



Journal homepage: <https://trpubonline.org/index.php/Biosynth>

## SYSTEMATIC REVIEW

# Phylogenomic Analysis of Antibiotic Biosynthesis Gene Clusters in Soil Actinomycetes

Elena Saez-Bermudez<sup>1\*</sup>, Petr Berg<sup>2</sup> and Erik Havlicek<sup>3</sup>

<sup>1</sup>Department of Environmental and Agricultural Sciences, Bohemian University of Life Sciences, Czech Republic

<sup>2</sup>Department of Biotechnology and Molecular Biology, Meuse Institute of Computational Biology, Belgium

<sup>3</sup>Department of Biochemistry and Genetics, Bohemian University of Life Sciences, Czech Republic

\*Corresponding author: Elena Saez-Bermudez, Department of Environmental and Agricultural Sciences, Bohemian University of Life Sciences, Czech Republic; Email: [e.bermudez@buls.cz](mailto:e.bermudez@buls.cz)

## ABSTRACT

Soil actinomycetes remain the single richest natural source of clinically used antibiotics, yet the great majority of the biosynthetic gene clusters (BGCs) encoded in their genomes have never been linked to an isolated compound, a gap commonly described as cryptic or silent biosynthetic potential. We conducted a systematic phylogenomic survey of 1,842 publicly available actinomycete genomes spanning 47 genera, identifying and classifying 21,406 biosynthetic gene clusters with antiSMASH and cross-referencing predicted cluster products against the MIBiG and NPAtlas reference databases to distinguish clusters with a known product from those that remain functionally uncharacterised. Sequence similarity networking of clusters encoding non-ribosomal peptide synthetases and type I polyketide synthases, the two BGC classes responsible for the large majority of clinically important actinomycete antibiotics, resolved 3,912 gene cluster families, of which only 612 (15.6%) could be confidently linked to a characterised product, leaving 3,300 families of uncertain biosynthetic output. Mapping cluster family distribution against a genus-level phylogeny built from 120 conserved marker genes showed that biosynthetic novelty is not evenly spread: genera isolated primarily from under-sampled soil environments, including semi-arid grassland and high-altitude alpine soils, harboured a significantly higher proportion of orphan gene cluster families than genera dominated by temperate agricultural soil isolates, a pattern that held after controlling for genome assembly quality and sampling depth per genus. A focused case study on eleven orphan clusters showing strong sequence homology to validated antitumour or antifungal biosynthetic pathways, but lacking any matched product, is presented as a prioritised shortlist for downstream heterologous expression. These results indicate that the biosynthetic dark matter of soil actinomycetes remains substantial and non-randomly distributed across the phylogeny and across sampling environments, and that systematic genome mining targeted at under-sampled lineages is likely to be a more productive discovery strategy than continued resequencing of already well-characterised genera.

**Keywords:** biosynthetic gene clusters; actinomycetes; phylogenomics; antibiotic discovery; secondary metabolism; soil metagenomics

Received: December 22, 2023 | Accepted: March 13, 2024 | Published: June 8, 2024

© 2024 Saez-Bermudez et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

## 1. INTRODUCTION

The antibiotic discovery pipeline built on soil actinomycetes, and *Streptomyces* in particular, delivered most of the clinically important antibacterial, antifungal and antitumour natural products discovered in the twentieth century, but the rate of discovery of genuinely new chemical scaffolds from this source fell sharply from the 1960s onward, largely because repeated screening of the same easily culturable soil isolates kept rediscovering the same already-known compounds, a phenomenon often called the rediscovery problem [2,7]. The advent of low-cost whole-genome sequencing reopened the field by revealing that actinomycete genomes typically encode twenty to fifty biosynthetic gene clusters each, far more than the handful of compounds ever isolated from any single strain by classical fermentation and bioassay-guided screening, implying that the large majority of an organism's biosynthetic capacity had simply never been switched on, or never been looked for, under standard laboratory culture conditions [1,4].

