7–10].

Nonetheless, genomics-based approaches to NP discovery[11–13] have revealed a largely untapped and much more diverse source of biosynthetic potential within genomes[5,14,15]. 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[14,16] and metabolomic data indicate that some biosynthetic pathways are unique to specific taxa[17], 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[16], methodical comparisons across taxa are lacking. Because of this, the scientific community appears undecided on the best strategy for natural product discovery: should the established known NP producers be studied further or should the community be investigating underexplored taxa[14,18]?

Here, we harnessed recent advances in computational genomic analysis of BGCs to survey the enormous amount of genome 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 NP Atlas[19]. This facilitated an analysis of the variance of diversity across major taxonomic ranks, which showed the genus rank to be the most appropriate to compare 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.

Results

Biosynthetic diversity of the bacterial kingdom

To assess the global number of Gene Cluster Families found in sequenced bacterial strains, we ran AntiSMASH on ~170,000 genomes from the NCBI RefSeq database[20] (Supplementary Table 1), spanning 48 bacterial phyla containing 464 families (according to the Genome Taxonomy DataBase classification - GTDB[21]). We also included almost 50,000 bacterial Metagenome Assembled Genomes (MAGs) from 6 metagenomic projects of various origins[22–27] (Table 1 and Supplementary Table 1). To accurately group similar BGCs – which likely encode pathways towards the production of similar compounds – into Gene Cluster Families (GCFs) across such a large dataset, we used a slightly modified version of the BiG-SLiCE tool[28], which has been calibrated to output GCFs that match the grouping of known compounds in the NP Atlas database[19] (see Methods: Quantification of biosynthetic diversity with BiG-SLiCE). The resulting GCFs were then used to measure biosynthetic diversity across taxa.

Table 1. Overview of input datasets and overall biosynthetic diversity under various BiG-SLiCE thresholds. 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 thresholds; BGC to GCF assignment for each threshold can be found in Supplementary Tables 2-5. *MAG sources: bovine rumen[22], chicken caecum[23], human gut[24], ocean[25], uncultivated bacteria[26], various sources[27].

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*	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 number of GCFs in RefSeq ranged from 19,152 to 51,052 depending on the threshold used by BiG-SLiCE (Table 1). While, as expected, the pure numbers of the analysis changed based on the threshold, the overall tendencies observed remained the same (Figure 1a, Supplementary Figure 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[14,16], 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 (Supplementary Figure 2). Nevertheless, linear relationships are not always applicable, as shown previously[29], 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[19]. As many as 947 compounds in NPAtlas are mapped to a known BGC in MIBiG repository[30], 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 NPAtlas (Supplementary Figure 3), we chose a threshold of 0.4, as it provided the most congruent agreements between the two groupings, with v-score=0.94 (out of 1.00) and ΔGCF=-17.

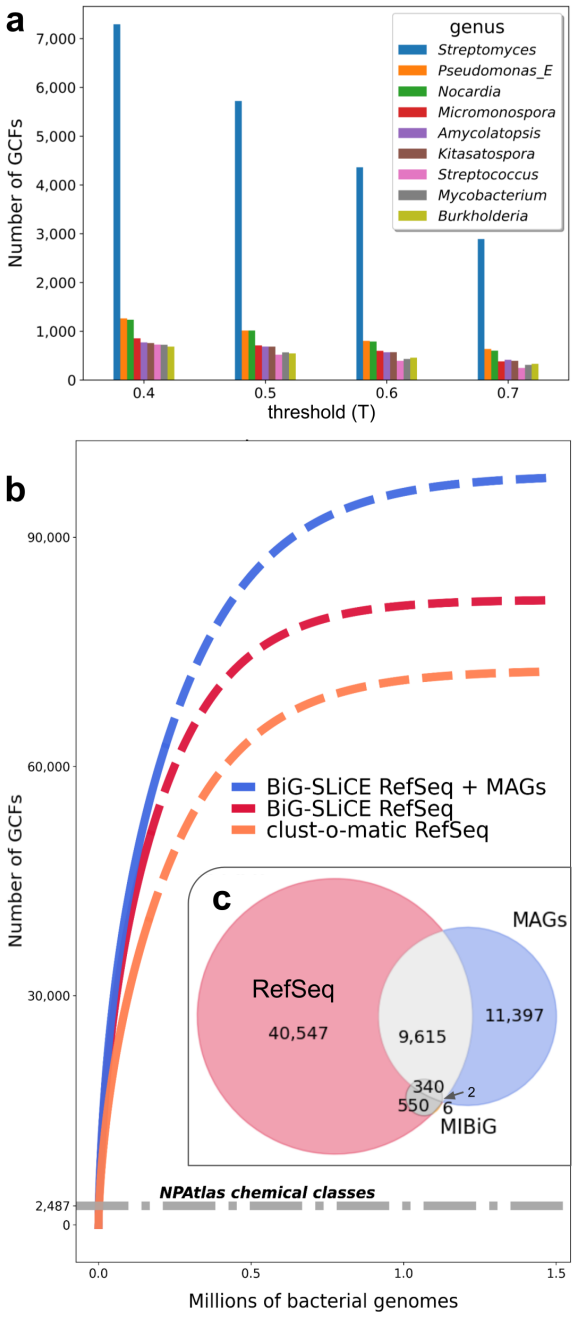

Figure 1. Biosynthetic diversity of the sequenced bacterial kingdom. Panel **a**: Bar plots of Gene Cluster Families (GCFs, as defined by BiG-SLiCE) of nine most biosynthetically diverse genera using different thresholds (T). The absolute number of GCFs changes from threshold to threshold, but the general tendencies (highest to lowest GCF count) are consistent between them. Panel **b**: Rarefaction curves of all RefSeq bacteria based on BiG-SLiCE (red) and based on 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 dotted lines represent extrapolated data. The number of chemical classes documented in NPAtlas[19], which come from bacterial producers (gray dotted 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). Panel **c**: Venn Diagram of GCFs (as defined by BiG-SLiCE, T=0.4) of the bacterial RefSeq, Minimum Information about a Biosynthetic Gene cluster (MIBiG[30]) 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.

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 (Figure 1b). The curve appears far from saturated, while the slope is steeper still if the bacterial MAGs are included in the analysis. When compared to the number of chemical classes documented in the NPAtlas[19] database (Figure 1b), it appears that, to date, only about 3% of the kingdom's biosynthetic diversity has been experimentally accessed.

In an attempt to evaluate the potential contribution of metagenomic data to Natural Product (NP) discovery, we studied how many of the GCFs found in the MAGs datasets were unique to this dataset (Figure 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[30]). Paradoxically, in Figure 1b, the contribution of MAGs does not reflect this finding, but this is most likely 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 only limited to one[27] 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 (Supplementary Figure 4, Supplementary Table 7).The latter finding is concordant with recent proof that most genes have a strong biogeography signal[31].

Genus as the most appropriate taxonomic rank to compare biosynthetic diversity

To identify the most promising bacterial producers, it is important to compare them at a specific taxonomic level. Several studies indicate that there is significant discontinuity in how BGCs are distributed across taxonomy: 'lower' taxonomic ranks like species within a genus carry more similar biosynthetic diversity, than 'higher' taxonomic ranks like 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[21]) was available only for the latter dataset(Table 1).

