Published in final edited form as:
Nat Chem Biol. 2014 November ; 10(11): 963–968. doi:10.1038/nchembio.1659.

A Roadmap for Natural Product Discovery Based on Large-Scale Genomics and Metabolomics

James R. Doroghazi[1,†], **Jessica C. Albright**[2,†], **Anthony W. Goering**[2], **Kou-San Ju**[1], **Robert R. Haines**[3], **Konstantin A. Tchalukov**[3], **David P. Labeda**[4], **Neil L. Kelleher**[2,*], and **William W. Metcalf**[1,3,*]

[1] Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

[2] Departments of Chemistry, Molecular Biosciences, and the Feinberg School of Medicine, Northwestern University, 2170 Campus Drive, Evanston, IL, 60208, USA

[3] Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

[4] Bacterial Foodborne Pathogens and Mycology Research, USDA, Agricultural Research Service, National Center for Agricultural Utilization Research, Peoria, IL, 61604, USA

Abstract

Actinobacteria encode a wealth of natural product biosynthetic gene clusters (NPGCs), whose systematic study is complicated by numerous repetitive motifs. By combining several metrics we developed a method for global classification of these gene clusters into families (GCFs) and analyzed the biosynthetic capacity of *Actinobacteria* in 830 genome sequences, including 344 obtained for this project. The GCF network, comprised of 11,422 gene clusters grouped into 4,122 GCFs, was validated in hundreds of strains by correlating confident mass spectrometric detection of known small molecules with the presence/absence of their established biosynthetic gene clusters. The method also linked previously unassigned GCFs to known natural products, an approach that will enable *de novo*, bioassay-free discovery of novel natural products using large data sets. Extrapolation from the 830-genome dataset reveals that *Actinobacteria* encode hundreds of thousands of future drug leads, while the strong correlation between phylogeny and GCFs frames a roadmap to efficiently access them.

[*] Corresponding authors: William W. Metcalf, metcalf@illinois.edu. Neil L. Kelleher, n-kelleher@northwestern.edu.
[†] Authors contributed equally to this study

Author Contributions JRD designed and performed bioinformatic analyses. JRD, RRH and KAT produced microbial extracts. JCA designed LC-MS experiments and JCA and AWG collected and analyzed LC-MS data. KSJ received and processed germplasm and contributed to genomic library preparation. DPL selected and provided germplasm from the ARS Culture Collection. WWM and NLK designed and directed the work. JRD, JCA, WWM and NLK wrote the manuscript.

Author Information Draft genomes sequenced as part of this project are available through NCBI BioProject PRJNA238534.

Supplementary Information is submitted as file "Metcalf SI.pdf", "Metcalf Suppl Note.pdf" "Metcalf_supplementary_dataset1.xlsx", "Metcalf_supplementary_dataset2.xlsx", and "Metcalf_supplementary_dataset3.xlsx". An interactive database of the GCF networks is available online at www.igb.illinois.edu/labs/metcalf/gcf.

Competing Financial Interests
The authors declare no competing financial interests.

Introduction

Natural products from actinomycetes have been the source or inspiration for the majority of clinically useful antibiotics, along with numerous other pharmaceutically useful compounds including immunosuppressive, antiproliferative, herbicidal, insecticidal, fungicidal and antiparasitic drugs [1]. Yet, despite the historical importance and numerical dominance of natural product-based medicines, many pharmaceutical companies have replaced their natural product discovery efforts with target-based screening of synthetic compound libraries [2]. Ironically, the coincident rise of rapid and inexpensive DNA sequencing technology has revealed a wealth of novel natural product biosynthetic gene clusters in actinomycete genomes: typically ten-fold higher than the number of molecules discovered by traditional approaches in each organism [3]. Thus, the current pharmaceutical industry is built upon <10% of the biosynthetic capacity of the microbial world. These observations have engendered the idea that “genome-mining” will lead to a renaissance in natural product discovery that could revitalize the pharmaceutical industry. Although genome-mining has allowed some notable natural product discoveries [4,5], its promise has yet to be fulfilled on a large scale due to bioinformatics hurdles related to the complex and repetitive nature of the biosynthetic genes involved and the need to have specific and sensitive assays for each new compound.

Systematic, large-scale genome-mining will require automated methods for recognition and classification of units that generate new compounds – namely biosynthetic gene clusters. Toward this end, numerous software tools have been developed for the automated recognition and database storage of NPGCs. Some of these tools predict the structure of the putative natural products, while others perform simple comparisons between the clusters found in various organisms [6-14]. However, to our knowledge, none perform a global classification of all gene clusters, which is a computationally difficult problem due to ubiquitous features found in the common natural product biosynthetic enzymes, polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs). Individual NRPS and PKS proteins are comprised of repeating modules that add or modify the chemical subunits of growing natural products (see [15] for review). These include derivatives of at least ten discrete modules that catalyze substrate activation and tethering, condensation with the next subunit, chemical modification of the growing chain and release of the final product. Over time, these repeating modules have recombined with other NPGCs, both within and between organisms, to produce a complex and diverse set of biosynthetic gene clusters that are difficult to functionally dissect due to shared sequences in genes that direct the synthesis of very different products and rearranged gene clusters that direct synthesis of the same product. We report a systematic bioinformatics framework for the study of natural product gene clusters. We used mass spectrometry data to verify gene cluster family designations and to demonstrate utility for *de novo* correlation of natural products and biosynthetic genes.