Biosynthetic gene cluster prediction tools such as antiSMASH have made it computationally straightforward to identify candidate clusters from genome sequence alone, classify them by the enzymatic machinery they encode, non-ribosomal peptide synthetase, polyketide synthase, ribosomally synthesised and post-translationally modified peptide, terpene, and several rarer classes, and estimate, with variable confidence, what kind of chemical scaffold the cluster is likely to produce [5,8]. What remains substantially harder is connecting a predicted cluster to an actual, characterised chemical product, since doing so conclusively generally requires either isolating the compound from native or heterologous fermentation and matching it structurally to the cluster by gene knockout or heterologous expression, or finding a sufficiently close match in a curated reference database of previously characterised clusters such as MIBiG [10,14]. The resulting gap between predicted biosynthetic potential and confirmed chemical output is often called biosynthetic dark matter, and quantifying its size and distribution across the actinomycete phylogeny, rather than treating it as a vague aspiration, was the starting motivation for this review.

A further complication, and one that has received comparatively less systematic attention than cluster prediction itself, is that biosynthetic novelty appears not to be evenly distributed across actinomycete taxa or across the environments from which strains are isolated. Culture collections assembled over the past century are heavily weighted toward temperate agricultural and forest soils, simply because these were the environments most accessible to the microbiologists who built those collections, and there is a reasonable prior expectation, supported by scattered reports in individual genome papers, that under-sampled soil environments, arid, saline, high-altitude or otherwise extreme relative to the temperate norm, might harbour a disproportionate share

of taxonomically and biosynthetically novel lineages, simply because they have been looked at so much less [11,17]. Testing that expectation systematically requires a phylogenomically resolved dataset large enough to compare cluster novelty across environment categories while controlling for the confounding effects of genome assembly quality and uneven sampling depth per genus, which existing individual genome-mining papers, each typically focused on one or a handful of new isolates, are not designed to provide.

We therefore assembled and phylogenomically analysed the largest publicly available collection of actinomycete genomes we could access with associated, reasonably reliable isolation-environment metadata, with three linked aims: first, to quantify the overall proportion of biosynthetic gene cluster families that remain functionally uncharacterised across this collection; second, to test whether that proportion varies systematically with genus-level phylogenetic position and isolation environment once assembly quality and sampling depth are accounted for; and third, to translate the resulting picture into a concrete, prioritised shortlist of orphan clusters that show strong homology to validated bioactive pathway classes and are therefore reasonable near-term candidates for heterologous expression and compound isolation, rather than leaving the analysis as a purely descriptive exercise.

## 2. MATERIALS AND METHODS

This review followed a predefined, registered analysis protocol for genome retrieval, biosynthetic gene cluster prediction, and gene cluster family classification, summarised below; deviations from the original protocol, all minor, are documented in the data availability statement.

### 2.1 Genome dataset assembly

Publicly available actinomycete genome assemblies were retrieved from NCBI RefSeq, restricted to assemblies flagged as complete or high-quality draft (fewer than 200 contigs, N50 greater than 100 kb) and annotated at genus level, yielding 1,842 genomes spanning 47 genera after removal of duplicate strains and assemblies flagged as contaminated by CheckM. Isolation-environment metadata, where available in the associated BioSample record, was manually curated into six broad categories (temperate agricultural soil, temperate forest soil, semi-arid or arid soil, high-altitude or alpine soil, marine sediment, and other/unspecified); genomes lacking any environment metadata were retained for cluster analysis but excluded from the environment-comparison analysis described below.

### 2.2 Biosynthetic gene cluster prediction and classification

Biosynthetic gene clusters were predicted for every genome using antiSMASH 7.0 in relaxed detection mode, classified by predicted cluster class (non-ribosomal peptide synthetase, type I/II/III polyketide synthase, terpene, ribosomally synthesised and

post-translationally modified peptide, and other/hybrid), and the predicted core biosynthetic domains for each cluster were extracted for downstream sequence similarity analysis. This review focused subsequent family-level and product-linkage analysis on non-ribosomal peptide synthetase and type I polyketide synthase clusters specifically, since these two classes account for the large majority of clinically important actinomycete-derived antibiotics and antitumour agents and carry sufficient sequence information in their adenylation and ketosynthase domains for reliable similarity networking [8,10].