We first decorated the GTDB[21] bacterial tree with GCF values from the BiG-SLiCE analysis (Figure 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[14,32]. As expected for known NP producers, the phyla Proteobacteria and Actinobacteria appeared particularly diverse[16,33,34]. However, these phyla are

amongst the most studied and therefore the most sequenced[16,33,34], a bias that was 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 (Figure 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 HGT events, although relatively frequent for BGCs[35], are much more common among closely related taxa[36–39].

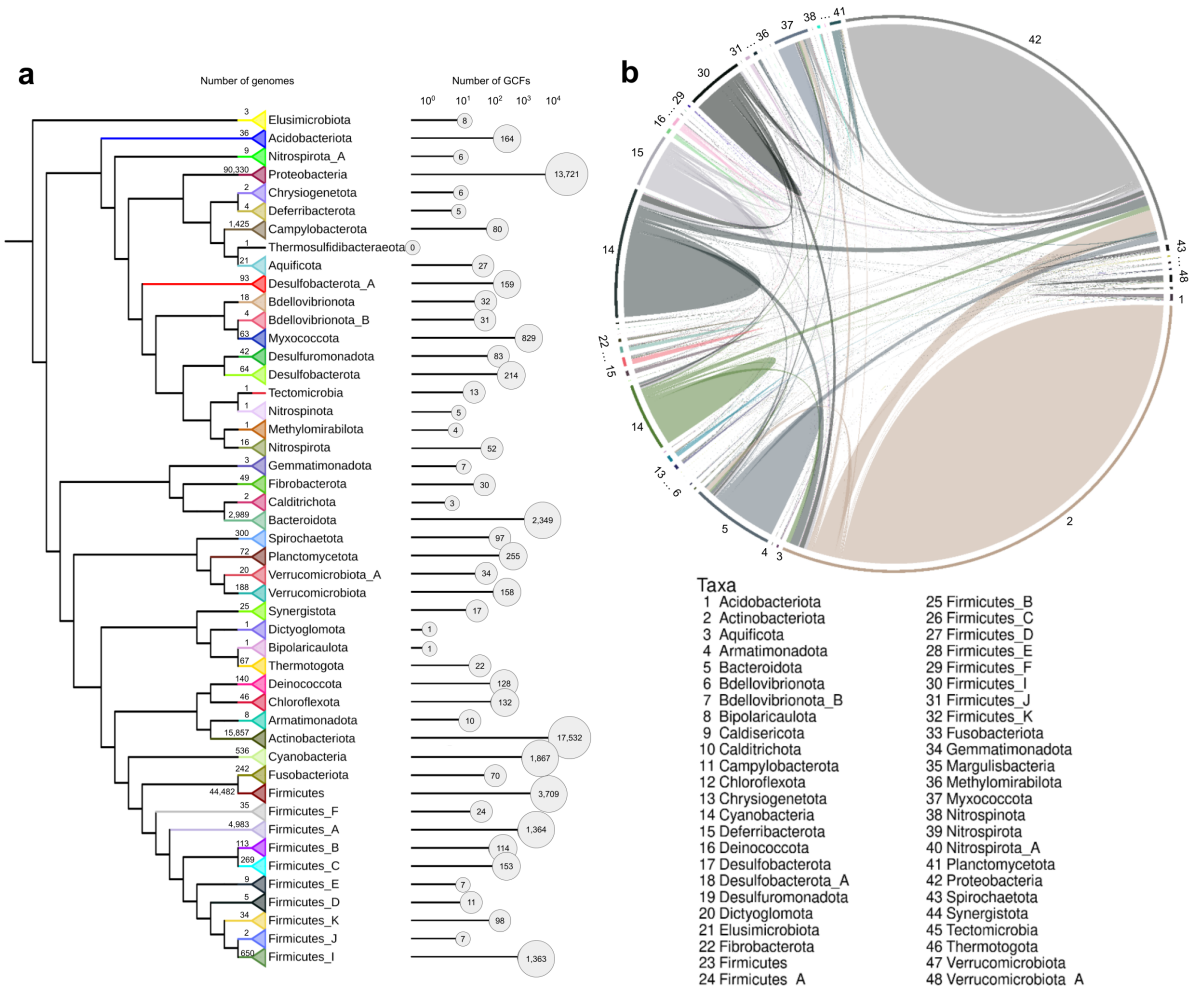

Figure 2. Comparison of biosynthetic diversity among phyla. Panel **a**: The Genome Taxonomy DataBase (GTDB[21]) bacterial tree was visualized with iTOL[40], decorated with Gene Cluster Families (GCFs) values (as defined by BiG-SLiCE at T=0.4), collapsed at the phylum rank and accompanied by bar plot 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. Panel **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[41]. Each phylum has a distinct color. Actinobacteriota (2) and Proteobacteria (40) seem particularly rich in unique GCFs.

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 from 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 based on the #GCFs values of immediately lower-ranked taxa (see Methods: Variance Analysis). The distribution of these variance values for each rank is visualized in Figure 3a.

There is a noticeable drop in the range of variance values for each rank, while diversity becomes highly homogeneous at the species level (Figures 3a,b). The plunge is most striking from the family to the genus level (Figure 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 (Figure 3b). Additional statistical analysis confirmed the significance of this observation (Supplementary Figure 5) thus pinpointing, for the first time, the genus rank as the most appropriate for comparative analyses.