Results

Creation and analysis of a GCF network

We created a dataset comprised of NPGCs involved in the synthesis of NRPS, type I and type II PKS, NRPS-independent siderophores (NISs), lanthipeptides, and thiazole-oxazole modified microcins (TOMMs). These include NPGCs identified in 830 actinomycete genomes, 344 of which were sequenced for this project (Supplementary Data Set 1), as well as 412 gene clusters with established natural products from Genbank. Three distance metrics were calculated for every NPGC pair: (i) the number of homologous genes shared (Fig. 1a); (ii) the proportion of nucleotides involved in a pairwise alignment (Fig. 1b); and (iii) the amino acid sequence identity between the domains of repeated protein modules (Fig. 1c). When used alone, each of these metrics has limited ability to separate homologous from paralogous gene clusters. For example, BLAST-based ortholog detection and clustering as reported previously [16] is especially limited by the repeated, modular structure of large PKS and NRPS genes, which produce very low e-values for genes in different gene cluster families. However, a combined score that incorporates all three metrics produces coherent GCFs that include only highly related NPGCs, as assessed by manual inspections of gene cluster diagrams and GCF network visualizations (Supplementary Results, Supplementary Fig. 1; an interactive version of Supplementary Figure 1 is available at www.igb.illinois.edu/labs/metcalf/gcf). The all-v-all nature of this approach was very computationally expensive, taking several months on a server utilizing 40 processors, mostly dedicated to PROmer comparisons. The combined scoring metric was not systematically compared to the BLAST only clustering because the families produced by this metric clearly produced families containing unrelated gene clusters (data not shown). To further validate the approach, we examined the gene cluster families that contain multiple NPGCs with established natural products. (Note that characterized gene clusters that are the sole members of GCFs are not informative in this regard and were not included in the analysis.) The combined metric correctly grouped 103 characterized gene clusters into forty-one GCFs that direct synthesis of highly similar natural products (Supplementary Note and Supplementary Fig. 2). The single known NPGC that was not grouped with relatives making highly similar compounds is an unusual actinomycin gene cluster with a large internal duplication, previously reported to be divergent from other actinomycin gene clusters [17].

The GCF network created using these methods from the 830 actinobacterial genomes was comprised of 140,986 genes from 11,422 NPGCs grouped into 4,122 GCFs (Fig. 2a). There were 77 GCFs that contain at least one characterized NPGC along with related, as yet uncharacterized, gene clusters in the newly sequenced genomes. These “anchored” GCFs provided 1193 new NPGCs that are likely to direct the synthesis of novel derivatives of known compounds (Supplementary Data Set 2), each with a high expectation of exhibiting a targeted bioactivity. We are currently developing functionality that will allow users to submit their own genomic data into this framework.

The ability of the GCF approach to recognize whether one NPGC is the same or different from another allows the use of rarefaction analysis to estimate the total number of natural product scaffolds made by actinomycetes and the number of strains that would need to be screened to find them all. This statistical method examines the rate at which new GCFs accumulate as new genomes are randomly added to the data set. This process is somewhat complicated by the non-random nature of the data; over 40% of the genomes in our dataset were derived from a few medically relevant genera: *Mycobacterium* (142 genomes); *Corynebacterium* (68 genomes); *Propionibacterium* (89 genomes); *Actinomyces* (24 genomes); and *Gordonia* (20 genomes (Fig. 2a). To correct for these phylogenetic biases, we randomly sampled our data from phylogenetic bins (operational taxonomic units, or OTUs) for extrapolation analyses and to produce a rarefaction curve (Fig. 2b). Three different estimation methods, Chao1, ACE and extrapolation, implemented within the software EstimateS [18], were used to analyze the data; all gave similar results (Supplementary Table 1). Although the NRPS GCFs were the most abundant in our current data, the extrapolation predicts that type I PKS will become more abundant as more genomes are sampled. These classes were followed, in order, by type II PKSs, lanthipeptides, NRPS-independent siderophores (NISs), and TOMMs. Collectively, we estimated that the actinomycetes encode ~17,350 GCFs in these six classes. Because each GCF represents the biosynthetic capacity to produce a suite of related molecules, these GCFs are predicted to direct the synthesis of hundreds of thousands, but not millions, of natural products. The extrapolations predict that essentially all actinomycete NPGCs will be identified after ca. 15,000 genomes are sequenced from different OTUs, an achievable proposition given the decreasing cost of sequencing. Considerably fewer genome sequences will be required to access the full repertoire of the less abundant classes (Fig. 2b).

Our analysis of the abundance and distribution of the GCFs in the dataset provided an efficient path for selection and prioritization of actinobacterial strains for future natural product discovery efforts. Consistent with previous reports[19], our data suggested that certain phylogenetic groups are gifted with an abundance and diversity of NPGCs (taxa between *Streptomycetales* to *Pseudonocardiales*, inclusive, in Fig. 2a). In total, these gifted groups encoded 73.6% of the GCFs present in our dataset (Table 1). Based on these finding, we asked whether inclusion of the non-gifted genomes biased the diversity estimates. However, when only the gifted taxa were analyzed, the predicted numbers of GCFs were essentially the same as presented in the previous section (Supplementary Table 2). Importantly, our data showed a strong correlation between phylogenetic distance and the GCF complement of individual strains, with the degree of conservation varying according to biosynthetic class and phylogenetic group (Fig. 3, Supplementary Fig. 3). Among the gifted taxa, related strains share 80% of their NRPS and 73% of their type I PKS GCFs at the 1% divergence level (based on a concatenated ribosomal protein phylogeny), These values drop sharply beyond the 1% cut-off, such that strains share only 6% and 3% of their respective NRPS and PKS I GCFs at a ribosomal protein divergence of 4-5%. We observed similar trends for most other natural product classes, although the rate of decline in GCF conservation across genetic distance for lanthipeptides and type II PKS clusters was not as pronounced (Supplementary Fig. 3). In contrast, NRPS-independent siderophores were conserved across a much larger phylogenetic distance: 84% and 66% conservation at 0-1% and 4-5% ribosomal protein divergence, respectively. Importantly, we observed the same correlations between phylogeny and GCF with the use of a simple 297 bp fragment of the *rpoB* gene suitable for amplicon sequencing (Fig. 3b). Significantly, 1% ribosomal protein divergence corresponds to a suggested “species-level” cutoff within *Streptomyces* (Supplementary Fig. 4), consistent with the idea that natural product biosynthetic capacity drives speciation within this genus[20].

