Title

Expanding the genomic encyclopedia of Actinobacteria with 824 isolate reference genomes

Journal

Cell Genomics, 2(12)

ISSN

2666-979X

Authors

Seshadri, Rekha

Roux, Simon

Huber, Katharina J

et al.

Publication Date

2022-12-01

DOI

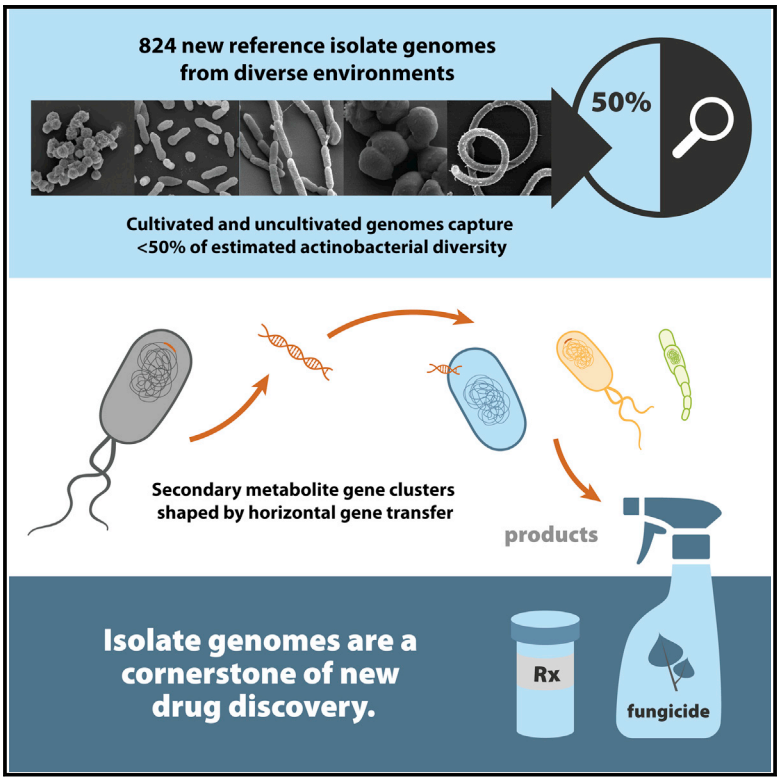
10.1016/j.xgen.2022.100213

Cell Genomics

Article

Expanding the genomic encyclopedia of *Actinobacteria* with 824 isolate reference genomes

Graphical abstract



Highlights

- 824 new actinobacterial isolate genomes from diverse environments
- Only a third of actinobacterial diversity has genome representation
- New niche-specific gene determinants highlighted, such as new antimicrobial peptides
- Secondary metabolite gene clusters shaped by horizontal gene transfer

Authors

Rekha Seshadri, Simon Roux, Katharina J. Huber, ..., Markus Göker, Nikos C. Kyrpides, Natalia N. Ivanova

Correspondence

rseshadri@lbl.gov (R.S.), markus.goeker@dsmz.de (M.G.), nnivanova@lbl.gov (N.N.I.)

In brief

Seshadri et al. contribute 824 new genomes of cultivated *Actinobacteria*, which are important for drug discovery. They observe that the genes responsible for producing such compounds often move around between microbes, making them harder to capture without high-quality genomes. They highlight interesting adaptations such as an experimentally verified antimicrobial peptide.

Seshadri et al., 2022, Cell Genomics *2*, 100213
December 14, 2022
https://doi.org/10.1016/j.xgen.2022.100213

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Expanding the genomic encyclopedia of *Actinobacteria* with 824 isolate reference genomes

Rekha Seshadri,[1,13,*] Simon Roux,[1] Katharina J. Huber,[2] Dongying Wu,[1] Sora Yu,[3] Dan Udwary,[1,3] Lee Call,[1] Stephen Nayfach,[1] Richard L. Hahnke,[2] Rüdiger Pukall,[2] James R. White,[4] Neha J. Varghese,[1] Cody Webb,[1] Krishnaveni Palaniappan,[1] Lorenz C. Reimer,[2] Joaquim Sardà,[2] Jonathon Bertsch,[1] Supratim Mukherjee,[1] T.B.K. Reddy,[1] Patrick P. Hajek,[1] Marcel Huntemann,[1] I-Min A. Chen,[1] Alex Spunde,[1] Alicia Clum,[1] Nicole Shapiro,[1] Zong-Yen Wu,[3] Zhiying Zhao,[1] Yuguang Zhou,[5] Lyudmila Evtushenko,[6] Sofie Thijs,[7] Vincent Stevens,[7] Emiley A. Eloe-Fadrosh,[1,3] Nigel J. Mouncey,[1,3] Yasuo Yoshikuni,[1,3,10,11,12] William B. Whitman,[8] Hans-Peter Klenk,[9] Tanja Woyke,[1,3] Markus Göker,[2,*] Nikos C. Kyrpides,[1,3] and Natalia N. Ivanova[1,3,*]

[1]US Department of Energy Joint Genome Institute, Berkeley, CA, USA
[2]Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany
[3]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[4]Resphera Biosciences, Baltimore, MD, USA
[5]China General Microbiological Culture Collection Center, Beijing, China
[6]Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences, All-Russian Collection of Microorganisms (VKM), Pushchino, Russia
[7]Center for Environmental Sciences, Environmental Biology, Hasselt University, Diepenbeek, Belgium
[8]Department of Microbiology, University of Georgia, Athens, GA, USA
[9]School of Biology, Newcastle University, Newcastle upon Tyne, UK
[10]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[11]Center for Advanced Bioenergy and Bioproducts Innovation, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[12]Global Institution for Collaborative Research and Education, Hokkaido University, Hokkaido 060-8589, Japan
[13]Lead contact
*Correspondence: rseshadri@lbl.gov (R.S.), markus.goeker@dsmz.de (M.G.), nnivanova@lbl.gov (N.N.I.)
https://doi.org/10.1016/j.xgen.2022.100213

SUMMARY

The phylum *Actinobacteria* includes important human pathogens like *Mycobacterium tuberculosis* and *Corynebacterium diphtheriae* and renowned producers of secondary metabolites of commercial interest, yet only a small part of its diversity is represented by sequenced genomes. Here, we present 824 actinobacterial isolate genomes in the context of a phylum-wide analysis of 6,700 genomes including public isolates and metagenome-assembled genomes (MAGs). We estimate that only 30%–50% of projected actinobacterial phylogenetic diversity possesses genomic representation via isolates and MAGs. A comparison of gene functions reveals novel determinants of host-microbe interaction as well as environment-specific adaptations such as potential antimicrobial peptides. We identify plasmids and prophages across isolates and uncover extensive prophage diversity structured mainly by host taxonomy. Analysis of >80,000 biosynthetic gene clusters reveals that horizontal gene transfer and gene loss shape secondary metabolite repertoire across taxa. Our observations illustrate the essential role of and need for high-quality isolate genome sequences.

INTRODUCTION

Actinobacteria is a large and diverse phylum comprising Gram-positive bacteria with high guanine-plus-cytosine (G + C) genome content and genome sizes ranging from <0.5 to 15.0 Mbp. Members of this phylum exhibit varying morphological and physiological features, including multicellularity and complex differentiation and are widely (and abundantly) distributed in diverse ecosystems.[1,2] Famous *Actinobacteria* include the causative agents of tuberculosis and diphtheria, some of the most devastating diseases in human history.[3] Others play key ecological roles in carbon cycles of soil and aquatic environments or are widespread as mutualistic symbionts of plants and animals, synthesizing natural products for host benefit or helping herbivores digest plant biomass. As renowned producers of diverse secondary metabolites including over two-thirds of all antibiotics in current clinical use and other compounds of clinical or agricultural importance, they are the subject of numerous natural product discovery efforts.[1,4–7]

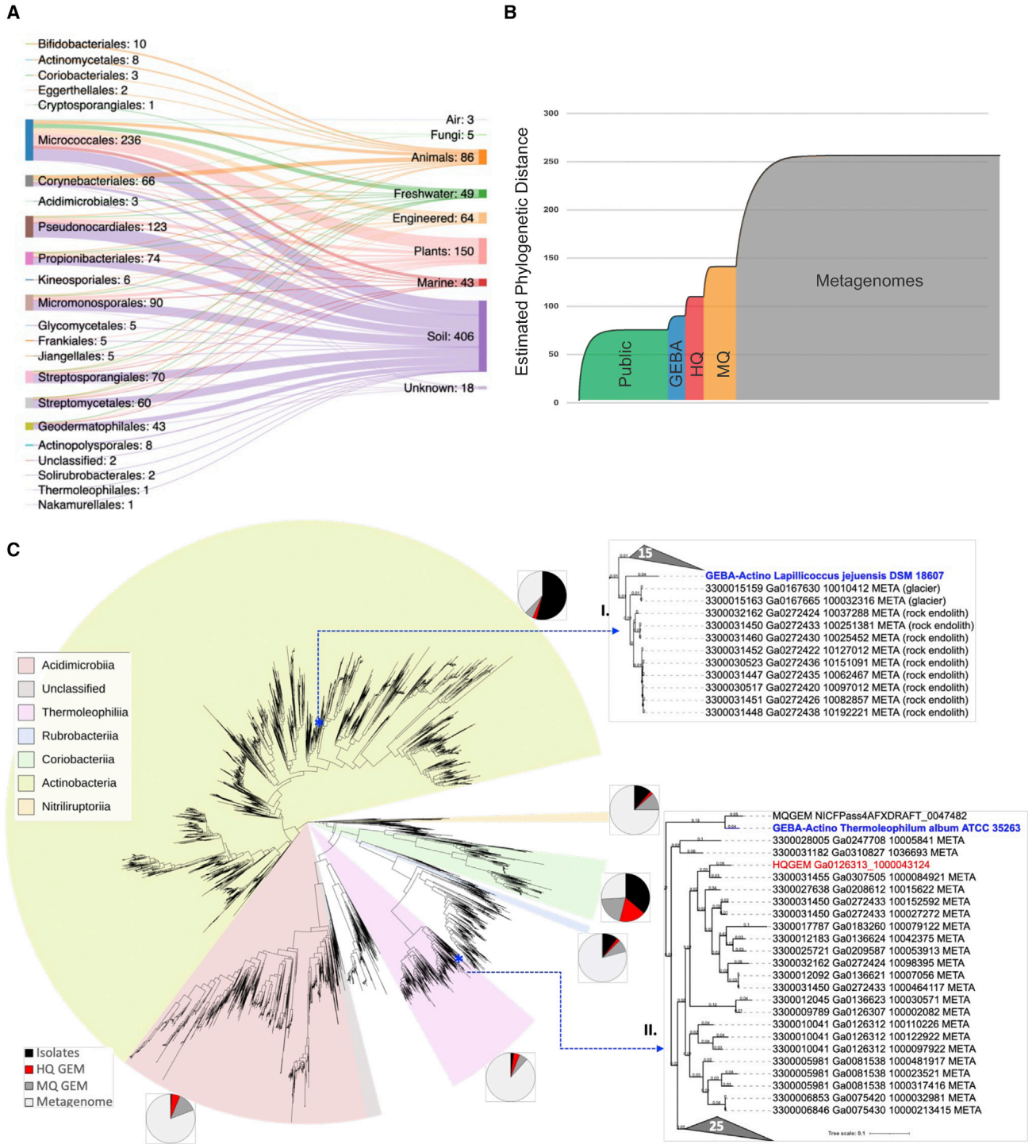


Figure 1. Phylogenetic diversity (PD) of phylum *Actinobacteria*

(A) A total of 824 isolate genomes were sequenced from diverse taxa and habitats. Snapshot of taxonomic (order level) composition and isolation source of the 824 GEBA-Actino genomes is presented. Number of genomes attributed to each taxon or isolation source is shown next to each label.

(B) PD accumulation curve depicting incremental increase in PD inferred from computed branch lengths of RpoB tree. The units on the x axis represent individual taxa or their equivalents (arising from metagenomes) ordered by genome category as the "accumulation units": isolates (Public in green and GEBA in blue), MAGs (HQ in red and MQ in orange), and metagenomic sequences in gray. PD score based on summed branch lengths is shown on the y axis.

(C) RpoB gene-based maximum likelihood phylogenetic tree used for PD calculation. The tree was rooted based on a representative set of archaeal RpoB sequences. For visualization purposes, clades with zero branch lengths were collapsed, and a single clade representative was retained. Individual actinobacterial

(legend continued on next page)

Despite their significance, *Actinobacteria* represent <10% of the 200,000+ publicly available genomes to date, and even these belong primarily to organisms relevant to human and veterinary medicine.[8] As of January 2020 (analysis start date), 18,411 actinobacterial isolate genomes were available in public databases, although a considerable proportion belonged to multiple strains of human pathogens like *Mycobacterium tuberculosis* and *Mycobacteroides abscessus*.

In this study, we report the genomes of 824 actinobacterial isolates sequenced under the auspices of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) initiative,[9] mostly of type strains from the Leibniz Institute DSMZ culture collection sourced from diverse habitats. Type strains are permanently attached to the names of species and subspecies as regulated by the International Code of Nomenclature of Prokaryotes (ICNP),[10] are well characterized with regard to phenotype, isolation sources, and other criteria, and have been made available to the worldwide scientific community via at least two different culture collections. A saturated collection of reference genomes of such isolates with pre-existing biochemical and genetic characterization (e.g., BacDive[11]) serves as a solid foundation for an array of experiments, including the development of microbial model systems and analyses of biotechnologically relevant pathways. Also, new opportunity for comparisons with non-pathogenic relatives could yield new insights and gene targets, expanding our understanding of important actinobacterial pathogens.