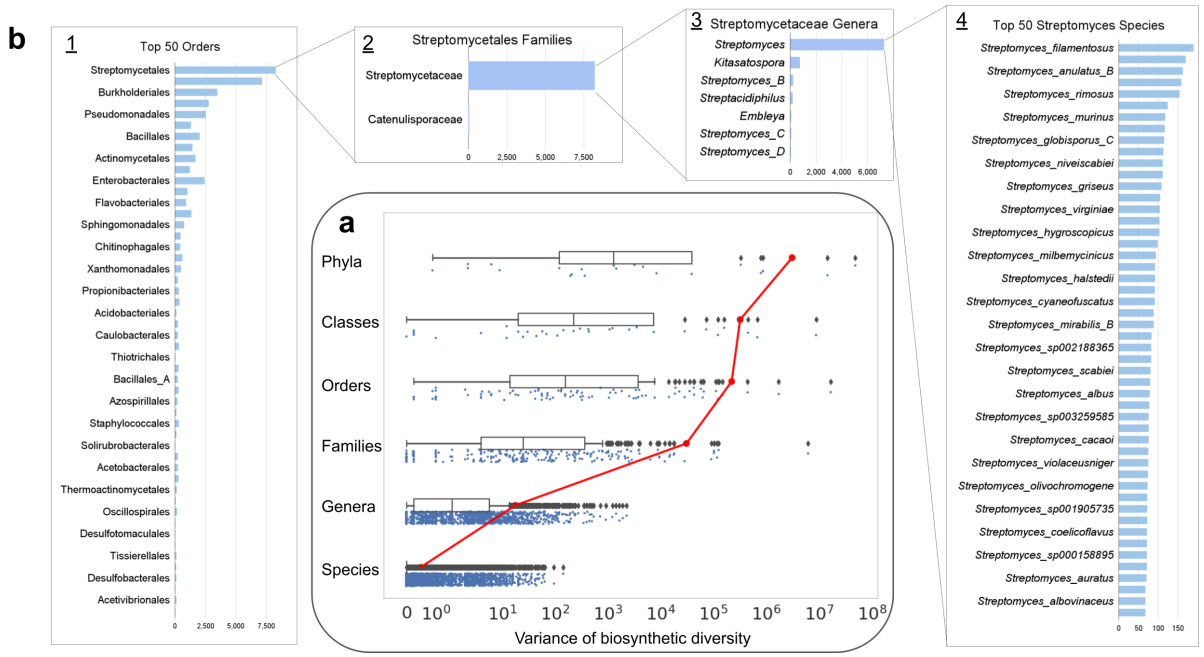

Figure 3. Relations of taxonomic levels to variability in biosynthetic diversity. Panel **a**: Modified "raincloud plots"[42] of major taxonomic ranks (X axis in logarithmic scale). Each boxplot represents the dispersion of variance values of a certain taxonomic rank, computed from the number of Gene Cluster Families (GCFs as 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 a 1.5x interquartile range. Points outside of the whiskers are outliers. 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 Figure 5), indicating that the genera are suitable taxonomic groups to be characterised as diverse and be compared to each other. Panel **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 (1), Streptomycetales families (2), Streptomycetaceae genera (3), top 50 most diverse *Streptomyces* species (4). The difference in variance is visible in the graphs 1,2,3, but becomes homogeneous at the species level as is shown in graph 4.

Well-known as well as overlooked taxa as sources of 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[21] 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 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 Figure 6). Some of the top groups (Supplementary Table 8) included known rich NP producers, such as *Streptomyces*, *Pseudomonas_E* and *Nocardia*[30,33,34,43].

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

overcome this bias, rarefaction analyses were conducted for each REDgroup (Figure 4b, Supplementary Table 8), as performed in previous studies[44,45]. 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 in order 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.*[21], as shown in Figure 4a. 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[16,33,34,46] and they remain in the top (Extended Data Table 1, Supplementary Table 1), seemingly having much unexplored biosynthetic potential.

To ensure that our conclusions are not the product of algorithmic artifacts, we reran 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 in order to form GCFs (see Methods: Random sampling). Like for BiG-SLiCE, we calibrated the threshold for clust-o-matic based on NP Atlas clusters. When comparing the results (Figure 4c,d, Supplementary Table 8), despite slight differences in absolute numbers, the two algorithms appeared to identify very similar trends.

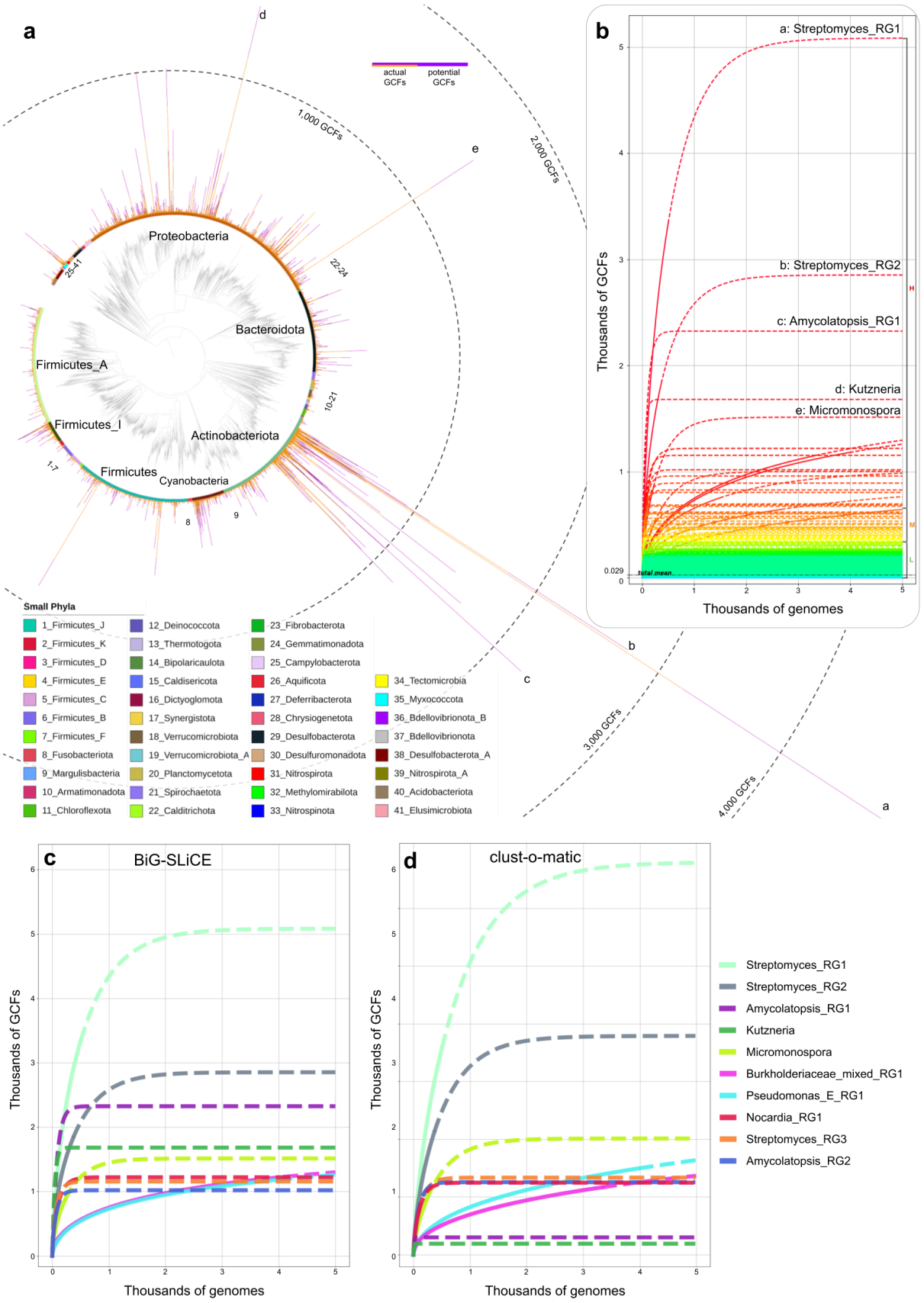
a
b
c
d
e
actual GCFs
potential GCFs
1,000 GCFs
2,000 GCFs
3,000 GCFs
4,000 GCFs
Proteobacteria
Bacteroidota
Actinobacteriota
Cyanobacteria
Firmicutes
Firmicutes_A
Firmicutes_I
25-41
22-24
10-21
1-7
8
9
a: Streptomyces_RG1
b: Streptomyces_RG2
c: Amycolatopsis_RG1
d: Kutzneria
e: Micromonospora
Thousands of GCFs
Thousands of genomes
total mean
0.029
H
M
L
Small Phyla
1_Firmicutes_J
2_Firmicutes_K
3_Firmicutes_D
4_Firmicutes_E
5_Firmicutes_C
6_Firmicutes_B
7_Firmicutes_F
8_Fusobacteriota
9_Margulisbacteria
10_Armatimonadota
11_Chloroflexota
12_Deinococcota
13_Thermotogota
14_Bipolaricaulota
15_Caldisericota
16_Dictyoglomota
17_Synergistota
18_Verrucomicrobiota
19_Verrucomicrobiota_A
20_Planctomycetota
21_Spirochaetota
22_Calditrichota
23_Fibrobacterota
24_Gemmatimonadota
25_Campylobacterota
26_Aquificota
27_Deferribacterota
28_Chrysiogenetota
29_Desulfobacterota
30_Desulfuromonadota
31_Nitrospirota
32_Methylomirabilota
33_Nitrospinota
34_Tectomicrobia
35_Myxococcota
36_Bdellovibrionota_B
37_Bdellovibrionota
38_Desulfobacterota_A
39_Nitrospirota_A
40_Acidobacteriota
41_Elusimicrobiota
BiG-SLiCE
clust-o-matic
Streptomyces_RG1
Streptomyces_RG2
Amycolatopsis_RG1
Kutzneria
Micromonospora
Burkholderiaceae_mixed_RG1
Pseudomonas_E_RG1
Nocardia_RG1
Streptomyces_RG3
Amycolatopsis_RG2