Correlation of GCFs with secondary metabolite production

Our analysis of GCFs substantially reinforces the idea that *Actinobacteria* represent a major source of new natural products; however, to move beyond the concept, the molecules produced by each GCF must be identified. With this in mind, we developed a liquid chromatography-high resolution mass spectrometry (LC-HRMS) method to quantitatively measure the exported metabolome of the strains being screened. We tested this method on pooled extracts produced by 178 of the strains included in our GCF data set after growth in four different media. After background subtraction, we detected an average of 105 compounds produced by each strain, many of which were found in multiple strains/samples. In total, 2,521 unique compounds (intact masses) were identified in the MS^1 data. Among these, we identified 110 previously characterized natural products after automated search of a database consisting of 9,817 known actinomycete natural products. After we collapsed highly similar molecules (e.g. actinomycin A and actinomycin D) into groups, we experimentally linked 27 known natural products to nine GCFs in our network. In the samples we analyzed, these 27 compounds were observed and verified by high-resolution tandem MS 268 times (representative data shown in Supplementary Fig.5). Significantly, we identified the correct GCF in the genome of the source organism in 92% of these cases. Thus, the method has a relatively low rate of false positives. Interestingly, these same gene clusters could be found in many strains that did not produce the compounds in question, i.e. the gene clusters were cryptic in these cases. In total, known GCFs were cryptic in 77% of the strains we examined. This observation belies the idea that NPGCs are usually cryptic, with major implications for discovery efforts. Accordingly, one may need to screen on average only four isolates with a given NPGC to find one in which the corresponding natural product is produced at levels sufficient to allow mass spectrometric detection. We wish to emphasize that our assessment of cryptic NPGCs applies only to the known compounds that we identified. It remains to be seen whether this trend will apply to all NPGCs.

We were particularly interested in whether an automated method could be developed for linking specific gene clusters to specific molecules (i.e. specific exact masses from the MS data) based on the simple fact that a molecule cannot be produced without the biosynthetic genes also being present. To achieve this, we performed a binary correlation between subsets of each GCF - based on clustered domain sequences - present in each genome and the MS^1 intact masses found in each extract. Without manual intervention, the binary correlation of the GCF and MS data sets associated experimentally established biosynthetic gene clusters with production of oxytetracycline, benarthin, nonactin, and enterocin, with correlation scores at the extreme high end of the overall distribution (Fig. 4a, Table 2). Further, manual searches revealed the presence of fragmentary gene clusters (due to the draft nature of the genome sequences) that had been omitted from our automated GFC network in some of the strains that produced MS^2 verified natural products. This improved the correlation scores for actinomycin D and pyridomycin, as did the use of the full GCF for proferrioxamine D2, such that the manual correlation score was actually higher than the best automated correlation score. Application of the automated method to the full data set also revealed a strong correlation between a previously uncharacterized GCF and the related compounds desertomycin A, B, D, and E and oasamycin A, D, and E (Fig. 4b, Supplementary Fig. 6). This cluster encodes PKS modules consistent with the known molecule, as well as proteins with predicted functions that match the unique structural features of the molecule, specifically N-methylation, and mannosylation. Thus, we believe it highly likely that this previously unassigned GCF is responsible for the synthesis of both desertomycin and oasamycin. We also observed a highly significant correlation between the griseobactin GCF and benarthin. It had previously been speculated, but never proven, that the two molecules share a biosynthetic pathway[21]. Our correlation data provide additional support for this hypothesis, and show that this method can also be used to find biosynthetic intermediates.

Discussion

Over the past several decades, pharmaceutical companies have cultivated millions of actinomycetes searching for novel bioactive compounds; yet, efforts to discover natural products within this vast and valuable resource is like exploring a continent without a map. As a result, high rates of rediscovery have plagued the natural product discovery arena. Here, we have developed a global, genome- and mass-spectrometric-enabled framework for the discovery of natural products that alleviates this blind screening approach. This discovery roadmap will enable researchers to explore the complete diversity of NPGCs in a systematic way.

While their potential is high, our analyses show that actinomycete natural products represent estimable number of scaffolds that reaches into the thousands rather than the millions. We note, however, that each GCF corresponds to a natural product scaffold, rather than a discrete molecule and that we have not estimated the number of possible modifications within each scaffold family. Given the number of total GCFs in each class of natural products, it is possible to envision research consortia focused on natural product discovery that will aim to uncover the majority of chemical diversity within the actinomycetes for each of the smaller classes.

Despite clear evidence of horizontal gene transfer of biosynthetic genes, our data reveal a strong phylogenetic signal to the genomic catalog of NPGCs. Thus, two strains that are separated by a ribosomal protein distance of 0.5% will likely share almost all of their natural product gene clusters, while two strains separated by 7% will share almost none but the most common. Accordingly, future discovery efforts would benefit greatly from a focus on gifted phylogenetic groups coupled with prescreening to reduce the number of close relatives. This places a much greater importance on knowledge of microbial diversity than if GCFs were randomly distributed. Because different species, and perhaps even supraspecific clades, have habitat preferences, GCFs may also be viewed as having “habitats”; an idea that is supported by recent metagenomic NPGC surveys [22]. The current state of knowledge regarding microbial ecology and phylogenetic diversity within these groups currently limits our ability to assess the degree to which these unknowns will affect our diversity estimates. For example, if novel cultivation methods or culture-independent sequencing uncover novel natural product rich actinobacterial families not currently covered by culture collections, then sampling 15000 genomic species may be insufficient to cover all natural product diversity. To paraphrase [23], our extrapolation to 15000 genomic species is performed under the assumption that the additional samples are collected under the same conditions and protocols as our existing data set. Despite these limitations, our observation that PCR-based screening of *rpoB* alleles provides an accurate reflection of NPGC content suggests a rapid and inexpensive phylogenetic typing method to accelerate this research.

Finally, we note that these GCF/NP linkages were the product of a small data set consisting of data from 178 strains grown under only four conditions, harvested at a single time point and restricted to molecules retained by a single purification methodology. There is every reason to believe that correlation scores will improve dramatically with larger, more inclusive datasets. Moreover, the method reported herein could be further strengthened by expanding to include MS-MS spectral networking as reported previously [24], which has yet to be coupled with genomic data that could be used to constrain the analysis. If this expectation is met, it is not unreasonable to believe that this correlative approach will lead to the discovery of nearly all natural products produced by actinomycetes (or any other group for which a sufficiently large dataset can be generated). In this regard, our analyses revealed numerous correlation scores in the same range as those for the established knowns. Experiments are currently in progress to validate these linkages and to determine the structure of the molecules in question.