Here, we undertook a phylum-wide comparative analysis combining the 824 newly sequenced genomes with 5,922 non-redundant public actinobacterial genomes to explore (1) the overall phylogenetic diversity and cultivation status of the phylum, (2) niche-specific functional adaptations of different representatives, and (3) a compendium of natural product-encoding biosynthetic gene clusters (BGCs) and the drivers of that diversity. The data and comprehensive analyses generated herein are of broad utility in the fields of biological, biomedical, agricultural, and environmental sciences.

RESULTS AND DISCUSSION

Description of study datasets

A total of 824 high-quality draft genomes of isolates of the phylum *Actinobacteria*[12] were sequenced, assembled, and annotated (>99.33% average [avg.] completeness, <1.36% avg. contamination, 1.88 Mbp avg. scaffold N50; see STAR Methods and Table S1). We chose to retain the phylum name *Actinobacteria* due to its familiarity to a broad readership but revised phylum names include *Actinobacteriota* and *Actinomycetota,* with this latter name being recently validly published.[13] These genomes (hereafter referred to as "GEBA-Actino") were processed using the IMG annotation pipeline,[14] resulting in 4,569,551 predicted coding sequences from over 4.9 Gbp assembled sequence data (see Table S1 for complete list with metadata).

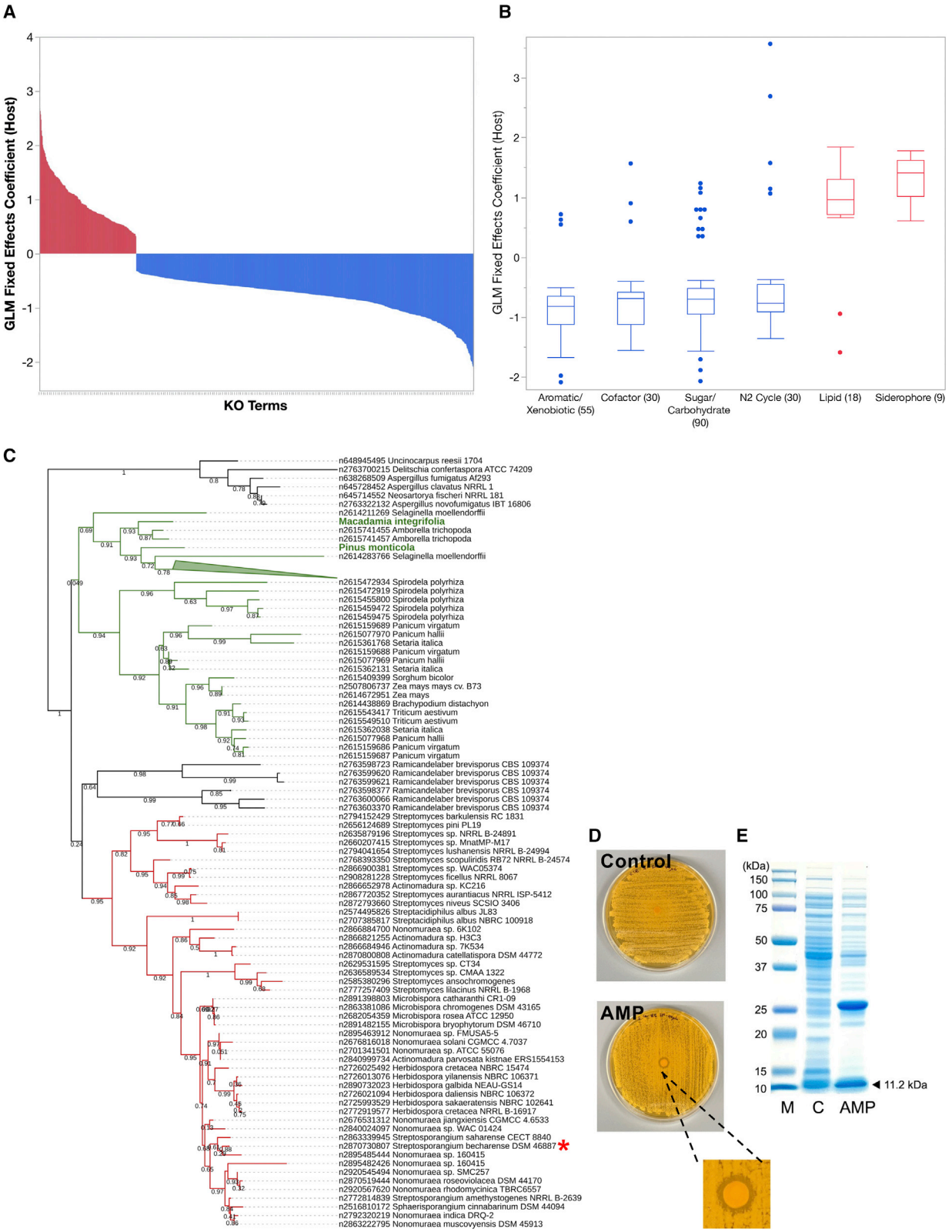
The investigated GEBA-Actino genomes represent 230 genera (54 families, 24 orders) from 4 classes: *Actinobacteria*, *Coriobacteriia*, *Acidimicrobiia*, and *Thermoleophilia*. Compared with other classes, which may be somewhat niche restricted, the class *Actinobacteria* is the largest and most diverse. The dataset includes the first sequenced representatives of 81 genera, expanding diversity in three unrepresented families (*Thermoleophilaceae*, *Rarobacteraceae*, *Motilibacteraceae*) as well as unclassified ones. *Thermoleophilum album* is the first sequenced isolate of the order *Thermoleophilales*. The overall taxonomic composition and isolation sources of the GEBA-Actino genomes are shown in Figure 1A and Table S1. GEBA-Actino type strains originate mainly from terrestrial and plant-associated habitats (Figure 1A), including some from extreme or unusual environments (e.g., alkaline, arid, permafrost, hypersaline, deep marine sediment) and non-human hosts such as sponges, fungi, and insects. These non-model microbes from environments posing unique metabolic challenges are of particular interest for the discovery of novel secondary metabolite prospects such as those with low toxicity to animals[1,15–17] and also enable inquiry into habitat-specific adaptations through comparative genomics.

For comparative analysis purposes, a dereplicated set of 4,824 publicly available isolate genomes (referred to as "Public") and 1,098 metagenome-assembled genomes (MAGs) from the comprehensive genomic catalog of Earth's microbiomes[19] was included (see STAR Methods and additional worksheets in Table S1). MAGs contributed significantly to the diversity of taxa, especially for classes underrepresented by isolates (*Coriobacteriia*, *Acidimicrobiia*, *Thermoleophilia*) (Figure S1). Notably, MAGs have 2.7 Mbp avg. genome size compared with 5.48 Mbp for isolates (Figure S2A). While this may be a potential bias due to lower completeness of MAGs or the difficulty of assembling larger genomes from metagenomics data, it may also reflect biases in phylogenetic and sample habitat composition and speak to reasons for their relative un-cultivability. MAGs also tend to be more fragmented with avg. scaffold length N50 of 131 Kbp (for MAGs) compared with over 1.88 Mbp avg. for GEBA-Actino (or 1.4 Mbp for all isolates). These differences are highlighted here since they impact downstream choices for analytical methods as well as results and biological inferences (Figure S2). More importantly, they emphasize the unique value of isolate genome sequences, particularly in the case of large and complex genomes of *Actinobacteria*.

Status of the "uncultivated iceberg" for *Actinobacteria*

The "great plate count anomaly experiment"[20] revealed that the vast majority (>99%) of microbial lineages were uncultivated and, consequently, unstudied. This concept is frequently illustrated by the disproportionately larger mass of submerged ice in the metaphorical iceberg. Given the multitude of recently sequenced genomes from both cultivated and uncultivated sources (due to innovations in metagenome assembly and binning methodologies), we revisited this precept as it pertains

classes are colored as indicated using the iToL interface.[18] Uncolored sectors indicate operational taxonomic units (OTUs) composed entirely of uncultivated (metagenome and MAG) signatures. Pie charts indicate the proportion of isolate versus uncultivated sequences contributing leaves to each designated class. Inset trees show clades within class *Actinobacteria* (inset I) or class *Thermoleophilia* (inset II), highlighting GEBA type strains that could inform cultivation of members of adjoining uncultivated clades.



(legend on next page)

to members of the phylum *Actinobacteria*. We estimated the phylogenetic diversity (PD) of actinobacterial taxa, a simple and effective measure of biodiversity based on summing the branch lengths connecting those taxa on a phylogenetic tree.[21,22] The maximum likelihood tree was generated based on universal single-copy marker genes identified from 5,648 isolate genomes (GEBA-Actino and Public), 3,321 MAGs (high quality [HQ] plus medium quality [MQ]), and over 20,000 metagenomes from diverse environmental samples (see STAR Methods). This analysis revealed that *Actinobacteria* isolate genomes account for only 34.68% of the total estimated diversity of the phylum (Figure 1B). While the contribution of HQ MAGs is relatively minor, including MQ MAGs boosts the coverage to 54.72% of total PD. This leaves close to 50% of actinobacterial diversity without any genome representation, highlighting the difficulty of genome recovery from metagenomics datasets. At the class level, isolates account for 60.25% of total PD of class *Actinobacteria* (Figure S3A), the largest and most diverse class within the phylum, and to which most isolates belong (Figure S1). There is a negligible boost from HQ MAGs, again pointing to possible difficulties in recovering such MAGs for large and complex actinobacterial genomes. For class *Coriobacteriia*, >45.31% is captured by isolates, while HQ MAGs boost coverage to well over 83.55% (Figure S3B) of this primarily host-associated taxonomic group with smaller genomes (Figure S2C).

Several clades of *Actinobacteria* were almost exclusively represented by metagenomic signatures or MAGs (Figure 1C). An examination of a sample source of these enigmatic clades reveals that new diversity arises from aquatic and terrestrial environments and notably, extreme, or nutrient-limited environments like sulfur acidic soils, peat permafrost, rocks, polar desert, and uranium-contaminated soils (Table S2). These clades include divergent members of classes with few to no isolate representatives (e.g., *Acidimicrobiia*, *Thermoleophilia*, *Rubrobacteriia*), as well as potentially new unclassified taxonomic groups (Figure 1C). Targeting extreme or nutrient-limited environments using standard or high-throughput cultivation strategies may result in the capture of these unrepresented lineages.[23] Where related, GEBA type strains can help guide cultivation of specific uncultivated subclades (Figure 1C, insets) since their phenotypic, growth, and other requirements are well documented within curated databases like BacDive.[11] For example, *Lapillicoccus jejuensis* DSM 18607, a well-characterized stone isolate,[24] may serve as an appropriate reference for an uncultivated clade of rock-dwelling endoliths within the family *Intrasporangiaceae* (Figure 1C, inset I).

Adaptations to the host or other environment

We compared genomes of host-associated (2,650 genomes including 678 MAGs) versus environmental (2,306 including 284 MAGs) organisms to identify novel pathways or factors that may be attributed to adaptation to different lifestyles. Using a phylogeny-normalized generalized linear model approach, we identified protein families (Pfams, or KEGG Orthology [KO] terms) that were overrepresented in host-associated or environmental groups (Table S3). For example, out of 6,546 KO terms captured by 4,956 genomes, 1,100 were significantly (false discovery rate [FDR]-adjusted $p < 0.005$) overrepresented in either group (Figure 2A). Environmental genomes were notably enriched in functions related to the degradation of various aromatic or xenobiotic compounds, uptake and utilization of sugars, and carbohydrate-active enzymes (CAZYmes) for degradation of plant lignocellulose (e.g., cellulose, hemicellulose, pectin) (Figure 2B). These results could largely be attributed to many soil-dwelling terrestrial isolates in this group (Figure S1B). Similar observations were made using Pfams (Table S3). Other overrepresented functions include nitrogen cycling, cofactor biosynthesis, various transporters and regulators, and, interestingly, known determinants of plant growth promotion like pyrroloquinoline (PQQ) synthesis,[25] 1-aminocyclopropane-1-carboxylate deaminase (ACCase),[26] and phytase.[27]

Conversely, about 238 KO terms were overrepresented in the host-associated group—this relatively smaller number of enriched KO terms may reflect the smaller genome sizes (and consequently smaller functional repertoire) of host-associated genomes (Figure S2E). Among the enriched functions were known determinants of pathogenesis or host interaction like adhesins, siderophores, lactocepin, lysozyme inhibitor, and steroid degradation enzymes.[28–30] Additionally, we found several potential markers of adaptation to anaerobic conditions, including the FeoABC system for ferrous iron uptake, anaerobic ribonucleoside-triphosphate reductase, and C4-dicarboxylate membrane transporter.[31] More than 15 KO terms for lipid metabolism are noteworthy (Figure 2B) and may play a role in host-derived fatty acid utilization—e.g., fatty acid coenzyme A (CoA) ligases (K12421, K12422, K12423, K12427, K12428, K01909), acyltransferases, acyl-coA synthetase, and others[32,33] (Table S3).