Figure 4. Overview of actual and potential biosynthetic diversity of bacterial kingdom, compared at REDgroup level. Panel **a:** GTDB[21] bacterial tree up to REDgroup level, visualized with iTOL[40], 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). Top REDgroups with most potential GCFs include the following: A: Streptomyces_RG1, B: Streptomyces_RG2, C: Amycolatopsis_RG1, D: Kutzneria, E: Pseudomonas_E. 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 significant biosynthetic diversity and potential. An interactive version of Figure 4a can be accessed online (Supplementary Figure 7). Panel **b:** Rarefaction curves of REDgroups (BiG-SLiCE T=0.4). The solid lines represent interpolated and actual data, while the dotted lines represent extrapolated data. The letters "L", "M" and "H" on the right edge of the figure correspond to Low- (0-389 pGCFs), Medium- (390-649 pGCFs) and High-diversity (more than 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, drawn in the "L" range of the graph). The genera included in the most promising REDgroups are indicated (the letters a-e correspond to the peaks in panel a. *Streptomyces* strains are included in several of the top most promising REDgroups. Panel **c:** Rarefaction curves of the most promising REDgroups (BiG-SLiCE T=0.4). The solid lines represent interpolated and actual data, while the dotted lines represent extrapolated data. Panel **d:** Rarefaction curves of the most promising REDgroups (clust-o-matic T=0.5). Again, the solid lines represent interpolated and actual data, while the dotted lines represent extrapolated data. Though the exact numbers differ, the similarities between the two methods are apparent.

Streptomyces, even when split into multiple REDgroups, is in the top groups both based on the known biosynthetic diversity and based on the estimated potential values. 5,908 (+103 Streptomyces_B, +39 Streptomyces_C, +16 Streptomyces_D) GCFs appear to be unique to the group, even among other phyla (Figure 5a). This is in agreement with previous studies investigating how much overlap there is among the main groups of producers[47]. What is more, streptomycetes appear to be the source of a good percentage of the biosynthetic diversity attributed to the Actinobacteria phylum, as seen in Figure 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[48] (Figure 5c). Among the 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, 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.

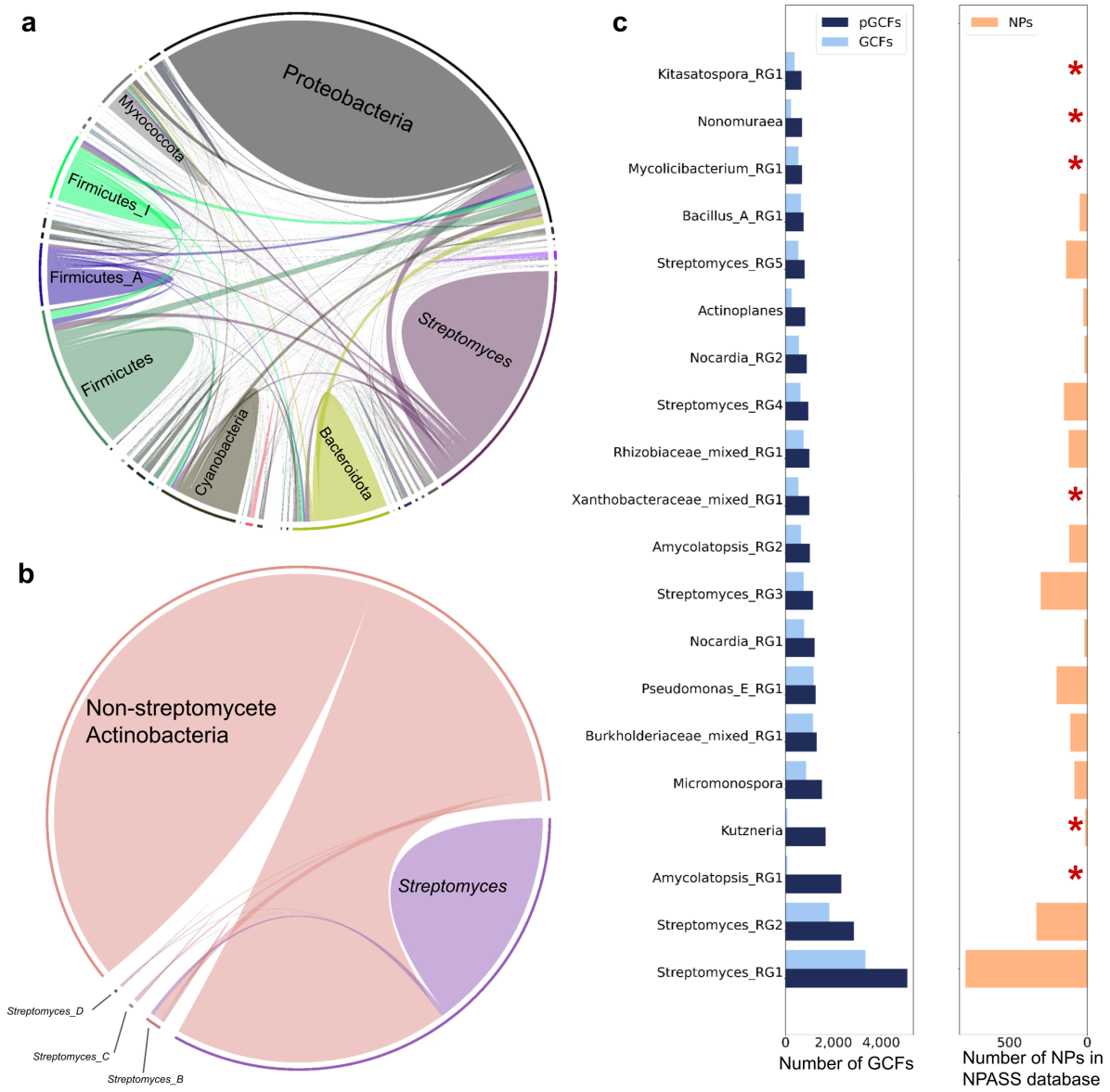
a
Proteobacteria
Myxococcota
Firmicutes_I
Firmicutes_A
Firmicutes
Cyanobacteria
Bacteroidota
Streptomyces
b
Non-streptomycete
Actinobacteria
Streptomyces
Streptomyces_D
Streptomyces_C
Streptomyces_B
c
pGCFs
GCFs
NPs
Kitasatospora_RG1
Nonomuraea
Mycolicibacterium_RG1
Bacillus_A_RG1
Streptomyces_RG5
Actinoplanes
Nocardia_RG2
Streptomyces_RG4
Rhizobiaceae_mixed_RG1
Xanthobacteraceae_mixed_RG1
Amycolatopsis_RG2
Streptomyces_RG3
Nocardia_RG1
Pseudomonas_E_RG1
Burkholderiaceae_mixed_RG1
Micromonospora
Kutzneria
Amycolatopsis_RG1
Streptomyces_RG2
Streptomyces_RG1
0
2,000
4,000
Number of GCFs
500
0
Number of NPs in
NPASS database