Methods

Genome Sequencing

All strains were received directly from the Agricultural Research Service (NRRL) culture collection, Peoria, IL. ATCC® Medium #172 broth [25] was used for liquid cultures (per liter: glucose, 10.0 g; soluble starch, 20.0 g; yeast extract, 5.0 g; N-Z amine type A, 5.0 g; $CaCO_3$, 1.0 g; agar, 15.0 g). Genomic DNA was prepared using the UltraClean Microbial DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA) and sequencing library preparation was performed with Nextera version 2 kits (Illumina, San Diego, CA). Genomes for the 343 strains reported here were sequenced at the University of Illinois at Urbana-Champaign Keck Sequencing Center and the University of Wisconsin Biotechnology Center. These genomes were sequenced in four separate batches. 87 were sequenced using version 1 Nextera library preparation kits. The 24 genomes handled at the University of Wisconsin Biotechnology Center were sequenced on the Illumina GAIIx platform. Sixty-four genomes, including one that had insufficient reads on the GAIIx, were sequenced on an Illumina HiSeq using v2 chemistry at the University of Illinois. For these 87 genomes, an initial set of genome assemblies was performed using Velvet version 1.0.15 [26], SOAPdenovo v1.04 [27], EULER-SR v1.1.2 [28]. Subsequently, faux reads created by breaking larger contigs from these assemblies into 1999 bp pieces and 400,000 reformatted Illumina paired-end reads were used as input for gsAssembler v2.5.3 [29]. The resulting contigs were then used as input, along with all Illumina reads for a final scaffolding step with SSPACE v1.1 [30]. For the remaining genomes, sequencing was performed on an Illumina HiSeq 2000 with version 3 chemistry and 24 samples pooled per lane. These genomes were assembled IDBA UD version 1.0.9 [31]. Gene prediction was performed with Prodigal version 2.5.0 [32]. The remaining 487 genomes were downloaded from NCBI. All available gene predictions were kept and not altered. NCBI taxonomy information was used in an attempt to include all *Actinomycetales* genomes in NCBI as of February 2013. A small number of these genomes were not carried all the way through the pipeline, primarily due to a lack of complete uniformity in locus tag and formatting conventions. All further processing was performed with Perl scripts interfaced with MySQL.

Bioinformatics

Gene clusters were identified as previously described [19]. Orthologs were defined using OrthoMCL to analyze an initial set of 231 genomes [33]. Genes included within gene cluster boundaries for additional genomes were recruited to this initial data set by defining orthologs based on best BLAST hits to the original 231 genome data set below an e-value cutoff of 1e-10.

Gene cluster families were created using three similarity measures. The first is the mean proportion of orthologs shared by two gene clusters, with each gene only being counted once to prevent inflated scores from repeated genes (e.g. polyketide synthases and nonribosomal peptide synthetases). The cutoff used for this score was 0.5 (50%). The second uses PROmer alignments of every possible pairwise combination of gene clusters that were scored as the mean proportion of each gene cluster that is part of an alignment [34]. The cutoff used for this score was 0.5 (50%). The third is based on similarity of one type of domain or full length gene for each biosynthetic class as follows: type I and II polyketide synthases, ketosynthase domain; nonribosomal peptide synthetases, adenylation domain; NISiderophores, IucA-IucC domains; lanthipeptides, lanthionine-cyclase containing genes; and microcins, dehydratase genes. The program uclust was used to group domains at a 70% similarity cutoff [35]. The cutoff used for this score was that half of the total number of key domains or genes present in a pair of clusters must be in the same 70% similarity groups. To be placed into a gene cluster family, all three of these cutoffs must be passed. These scores were then weighted, counting the domain similarity twice, and converted to a distance metric as $D = a + h + 2s$, where D is the distance metric, a is the alignment score, h is the proportion of shared homologous genes and s is the highest clustering threshold in uclust that groups at least half of the domains in a pair of clusters together. These distance scores were used as input for density-based clustering with DBSCAN [36] implemented in the R package fpc (http://cran.r-project.org/web/packages/fpc/) with the following parameters: eps, 0.3; and MinPts, 2. The clusters assignments produced using DBSCAN were used as input to color the gene cluster family networks in the program Cytoscape to aid manual auditing [37]. All GCFs were manually examined, with special attention paid to GCFs split into separate clusters by DBSCAN.

Lanthipeptides and TOMMs were subjected to additional scrutiny based on predicted precursor peptides. The lan-cyclase domain containing genes and the cyclodehydratase, previously known as the docking protein, genes were aligned with MUSCLE v3.7 [38] and used as input for FastTree version 2.1.5 SSE3, OpenMP [39] with option –gamma for rate optimization under the Gamma20 model. The resulting phylogenetic framework was used to examine precursor peptides, with special attention paid to likely secondary structure for lanthipeptides, which is tied to resulting function [40].

All GCFs were manually checked for similarity based on gene cluster diagrams highlighting orthologous genes. This visualization was performed with in-house Perl scripts interfaced with MySQL and output as HTML with JavaScript. The visualizations are provided in Database S1. The annotations presented of the website are based on the most frequently used annotation for each set of homologous genes, as found within the set of annotated genomes in Genbank. This generic method allows automated annotation based on existing data. We strongly recommend additional bioinformatics analyses for genes of interest.

Ribosomal proteins used to create the phylogenetic tree in Figure 2a were detected using pHMMs created with HMMER using the files developed in a previous study [41]. The proteins were aligned using ClustalW [42] and the tree was created with FastTreeMP [39] and visualized with the Python library ETE2 [43]. Genomic data was presented with Circos [44].

All sequence distance comparisons, for both amino acid and nucleotide sequences) were performed based solely on identity and represent the proportion of differences. These scale linearly and range from 0-1. The *rpoB* gene fragment used ranges from the *Streptomyces coelicolor* A3(2) gene SCO4654 coordinates 1451-1747.