Other significantly over- or underrepresented functions are potentially less well understood or characterized in bacteria—for

Figure 2. Functional adaptations of host versus environmental *Actinobacteria*

(A) Significantly over- or underrepresented functions (KO terms, FDR-adjusted [adj.] $p < 0.005$) in host-associated versus other environmental genomes are shown. The x axis shows individual KO terms, while the y axis shows the logistic regression coefficient from a fixed-effect generalized linear model. Positive values (in red) indicate overrepresentation in host-associated genomes, while negative values (in blue) indicate overrepresentation in environmental group genomes.

(B) Distribution of logistic regression coefficients (y axis) for individual KO function categories (x axis, discussed in the main text) is shown. Number of individual KO terms within each function category is shown in parentheses. Blue boxplots denote categories that are overrepresented in the environmental group, while red boxes denote categories in the host-associated group.

(C) Maximum likelihood tree of eukaryal and bacterial candidate sequences assigned to PF09117. Characterized plant reference sequences are highlighted with green text. Bacterial branches are colored red, plant branches are green, and fungal branches are black.

(D) Inhibition of *Saccharomyces cerevisiae* by AMP candidate of *Streptosporangium becharense* DSM 46887 overexpressed in *E. coli*.

(E) SDS-PAGE gel showing the overexpression of recombinant AMP in *E. coli*. Lanes are protein size marker (M), control strain (C), and AMP-producing strain (AMP), respectively. The expected 11.2 kDa band of the AMP is highlighted.

example, Pfams with limited phylogenetic distribution (LPD) or potential eukaryal origin within the host-overrepresented set are demarcated based on proportions of sequences recruited to individual Pfams from the 100,000+ isolate genomes of bacteria, archaea, or eukarya stored in the IMG database. For example, an arthropod defensin (PF01097, 91% eukaryal candidate sequences) from insects and scorpions with activity against Gram-positive bacterial pathogens may be similarly employed by members of *Actinomyces* spp.[34] The roles of other eukaryal-like Pfams may be more cryptic, like PF01490 (amino acid transporter, 94% eukaryal) found in *Corynebacterium* spp. and *Kocuria* spp., or PF05241 (expanded emopamil binding protein superfamily including characterized sterol isomerases, 84% eukaryal), which is restricted to several species of host-associated *Mycobacterium* spp., *Mycolicibacterium* spp., *Microbacterium* spp., and *Nocardia* spp., and are membrane bound (6 transmembrane regions on average). (Figure S4). A eukaryal phospholipase B (PF04916, 45% eukaryal) has remote homologs in *Bifidobacterium* spp., *Mycobacterium* spp., and *Adlercreutzia* spp.; horizontal gene transfer among members residing in a shared niche is conjectured (e.g., between *Bifidobacterium* sp. and *Lactobacillus* sp.) (Figure S5).

A potential novel antimicrobial peptide or AMP (PF09117, 96% eukaryal) is detected only in a small subset of soil- and plant-associated *Actinobacteria* outside of plant and fungal genomes (Figure 2C). We demonstrate inhibition of *Saccharomyces cerevisiae* by an AMP candidate from *Streptosporangium becharense* DSM 46887 cloned into *E. coli* (see STAR Methods; Figure 2D). A potential dimeric form of the AMP is suggested by the presence of a ~25 kDa band in addition to the expected 11.2 kDa product on an SDS-PAGE gel (Figure 2E). AMP dimerism has been previously reported.[35,36] The sequence lengths of 59 candidate actinobacterial AMPs varied from 101 to 121 amino acids with a median length of 102 residues. An N-terminal signal peptide was detected in every instance. A survey of gene neighborhoods revealed no conserved colocalized functions. AMPs are a promising new class of therapeutic antibiotics displaying broad-spectrum antimicrobial efficacy against bacteria, fungi, and viruses.[37–39]

LPD Pfams showing a discordant phylogenetic distribution within a narrow subset of bacterial lineages are also intriguing—e.g., DUF4300 (PF14133) was detected in known pathogenic or host-associated lineages within *Actinobacteria* and a few other bacteria phyla (Figure S4). This and other examples are described in Data S1. Many other comparisons are possible depending on the availability of underlying metadata, highlighting interesting targets for experimental investigation. For example, notable differences arising from genome comparisons of plant (195) versus animal (214) host isolates of the order *Micrococcales* include the uptake and utilization of known plant sugars like rhamnose or xylose and the utilization of GABA (a plant signal), ACCase (a well-recognized plant-growth-promoting factor), flagellar components, urate catabolism, etc. Similarly, for animal-associated isolates, enrichment of known virulence determinants like autotransporters and adhesins were found along with markers of anaerobiosis, antibiotic resistance, toxin/antitoxin systems, CRISPR-Cas systems, and many LPD families (Table S4).

Shaping of the secondary metabolite repertoire

Actinobacteria have been the focus of natural product or secondary metabolites (SMs) discovery for decades, and large-scale genomics has illuminated thousands of BGCs with the potential for new therapeutic and antimicrobial applications.[4,40–43] Beyond defense and competition, SMs can mediate diverse biotic interactions (including cooperative ones) like communication, nutrient acquisition, metal scavenging, stress protection, phage induction, and more, all of which can influence microbial fitness with impacts on microbial ecology and evolution. Here, we analyzed BGCs for SM production across all 5,648 isolate genomes using AntiSMASH 6.[44] A total of 80,947 BGCs were predicted from 5,194 genomes (out of 5,648) (Table S5; Data S2). These were assigned to 44,923 distinct gene cluster families (GCFs) using BiG-SLICE (Table S6), of which 32,570 were singletons, while the largest-sized GCFs with >100 BGCs included non-ribosomal peptide synthases (NRPSs) (1,040 BGCs), siderophores (523), RiPP-like (297), ectoine (259), terpene (193), etc. (Figure S6). The taxonomic composition of most of these GCFs was broad with a few exceptions like a siderophore (GCF ID 249228), RiPP-like (ID 249163), and terpene (ID 252912), restricted primarily to various *Streptomyces* spp., an ectoine (ID 251253) restricted to *Rhodococcus* spp., or a 98-member terpene GCF (ID 251612) from *Micromonospora* spp. A total of 6,939 GCFs were contributed exclusively by 744 GEBA-Actino genomes from the current study, 822 of which arise from 94 new genera. These results agree with the recent survey of BGCs by Gavriilidou et al. that highlight *Actinobacteria* (particularly *Streptomyces*, *Amycolatopsis*, *Kutzneria*, and *Micromonospora*) as top contributors of GCF diversity across all bacterial phyla.[43]

Overall, NRPS, terpenes, and type I polyketide synthase (T1PKS) were the most abundant SM classes, with terpenes (and, to a lesser extent, T3PKS, RiPP-like, and betalactones) widely distributed across genera. Other classes of SMs showed highly sporadic or phylogenetically incongruent distribution, alluding to widespread horizontal gene transfer of SMs, which is explored further below. Only 2,609 (3.2%) of the total BGCs had a significant ($\geq$80% identity over $\geq$80% of the reference sequence) hit to the manually curated MIBiG BGCs of known function.[45] At $\geq$90% identity, a mere 1,155 (1.4%) had hits, a low value similar to those reported in other studies,[19] since the vast majority of BGC products have not been chemically characterized or otherwise experimentally validated.

As expected, there was a positive trend between genome size and the number of BGCs[46,47] with an avg. of 15.58 BGCs detected per genome, comprising 8.05% of total genome length, hereafter referred to as %BGCs (Figures 3A and 3B). Host-associated genome sizes were smaller on avg. than environmental genomes (Figure 1B) and encoded fewer BGCs, which comprised, on avg., 7.15% BGCs compared with 9.09% BGCs for environmental genomes (Figure S7). *Kitasatospora kifunensis* DSM 41654, a soil isolate, displayed the top BGC commitment with 26.50% BGCs (Table S5). Other genomes with notable BGC commitment included several *Streptomyces* spp., *Nocardia* spp., and newly sequenced genera from the GEBA-Actino set (e.g., *Goodfellowiella coeruleoviolacea* DSM 43935, *Actinocrispum wychmicini* DSM 45934, *Labedaea rhizosphaerae*

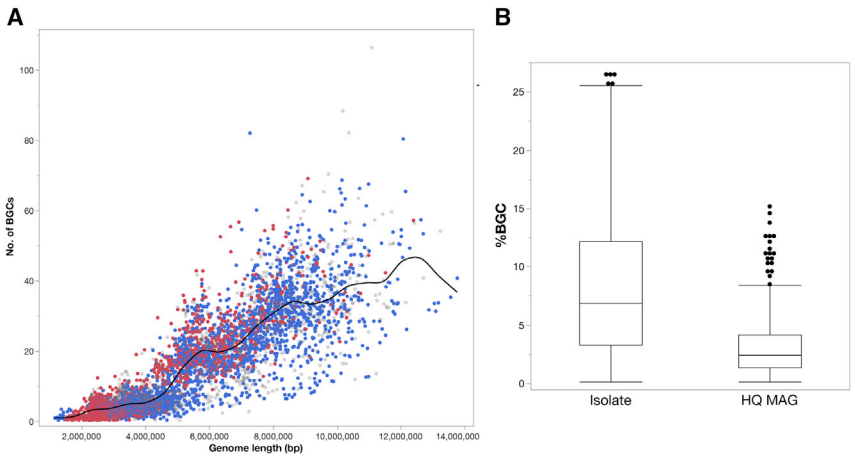


Figure 3. Overview of BGC abundances across actinobacterial genomes

(A) Relationship between genome size and total number of predicted BGCs per genome. Data points are colored based on isolation source (where available). The x axis is the genome size (in Mbp), and the y axis is the total number of BGCs.

(B) Distribution of percentage of BGCs (total BGC length as percentage of total genome length) for isolate genomes (including GEBA and public) compared with HQ MAGs.

DSM 45361). BGC commitment is summarized by various taxonomic levels in Figures S8A–S8C.

No BGCs could be predicted in 454 isolate genomes using AntiSMASH or an alternate machine-learning-based method, DeepBGC.[48] These were almost entirely small host-associated genomes (2.2 Mbp median length; Figure S9). The few exceptions included genomes of terrestrial *Nocardioides* spp. with genome size up to 5 Mbp (avg. > 99% completeness)—other *Nocardioides* spp. (from diverse environments) showed very low BGC commitment (avg. 2.63% BGCs). An inspection of individual genera that contained species both with and without BGCs demonstrated consistent patterns of BGC presence or absence in individual subclades—for example, the relative loss of the solitary type III polyketide synthase cluster in the last common ancestor of a subclade of *Gardnerella vaginalis* strains (Figure S10). A discontinuous distribution of individual SM classes in *Bifidobacterium* species again suggests relative gains and losses (Figure S11A)—for example, an interrupted pattern of lanthipeptide in *B. pseudocatenulatum* DSM 20438 and DC2A can be attributed to inactivation by truncation or point mutation of a lanthipeptide "hook" protein (Figure S11B), while no marker genes are detectable in strain L15. A phenazine-like BGC (containing PhzA/B but no other genes associated with a canonical phenazine operon)[49] detected in all strains of a discrete *B. thermophilum* clade, but with few instances elsewhere in the genus, suggest potential acquisition by a last common ancestor of this cohort (Figure S11A).

This pattern of sporadic distribution of SM type is the rule, rather than the exception, and is observed within every genus and most species, echoing individual reports.[50–52] The hypothesis follows that horizontal gene transfer (HGT) drives expansion of SM repertoires due to variable evolutionary pressures, even for narrow sublineages. To address this, we cross-referenced BGC-containing scaffolds with a list of scaffolds designated as putative plasmids by at least two independent prediction methods. A total of 936 plasmids bearing one or more BGC (1,119 total) were identified from 659 genomes belonging to 74 genera (11,999 plasmid scaffolds from 2,920 genomes from 240 genera were predicted with or without a BGC, and these are presented in Table S7). The length of BGC-encoding plasmids ranged from 2,535 (partial plasmid lengths due to higher fragmentation of some draft genomes is possible) up to 1,356,931 bp (Figure S12). All megaplasmids (>500 Kbp) are detected in terrestrial isolates, and some have been previously reported.[53,54]

Examining numbers of genomes with BGC-bearing plasmids at the genus level, *Streptomyces* spp., *Rhodococcus* spp., *Frankia* spp., and *Salinispora* spp. are examples employing plasmids as a prominent strategy for BGC expansion, in addition to *Pseudonocardia* spp., *Actinomadura* spp., *Mycobacterium* spp. etc. (Figure S13; Table S5). The role of plasmids in shaping the SM repertoire of a small subset of *Streptomyces* spp. and *Rhodococcus* spp. have been previously examined.[53,55] The contribution of plasmid-borne BGCs to the total number or total %BGCs per genome ranges from <1% to 66.6% (three genomes have a solitary BGC that is located on a plasmid). While there was no clear preponderance of plasmid-encoded SM classes, lanthipeptide-class-I, thioamitides, and butyrolactone were relatively overrepresented. Genera like *Mycobacteroides*, *Rathayibacter*, *Gordonia*, *Mycolicibacterium*, and others have above avg. %BGC commitment but do not appear to employ plasmids for BGC expansion.