Figure 5. Unique diversity in the known producer *Streptomyces* and promising potential of less popular taxa. Panel **a**: Unique Gene Cluster Families (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[41]. Each taxon has a distinct color. The smaller shapes and ribbons represent smaller phyla that can be seen in Supplementary Figure 8. The genus *Streptomyces* appears to have a very high amount of unique GCFs comparable to entire phyla, such as Proteobacteria. Panel **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[41]. The *Streptomyces* genus, only one of many belonging to the Actinobacteriota phylum, appears to be responsible for a big percentage of the phylum's unique diversity (see big pink ribbon). Panel **c**: Left: Potential (pGCFs) and actual (GCFs) number of Gene Cluster Families as defined by BiG-SLiCE (T=0.4), of top 20 most promising REDgroups. Right: number of Natural Products (NPs) found in the NPASS database[48], 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, Kitasatospora_RG1.

Discussion

Using two different independent algorithms, we mined the large amounts of available sequencing data to identify microbial Biosynthetic Gene Clusters (BGCs) and group them into gene cluster families (GCFs), making sure that they mirror chemical families of encoded compounds. We located the 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 not only the largest investigation of this sort including all available genomes so far, as well as published Metagenome Assembled Genomes (MAGs) from diverse environments, but it also provides a reproducible pipeline to determine the metabolic diversity of comparable groups of bacteria and propounds a rationale for drug discovery efforts.

The full biosynthetic capacity of the bacterial kingdom has been assessed in a previous study by Cimermancic et al.[14], already providing a glimpse into the huge untapped biosynthetic potential within bacterial genomes. However, besides their much smaller dataset (about 33,000 BGCs vs 1,185,995 BGCs here), they used ClusterFinder as an identification tool, a more exploratory software algorithm which can include false positive gene clusters. 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 - though potent as ever - still have some limitations[49]. For instance, the study of GCF uniqueness among taxa may be affected by antiSMASH's imperfect BGC boundary prediction[50]. Even though BiG-SLiCE converts BGCs into features based only on domains related to biosynthesis[28], genomic context unrelated to the biosynthetic pathway of a BGC

could still play a role in the GCF assignment; this issue cannot be fully addressed yet 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 so many different BGC types is unparalleled[51], as is apparent from its common use in Natural Product (NP) research[14,17,32,43,45,52,53]. What is more, the fact that it is rule-based[50] implies the possibility of undetected types of clusters and increases the likelihood that our calculations have underestimated the true biosynthetic potential of bacterial organisms, which could be even more impressive.

Furthermore, our pipeline was the first that used GTDB[21] taxonomy for studying bacterial biosynthetic diversity globally. This enabled us to avoid the established misclassifications of the NCBI taxonomic placement[54–57]. The use of rarefaction curves allowed us to infer the biosynthetic potential of bacterial groups as done in some smaller-scaled projects[14,16,44,45]. This method aims to enable fair comparisons among incomplete samples[58]. 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[58]. 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[24,27].

Our analysis identified a plethora of unexplored taxonomic groups with high biosynthetic potential, rendering their characterization essential[9,17,18,59–63]. At the same time, it revealed that undiscovered biosynthetic diversity hides among the popular targets for NP discovery. Indicatively, multiple Proteobacteria taxa were identified among the top producers: *Pseudomonas*, *Pseudoalteromonas*, *Paracoccus*, *Serratia* among others. This is in accordance with the known biosynthetic potential of the Proteobacteria phylum[46]. Furthermore, even undersequenced phyla included promising groups, like the myxobacterial genera *Cystobacter*, *Melittangium*, *Archangium*, *Vitiosangium*, *Sorangium* and *Myxococcus*[17,43,64], or the genera *Chryseobacterium* and *Chryseobacterium_A*[65–67] from the Bacteroidota phylum. However, the majority of most diverse groups comprise actinobacterial strains of well-known and well-studied NP producers from genera such as *Actinoplanes*, *Amycolatopsis*, *Micromonospora*, *Mycobacterium*, *Nocardia* and *Streptomyces*[16,33,34,60,68]. These bacteria produce the majority of Natural Products used as antibiotics[33,60] and our analysis confirms the notion that, contrary to some estimates, there is still much more natural product diversity to be discovered within this group as more diversified strains get sequenced[16,33,34].

Streptomyces is a genus of the Actinobacteria phylum that contains some of the most complex bacteria that we know of, though by far not the most sequenced in our

dataset (Supplementary Figure 9). These bacteria have been known as NP producers for a long time[68], as single strains containing a high number of BGCs have been discovered, taking up to 10% of their genome[69]. However, members of other genera contain comparable absolute numbers of BGCs. This is 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 to all others[68]. The factors that cause this taxonomic group to stand out are not completely clear but probably related to their sophisticated lifestyle. Many observations suggest that NP biosynthesis drives speciation within the *Streptomyces* genus[16]. 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 new compound though[70,71]. Sometimes, the bacterium in question cannot be grown or BGCs are not expressed in laboratory conditions[33,61,63,70,71]. 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[12,61,70,72]. However, efforts to decode biosynthetic mechanisms for the activation of silent clusters need to be tailored to specific producer groups[33,34,71,73], such as groups phylogenetically related to promising producers, e.g. members of the Pseudonocardiaceae family (Figure 4, Supplementary Table 8), partly on the grounds that each phylum has unique diversity (Figure 2b).

Original approaches to the prioritization issue of NP research continue to emerge, fueled by the advances in metagenomics and computational tools that enable the use of the biosynthetic potential of unculturable bacteria from environmental samples[74,75]. 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[76,77]. Although the reconstruction of genomes from metagenomes remains a challenge[78–80] and the assembly will often miss BGCs[81–84], which has indirectly prevented their comparison to the cultured bacteria in the current project, metagenomics is proving a promising source of information on NPs and their producers[14,47,61,74,76,85,86], as made apparent in the present investigation. We expect the effect of this field on NP research to become more evident in the following years.