Correlations were performed on subsets of each GCF to improve accuracy when looking for an exact mass. An entire GCF likely includes multiple unique structures that share a common core, and this diversity would not be reflected in our current analysis based on intact masses. The central domains or genes discussed above, e.g. ketosynthase domains or full length *pepM*, were clustered at a threshold of 0.9 using the program uclust[35]. Individual domains were given unique identifiers, such that some fragmented clusters that are closely related could be used for correlations that were left out of the larger GCF analysis. These subsets of each GCF were the correlated with the binary occurrence of each individual component ID in the output from SIEVE (see below). Scoring was applied in the following fashion: GCF present, component ID present = +10; GCF absent, component ID present = −10; GCF present, component ID absent = 0; GCF absent, component ID absent = +1.

Extrapolations were performed using the program estimateS [18]. Genomes were considered as sampling sites and occurrence of gene cluster families within the genomes as counts of individuals. Analyses were repeated for 100 runs. Classical measures were used for Chao1 and ACE estimates. Extrapolation was performed for a total of 750 knots out to 15000 samples.

Growth and Extraction

Seed cultures for production were grown in yeast-extract, malt-extract medium for 7 days at 30°C. 200 μl of this culture was used as inoculum for four solid media types (all ingredients given per liter, all pH should be adjusted to 7): (1) 12.5 g glycerol, 1.0 g arginine-HCl, 1.0 g NaCl, 1.0 g K_2HPO_4, 0.5 g $MgSO_4$-$7H_2O$, 0.01 g $Fe_2(SO_4)_3$ -$6H_2O$, 1 mg $FeSO_4$-$7H_2O$, 1 mg $MnCl_2$-$4H_2O$, 1 mg $ZnSO_4$-$7H_2O$, 15.0 g agar [45]; (2) 20.0 g mannitol, 20.0 g soya flour, 20.0 g agar; (3) 5 ml glycerol, 10.0 g sucrose, 5.0 g beef extract, 5.0 g casamino acids; (4)

ISP medium 4 [46] with 10 mM N-acetylglucosamine. Strains were grown on these solid media for 10 days. Plates were frozen and then manually extruded through filter paper to obtain a crude extract. Crude sample was acidified to a final concentration of 0.1% formic acid and then loaded onto an Oasis HLB extraction column (Waters, Milford, MA). Following equilibration, columns were washed with 2 mL of water, followed by 2 mL of 10% ACN. Metabolites were eluted in 1.2 ml of 80% ACN and the eluate was evaporated to dryness. All flow-through was performed by gravity-flow, and a vacuum manifold was used when gravity was not sufficient. Microcentrifuge tubes were weighed empty and after drying to obtain a sample weight.

LC-MS Analysis

Extracts were resuspended in 5% ACN with 0.2% formic acid to a final concentration of 2 μg/μL. For LC-MS analysis, 40 μg of sample was loaded onto a 150 mm × 2.1 mm i.d., 2 μm particle size Kinetex C18 RPLC column (Phenomenex, Torrance, CA). Analysis was performed using an Agilent 1150 LC system (Agilent, Santa Clara, CA) equipped with a photodiode array and placed in-line with a Q-Exactive mass spectrometer (Thermo Fisher Scientific, Waltham, MA). Chromatography was performed at a flow rate of 200 μL/min using water/0.1% formic acid (solvent A) and acetonitrile/0.1% formic acid (solvent B) with the following gradient: time 0 min., 2% B; 35 min., 60% B; 54 min., 98% B. UV spectra were acquired at a rate of 1 Hz. For every spectrum, the three most intense apices were recorded. The mass spectrometer instrument settings were as follows: capillary temperature 275 °C, sheath gas 8 (arbitrary units), spray voltage 4.2 kV. Full MS spectra were acquired at 35,000 resolution for the mass range *m/z* 250 to 3750 for all samples. This resulted in an average scan rate of 6 Hz. Following each full MS scan, the top 5 most intense ions were selected for a dependent MS^2 scan. MS^2 was conducted using HCD with a collisional energy of 25%.

Software, Informatics and Statistical Treatment of Data

SIEVE software was used for chromatographic alignment, component detection, removal of background, and relative quantification. Chromatographic alignment was performed for all samples using an initial tile size of 500 frames. Following alignment, feature detection was performed using an initial mass tolerance of 10 ppm and a retention time window of 3.0 minutes. Analysis parameters were optimized for low *m/z* (250-500), mid *m/z* (500-900), and high *m/z* (900-3,750). A minimum normalized intensity of 5×10^7, 1×10^7, or 5×10^6, was selected as the threshold for defining a peak as a feature in the low-, mid- and high-*m/z* regimes, respectively. Deisotoping and the summation of multiple adducts observed for a single species were performed to reduce data complexity. Following complexity reduction, a final list of components was output. For each component, a reconstructed ion chromatogram was created and the integrated intensity of the peak was calculated. An integrated intensity of 2×10^6, 1×10^6, or 7.5×10^5 was selected as the cut-off value for a component to be considered present in a given sample in the low-, mid- and high-*m/z* regimes, respectively.

Automated Dereplication

All component mass values for each of the components were searched against an accurate mass database consisting of known bacterial metabolites using a mass tolerance of only 3 parts-per-million (ppm). The database of 9,817 known natural products was prepared using Antibase (http://wwwuser.gwdg.de/~hlaatsc/antibase.htm), Dictionary of Natural Products, as well as additional bacterial natural products found in the literature. Manual analysis of fragmentation spectra was performed to confirm intact mass-based identifications. In addition, UV data was compared to literature values when available to offer additional and orthogonal confirmation of compound identity (Supplementary Table 3).