In *Pseudonocardia* spp., multiple plasmid-borne BGCs are in evidence (Figure 4A)—for example, an 800 kb megaplasmid in strain EC080610-09 results in eight new strain-specific BGCs (see subclade I). In subclade II, all possess a lassopeptide-bearing plasmid (Figure 4A) except for a near-identical strain, HH130629-09, that is missing this (or any other) plasmid but possesses additional strain-specific SMs. Examining their gene neighborhoods reveals they are flanked by transposases, integrases, recombinases, etc., suggesting that other means of HGT may have been employed (Figure 4B). BGCs for nucleoside and NRPS + other appear to be inserted at tRNA genes, suggesting they may have been borne on integrative and mobilizable or conjugative elements (IMEs or ICEs, respectively).[56,57]

To survey this more systematically across the entire dataset, we cross-referenced BGCs against HGT-derived genes predicted by HGTector[58] and found that 28,913 BGCs from 4,776 genomes have predicted HGT genes. 457 of these BGCs were found on predicted plasmids. This implies that most genomes may possess at least one horizontally acquired BGC. The proportion of such HGT BGCs ranges from 2.85% to 100% (median 38%) of the total number of BGCs per genome. 178 genomes

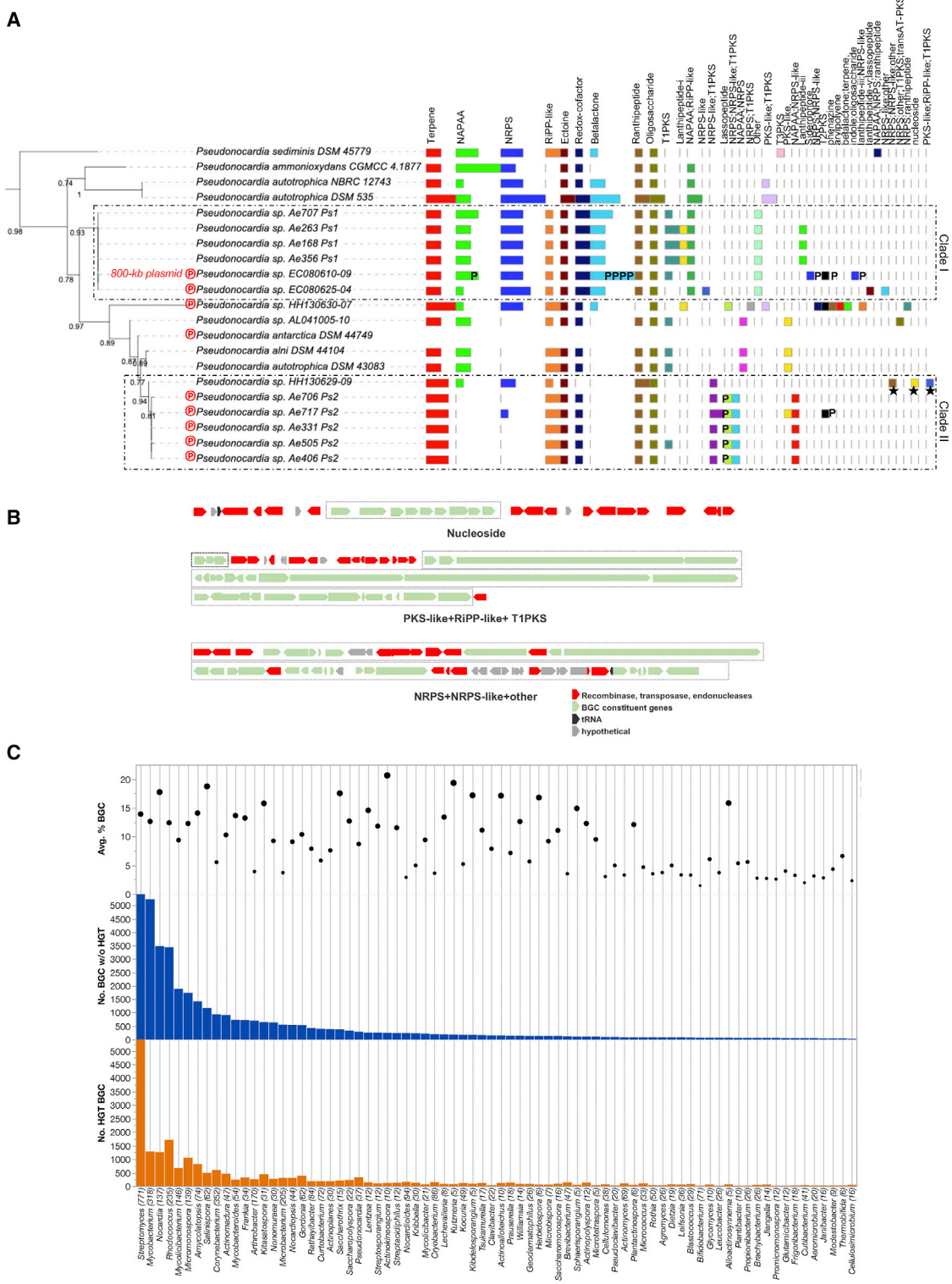


Figure 4. Horizontal gene transfer of BGCs

(A) Examining the role of plasmid-mediated HGT within a closely related subset of *Pseudonocardia* spp. Maximum likelihood tree based on RpoB gene alignments of selected strains is annotated with bar charts depicting the number of BGCs for each class predicted by AntiSMASH (drawn using iToL). Bootstrap support is included. "P" is used to indicate a BGC detected on a plasmid scaffold predicted in that genome. "Ⓟ" adjacent to a genome label indicates a BGC-bearing

(legend continued on next page)

showed 100% HGT rate; however, most of these belonged to small host-associated genera with a single BGC—such as several species of *Actinomyces*, *Bifidobacterium*, *Cutibacterium*, *Candidatus Planktophila*, etc. Other genera with a striking proportion of HGT-derived BGCs as well as high BGC commitment included *Streptomyces* (41.5% of total BGCs from 770 genomes), *Rhodococcus* (33% from 235 genomes), *Micromonospora* (38% from 139 genomes), *Kitasatospora* (41% from 31 genomes), *Gordonia* (42% from 61 genomes), *Pseudonocardia* (54% from 37 genomes), etc. (Figure 4C). In *Streptomyces* spp., BGC flux mediated by plasmids or actinomycete IMEs and ICEs has previously been recognized.[59]

SM classes that are notably overrepresented in this HGT subset include terpene, RiPP-like, siderophore, ectoine, butyrolactone, redox-cofactor, melanin, etc. (Figure S14). A total of 313 out of 646 *Pseudonocardia* spp. BGCs appear in this list, including the strain-specific ones highlighted above. Similarly, in *Bifidobacterium* spp., 23 BGCs in 21 genomes may have been recently acquired (Figure S11A). Overall, subclades within each lineage are likely under different ongoing selective pressures driving the highly dissimilar BGC composition, facilitated by various HGT strategies as well as deletion events.[60] This evidence of relatively recent acquisition may be used as a strategy for prioritizing characterization of specific BGCs in addition to previously suggested ones.[61]

Horizontal transfer may impact the detection of BGCs even in HQ MAGs—for example, MAGs encoded only 3 BGCs/genome or 2.39% BGCs (Figure 3B). These are possible underpredictions since the metagenome-binning process is expected to be biased against HGT regions due to their deviant nucleotide composition and/or coverage (plasmid copy-number effects), compared with the main chromosome.[62] Furthermore, the higher relative fragmentation of MAGs (avg. scaffold N50 of 141.6 Kbp for HQ MAGs versus 1.4 Mbp for all isolates) can also contribute to false negatives since BGC lengths avg. > 33 kb (based on MiBIG[45] and GenBank entries). This further underscores the need for HQ isolate genome sequences for continued SM gene discovery efforts.[63,64]

Prophages and host-virus interactions

Prophages are phage genomes residing in bacterial cells, often integrated into their host chromosome, during latent phases of their infection cycles. In addition to contributing to HGT, phage-host interactions may also play a role in iterative genome evolution and possibly contribute to host fitness by conferring resistance mechanisms or metabolic advantages. Identifying prophages from whole-genome data provides a unique opportunity to better understand the prevalence, diversity, host range, and gene content of phages infecting *Actinobacteria*.

We applied VirSorter2[65] and CheckV[66] to automatically detect, curate, and identify (near-)complete prophage sequences in *Actinobacteria* isolate genomes (see STAR Methods; Figure S15). After quality filtering and dereplication, a final dataset of 4,831 distinct prophages from 2,756 genomes was obtained, including 3,393 estimated to be (near-)complete from 2,244 genomes. We then mapped predicted proteins from all *Actinobacteria* isolate genomes to this non-redundant catalog of *Actinobacteria* prophages to establish a global picture of prophage prevalence and distribution across *Actinobacteria*.

Overall, 60.4% of *Actinobacteria* isolate genomes (n = 3,412) included at least one prophage-like region (Table S8), while a complete or near-complete prophage could be detected in 45.4% of the genomes. This difference is likely due to the presence of inactive and/or decayed prophages and to challenges in assembling variable genome regions, including prophages, from short reads. The relatively high frequency of genomes without any detectable prophages across *Actinobacteria* (~40%) is in line with previous observations[67–69] and seems to be consistent across taxa within the *Actinobacteria* phylum (Figure 5A). Overall, the number of prophages detected per genus scaled with the number of genomes sequenced within this genus (Pearson correlation coefficient = 0.89) with a handful of outliers. First, strains in the *Mycobacteroides*, *Bifidobacterium*, and *Leifsonia* yielded a disproportionately large number of prophages and consistently displayed a lower percentage of genomes without any trace of prophage compared with other genera (19%, 26%, and 31% respectively). In the case of *Mycobacteroides*, this may be due in part to the large collection of phages isolated from strains in this genus,[70] which may help with the identification of (HQ) prophages. On the other end of the spectrum, *Clavibacter* strains included 90% of genomes without any trace of a prophage. Since *Clavibacter* genomes are relatively compact (~3 Mb), they may include less prophages than other larger *Actinobacteria* genomes; however, it is also possible that *Clavibacter* prophages are simply more distant from references and more challenging to detect than other *Actinobacteria* phages.

When present however, it is not infrequent to observe multiple distinct prophages in the same host genome (17% overall), which could provide opportunities for recombination and HGT between unrelated phages. This is consistent with some temperate actinophages having been identified as exhibiting "high lateral gene flow" pattern, i.e., subject to a higher rate of horizontal gene exchange than most other phages.[71] Among various genome features including genome size, isolation source, host taxonomy at the genus, family, or order rank, number of tRNAs, and presence of CRISPR-CAS systems and BGCs, only taxonomy was detected as significantly associated with the number of prophages detected (ANOVA p value $< 2 \times 10^{-16}$ at all ranks tested). This indicates that the variation in prophage presence is not directly linked to a general environment or lifestyle (e.g., SM producer) but instead is likely due to differences in

plasmid. Other genomes may have plasmids, but BGCs were not encoded on those plasmids. Subclades are highlighted as discussed in the manuscript. Black stars mark further instances of HGT as illustrated in (B).

(B) Schematic of BGC examples in strain HH130629-09 that may have been acquired via alternative means of HGT such as ICEs. The core genes for the BGC are colored green, while red indicates hallmark genes for integration or transposition. tRNA genes are shown in black.

(C) Genera encoding the highest numbers of HGT BGCs (orange bars) are contrasted with non-HGT BGCs (blue bars). Bars for *Streptomyces* are truncated for better display and total 11,018 HGT BGCs versus 15,507 non-HGT BGCs. Top panel with weighted points is the average percentage of BGCs of genomes in each genus. On the x axis, genera are ordered by descending order of total number of BGCs without HGT. Number of genomes for each genus is shown in parentheses.

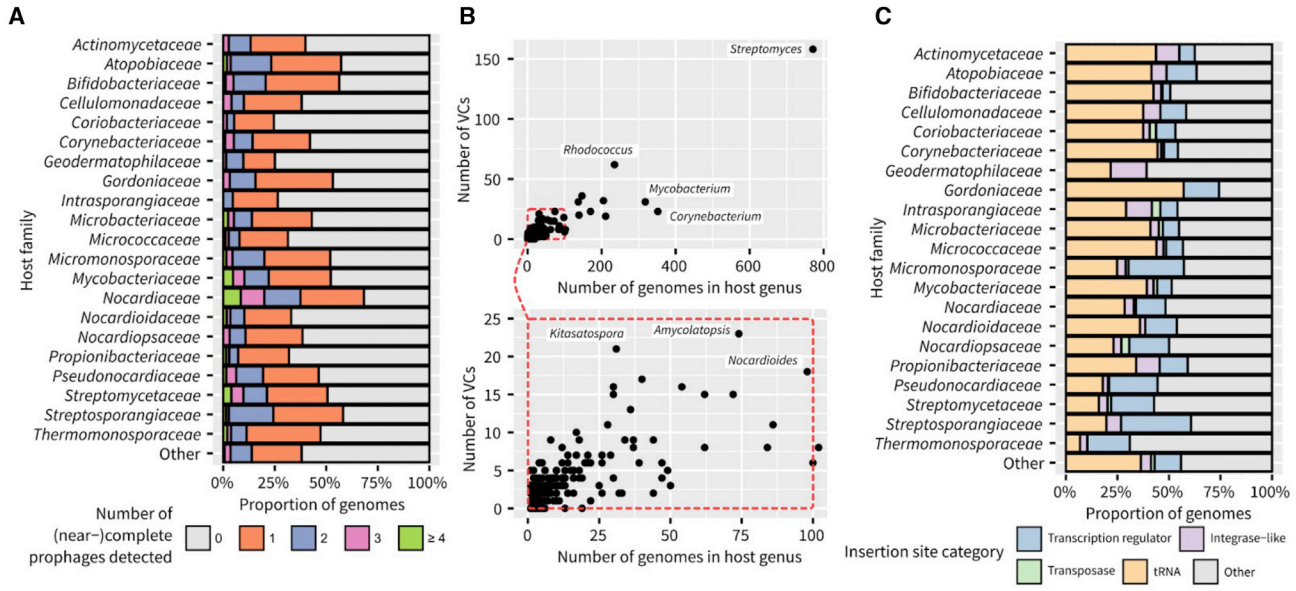


Figure 5. Overview of prophage content across *Actinobacteria* genomes

(A) Number of complete and near-complete prophages detected by genomes across major families. Families with ≤50 genomes are gathered in the "other" category.