The collection of microbial data from increasingly new 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[31,32,45,46,63] 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 Product discovery: the prioritization of producers for research[74].

Methods

BGC data set

We obtained 170,585 complete and draft bacterial genomes (Table 1) from RefSeq[20] 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 analyzed with antiSMASH (version 5)[50], which identified their BGCs (Extended Data Table 1, Supplementary Table 1).

Taxonomic classification

Due to multiple indications regarding a lack of accuracy of NCBI's taxonomic classification of bacterial genomes[54–57], we chose to use the Genome Taxonomy Database (GTDB[21]) 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 important, but rather how different these BGCs are to each other. In order to quantify this diversity, we analyzed all BGCs with the new BiG-SLiCE tool[28], which groups similar clusters into Gene Cluster Families (GCFs). However, the first version of this tool has an inherent bias towards multi-protein families BGCs, producing uneven coverage between BGCs of different classes (i.e., 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 will produce a cosine-like distance when processed by the Euclidean-based BIRCH algorithm. This usage of cosine-like distance will virtually balance the measured distance between BGCs with “high” and “low” feature counts (Supplementary Figure 10a), in the end providing an improved clustering performance when measured using the reference data of manually-curated MIBiG GCFs (Supplementary Figure 10b).

The GTDB[21] (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 #GCFs 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 taxonomy richness analysis, we compared BiG-SLiCE results on 947 MIBiG BGCs versus the compound-based clustering provided by the NPAtlas database[19] (Supplementary Figure 3). A final threshold of T=0.4 was chosen based on its similarity to 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 that instead of grouping BGCs into GCFs based on 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[87] 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 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.

Biogeography Analysis

One[27] 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, 4 of this project and from the Nayfach paper[27] Supplementary Information. This led to the creation of Supplementary Table 7. Then, the largest occurring intersections were computed and visualised in Supplementary Figure 4 using the UpSet[88] visualisation technique.

Variance Analysis

In order 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[89]) based on the #GCFs values of immediately lower-ranked taxa, as long as there were at least two such taxa. For example, a phylum that includes only one class in our dataset was omitted from this

computation. But a phylum with two or more classes would be assigned a variance value computed from its classes' #GCFs values. The distribution of these variance values was plotted for each rank in Figure 3a. We noticed a significant reduction in variance from the family to the genus rank, which was confirmed with an additional statistical test (Supplementary Figure 5). A similar variance analysis was performed to compare genera and REDgroups (Supplementary Figure 6) but in this case variance was calculated based on the strains' biosynthetic diversity.

Definition of REDgroups

To study the biosynthetic diversity of genera, we attempted to achieve uniform taxa. The creators of GTDB used Relative Evolutionary Divergence (RED) for taxonomic rank normalization[21]; it 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[21], we chose the median of GTDB genus RED values, namely 0.934, as a cutoff threshold. Any clade in the GTDB bacterial tree with an assigned RED value higher than the threshold was considered one group (Supplementary Figure 11) that we named "REDgroup". For REDgroup naming conventions, see Supplementary Figure 11.

Rarefaction analysis

The extrapolation of potential #GCFs values was achieved by conducting rarefaction analyses, by use of the iNEXT R package[90]. 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 5000 for the REDgroups, and of 1.6 million (about 8 times the number of strains in the Complete Dataset) in each group for the RefSeq analyses (where 2000 points were inter- or extrapolated). By default, the number of bootstrap replications is 50.

Random sampling

In order to test whether the above methods (creation of REDgroups and the subsequent rarefaction analyses) overcome the inherent sequencing bias in our dataset, a random sampling technique was used. A reduced dataset was tested that included only those REDgroups containing at least 20 members. 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 potential GCFs (pGCFs) value for each REDgroup from all iterations was calculated and 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[48] (accessed on 15 October 2020), taking care to match the GTDB-given names to the NCBI-given names that the database uses.

References

1. O'Connor, S. E. Engineering of Secondary Metabolism. *Annu. Rev. Genet.* **49**, 71–94 (2015).
2. Wright, G. D. Unlocking the potential of natural products in drug discovery. *Microb. Biotechnol.* **12**, 55–57 (2019).
3. 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).
4. Fedorenko, V. *et al.* Antibacterial Discovery and Development: From Gene to Product and Back. *Biomed Res. Int.* **2015**, 591349 (2015).
5. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).
6. 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).
7. Michael, C. A., Dominey-Howes, D. & Labbate, M. The antimicrobial resistance crisis: causes, consequences, and management. *Front Public Health* **2**, 145 (2014).
8. Lyddiard, D., Jones, G. L. & Greatrex, B. W. Keeping it simple: lessons from the golden era of antibiotic discovery. *FEMS Microbiol. Lett.* **363**, (2016).
9. Pye, C. R., Bertin, M. J., Lokey, R. S., Gerwick, W. H. & Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl.*

Acad. Sci. U. S. A. **114**, 5601–5606 (2017).

10. Hutchings, M. I., Truman, A. W. & Wilkinson, B. Antibiotics: past, present and future. *Current Opinion in Microbiology* vol. 51 72–80 (2019).
11. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).
12. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).
13. Medema, M. H., de Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* (2021) doi:10.1038/s41576-021-00363-7.
14. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
15. Chevrette, M. G. *et al.* Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.* **37**, 566–599 (2020).
16. Doroghazi, J. R. *et al.* Aroadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
17. Hoffmann, T. *et al.* Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nat. Commun.* **9**, 803 (2018).
18. Lewis, K. The Science of Antibiotic Discovery. *Cell* vol. 181 29–45 (2020).
19. 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).
20. Haft, D. H. *et al.* RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).
21. 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).
22. 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).
23. 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).

24. Almeida, A. *et al.* A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. doi:10.1101/762682.
25. 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).
26. 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).
27. Nayfach, S. *et al.* Author Correction: A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 521 (2021).
28. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters. *Bioinformatics* (2020).
29. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
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. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* (2021) doi:10.1038/s41586-021-04233-4.
32. Sharrar, A. M. *et al.* Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. *MBio* **11**, (2020).
33. Barka, E. A. *et al.* Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol. Mol. Biol. Rev.* **80**, 1–43 (2016).
34. Genilloud, O. Actinomycetes: still a source of novel antibiotics. *Nat. Prod. Rep.* **34**, 1203–1232 (2017).
35. Chevrette, M. G. *et al.* The confluence of big data and evolutionary genome mining for the discovery of natural products. *Nat. Prod. Rep.* (2021) doi:10.1039/d1np00013f.
36. McDonald, B. R. & Currie, C. R. Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus. *MBio* **8**, (2017).