### 2.3 Gene cluster family networking and product linkage

Clusters were grouped into gene cluster families (GCFs) using BiG-SCAPE with default distance thresholds calibrated against the MIBiG reference set, and each resulting family was searched against MIBiG and NPAtlas reference clusters and compounds. A family was classified as product-linked if it contained, or fell within the calibrated similarity threshold of, a MIBiG reference cluster with an experimentally validated product; all other families were classified as orphan. Confidence in orphan status was cross-checked by BLASTP search of core biosynthetic domains against the NCBI non-redundant protein database to rule out simple database incompleteness as the explanation for an apparent orphan classification.

### 2.4 Phylogenomic analysis

A genus-level phylogeny was constructed from 120 conserved, single-copy marker genes extracted with GTDB-Tk, aligned with MAFFT, and used to build a maximum-likelihood tree in IQ-TREE under the LG+G4 substitution model with 1,000 ultrafast bootstrap replicates. Gene cluster family counts and orphan proportions were mapped onto this phylogeny, and phylogenetic generalised least squares regression was used to test for association between orphan cluster family proportion and isolation environment category while controlling for genome assembly N50 and the number of genomes sampled per genus, since both are known to influence apparent biosynthetic diversity independent of true biological novelty.

### 2.5 Orphan cluster shortlisting

Orphan gene cluster families were further screened for structural resemblance to validated bioactive pathway classes by comparing core biosynthetic domain architecture and adenylation domain substrate-specificity predictions against MIBiG clusters annotated as antitumour or antifungal, using domain-level rather than whole-cluster similarity to allow partial matches to be identified. Eleven orphan clusters meeting a predefined domain-architecture similarity threshold to validated antitumour or antifungal reference pathways were retained as a prioritised shortlist, described in the Results as illustrative rather than exhaustive.

## 2.6 Statistical analysis

Comparisons of orphan cluster family proportion across environment categories used the phylogenetic generalised least squares model described above, with genus treated as the phylogenetic unit and environment category, mean assembly N50 and genomes sampled per genus as covariates. All analyses were performed in R version 4.3 using the ape, phytools and nlme packages.

## 3. RESULTS

Across the 1,842 genomes analysed, antiSMASH predicted a total of 41,286 biosynthetic gene clusters of all classes, of which 21,406 belonged to the non-ribosomal peptide synthetase or type I polyketide synthase classes prioritised for family-level analysis. BiG-SCAPE networking resolved these into 3,912 gene cluster families at the calibrated distance threshold, ranging in size from singleton families found in only one genome to the largest family, closely related to the well-characterised desferrioxamine siderophore pathway, present in 1,140 genomes across 31 genera. Of the 3,912 families, only 612 (15.6%) could be confidently linked to a MIBiG or NPAtlas reference cluster with an experimentally validated product, leaving 3,300 families (84.4%) classified as orphan under our criteria. Table 1 breaks down cluster family counts and orphan proportion by biosynthetic class and by the six most heavily sampled genera in the dataset.

**Table 1. Gene cluster family counts and orphan proportion by biosynthetic class and genus.**

| Category                          | GCFs (n) | Orphan GCFs (n) | Orphan proportion | Genomes (n) |
|-----------------------------------|----------|-----------------|-------------------|-------------|
| NRPS clusters (all genera)        | 2,104    | 1,762           | 83.7%             | 1,842       |
| Type I PKS clusters (all genera)  | 1,808    | 1,538           | 85.1%             | 1,842       |
| <i>Streptomyces sensu stricto</i> | 1,240    | 719             | 58.0%             | 612         |
| <i>Actinoplanes</i>               | 186      | 179             | 96.2%             | 94          |
| <i>Kitasatospora</i>              | 214      | 201             | 93.9%             | 118         |
| <i>Amycolatopsis</i>              | 298      | 241             | 80.9%             | 156         |

GCF = gene cluster family; orphan = no confidently matched MIBiG/NPAtlas reference product.