(B) Number of distinct viral clusters detected by host genus, relative to the number of genomes screened in the genus. The bottom panel shows a zoomed-in version of the data for genera with ≤105 genomes. Individual genera with the most VCs and/or genomes mined are named on each plot.

(C) Prophage insertion site across major *Actinobacteria* families. For each prophage, the host genomic regions immediately 1 kb upstream and downstream of the 5′ and 3′ ends, respectively, were screened for the detection of tRNA, integrase-like genes, or transposases belonging to other mobile genetic elements (i.e., not the prophage currently considered), and transcription regulators. Families with ≤50 genomes are gathered in the "other" category.

life-history traits between strains, which are likely best captured as taxonomic classification in this dataset.[72]

Next, we evaluated the diversity of prophages recovered across *Actinobacteria* genomes through automated phage genome network analysis implemented in vContact 2.[73] Clustering all (near-)complete *Actinobacteria* prophages along with 14,256 reference genomes from the INPHARED database[74] yielded a total of ~1,837 genus-level groups (i.e., viral clusters [VCs]), including 365 with ≥2 phages. Almost half (46%) of host genera were associated with 2 or more VCs, and the number of VCs detected per genus was clearly increasing with the number of genomes sampled in the host genus (Figure 5B). This illustrates how *Actinobacteria* within individual genera can be infected by a broad range of phages and how whole-genome shotgun sequencing of many members within a given genus can shed light on this extensive prophage diversity.

Given this broad phage diversity, we next evaluated the distribution of individual prophages across host diversity. Prophages were typically (78%) detected in a single genome and, when detected in multiple genomes, were majorly associated with a single genus (85%; Figure S16). When detected across multiple genera, however, the host genera tended to be in different families (58%) and order (45%; Figure S16). This suggests that while most *Actinobacteria* prophages are "specialists," i.e., have a narrow host range, the host range of "generalist" prophages does not closely reflect host taxonomy beyond the genus rank. Conversely, individual VCs were much more frequently detected across diverse hosts (Figure S16). Among VCs with 2 or more prophages, 50% were associated with more than one host genus and 25% with multiple host families. Several VCs also included members infecting multiple classes of *Actinobacteria*, suggesting that these either reflected ancient groups of phages predating the divergence of these different classes or, more likely, that some prophages were able to "jump" from one host to another in a different class.

Finally, we explored the gene content of *Actinobacteria* prophages to evaluate the potential impact of prophage on host cell functioning. As is typical in phage genomes, most genes (60%–80% depending on the host genus) could not be functionally annotated, while the annotated functions were mostly directly related to phage replication and capsid production, e.g., integrases, major capsid proteins, or tail proteins. However, one exception, a gene encoding a component of predicted Mn/Zn uptake complex, was identified in 3 *Atopobium* prophages (Figure S17). Zn uptake can play a critical role in the pathogenicity of some bacteria,[75] and the presence of phage-encoded Mn/Zn transporters suggests that some *Actinobacteria* prophages may directly increase their host's fitness by providing additional resources for acquiring these nutrients. Beyond phage-encoded genes, however, prophage integration can also influence host cell functioning by disrupting neighboring genes.

Since the vast majority (94.7%) of *Actinobacteria* prophages were detected as integrated in the host chromosome, we also explored the function of genes found near insertion sites. For 65% of integrated prophages, an integrase-like and/or tRNA

gene could be identified on the phage side of the integration sites at the 5′ or 3′ end, as is typical for *Caudoviricetes* (Table S8). In contrast, the genes found immediately outside of prophage insertion sites were much more variable (Figure 5C). Notably, these included a substantial number of transposases and integrases, distinct from the ones identified within the prophage and often belonging to other mobile genetic elements integrated immediately upstream/downstream of the prophage. This suggests that a number of *Actinobacteria* prophages may be integrated in integration hotspots, likely representing hypervariable regions of the *Actinobacteria* genome. The other common functional category identified immediately next to insertion sites was transcriptional regulators, suggesting that some prophage integration events may impact regulatory pathways within the host cell.[76]

Conclusions

Microbial genomics has come a long way since the first bacterial whole-genome sequence of *Haemophilus influenzae* published in 1995[77,78]—as of March 2021, over 220,700 genomes of bacteria and archaea are listed in RefSeq.[79] These numbers are of course dwarfed by those of uncultivated genome equivalents (MAGs and single-amplified genomes [SAGs]) derived from environmental samples. Many of these genomes from "dark matter" lineages like the candidate phylum radiation (CPR) and others upend microbial precepts arising from the study of experimentally tractable lineages and model organisms like *E. coli*.[80–83] While this data deluge is impressive, the role and importance of HQ genomes of isolates is undeniable, not only in serving as a reference point for the interpretation of uncultivated sequences but also as an experimentally tractable resource in the laboratory.

Here, we explore *Actinobacteria*, a large and ancient phylum renowned for the richness and diversity of its natural products, by first producing HQ draft genomes of 824 isolates of primarily type strains. Comparative analyses with public genomes (both isolates and MAGs) revealed that only half of total actinobacterial PD is represented by a genome (even if including MQ MAGs with >50% incompleteness). A large portion of the remaining diversity can be attributed to underrepresented or new lineages arising from poorly accessible or extreme environments. Isolation efforts concentrated on such understudied or rare samples could result in the capture of a significant portion of this unrepresented diversity. The inherent value of well characterized type strains in informing cultivation of novel or unrepresented clades is also underscored with some examples.

The term "dark matter" may also be applied to the functionally unknown content within genomes (e.g., orphan genes, intergenic regions, proto-genes, etc.), which is even more extensive and intractable than the taxonomic dark matter. So, while the fraction of inaccessible taxa may diminish, the functional characterization lags far behind.[84] Here, again, the value of type strains as accessible standardized material is obvious. With greater statistical power achievable due to increased numbers of genomes of *Actinobacteria* from diverse lineages and environments, robust genome-wide comparisons are feasible toward identifying adaptations specific to a lineage, environment, or observed genotype or phenotype. We identify new and uncharacterized functions involved in niche adaptation by comparing host-associated versus environmental genomes. For example, several enriched Pfams for lipid metabolism may represent new or overlooked determinants of host-microbe interaction and possibly virulence, even in well-studied human pathogens. Functions with restricted taxonomic distribution are highlighted, and a previously uncharacterized antimicrobial peptide family enriched in soil- and plant-associated *Actinobacteria* is preliminarily characterized. Much more is possible with this expanded set of genomes accounting for almost 80% of projected diversity for the class *Actinobacteria* (the largest class)—underpinnings of phenotypes such as sporulation, cell shape, multicellularity, DNA topology, etc., await discovery (the only constraint being the availability of reliable metadata).

We also analyze an inventory of >80,000 BGCs predicted from isolates and examine the widespread role of HGT in shaping the repertoire across taxa. The ubiquity of this phenomenon and the highly fragmented nature of HQ MAGs results in a potential bias in BGC discovery, again reiterating the need for reference isolate genome sequences. However, the sequence itself is merely a starting point, and unfortunately only an insignificant fraction of over a million BGCs have any confirmed bioactivity, so the need for targeted efforts is great. To this end, the evolutionary and ecological history of a BGC, such as recent HGT events and its distribution in different environments, could provide an additional line of reasoning in prioritization of BGCs for biochemical characterization. Overall, our findings emphasize the essential role- and unique value of reference isolate genomes and present a compelling case for the continued sequencing of extant strains of isolates.

Limitations of the study

While we have emphasized the value of HQ genomes of cultured species as a reference point for various analyses and experimentation, the procurement of such actinobacterial cultures is non-trivial—a cultivation bias due to predicted slow growth rates is likely for many, as recently indicated for marine actinobacteria.[85] *Actinobacteria* are also known to have very large and highly repetitive genomes, which prevents their recovery from metagenomes. Furthermore, while existing isolate genomes are a notable resource, almost a quarter of coding sequences (CDSs) elicit no functional annotation, and the vast majority are uncharacterized.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Sequence, assembly and annotation
  - Curating public genomes for comparative analyses
  - Phylogenetic diversity (PD) estimation
  - Biosynthetic gene cluster analysis
  - Genome comparisons

- Antimicrobial peptide cloning and inhibition assay
- Prophage detection

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100213.

ACKNOWLEDGMENTS

The work (proposal DOI[s]: https://doi.org/10.46936/10.25585/60001024; https://doi.org/10.46936/10.25585/60000886; https://doi.org/10.46936/10.25585/60001401; https://doi.org/10.46936/10.25585/60001087; https://doi.org/10.46936/10.25585/60001079; https://doi.org/10.46936/10.25585/60001044) conducted by the US Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under contract no. DE-AC02-05CH11231. Design and synthesis of peptide constructs were supported by the Biosystems Design program (DOE Office of Science contract DE-SC0018260) and the US Department of Energy Joint Genome Institute (DOE Office of Science contract DE-AC02-05CH1123), respectively. Characterization of the peptide was supported by the Secured Biosystem Design project entitled "Rapid Design and Engineering of Smart and Secure Microbiological Systems" (DOE Office of Science contract DE-AC02-05CH1123). We thank the following researchers for their support of this study by providing free use of their public genome data: Kristen De-Angelis, Grace Pold, Mallory Choudoir, Camila Carlos-Shanley, Paul Carini, H. Corby C. Kistler, James Elkins, Javier A. Izquierdo, Dimitris Hatzinikolaou, Daniel Schachtman, Paul R. Jensen, Aindrila Mukhopadhyay, John Vogel, Carolin Frank, Paul M. D'Agostino, Ann M. Hirsch, Satoshi Yuzawa, Regina Lamendella, Bernhard Fuchs, Dale Pelletier, Laila P. Partida-Martinez, Cameron Currie, Seth De-Bolt, Jeff Dangl, David Mead, Susannah Tringe, David A. Baltrus, Seung Bum Kim, Linda Kinkel, Kelly Wrighton, William Mohn, Ludmila Christoserdova, Sarah Lebeis, Janet Janssen, Sandra Baena Garzon, and Nicholas Coleman. Special thanks to Marie Louise Ballon at the JGI Communications and Outreach office for all her help with the graphical abstract and many other details.

AUTHOR CONTRIBUTIONS

R.S., N.N.I., N.C.K., M.G., H.-P.K., and W.B.W. conceived the study. R.S., L.C.R., J.S., J.B., and S.M. acquired and curated the data. K.J.H., R.P., R.L.H., Z.Y., L.E., V.S., and S.T. provided DNA for sequencing. R.S., S.R., D.W., N.N.I., D.U., L.C., J.R.W., and N.J.V. analyzed and interpreted the data. S.Y., Z.-Y.W., Z.Z., and Y.Y. conducted experimental validation of antimicrobial peptide. I.-M.A.C., P.P.H., M.H., C.W., K.P., J.B., S.M., and T.B.K.R. provided support for data through IMG/M and GOLD. A.C. and A.S. performed genome sequencing and assembly. N.S., R.S., and T.W. provided project management and coordination. R.S. and N.N.I. designed and wrote the manuscript with feedback from S.N., T.W., E.A.E.-F., N.J.M., and N.C.K. All authors reviewed and corrected the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: January 14, 2022
Revised: July 19, 2022
Accepted: October 16, 2022
Published: November 11, 2022