MS/MS Verification Procedure

All compounds putatively identified via accurate intact mass were confirmed using accurate mass, tandem MS (MS^2) data. To ensure that low-quality spectra were not included, MS^2 spectra containing less than 5 peaks at >1% relative abundance were excluded from analysis. Additionally, spectra containing more than 100 peaks at >1% abundance were included only if >20% of the peaks appeared in the higher *m/z* half of the spectrum. A large number of very low *m/z* fragment ions in the absence of high molecular weight fragments was found to be characteristic of over-fragmentation, and had the potential to lead to false identifications. Structures for putatively identified compounds were fragmented *in silico* using the software suite Mass Frontier (Thermo Fisher Scientific, Waltham, MA). Both general fragmentation rules and fragmentation library modes were used. The fragmentation library was composed of the HighChem ESI Positive 2008 library as well as annotated spectra from in-house compound libraries. General fragmentation rules were bypassed for library reactions. For general fragmentation rules, electron sharing and charge stabilization resonance reactions were allowed. Ionization, stabilization, and cleavage were allowed on aromatic systems. For compounds with a molecular mass less than 500 Da, 8 reaction steps were allowed with an upper limit of 10,000 reactions. For compounds with a molecular mass greater than 500 Da, 12 reaction steps were allowed with an upper limit of 20,000 reactions. Confirmation required that at least 4 of the top 5 most intense peaks were consistent with theoretical fragment ions to within 5 ppm. On average, 8.4 out of the top 10 most abundant peaks were consistent with theoretical fragment ions among all confirmed species in the dataset.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

JRD was funded through an Institute for Genomic Biology fellowship. This work was supported in part by the National Institutes of Health GM PO1 GM077596, GM 067725 (NLK) and an Institute for Genomic Biology Proof of Concept grant. DPL and the ARS Culture Collection CRIS project is funded through ARS National Program 301.

References

1. Bérdy J. Bioactive microbial metabolites. J. Antibiot. 2005; 58:1–26. [PubMed: 15813176]

2. Bérdy J. Thoughts and facts about antibiotics: Where we are now and where we are heading. J. Antibiot. 2012; 65:385–395. [PubMed: 22511224]

3. Bentley SD, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature. 2002; 417:141–147. [PubMed: 12000953]

4. Lautru S, Deeth RJ, Bailey LM, Challis GL. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat. Chem. Biol. 2005; 1:265–269. [PubMed: 16408055]

5. Kersten RD, et al. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. Nat. Chem. Biol. 2011; 7:794–802. [PubMed: 21983601]

6. Ziemert N, et al. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS ONE. 2012; 7:e34064. [PubMed: 22479523]

7. Medema MH, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011; 39:W339–W346. [PubMed: 21672958]

8. Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res. 2013; 41:D402–D407. [PubMed: 23104377]

9. Diminic J, et al. Databases of the thiotemplate modular systems (CSDB) and their in silico recombinants (r-CSDB). J. Ind. Microbiol. Biotechnol. 2013; 40:653–659. [PubMed: 23504028]

10. Yadav G, Gokhale RS, Mohanty D. SEARCHPKS: a program for detection and analysis of polyketide synthase domains. Nucleic Acids Res. 2003; 31:3654–3658. [PubMed: 12824387]

11. Tae H, Kong E-B, Park K. ASMPKS: an analysis system for modular polyketide synthases. BMC Bioinformatics. 2007; 8:327. [PubMed: 17764579]

12. Ichikawa N, et al. DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2013; 41:D408–D414. [PubMed: 23185043]

13. Caboche S, et al. NORINE: a database of nonribosomal peptides. Nucleic Acids Res. 2008; 36:D326–D331. [PubMed: 17913739]

14. Kim J, Yi G-S. PKMiner: a database for exploring type II polyketide synthases. BMC Microbiol. 2012; 12:169. [PubMed: 22871112]

15. Fischbach M, Walsh C. Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. Chem. Rev. 2006; 106:3468–3564. [PubMed: 16895337]

16. Raghupathy N, Durand D. Gene cluster statistics with gene families. Mol. Biol. Evol. 2009; 26:957–968. [PubMed: 19150803]

17. Wang X, et al. Identification and characterization of the actinomycin G gene cluster of *Streptomyces iakyrus*. Mol. Biosyst. 2013; 9:1286–1289. [PubMed: 23567908]

18. Colwell RK, et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J. Plant Ecol. 2012; 5:3–21.

19. Doroghazi JR, Metcalf WW. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. BMC Genomics. 2013; 14:611. [PubMed: 24020438]

20. Jensen PR, Williams PG, Oh DC, Zeigler L, Fenical W. Species-Specific Secondary Metabolite Production in Marine Actinomycetes of the Genus Salinispora. Appl. Environ. Microbiol. 2006; 73:1146–1152. [PubMed: 17158611]

21. Dunbar KL, Melby JO, Mitchell DA. YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. Nat. Chem. Biol. 2012; 8:569–575. [PubMed: 22522320]

22. Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Brady SF. Chemical-biogeographic survey of secondary metabolism in soil. Proc. Natl. Acad. Sci. U. S. A. 2014; 111:3757–3762. [PubMed: 24550451]

23. Bunge J, Willis A, Walsh F. Estimating the Number of Species in Microbial Diversity Studies. Annual Review of Statistics and Its Application. 2014; 1:427–445.

24. Nguyen DD, et al. MS/MS networking guided analysis of molecule and gene cluster families. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:E2611–E2620. [PubMed: 23798442]

25. Cote, R. ATCC Bacteria and Bacteriophages, 19th edn. Pienta, P.; Tang, J.; Cote, R., editors. American Type Culture Collection; 1996. p. 484

26. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]

27. Li R, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20:265–272. [PubMed: 20019144]

28. Chaisson MJ, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 2009; 19:336–346. [PubMed: 19056694]

29. Margulies M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]

30. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding preassembled contigs using SSPACE. Bioinformatics. 2011; 27:578–579. [PubMed: 21149342]

31. Peng Y, Leung HC, Yiu S-M, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012; 28:1420–1428. [PubMed: 22495754]

32. Hyatt D, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010; 11:119. [PubMed: 20211023]

33. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13:2178–2189. [PubMed: 12952885]

34. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]

35. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26:2460–2461. [PubMed: 20709691]

36. Ester, M.; Kriegel, H-P.; Sander, J.; Xu, X. KDD. Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama, editors. Vol. 96. 1996. p. 226-231.