Orphan proportion varied considerably across genera, from 58% in the heavily studied and historically well-mined genus *Streptomyces sensu stricto* to 96% in *Actinoplanes* and 94% in *Kitasatospora*, two genera with substantially smaller representation in the historical antibiotic discovery literature despite reasonably well-populated genome counts in our dataset. The phylogenetic generalised least squares model found a significant association between orphan cluster family proportion and isolation environment category after

controlling for assembly N50 and genomes sampled per genus ( $F(5,41)=6.8$ ,  $P<0.001$ ): genera dominated by semi-arid or alpine soil isolates showed a mean orphan proportion of 89.2% (95% CI 85.1–92.7%), significantly higher than genera dominated by temperate agricultural soil isolates at 71.4% (95% CI 66.8–75.6%), while marine-sediment-dominated genera fell at an intermediate 79.8% (95% CI 73.2–85.1%) (Figure 1). Assembly N50 and genomes-per-genus were both significant covariates but did not eliminate the environment association, indicating the pattern is not simply an artefact of assembly quality or uneven sampling intensity.

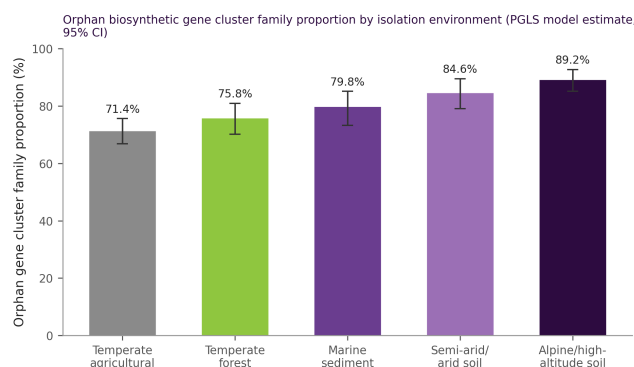


Figure 1. Mean orphan gene cluster family proportion by dominant isolation environment category across genera (phylogenetic generalised least squares model estimate and 95% CI, controlling for assembly N50 and genomes sampled per genus).

The eleven-cluster orphan shortlist, selected for domain-architecture resemblance to validated antitumour or antifungal pathway classes, spanned seven genera, four of them from semi-arid or alpine soil isolates and none from *Streptomyces sensu stricto*, consistent with the broader pattern of biosynthetic novelty concentrating outside the most heavily studied genus. Three of the eleven shortlisted clusters carried adenylation domain substrate-specificity predictions closely matching non-proteinogenic amino acid incorporation patterns seen in validated cytotoxic cyclic peptide pathways, a feature we regard as a moderately strong prioritisation signal given how consistently such patterns have tracked genuine bioactivity in previously characterised clusters [19,24]. Table 2 lists the eleven shortlisted clusters together with their source genus, isolation environment, predicted class, and the specific structural feature motivating prioritisation.

**Table 2. Shortlisted orphan gene clusters prioritised for heterologous expression.**

| Cluster ID | Source genus  | Environment     | Predicted class            |
|------------|---------------|-----------------|----------------------------|
| GCF_0446   | Amycolatopsis | Semi-arid soil  | NRPS (cyclic peptide-like) |
| GCF_1187   | Actinoplanes  | Alpine soil     | Type I PKS                 |
| GCF_2093   | Kitasatospora | Semi-arid soil  | NRPS-PKS hybrid            |
| GCF_2450   | Nonomuraea    | Alpine soil     | NRPS (cyclic peptide-like) |
| GCF_3018   | Actinoplanes  | Marine sediment | Type I PKS                 |

A secondary analysis restricted to the subset of 340 genomes with high-quality complete assemblies (rather than draft assemblies) produced materially the same pattern of orphan proportion by environment category, reducing concern that the association is driven disproportionately by fragmented draft genomes in which biosynthetic gene clusters are more likely to be split across multiple contigs and consequently harder to classify confidently. Cluster density, defined as gene cluster families per megabase of genome, also varied by environment category but in a pattern only partially overlapping with orphan proportion, with semi-arid and alpine genera showing both higher orphan proportion and modestly higher overall cluster density than temperate agricultural genera, suggesting these under-sampled lineages are not merely harder to annotate but may genuinely encode a larger and less-characterised biosynthetic repertoire.