REFERENCES

1. van Bergeijk, D.A., Terlouw, B.R., Medema, M.H., and van Wezel, G.P. (2020). Ecology and genomics of Actinobacteria: new concepts for natural product discovery. Nat. Rev. Microbiol. *18*, 546–558. https://doi.org/10.1038/s41579-020-0379-y.
2. Barka, E.A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Meier-Kolthoff, J.P., Klenk, H.P., Clément, C., Ouhdouch, Y., and van Wezel, G.P. (2016). Taxonomy, physiology, and natural products of Actinobacteria. Microbiol. Mol. Biol. Rev. *80*, 1–43. https://doi.org/10.1128/MMBR.00019-15.
3. Lewin, G.R., Carlos, C., Chevrette, M.G., Horn, H.A., McDonald, B.R., Stankey, R.J., Fox, B.G., and Currie, C.R. (2016). Evolution and ecology of Actinobacteria and their bioenergy applications. Annu. Rev. Microbiol. *70*, 235–254. https://doi.org/10.1146/annurev-micro-102215-095748.
4. Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L.C., Yeong, M., Cruz-Morales, P., Abubucker, S., et al. (2020). A computational framework to explore large-scale biosynthetic diversity. Nat. Chem. Biol. *16*, 60–68. https://doi.org/10.1038/s41589-019-0400-9.
5. Prudence, S.M.M., Addington, E., Castaño-Espriu, L., Mark, D.R., Pintor-Escobar, L., Russell, A.H., and McLean, T.C. (2020). Advances in actinomycete research: an ActinoBase review of 2019. Microbiology *166*, 683–694. https://doi.org/10.1099/mic.0.000944.
6. Pan, G., Xu, Z., Guo, Z., Hindra, Ma, M., Yang, D., Zhou, H., Gansemans, Y., Zhu, X., Huang, Y., et al. (2017). Discovery of the leinamycin family of natural products by mining actinobacterial genomes. Proc. Natl. Acad. Sci. USA *114*, E11131–E11140. https://doi.org/10.1073/pnas.1716245115.
7. Belknap, K.C., Park, C.J., Barth, B.M., and Andam, C.P. (2020). Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. Sci. Rep. *10*, 2003. https://doi.org/10.1038/s41598-020-58904-9.
8. Kalkreuter, E., Pan, G., Cepeda, A.J., and Shen, B. (2020). Targeting bacterial genomes for natural product discovery. Trends Pharmacol. Sci. *41*, 13–26. https://doi.org/10.1016/j.tips.2019.11.002.
9. Kyrpides, N.C., Hugenholtz, P., Eisen, J.A., Woyke, T., Göker, M., Parker, C.T., Amann, R., Beck, B.J., Chain, P.S.G., Chun, J., et al. (2014). Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. PLoS Biol. *12*, e1001920. https://doi.org/10.1371/journal.pbio.1001920.
10. International Code of (2019). Nomenclature of Prokaryotes Int. J. Syst. Evol. Microbiol. *69*, S1–S111. https://doi.org/10.1099/ijsem.0.000778.
11. Reimer, L.C., Vetcininova, A., Carbasse, J.S., Söhngen, C., Gleim, D., Ebeling, C., and Overmann, J. (2019). BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. Nucleic Acids Res. *47*, D631–D636. https://doi.org/10.1093/nar/gky879.
12. Nouioui, I., Carro, L., García-López, M., Meier-Kolthoff, J.P., Woyke, T., Kyrpides, N.C., Pukall, R., Klenk, H.P., Goodfellow, M., and Göker, M. (2018). Genome-based taxonomic classification of the phylum Actinobacteria. Front. Microbiol. *9*, 2007. https://doi.org/10.3389/fmicb.2018.02007.
13. Oren, A., and Garrity, G.M. (2021). Valid publication of the names of forty-two phyla of prokaryotes. Int. J. Syst. Evol. Microbiol. *71*. https://doi.org/10.1099/ijsem.0.005056.
14. Chen, I.M.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S., Varghese, N., Seshadri, R., et al. (2021). The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res. *49*, D751–D763. https://doi.org/10.1093/nar/gkaa939.
15. Schorn, M.A., Alanjary, M.M., Aguinaldo, K., Korobeynikov, A., Podell, S., Patin, N., Lincecum, T., Jensen, P.R., Ziemert, N., and Moore, B.S. (2016). Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. Microbiology *162*, 2075–2086. https://doi.org/10.1099/mic.0.000386.
16. Sayed, A.M., Hassan, M.H.A., Alhadrami, H.A., Hassan, H.M., Goodfellow, M., and Rateb, M.E. (2020). Extreme environments: microbiology

leading to specialized metabolites. J. Appl. Microbiol. *128*, 630–657. https://doi.org/10.1111/jam.14386.

17. Subramani, R., and Aalbersberg, W. (2013). Culturable rare actinomycetes: diversity, isolation and marine natural product discovery. Appl. Microbiol. Biotechnol. *97*, 9291–9321. https://doi.org/10.1007/s00253-013-5229-7.
18. Letunic, I., and Bork, P. (2021). Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296. https://doi.org/10.1093/nar/gkab301.
19. Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.M., Huntemann, M., et al. (2020). A genomic catalog of Earth's microbiomes. Nat. Biotechnol. *39*, 499–509. https://doi.org/10.1038/s41587-020-0718-6.
20. Staley, J.T., and Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. *39*, 321–346. https://doi.org/10.1146/annurev.mi.39.100185.001541.
21. Faith, D.P. (2013). Biodiversity and evolutionary history: useful extensions of the PD phylogenetic diversity assessment framework. Ann. N. Y. Acad. Sci. *1289*, 69–89. https://doi.org/10.1111/nyas.12186.
22. Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V., Goodwin, L., Wu, M., Tindall, B.J., et al. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature *462*, 1056–1060. https://doi.org/10.1038/nature08656.
23. Lewis, W.H., Tahon, G., Geesink, P., Sousa, D.Z., and Ettema, T.J.G. (2021). Innovations to culturing the uncultured microbial majority. Nat. Rev. Microbiol. *19*, 225–240. https://doi.org/10.1038/s41579-020-00458-8.
24. Lee, S.D., and Lee, D.W. (2007). Lapillicoccus jejuensis gen. nov., sp. nov., a novel actinobacterium of the family Intrasporangiaceae, isolated from stone. Int. J. Syst. Evol. Microbiol. *57*, 2794–2798. https://doi.org/10.1099/ijs.0.64911-0.
25. Choi, O., Kim, J., Kim, J.G., Jeong, Y., Moon, J.S., Park, C.S., and Hwang, I. (2008). Pyrroloquinoline quinone is a plant growth promotion factor produced by Pseudomonas fluorescens B16. Plant Physiol. *146*, 657–668. https://doi.org/10.1104/pp.107.112748.
26. Sun, Y., Cheng, Z., and Glick, B.R. (2009). The presence of a 1-aminocyclopropane-1-carboxylate (ACC) deaminase deletion mutation alters the physiology of the endophytic plant growth-promoting bacterium Burkholderia phytofirmans PsJN. FEMS Microbiol. Lett. *296*, 131–136. https://doi.org/10.1111/j.1574-6968.2009.01625.x.
27. Shulse, C.N., Chovatia, M., Agosto, C., Wang, G., Hamilton, M., Deutsch, S., Yoshikuni, Y., and Blow, M.J. (2019). Engineered root bacteria release plant-available phosphate from phytate. Appl. Environ. Microbiol. *85*, e01210-19. https://doi.org/10.1128/AEM.01210-19.
28. van der Geize, R., Grommen, A.W.F., Hessels, G.I., Jacobs, A.A.C., and Dijkhuizen, L. (2011). The steroid catabolic pathway of the intracellular pathogen Rhodococcus equi is important for pathogenesis and a target for vaccine development. PLoS Pathog. *7*, e1002181. https://doi.org/10.1371/journal.ppat.1002181.
29. Sethi, D., Mahajan, S., Singh, C., Lama, A., Hade, M.D., Gupta, P., and Dikshit, K.L. (2016). Lipoprotein Lprl of Mycobacterium tuberculosis acts as a lysozyme inhibitor. J. Biol. Chem. *291*, 2938–2953. https://doi.org/10.1074/jbc.M115.662593.
30. Ragas, A., Roussel, L., Puzo, G., and Rivière, M. (2007). The Mycobacterium tuberculosis cell-surface glycoprotein Apa as a potential adhesin to colonize target cells via the innate immune system pulmonary C-type lectin surfactant protein A. J. Biol. Chem. *282*, 5133–5142. https://doi.org/10.1074/jbc.M610183200.
31. Lau, C.K.Y., Krewulak, K.D., and Vogel, H.J. (2016). Bacterial ferrous iron transport: the Feo system. FEMS Microbiol. Rev. *40*, 273–298. https://doi.org/10.1093/femsre/fuv049.
32. Casabon, I., Swain, K., Crowe, A.M., Eltis, L.D., and Mohn, W.W. (2014). Actinobacterial acyl coenzyme A synthetases involved in steroid side-chain catabolism. J. Bacteriol. *196*, 579–587. https://doi.org/10.1128/JB.01012-13.
33. Daniel, J., Sirakova, T., and Kolattukudy, P. (2014). An acyl-CoA synthetase in Mycobacterium tuberculosis involved in triacylglycerol accumulation during dormancy. PLoS One *9*, e114877. https://doi.org/10.1371/journal.pone.0114877.
34. Sugrue, I., O'Connor, P.M., Hill, C., Stanton, C., and Ross, R.P. (2020). Actinomyces produces defensin-like bacteriocins (actifensins) with a highly degenerate structure and broad antimicrobial activity. J. Bacteriol. *202*, 005299-19. https://doi.org/10.1128/JB.00529-19.
35. Verly, R.M., Resende, J.M., Junior, E.F.C., de Magalhães, M.T.Q., Guimarães, C.F.C.R., Munhoz, V.H.O., Bemquerer, M.P., Almeida, F.C.L., Santoro, M.M., Piló-Veloso, D., and Bechinger, B. (2017). Structure and membrane interactions of the homodimeric antibiotic peptide homotarsinin. Sci. Rep. *7*, 40854. https://doi.org/10.1038/srep40854.
36. Sathoff, A.E., and Samac, D.A. (2019). Antibacterial activity of plant defensins. Mol. Plant Microbe Interact. *32*, 507–514. https://doi.org/10.1094/MPMI-08-18-0229-CR.
37. Brogden, K.A. (2005). Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? Nat. Rev. Microbiol. *3*, 238–250. https://doi.org/10.1038/nrmicro1098.
38. Rima, M., Rima, M., Fajloun, Z., Sabatier, J.M., Bechinger, B., and Naas, T. (2021). Antimicrobial peptides: a potent alternative to antibiotics. Antibiotics *10*, 1095. https://doi.org/10.3390/antibiotics10091095.
39. Zhang, L.J., and Gallo, R.L. (2016). Antimicrobial peptides. Curr. Biol. *26*, R14–R19. https://doi.org/10.1016/j.cub.2015.11.017.
40. Hu, D., Sun, C., Jin, T., Fan, G., Mok, K.M., Li, K., and Lee, S.M.Y. (2020). Exploring the potential of antibiotic production from rare Actinobacteria by whole-genome sequencing and guided MS/MS analysis. Front. Microbiol. *11*, 1540. https://doi.org/10.3389/fmicb.2020.01540.
41. Niu, G. (2018). Genomics-driven natural product discovery in actinomycetes. Trends Biotechnol. *36*, 238–241. https://doi.org/10.1016/j.tibtech.2017.10.009.
42. Bérdy, J. (2012). Thoughts and facts about antibiotics: where we are now and where we are heading. J. Antibiot. *65*, 441. https://doi.org/10.1038/ja.2012.54.
43. Gavriilidou, A., Kautsar, S.A., Zaburannyi, N., Krug, D., Müller, R., Medema, M.H., and Ziemert, N. (2022). Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nat. Microbiol. *7*, 726–735. https://doi.org/10.1038/s41564-022-01110-2.
44. Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., and Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. *49*, W29–W35. https://doi.org/10.1093/nar/gkab335.
45. Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., et al. (2020). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. *48*, D454–D458. https://doi.org/10.1093/nar/gkz882.
46. Baltz, R.H. (2017). Gifted microbes for genome mining and natural product discovery. J. Ind. Microbiol. Biotechnol. *44*, 573–588. https://doi.org/10.1007/s10295-016-1815-x.
47. Mukherjee, S., Seshadri, R., Varghese, N.J., Eloe-Fadrosh, E.A., Meier-Kolthoff, J.P., Göker, M., Coates, R.C., Hadjithomas, M., Pavlopoulos, G.A., Paez-Espino, D., et al. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat. Biotechnol. *35*, 676–683. https://doi.org/10.1038/nbt.3886.
48. Hannigan, G.D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., et al. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster

prediction. Nucleic Acids Res. *47*, e110. https://doi.org/10.1093/nar/gkz654.

49. Mavrodi, D.V., Peever, T.L., Mavrodi, O.V., Parejko, J.A., Raaijmakers, J.M., Lemanceau, P., Mazurier, S., Heide, L., Blankenfeldt, W., Weller, D.M., and Thomashow, L.S. (2010). Diversity and evolution of the phenazine biosynthesis pathway. Appl. Environ. Microbiol. *76*, 866–879. https://doi.org/10.1128/AEM.02009-09.

50. Seipke, R.F. (2015). Strain-level diversity of secondary metabolism in Streptomyces albus. PLoS One *10*, e0116457. https://doi.org/10.1371/journal.pone.0116457.

51. Choudoir, M.J., Pepe-Ranney, C., and Buckley, D.H. (2018). Diversification of secondary metabolite biosynthetic gene clusters coincides with lineage divergence in Streptomyces. Antibiotics *7*, E12. https://doi.org/10.3390/antibiotics7010012.

52. Doroghazi, J.R., and Metcalf, W.W. (2013). Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. BMC Genom. *14*, 611. https://doi.org/10.1186/1471-2164-14-611.

53. Medema, M.H., Trefzer, A., Kovalchuk, A., van den Berg, M., Müller, U., Heijne, W., Wu, L., Alam, M.T., Ronning, C.M., Nierman, W.C., et al. (2010). The sequence of a 1.8-mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. Genome Biol. Evol. *2*, 212–224. https://doi.org/10.1093/gbe/evq013.

54. Pathak, A., Chauhan, A., Blom, J., Indest, K.J., Jung, C.M., Stothard, P., Bera, G., Green, S.J., and Ogram, A. (2016). Comparative genomics and metabolic analysis reveals peculiar characteristics of Rhodococcus opacus strain M213 particularly for naphthalene degradation. PLoS One *11*, e0161032. https://doi.org/10.1371/journal.pone.0161032.