37. Shannon P, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]

38. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

39. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010; 5:e9490. [PubMed: 20224823]

40. Zhang Q, Yu Y, Vélasquez JE, van der Donk WA. Evolution of lanthipeptide synthetases. Proc. Natl. Acad. Sci. U. S. A. 2012; 109:18361–18366. [PubMed: 23071302]

41. Yutin N, Puigbò P, Koonin EV, Wolf YI. Phylogenomics of prokaryotic ribosomal proteins. PLoS ONE. 2012; 7:e36972. [PubMed: 22615861]

42. Larkin MA, et al. Clustal W and clustal X version 2.0. Bioinformatics. 2007; 23:2947–2948. [PubMed: 17846036]

43. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010; 11:24. [PubMed: 20070885]

44. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19:1639–1645. [PubMed: 19541911]

45. El-Nakeeb MA, Lechevalier HA. Selective isolation of aerobic actinomycetes. Appl. Microbiol. 1963; 11:75–77. [PubMed: 13937509]

46. Smith SE, et al. Comparative Genomic and Phylogenetic Approaches to Characterize the Role of Genetic Recombination in Mycobacterial Evolution. PLoS ONE. 2012; 7:e50070. [PubMed: 23189179]

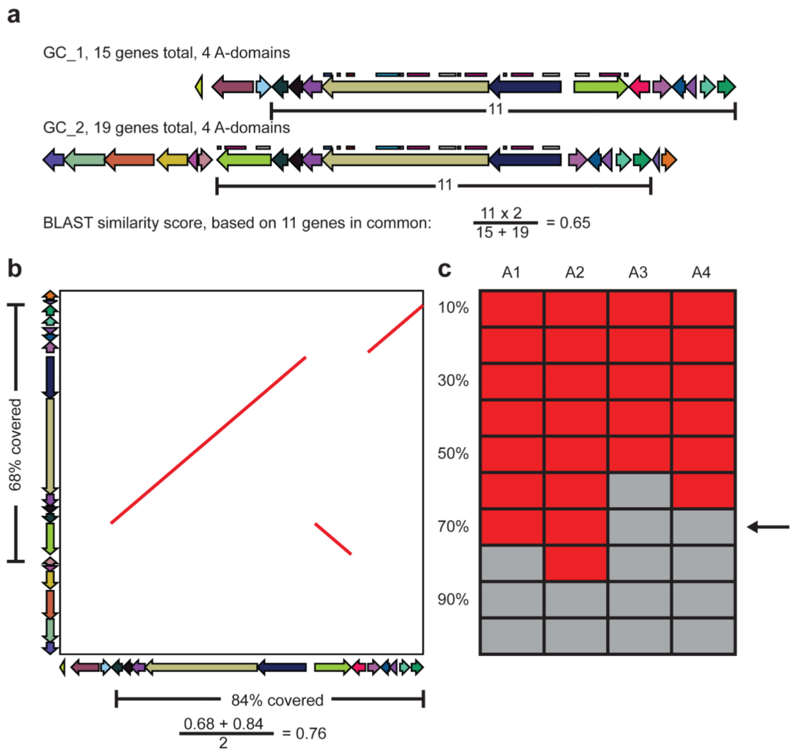
a
GC_1, 15 genes total, 4 A-domains
11
GC_2, 19 genes total, 4 A-domains
11
BLAST similarity score, based on 11 genes in common: $\frac{11 \times 2}{15 + 19} = 0.65$
b
68% covered
84% covered
$\frac{0.68 + 0.84}{2} = 0.76$
c
A1
A2
A3
A4
10%
30%
50%
70%
90%

Figure 1. Similarity metrics for NPGC comparisons

Three similarity metrics were created for comparison of NPGCs. (a) The number of orthologous genes shared by the two clusters divided by the total number of genes in both clusters. Each gene is scored only once. (b) The total amount of each cluster involved in a PROmer alignment. (c) For the core biosynthetic domains or genes described in the Materials and Methods, corresponding domains/genes from GC_1 are found in GC_2 based on whether they are clustered together with the program uclust at clustering thresholds that increase in steps of 10%. Red indicates that A1-A4 from GC_1 are clustered together with an adenylation domain from GC_2 at a given clustering threshold. Gray indicates that there is no corresponding adenylation domain from GC_2 at a given clustering threshold. The third score used is the highest clustering threshold in which half of the domains/genes in GC_1 have a corresponding domain/gene in GC_2. The arrow indicates the maximum score for GC_1 and GC_2 of 70%, or 0.7, where half of the GC_1 A-domains are present in GC_2.

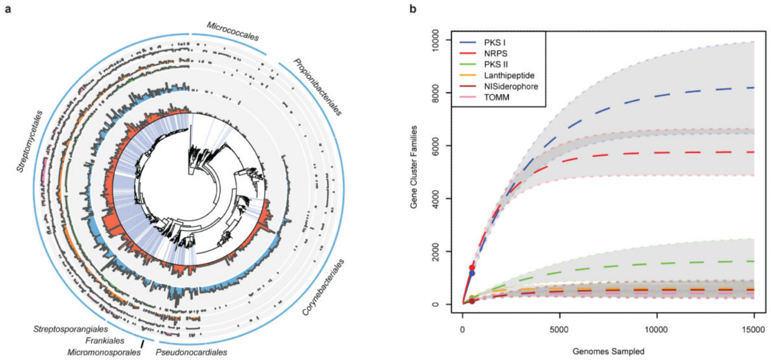

Figure 2. Genomic NPGC content and extrapolation

(a) A phylogenetic tree for the sampled organisms is shown surrounded by natural product gene cluster content of each genome. Blue shading indicates genomes sequenced for this project. Concentric rings, from the inside out, show counts of NRPS, type I PKS, type II PKS, NISiderophore, lanthipeptide, and TOMM gene clusters. The names of the most abundant taxonomic families are shown in the outer ring. (b) Extrapolation of the number of GCFs encoded by *Actinobacteria*, with 95% confidence intervals indicated as the grey area inside of dashed lines. Extrapolation was performed out to 15,000 genomes. Filled circles indicate the current extent of our sampling.