37. Chase, A. B., Sweeney, D., Muskat, M. N., Guillén-Matus, D. & Jensen, P. R. Vertical inheritance governs biosynthetic gene cluster evolution and chemical diversification. doi:10.1101/2020.12.19.423547.
38. Avni, E. & Snir, S. A New Phylogenomic Approach For Quantifying Horizontal Gene Transfer Trends in Prokaryotes. *Sci. Rep.* **10**, 1–14 (2020).
39. Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* **36**, 6688–6719 (2008).
40. 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).
41. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* vol. 30 2811–2812 (2014).
42. 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).
43. Männle, D. *et al.* Comparative Genomics and Metabolomics in the Genus Nocardia. *mSystems* **5**, (2020).
44. Ziemert, N. *et al.* Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1130–9 (2014).
45. Adamek, M. *et al.* Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics* **19**, 426 (2018).
46. 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).
47. Bérdy, J. Bioactive microbial metabolites. *J. Antibiot.* **58**, 1–26 (2005).
48. 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).
49. 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).

50. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
51. Chavali, A. K. & Rhee, S. Y. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Brief. Bioinform.* **19**, 1022–1034 (2018).
52. Ueoka, R. *et al.* Genome-Based Identification of a Plant-Associated Marine Bacterium as a Rich Natural Product Source. *Angew. Chem. Int. Ed Engl.* **57**, 14519–14523 (2018).
53. Adamek, M., Alanjary, M. & Ziemert, N. Applied evolution: Phylogeny-based approaches in natural products research. *Natural Product Reports* vol. 36 1295–1312 (2019).
54. 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).
55. 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).
56. 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).
57. 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).
58. Chao, A. *et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies*. vol. 84 45–67 http://purl.oclc.org/estimates (2014).
59. Zheng, Y. *et al.* Genome Features and Secondary Metabolites Biosynthetic Potential of the Class Ktedonobacteria. *Front. Microbiol.* **10**, 893 (2019).
60. Amin, D. H., Abdallah, N. A., Abolmaaty, A., Tolba, S. & Wellington, E. M. H. Microbiological and molecular insights on rare Actinobacteria harboring bioactive

prospective. *Bulletin of the National Research Centre* **44**, 1–12 (2020).

61. Hug, J. J., Bader, C. D., Remškar, M., Cirnski, K. & Müller, R. Concepts and Methods to Access Novel Antibiotics from Actinomycetes. *Antibiotics (Basel)* **7**, (2018).
62. Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
63. Subramani, R. & Sipkema, D. Marine rare actinomycetes: A promising source of structurally diverse and unique novel natural products. *Marine Drugs* vol. 17 (2019).
64. Weissman, K. J. & Müller, R. Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.* **27**, 1276–1295 (2010).
65. Cimmino, T. & Rolain, J.-M. Whole genome sequencing for deciphering the resistome of Chryseobacterium indologenes, an emerging multidrug-resistant bacterium isolated from a cystic fibrosis patient in Marseille, France. *New Microbes New Infect* **12**, 35–42 (2016).
66. Kang, D., Shoaie, S., Jacquiod, S., Sørensen, S. J. & Ledesma-Amaro, R. Comparative genomics analysis of Chryseobacterium sp. KMC2 reveals metabolic pathways involved in keratinous utilization and natural product biosynthesis. *bioRxiv* (2021) doi:10.1101/2021.02.23.432615.
67. 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. Antibiot.* **74**, 115–123 (2021).
68. Berdi, J. Bioactive Microbial Metabolites: A Personal View. *Journal of Antibiotics. Antibiotics* **58**, 1–26 (2005).
69. Baltz, R. H. Gifted microbes for genome mining and natural product discovery. *J. Ind. Microbiol. Biotechnol.* **44**, 573–588 (2017).
70. Seyedsayamdost, M. R. Toward a global picture of bacterial secondary metabolism. *Journal of Industrial Microbiology and Biotechnology* vol. 46 301–311 (2019).
71. Wohlleben, W., Mast, Y., Stegmann, E. & Ziemert, N. Antibiotic drug discovery. *Microb. Biotechnol.* **9**, 541–548 (2016).

72. 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).
73. Kalkreuter, E., Pan, G., Cepeda, A. J. & Shen, B. Targeting Bacterial Genomes for Natural Product Discovery. *Trends Pharmacol. Sci.* **41**, 13–26 (2020).
74. Tracanna, V., de Jong, A., Medema, M. H. & Kuipers, O. P. Mining prokaryotes for antimicrobial compounds: From diversity to function. *FEMS Microbiology Reviews* vol. 41 417–429 (2017).
75. Gluck-Thaler, E. *et al.* The Architecture of Metabolism Maximizes Biosynthetic Diversity in the Largest Class of Fungi. *Mol. Biol. Evol.* **37**, 2838–2856 (2020).
76. Chen, R. *et al.* Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats. *Front. Microbiol.* **11**, 1950 (2020).
77. Milshteyn, A., Colosimo, D. A. & Brady, S. F. Accessing Bioactive Natural Products from the Human Microbiome. *Cell Host Microbe* **23**, 725–736 (2018).
78. Vollmers, J., Wiegand, S. & Kaster, A.-K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLoS One* **12**, e0169662 (2017).
79. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Applied and Environmental Microbiology* vol. 87 (2021).
80. Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic Assembly: Overview, Challenges and Applications. *Yale J. Biol. Med.* **89**, 353–362 (2016).
81. 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).
82. Pereira-Flores, E. *et al.* Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultrafast techniques. *bioRxiv* 2021.01.20.427441

(2021) doi:10.1101/2021.01.20.427441.

83. Altamia, M. A. *et al.* Secondary Metabolism in the Gill Microbiota of Shipworms (Teredinidae) as Revealed by Comparison of Metagenomes and Nearly Complete Symbiont Genomes. *mSystems* **5**, (2020).
84. Chanson, A., Moreau, C. S. & Duplais, C. Assessing Biosynthetic Gene Cluster Diversity of Specialized Metabolites in the Conserved Gut Symbionts of Herbivorous Turtle Ants. *Front. Microbiol.* **0**, (2021).
85. Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15**, 141–161 (2015).
86. Shen, B. A New Golden Age of Natural Products Drug Discovery. *Cell* **163**, 1297–1300 (2015).
87. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
88. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
89. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
90. Hsieh, T. C., Ma, K. H. & Chao, A. iNEXT: an R package for rarefaction and extrapolation of species diversity (H ill numbers). *Methods Ecol. Evol.* **7**, 1451–1456 (2016).
91. Zielezinski, A. *et al.* Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **20**, 1–18 (2019).

Fundings

A.G. is grateful for the support of the Deutsche Forschungsgemeinschaft (DFG; Project ID # 398967434-TRR 261). N.Zi. is supported by the German Center for Infection Research (DZIF) (TTU 09.716). M.H.M. is supported by an 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. is supported by BMBF, DFG and DZIF.

Acknowledgements

A.G. and N.Zi. thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2124 – 390838134 for the infrastructural support. A.G. thanks Direnc M. Mungan for valued discussions on optimizing the analysis, as well as Caner Bagci for his imaginative suggestion on dealing with large data. We also thank Dr. Libera do Presti for invaluable comments on the manuscript.

Author contributions

A.G., S.A.K., N.Za. and D.K. have performed the analysis. S.A.K. and N.Za. have contributed analysis tools. A.G., D.K., R.M., M.H.M. and N.Zi. have written the paper. All authors have contributed to the conception and design of the analysis. All authors have 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.