55. Ceniceros, A., Dijkhuizen, L., Petrusma, M., and Medema, M.H. (2017). Genome-based exploration of the specialized metabolic capacities of the genus Rhodococcus. BMC Genom. *18*, 593. https://doi.org/10.1186/s12864-017-3966-1.

56. Bellanger, X., Payot, S., Leblond-Bourget, N., and Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiol. Rev. *38*, 720–760. https://doi.org/10.1111/1574-6976.12058.

57. Guédon, G., Libante, V., Coluzzi, C., Payot, S., and Leblond-Bourget, N. (2017). The obscure world of integrative and mobilizable elements, highly widespread elements that pirate bacterial conjugative systems. Genes *8*, E337. https://doi.org/10.3390/genes8110337.

58. Zhu, Q., Kosoy, M., and Dittmar, K. (2014). HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. BMC Genom. *15*, 717. https://doi.org/10.1186/1471-2164-15-717.

59. Tidjani, A.R., Lorenzi, J.N., Toussaint, M., van Dijk, E., Naquin, D., Lespinet, O., Bontemps, C., and Leblond, P. (2019). Massive gene flux drives genome diversity between sympatric Streptomyces conspecifics. mBio *10*, 015333-19. https://doi.org/10.1128/mBio.01533-19.

60. Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W., and Crook, D.W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol. Rev. *33*, 376–393. https://doi.org/10.1111/j.1574-6976.2008.00136.x.

61. Tran, P.N., Yen, M.R., Chiang, C.Y., Lin, H.C., and Chen, P.Y. (2019). Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. Appl. Microbiol. Biotechnol. *103*, 3277–3287. https://doi.org/10.1007/s00253-019-09708-z.

62. Nelson, W.C., Tully, B.J., and Mobberley, J.M. (2020). Biases in genome reconstruction from metagenomic data. PeerJ *8*, e10119. https://doi.org/10.7717/peerj.10119.

63. Baltz, R.H. (2019). Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. J. Ind. Microbiol. Biotechnol. *46*, 281–299. https://doi.org/10.1007/s10295-018-2115-4.

64. Baltz, R.H. (2021). Genome mining for drug discovery: progress at the front end. J. Ind. Microbiol. Biotechnol. *48*, kuab044. https://doi.org/10.1093/jimb/kuab044.

65. Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D., Sullivan, M.B., and Roux, S. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome *9*, 37. https://doi.org/10.1186/s40168-020-00990-y.

66. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585. https://doi.org/10.1038/s41587-020-00774-7.

67. Kim, M.S., and Bae, J.W. (2018). Lysogeny is prevalent and widely distributed in the murine gut microbiota. ISME J. *12*, 1127–1141. https://doi.org/10.1038/s41396-018-0061-9.

68. Mavrich, T.N., Casey, E., Oliveira, J., Bottacini, F., James, K., Franz, C.M.A.P., Lugli, G.A., Neve, H., Ventura, M., Hatfull, G.F., et al. (2018). Characterization and induction of prophages in human gut-associated Bifidobacterium hosts. Sci. Rep. *8*, 12772. https://doi.org/10.1038/s41598-018-31181-3.

69. Penn, K., Jenkins, C., Nett, M., Udwary, D.W., Gontang, E.A., McGlinchey, R.P., Foster, B., Lapidus, A., Podell, S., Allen, E.E., et al. (2009). Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J. *3*, 1193–1203. https://doi.org/10.1038/ismej.2009.58.

70. Jordan, T.C., Burnett, S.H., Carson, S., Caruso, S.M., Clase, K., DeJong, R.J., Dennehy, J.J., Denver, D.R., Dunbar, D., Elgin, S.C.R., et al. (2014). A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. mBio *5*, e01051-e01013. https://doi.org/10.1128/mBio.01051-13.

71. Mavrich, T.N., and Hatfull, G.F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. Nat. Microbiol. *2*, 17112. https://doi.org/10.1038/nmicrobiol.2017.112.

72. Touchon, M., Bernheim, A., and Rocha, E.P. (2016). Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME J. *10*, 2744–2754. https://doi.org/10.1038/ismej.2016.47.

73. Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. *37*, 632–639. https://doi.org/10.1038/s41587-019-0100-8.

74. Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D.J., Hobman, J., Jones, M.A., and Millard, A. (2021). INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. Phage *2*, 214–223. https://doi.org/10.1089/phage.2021.0007.

75. Quan, G., Xia, P., Lian, S., Wu, Y., and Zhu, G. (2020). Zinc uptake system ZnuACB is essential for maintaining pathogenic phenotype of F4ac(+) enterotoxigenic E. coli (ETEC) under a zinc restricted environment. Vet. Res. *51*, 127. https://doi.org/10.1186/s13567-020-00854-1.

76. Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., and Herskovits, A.A. (2015). A new perspective on lysogeny: prophages as active regulatory switches of bacteria. Nat. Rev. Microbiol. *13*, 641–650. https://doi.org/10.1038/nrmicro3527.

77. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496–512. https://doi.org/10.1126/science.7542800.

78. Kyrpides, N.C. (2009). Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. Nat. Biotechnol. *27*, 627–632. https://doi.org/10.1038/nbt.1552.

79. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.

80. Giovannoni, S.J. (2017). SAR11 bacteria: the most abundant plankton in the oceans. Ann. Rev. Mar. Sci *9*, 231–255. https://doi.org/10.1146/annurev-marine-010814-015934.

81. Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. Nat. Microbiol. *1*, 16048. https://doi.org/10.1038/nmicrobiol.2016.48.

82. Liu, Y., Makarova, K.S., Huang, W.C., Wolf, Y.I., Nikolskaya, A.N., Zhang, X., Cai, M., Zhang, C.J., Xu, W., Luo, Z., et al. (2021). Expanded diversity of Asgard archaea and their relationships with eukaryotes. Nature *593*, 553–557. https://doi.org/10.1038/s41586-021-03494-3.

83. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.

84. Thomas, A.M., and Segata, N. (2019). Multiple levels of the unknown in microbiome research. BMC Biol. *17*, 48. https://doi.org/10.1186/s12915-019-0667-z.

85. Long, A.M., Hou, S., Ignacio-Espinoza, J.C., and Fuhrman, J.A. (2021). Benchmarking microbial growth rate predictions from metagenomes. ISME J. *15*, 183–195. https://doi.org/10.1038/s41396-020-00773-1.

86. Mavromatis, K., Land, M.L., Brettin, T.S., Quest, D.J., Copeland, A., Clum, A., Goodwin, L., Woyke, T., Lapidus, A., Klenk, H.P., et al. (2012). The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. PLoS One *7*, e48837. https://doi.org/10.1371/journal.pone.0048837.

87. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138. https://doi.org/10.1126/science.1162986.

88. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. *37*, 540–546. https://doi.org/10.1038/s41587-019-0072-8.

89. Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. *18*, 810–820. https://doi.org/10.1101/gr.7337908.

90. Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569. https://doi.org/10.1038/nmeth.2474.

91. Huntemann, M., Ivanova, N.N., Mavromatis, K., Tripp, H.J., Paez-Espino, D., Tennessen, K., Palaniappan, K., Szeto, E., Pillay, M., Chen, I.M.A., et al. (2016). The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). Stand. Genomic Sci. *11*, 17. https://doi.org/10.1186/s40793-016-0138-x.

92. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

93. Pati, A., Ivanova, N.N., Mikhailova, N., Ovchinnikova, G., Hooper, S.D., Lykidis, A., and Kyrpides, N.C. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nat. Methods *7*, 455–457. https://doi.org/10.1038/nmeth.1457.

94. Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2014). Nucleic Acids Res. *42*, D32–D37. https://doi.org/10.1093/nar/gkt1030.

95. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

96. Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Res. *46*, W200–W204. https://doi.org/10.1093/nar/gky448.

97. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. *41*, e121. https://doi.org/10.1093/nar/gkt263.

98. Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650. https://doi.org/10.1093/molbev/msp077.

99. Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinf. *11*, 538. https://doi.org/10.1186/1471-2105-11-538.

100. Kautsar, S.A., van der Hooft, J.J.J., de Ridder, D., and Medema, M.H. (2021). BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience *10*, giaa154. https://doi.org/10.1093/gigascience/giaa154.

101. Antipov, D., Raiko, M., Lapidus, A., and Pevzner, P.A. (2019). Plasmid detection and assembly in genomic and metagenomic data sets. Genome Res. *29*, 961–968. https://doi.org/10.1101/gr.241299.118.

102. Krawczyk, P.S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res. *46*, e35. https://doi.org/10.1093/nar/gkx1321.

103. Pellow, D., Mizrahi, I., and Shamir, R. (2020). PlasClass improves plasmid sequence classification. PLoS Comput. Biol. *16*, e1007781. https://doi.org/10.1371/journal.pcbi.1007781.

104. Shintani, M., Sanchez, Z.K., and Kimbara, K. (2015). Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front. Microbiol. *6*, 242. https://doi.org/10.3389/fmicb.2015.00242.

105. R Core Team, A.C.W. (2002). The R Stats Package (R Core Team).

106. Roux, S., Krupovic, M., Daly, R.A., Borges, A.L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P.B., Cheng, J.F., Ivanova, N.N., et al. (2019). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. Nat. Microbiol. *4*, 1895–1906. https://doi.org/10.1038/s41564-019-0510-x.

107. Buchfink, B., Reuter, K., and Drost, H.G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

108. Team, R.C. (2021). R: A Language and Environment for Statistical Computing. https://www.R-project.org/.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Escherichia. coli BL21(DE3)	ThermoFisher Scientific	Cat#EC0114
Saccharomyces cerevisiae CEN.PK2	EUROSCARF Institute for Molecular Biosciences	54398
Chemicals, peptides, and recombinant proteins		
LB (Miller's) broth	Growcells	Cat#MBLE-7030
Carbenicillin	Sigma-Aldrich	Cat#C1613
Isopropyl β-D-1-thiogalactopyranoside (IPTG)	Sigma-Aldrich	Cat#I6758
BugBuster® Protein Extraction Reagent	Millpore	Cat#70584
YPD Agar Plates	TEKNOVA	Cat#Y5131
Fluconazole	Cerilliant®	Cat#F-097
Coomassie Brilliant Blue R-250 Staining Solution	Bio-Rad	Cat#1610436
Deposited data		
GEBA-Actino genomes generated by this study	This study	GenBank accessions provided in Table S1 https://www.ncbi.nlm.nih.gov/genbank/
Metagenome-assembled genomes (MAGs) from GEM dataset	Nayfach et al.[19]	https://genome.jgi.doe.gov/GEMs
Recombinant DNA		
pET-21(+)_AMP_S.be: pET21-(+) plasmid containing AMP sequences originating from Streptosporangium becharense DSM 46887	This study	N/A
Software and algorithms		
Interactive tree of like (iTOL)	Letunic and Bork[18]	https://itol.embl.de/
PD estimation	This paper	https://doi.org/10.5281/zenodo.7058177
Generalized linear model analysis	This paper	https://doi.org/10.5281/zenodo.7058201

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and analyses should be directed to Rekha Seshadri (rseshadri@lbl.gov).

Materials availability

This study did not generate new materials.

Data and code availability

- All genome data generated in this study are publicly available in GenBank and IMG (individual accession numbers are listed in Table S1).
- All original code has been deposited at Zenodo and is publicly available. DOIs are listed in the key resources table.
- Alignment and tree files used for PD estimation have been deposited in Treebase: http://purl.org/phylo/treebase/phylows/study/TB2:S29629.

METHOD DETAILS

Sequence, assembly and annotation

All GEBA-Actino genomes were sequenced at the DOE Joint Genome Institute (JGI) using Illumina technology[86] or Pacific Biosciences (PacBio) RS technology.[87] For all genomes, we either constructed and sequenced an Illumina short-insert paired-end library with an average insert size of 270 bp, or a Pacbio SMRTbell library. Genomes were assembled using Flye v. 2.6,[88] ALLPATHS[89] or

Hierarchical Genome Assembly Process (HGAP)[90] assembly methods (specifics provided in Table S1). Genomes were annotated by the DOE–JGI genome annotation pipeline.[91] Briefly, protein-coding genes (CDSs) were identified using Prodigal[92] followed by a round of automated and manual curation using the JGI GenePrimp pipeline.[93] Functional annotation and additional analyses were performed within the Integrated Microbial Genomes (IMG-ER) platform.[14] All GEBA-Actino data are available through the Integrated Microbial Genomes with Microbiomes (IMG/M) system[14] and GenBank,[94] and the corresponding type strains through the respective culture collections (Table S1). All data including detailed sequencing and assembly reports can be downloaded from GenBank and JGI Genome portal: https://genome.jgi.doe.gov/portal/

Curating public genomes for comparative analyses

For various comparisons described in the study, a total of 4,824 good quality phylum *Actinobacteria* isolate genomes were curated from the complete set of available public genomes (at the inception of this analysis in Jan 2020). "High quality" public genomes are designated by the IMG quality control pipeline (based on phylum-level taxonomic assignment or if the coding density is >70% or <100%, or the number of genes per million base pair is >300 or <1,200.[91] CheckM completeness/contamination criteria were also applied with some exceptions such as highly reduced genomes of *Tropheryma* spp. (~0.83 Mbp) that are likely underestimated by checkM due to loss of marker genes.[95] Isolate genomes dataset was partially de-duplicated by removing multiple strains of *Mycobacterium tuberculosis, Mycobacteroides abscessus* (for example) after assessing the average nucleotide identity (ANI) of total best bidirectional hits and removing genomes sharing >99% ANI (alignment fraction of total CDS $\geq$ 90%) to another genome within that set. A total of 1,098 Actinobacterial MAGs were selected (based on $\geq$95% completeness and $\leq$5% contamination) from a recently published comprehensive catalog of MAGs recovered from over 10,000 public metagenomes representing the breadth of existing diversity of sampled environments (Table S1)[19] - referred to as HQ MAGs in this study. For PD estimation alone, 2,223 medium quality (MQ) MAGs was also included.