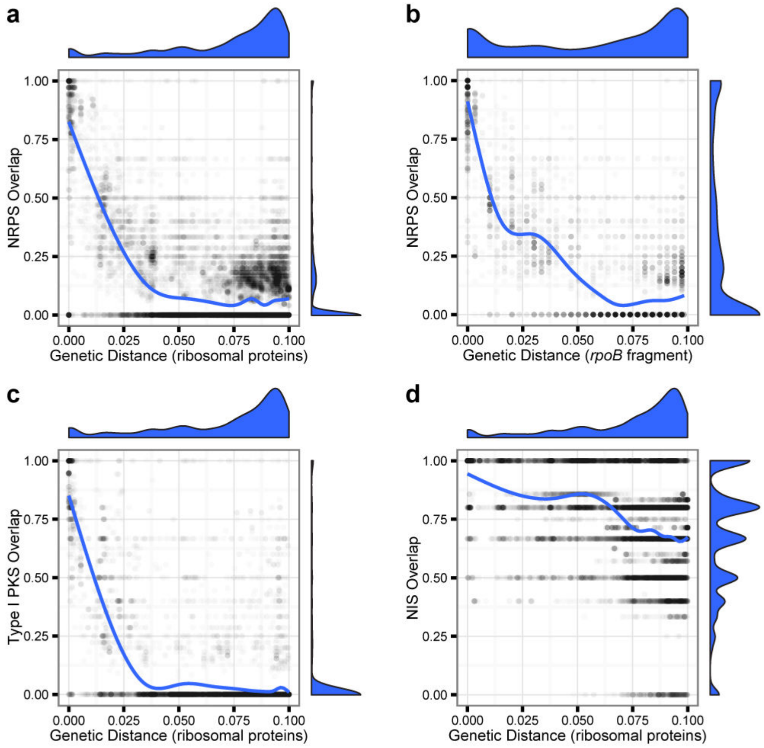

Figure 3. GCF conservation over genetic distance

For every pair of genomes highlighted that span the *Streptomycetales* to the *Pseudonocardiales* in Fig. 2a, the proportion of GCFs shared between them is plotted against their genetic distance. NRPS conservation plotted against (a) ribosomal protein distance, (b) *rpoB* gene fragment. Conservation of type I PKS clusters (c) and NRPS-independent siderophores (d) plotted against ribosomal protein distance. The density of points across both axes is shown beside all plots.

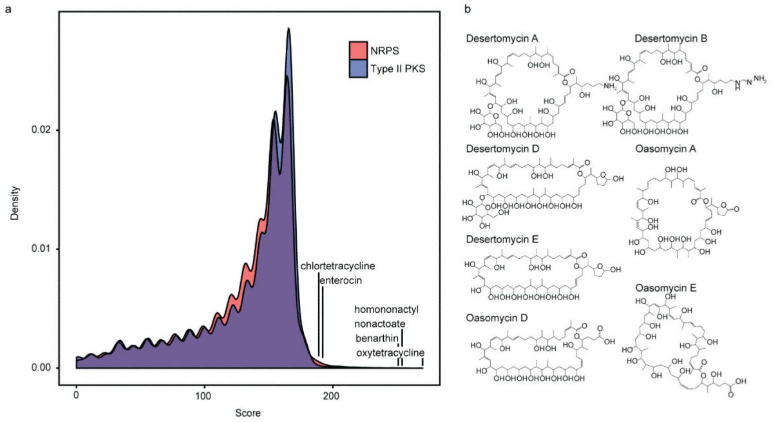

Figure 4. MS/GCF correlations

(a) The density distribution of the correlation scores for every GCF compound is shown for NRPS and type II PKS classes along with scores for selected known compound-gene cluster pairs. (b) Desertomycin and oasamycin compounds with the highest correlation scores (196) to a novel type I PKS gene cluster (PKS_I_18) are shown. Additional details are shown in Supplementary Fig. 6.

Table 1

NPGC abundance by taxonomic family

Taxon	PKS I	NRPS	PKS II	Lant	TOMMs	NIS	genomes	genome^{-1}
Streptomycetales	5.4	7.9	1.7	2.7	1.2	2.6	341	21.6 (3, 43)
Pseudonocardiales	7.5	7	1	2.6	1	0.6	40	19.8 (1, 44)
Streptosporangiales	2.9	5.8	0.5	3	1.6	1.1	14	15.0 (0, 26)
Micromonosporales	4.6	4	1.3	1.4	0.7	1.3	19	13.3 (2, 18)
Frankiales	4.9	1.3	1.6	1.5	0.4	0.5	11	10.2 (0, 17)
Corynebacteriales	4.1	3.8	0.2	0.1	0.1	0.2	238	8.4 (0, 31)
Micrococcales	0.1	0.3	0	0.2	0.1	0.3	67	1.1 (0, 5)
Propionibacteriales	0	0.2	0	0.2	0	0	81	0.4 (0, 8)

This table shows the total count of NPGCs of each biosynthetic class, the total number of genomes and the mean number of NPGCs in each genome, reported for the main taxonomic families shown. The range is shown in parentheses to the right of the per genome average.

Table 2

Gene clusters with MS characterized products

Natural Product	Mass (Da)	Error (ppm)	Corr. Score	Strains
proferrioxamine D2	600.3483	0.41	499[#]/268	47
oxytetracycline	460.1482	0.19	270/270	12
homononactyl nonactoate	400.2461	1.04	254/254	9
benarthin	411.1754	0.29	251/251	12
griseobactin	1179.495	1.27	162[*]/251	3
actinomycin D	1254.628	0.34	204[#]/177	3
rimocidin amide	766.4252	1.66	210/250	7
desertomycin A	1191.7492	0.76	196/196	2
pyridomycin	540.2220	1.60	195[#]/176	2
enterocin	444.1056	0.45	192/192	2
chlortetracycline	478.1143	0.08	189/199	2
kirromycin	796.4146	1.83	185/195	1

Natural products shown were found by searching the accurate mass for each SIEVE-identified compound against an *Actinobacteria* specific compound database. Hits were then verified manually using accurate mass MS^2 data and comparing previously characterized gene clusters against gene cluster family results. The correlation score is shown compared to the maximum automated score for all compounds correlated with the given gene cluster family subset containing the characterized gene cluster.

[*] indicates that the result was affected by manual verification of the mass spectrometry data,

[#] indicates that the result was affected by manual verification of the gene cluster family data set. Raw data for these compounds from all 178 strains are available in Supplementary Data Set 3.