Extended Data

Extended Data Table 1
Information on the most biosynthetically promising REDgroups (BiG-SLiCE T=0.4).

Supplementary information

Supplementary Figures
All Supplementary Figures and their labels are provided in an additional file.
Supplementary Table 1
Accession numbers, GTDB-based taxonomic information and BGC IDs of all genomes from all datasets used in the analysis. MAG datasets: mag_uba[26], mag_humangut[91], mag_chicken[23], mag_bovine[22], mag_ocean[25], GEMS[27].

Supplementary Table 2
BGC to BiG-SLiCE GCF assignment and centroid distance for T=0.4 (which proved to be the most suitable threshold)

Supplementary Table 3
BGC to BiG-SLiCE GCF assignment and centroid distance for T=0.5

Supplementary Table 4
BGC to BiG-SLiCE GCF assignment and centroid distance for T=0.6

Supplementary Table 5
BGC to BiG-SLiCE GCF assignment and centroid distance for T=0.7

Supplementary Table 6
BGC IDs, MiBIG IDs, producer GTDB-based taxonomic information and GCF assignment (for BiG-SLiCE T=0.4) for all MiBIG BGCs included in the creation of Figure 1C.

Supplementary Table 7
Biogeography analysis of the Nayfach MAGs dataset[27]. Number of genomes, BGCs, GCFs and unique GCFs per ecosystem type (as defined in the corresponding paper[27]).

Supplementary Table 8
REDgroup full metadata: Node IDs (can be used in the exploration of the tree in Supplementary Figure 7), labels, number of members, number of BGCs, number of GCFs and potential GCFs (pGCFs) as defined by BiG-SLiCE (T=0.4) and clust-o-matic (T=0.5), GTDB taxonomic information and number of products in the NPASS database whose producer is a member of the REDgroup (NPASS_hits).

Supplementary Table 9
Comparison of random sampling analysis results to original results. The table includes: Node IDs, labels, number of members, number of BGCs, number of GCFs and potential GCFs (pGCFs) as defined by BiG-SLiCE (T=0.4), the original ranking based on the pGCFs, the average pGCFs from all random sampling iterations, the ranking based on the random sampling and GTDB taxonomic information.

(Note to reviewers: These data in Supplementary Tables 1 to 5, which are especially large files, would be publicly released upon completion of peer review of this manuscript under the following link: https://doi.org/10.5281/zenodo.5159210)

Code availability

The clust-o-matic code is available here: 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) can be found in this link: https://github.com/medema-group/bigslice/blob/master/misc/useful_scripts/perform_l2norm_clustering.py

Figure Legends

Figure 1. Biosynthetic diversity of the sequenced bacterial kingdom. Panel **a**: Bar plots of Gene Cluster Families (GCFs, as defined by BiG-SLiCE) of 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. Panel **b**: Rarefaction curves of all RefSeq bacteria based on BiG-SLiCE (red) and based on 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 dotted lines represent extrapolated data. The number of chemical classes documented in NPAtlas[19], which come from bacterial producers (gray dotted 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). Panel **c**: Venn Diagram of GCFs (as defined by BiG-SLiCE, T=0.4) of the bacterial RefSeq, Minimum Information about a Biosynthetic Gene cluster (MIBiG[30]) 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.

Figure 2. Comparison of biosynthetic diversity among phyla. Panel **a**: The Genome Taxonomy DataBase (GTDB[21]) bacterial tree was visualized with iTOL[40], decorated with Gene Cluster Families (GCFs) values (as defined by BiG-SLiCE at T=0.4), collapsed at the phylum rank and accompanied by bar plot 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. Panel **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[41]. Each phylum has a distinct color. Actinobacteriota (2) and Proteobacteria (40) seem particularly rich in unique GCFs.

Figure 3. Relations of taxonomic levels to variability in biosynthetic diversity. Panel **a**: Modified "raincloud plots"[42] 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 Gene Cluster Families (GCFs as 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 a 1.5x interquartile range. Points outside of the whiskers are outliers. 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 Figure 5), indicating that the genera are suitable taxonomic groups to be characterised as diverse and be compared to each other. Panel **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 (1), Streptomycetales families (2), Streptomycetaceae genera (3), top 50 most diverse *Streptomyces* species (4). The difference in variance is visible in the graphs 1,2,3, but becomes homogeneous at the species level as is shown in graph 4.

Figure 4. Overview of actual and potential biosynthetic diversity of bacterial kingdom, compared at REDgroup level. Panel **a:** GTDB[21] bacterial tree up to REDgroup level, visualized with iTOL[40], 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). Top REDgroups with most potential GCFs include the following: A: Streptomyces_RG1, B: Streptomyces_RG2, C: Amycolatopsis_RG1, D: Kutzneria, E: Pseudomonas_E. 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 significant biosynthetic diversity and potential. An interactive version of Figure 4a can be accessed online (Supplementary Figure 7). Panel **b:** Rarefaction curves of REDgroups (BiG-SLiCE T=0.4). The solid lines represent interpolated and actual data, while the dotted lines represent extrapolated data. The letters "L", "M" and "H" on the right edge of the figure correspond to Low- (0-389 pGCFs), Medium- (390-649 pGCFs) and High-diversity (more than 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, drawn in the "L" range of the graph). The genera included in the most promising REDgroups are indicated (the letters a-e correspond to the peaks in panel a. *Streptomyces* strains are included in several of the top most promising REDgroups. Panel **c:** Rarefaction curves of the most promising REDgroups (BiG-SLiCE T=0.4). The solid lines represent interpolated and actual data, while the dotted lines represent extrapolated data. Panel **d:** Rarefaction curves of the most promising REDgroups (clust-o-matic T=0.5). Again, the solid lines represent interpolated and actual data, while the dotted lines represent extrapolated data. Though the exact numbers differ, the similarities between the two methods are apparent.

Figure 5. Unique diversity in the known producer *Streptomyces* and promising potential of less popular taxa. Panel **a**: Unique Gene Cluster Families (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[41]. Each taxon has a distinct color. The smaller shapes and ribbons represent smaller phyla that can be seen in Supplementary Figure 8. The genus *Streptomyces* appears to have a very high amount of unique GCFs comparable to entire phyla, such as Proteobacteria. Panel **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[41]. The *Streptomyces* genus, only one of many belonging to the Actinobacteriota phylum, appears to be responsible for a big percentage of the phylum's unique diversity (see big pink ribbon). Panel **c**: Left: Potential (pGCFs) and actual (GCFs) number of Gene Cluster Families as defined by BiG-SLiCE (T=0.4), of top 20 most promising REDgroups. Right: number of Natural Products (NPs) found in the NPASS database[48], 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, Kitasatospora_RG1.

Tables

Table 1. Overview of input datasets and overall biosynthetic diversity under various BiG-SLiCE thresholds. 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 thresholds; BGC to GCF assignment for each threshold can be found in Supplementary Tables 2-5. *MAG sources: bovine rumen[22], chicken caecum[23], human gut[24], ocean[25], uncultivated bacteria[26], various sources[27].

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*	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