Phylogenetic diversity (PD) estimation

Universally conserved single-copy marker proteins, RpoB and Ribosomal protein L1 were used for construction of a maximum likelihood phylogenetic tree and estimating total phylogenetic diversity of isolates, MAGs and metagenomes. Marker genes for RpoB were detected with multiple Pfam domains (pf04560 RNA_polymerase_Rpb2_domain_7, pf04561 RNA_polymerase_Rpb_domain_2, pf04563 RNA_polymerase_beta_subunit, pf04565 RNA_polymerase_Rpb2_domain_3 and pf00562 RNA_polymerase_Rpb2,_domain_6) assigned by the IMG annotation pipeline[14] (that employs hmmsearch[96]), aligned with hmmalign,[97] and individual domain alignments were concatenated into one cohesive RpoB alignment. Only sequences that covered $\geq$70% of the total model positions were included in tree building using Fasttree2 (LG model).[98] Markers for ribosomal protein L1 were similarly detected (with pf00687 Ribosomal_L1), aligned and treed. For markers arising from metagenomic sequences, a minimum scaffold length of 5 kb was imposed and Actinobacterial marker sequences were identified using pplacer[99] to place candidate sequences on a reference tree including non-actinobacterial marker genes for tree rooting and removing non-actinobacterial sequences. Using this protocol, a total of 15,114 RpoB and 16,302 ribosomal L1 genes respectively, were recovered from potentially uncultivated actinobacterial genomes from 20,100 metagenome samples from diverse environments housed within the IMG database.[14] The PD contribution of sequences from each group (public isolates, GEBA isolates, MAGs (MQ and HQ GEMs) and metagenomes) to the overall phylogenetic diversity was inferred from the ribosomal L1 and RpoB trees separately using methods described in Wu et al.[22] Consistent results were obtained with both markers. Original code for this analysis is publicly available: https://doi.org/10.5281/zenodo.7058177. Alignment and tree files for RpoB are available in Treebase (http://purl.org/phylo/treebase/phylows/study/TB2:S29629).

Biosynthetic gene cluster analysis

Secondary metabolite encoding BGC regions were identified using AntiSMASH (v6) with default settings,[44] and ignoring contigs with lengths shorter than 5 kb. Gene cluster family (GCF) assignment for each BGC region was determined using BiG-SLICE with default settings.[100] Potential HGT-derived BGCs were predicted by mapping genes against a list of HGT-derived genes predicted by HGTector that targets atypically distributed genes.[58] Other horizontally acquired BGCs were identified by their location on plasmid scaffolds. Three software prediction tools were utilized to identify putative plasmid scaffolds - plasmidVerify,[101] PlasFlow[102] and PlasClass.[103] These three tools employ different types of machine-learning-based classifiers (naïve Bayes, neural-network, or logistic regression, respectively) and were trained on two types of features - either plasmid-specific gene signatures (plasmidVerify) or nucleotide signatures (PlasFlow and PlasClass), thus using all three provided a robust way to identify a diverse set of plasmid scaffolds. The final set of predicted plasmid scaffolds was delineated based on overlapping predictions from at least two methods, and a minimum scaffold length of 2.5 kb (based on previous report of Actinobacterial plasmid lengths[104]).

Genome comparisons

For whole genome comparisons, isolate or MAG genomes were carefully selected from the IMG database using available metadata fields pertaining to isolation source or manually curated when possible. Comparisons of gene counts for individual Pfams, Tigrfams or KO terms between members of each set or group of isolates were performed. For host (2,650 genomes including 678 MAGs) versus environment (2,306 including 284 MAGs) comparisons, host genomes were smaller on average than environmental isolates (Figure S2D), therefore analyses were based on gene presence versus absence rather than gene copy number or relative abundances.

Since overall taxonomic composition of each set was also highly varied or relatively biased (such as a preponderance of members of class *Acidimicrobiia* in the environmental group or class *Coriobacteriia* in the host group) (Figure S1), a phylogeny-normalized generalized linear model (GLM) approach was employed[105] using a pairwise distance matrix based on phylogenetic distances computed from the RpoB marker tree (mentioned above), in order to minimize potentially confounding effects arising from a biased phylogenetic signal. Significant results from both a fixed and a mixed model were utilized, since occasionally, the mixed effects model fitting procedure failed to converge on a reasonable fit due to imbalances in the distribution of underlying taxa (throwing an error reported as "n/a " in the results table). Most significant features were delineated using a false discovery rate (FDR) adjusted p-value cutoff of <0.005, and positive or negative regression coefficients (which capture magnitude of the fold differences between each group as well as the phylogenetic regression), reflect overrepresentation in host group versus environmental group, respectively (Table S3). KO and Pfam (accounting for >80% coverage (on average) of total CDS) were used to complement and validate results arising from any single function annotation type, and to examine metabolic pathways more closely (KO pathways). Custom code used for this analysis is available: https://doi.org/10.5281/zenodo.7058201.

Antimicrobial peptide cloning and inhibition assay

A 102 amino acid AMP candidate from *Streptosporangium becharense* DSM 46887 (GenBank ID: MBB5817359.1) with a predicted molecular weight of 11.2 kDa, pI of 8.67 and charge of 3.5, was cloned in *E. coli* for functional validation. The sequence was codon-optimized and sent to Twist Biosciences for synthesis and cloned into a pET-21(+) expression vector. The plasmid DNA (pET-21(+)_AMP_S.be) was then transferred into *E. coli* BL21(DE3) strain with 100 μg/mL carbenicillin (MilliporeSigma, Burlington, MA, USA) selection. To overexpress the short peptide, 100 mL of *E. coli* BL21(DE3) harboring the recombinant plasmid was cultured in Luria-Bertani broth containing 100 μg/mL of carbenicillin at 37°C until the mid-exponential phase. The overexpression was induced by addition of 0.1 mM isopropyl-β-D-thiogalactopyranoside (MilliporeSigma) at 25°C and 120 rpm for 16 h. The cells were harvested by centrifugation at 6,000 rpm and 4°C for 10 min and protein was extracted with BugBuster® Protein Extraction Reagent (MilliporeSigma) following the protocol from the kit. The protein was concentrated using two different Amicon Ultra centrifugal filters (MWCO 3 kDa and 30 kDa, MilliporeSigma).

Yeast *Saccharomyces cerevisiae* was used for the antimicrobial activity assessment with proteins extracted from the recombinant *E. coli. S. cerevisiae* was streaked and cultivated on the YPD agar plate at 30°C one day prior to the assay. One or a few yeast colonies were resuspended in 2 mL of 0.85% sterile saline with a sterile inoculating loop. A sterile swab was dipped into the inoculum tube and was rotated to remove the excess fluid. The swab was then streaked on the YPD plate while rotating to distribute the inoculum evenly. 6-mm sterilized filter paper disk soaked with 20–40 μL of the protein extract was placed on the YPD agar plate. Fluconazole (25 μg/disk) was used for positive control. The protein extract obtained from *E. coli* BL21(DE3) harboring empty pET-21(+) plasmid was used as negative control. After incubation at 30°C for 24 h, the antimicrobial activity of the short peptide was determined by appearance of the zone of inhibition. To confirm the overexpression of the short peptide, the protein extract was analyzed using SDS-PAGE gel (12% Mini-PROTEAN® TGX™ Precast Gel, Bio-Rad Laboratories, Hercules, CA, USA). The same amount (20 μg) of each protein extract from *E. coli* BL21(DE3) harboring the recombinant plasmid or empty plasmid was loaded in a single well and the gel was run at 180 V for 40 min. The gel was stained using Coomassie Brilliant Blue R-250 Staining Solution (Bio-Rad Laboratories).

Prophage detection

The 5,648 isolate genomes were screened for potential prophages using VirSorter 2.1,[65] using the dsDNAphage, RNA, and ssDNA models, a minimum score of 0.7, and a minimum length of 1 kb, as well as the Inovirus Detector scripts (v2019-06-30) with default parameters.[106] Predicted prophages were then analyzed for completeness and contamination with CheckV 0.7.0[66] using the end_to_end option and default parameters otherwise. Prophage prediction corresponding to common contamination/errors were identified based on the CheckV results and the VirSorter2 functional annotations as follows: all genes from the predicted prophages are similar to gene from Type 6 Secretion Systems; the predicted prophages is only composed of ≥3 contiguous peptidase genes; CheckV detects ≥1 host marker and 0 viral marker in the prophage; CheckV detects ≥2 host markers and ≤1 viral marker in the prophage; CheckV detects a host region of ≥2 genes. For all potential contamination detected via CheckV, the predicted prophage region was trimmed to the CheckV-predicted viral region and/or the nearest integrase(-like) gene within 5 kb of the predicted ends of a CheckV-predicted viral region. The gene content and VirSorter2 annotation of all prophages predicted to be ≥150% complete and/or with a length ≥50 kb were inspected, and the prophage boundaries were adjusted when the initial prediction included two contiguous but distinct prophages. Finally, to avoid including partial prophages, only predicted prophages predicted to be ≥50% complete or ≥10 kb (or ≥ 1 kb for inoviruses) were retained.

Selected prophages were next clustered at 95% average nucleotide identity (ANI) and 85% alignment fraction (AF) using ClusterGenomes v5.1 (https://github.com/simroux/ClusterGenomes) and only the largest representative of each cluster was retained.[106] From this non-redundant set, predicted prophages with a CheckV completeness estimation of ≥75% were considered as "(near-)complete", along with predicted prophages with surrounding host regions of ≥ 5 kb on both 5′ and 3′ ends except if these had a high-confidence CheckV completeness estimation <75% or if CheckV provided no completeness estimation at all, the latter typically corresponding to partial prophages too short for CheckV to estimate completeness. This was based on the observation that predicted prophages with a high-confidence CheckV completeness estimation and with surrounding host regions of ≥ 5 kb on both 5′ and 3′ ends were overwhelmingly (89.6%) estimated to be ≥75% complete. To evaluate the distribution and diversity of

Actinobacteria prophages, predicted proteins from *Actinobacteria* genomes were compared to the predicted proteins from the non-redundant prophage set using Diamond v0.9.25 in "–sensitive" mode.[107] A prophage protein was considered as detected in an *Actinobacteria* genome if covered by a hit with ≥90% identity and ≥90% coverage. A prophage was considered as detected in an *Actinobacteria* genome if ≥ 50% of its predicted proteins were detected, and a (near-)complete prophage was considered as entirely detected in a genome if ≥ 90% of its predicted proteins were detected. Potential link between the detection of prophages and other genome features, including taxonomy, environment, presence of a BGC, and number of scaffolds in the genome (≤5 or >5) was evaluated via ANOVA performed in R v4.1 (function aov[108]). All genomes for which these features were unknown were excluded from the analysis. Each taxonomic rank (genus, family, order, and class) was evaluated separately, and each time only taxa with ≥20 genomes were included in the analysis.

The diversity of prophages identified in *Actinobacteria* genomes was evaluated through a vContact2 v 0.9.11 genome network clustering.[73] For this analysis, 14,256 reference phage genomes from NCBI RefSeq and GenBank collected using the INfrastructure for a PHAge Reference Database perl script (https://github.com/RyanCook94/inphared.pl; data downloaded on 01/21/2021) were clustered along with the 3,393 representative *Actinobacteria* prophages considered as "near-complete" (see above). The viral clusters ("VCs") identified by vContact2 and including at least one *Actinobacteria* prophage were then interpreted as "genus-level" groups when evaluating the diversity of prophages associated with different *Actinobacteria* host taxa.

Finally, the gene content of insertion sites/regions was evaluated in *Actinobacteria* prophages based on IMG genome annotation. The predicted functions of prophage-encoded genes situated within 1 kb of the prophage predicted boundaries, i.e., the last few genes encoded by the prophage before its 5′ and 3′ ends, were searched for canonical insertion sites and prophage edge genes including tRNA and integrase-like genes (annotated as "integrase", "recombinase", or "excisionase"). Prophages were classified as "tRNA and integrase-like", "tRNA", "integrase-like", or "other" ends based on the presence of a tRNA and an integrase-like gene, a tRNA gene only, an integrase-like gene only, or neither a tRNA nor an integrase-like gene within the 1 kb edge of a prophage. Similarly, regions within 1 kb of the prophage predicted boundaries, i.e., the 1 kb regions immediately outside of the prophage 5′ and 3′ ends, were searched for tRNA, DNA binding/transcriptional regulator genes (i.e., genes annotated "transcription regulator", "transcription activator", "gntR", "acrR", "tetR", "HTH", and "DNA-binding"), transposase genes, and integrase-like genes.