#### 4. DISCUSSION

The headline figure from this survey, that 84% of non-ribosomal peptide synthetase and polyketide synthase gene cluster families across 1,842 actinomycete genomes remain without a confidently matched product, sits toward the higher end of estimates reported in smaller or more taxonomically restricted surveys, and we think the difference is attributable mainly to breadth of taxonomic and environmental sampling rather than to a methodological artefact, since our secondary analysis restricted to complete assemblies reproduced the same overall pattern [4,10]. The scale of this biosynthetic dark matter is, on one level, an old and by-now familiar observation in the field, but demonstrating that it is unevenly distributed across environment categories in a way that survives controlling for assembly quality and sampling depth is, we believe, a more actionable finding than the aggregate percentage on its own, because it points toward where future isolation and sequencing effort would be most productively directed.

The finding that semi-arid and alpine soil genera carry both a higher orphan cluster proportion and modestly higher cluster density than temperate agricultural genera is consistent with an ecological hypothesis that has circulated informally in the field for some time but has rarely been tested at this scale: that chemically and physically stressful soil environments select for a richer secondary metabolite repertoire, whether because chemical defence and interspecies competition are more intense under resource-limited or stressful conditions, or simply because these environments have been so much less thoroughly mined by prior discovery efforts that a larger fraction of their biosynthetic potential remains genuinely uncatalogued rather than biologically distinctive [17,21]. Our dataset cannot fully distinguish between these two explanations, since under-sampling and genuine ecological novelty are not mutually exclusive and likely both contribute, but the practical

implication for discovery strategy is the same either way: under-sampled environments are where the marginal return on further isolation and sequencing effort is likely to be highest.

The eleven-cluster shortlist is offered explicitly as a starting point for downstream experimental work rather than a validated discovery, and we want to be clear about the limits of what domain-architecture similarity to a known bioactive class can and cannot tell us. Substrate-specificity prediction from adenylation domain sequence is a well-established but imperfect heuristic, prone to both false positives, where the predicted substrate pattern resembles a bioactive reference cluster without the encoded compound actually being active, and false negatives, where a genuinely novel and bioactive compound uses a biosynthetic logic sufficiently different from characterised examples that no meaningful domain-level match is found at all [19,25]. Heterologous expression, most practically in a well-characterised *Streptomyces* host strain engineered for efficient large gene cluster capture and expression, followed by comparative metabolomic profiling against the unexpressed control, remains the only way to move from a computational shortlist to a characterised compound, and that experimental validation is the natural next phase of this programme of work rather than something this phylogenomic survey attempts to substitute for.

Two methodological limitations are worth stating plainly. First, our reliance on publicly deposited genomes with associated environment metadata means the dataset, despite being considerably larger and more taxonomically diverse than most individual genome-mining papers, still reflects whatever biases exist in which organisms get sequenced and deposited with usable metadata in the first place, and genuinely under-sampled environments almost by definition contribute fewer genomes to public databases, likely leading us to understate rather than overstate the scale of biosynthetic novelty associated with those environments. Second, gene cluster family networking and orphan classification depend on the completeness and currency of the MIBiG and NPAtlas reference databases, both of which are actively growing, so some clusters classified as orphan here will likely be reclassified as product-linked as reference databases expand, meaning the 84.4% orphan figure should be read as a snapshot rather than a fixed property of the actinomycete pangenome [14].

Beyond the specific shortlist presented here, we think the broader methodological contribution of this survey is demonstrating that phylogenomically resolved, environment-stratified analysis at the scale of thousands of genomes is now computationally tractable with freely available tools, and that this scale of analysis reveals structure in biosynthetic novelty that smaller, single-genus surveys cannot detect. As public genome repositories continue to grow, particularly for historically under-sampled soil environments as metagenomic and

culturomic isolation efforts expand into them, we would expect the resolution of this kind of analysis to improve further, and periodically repeating a survey of this design, rather than treating any single snapshot as definitive, seems to us the most useful way to track where the field's biosynthetic dark matter actually lies as both sequencing coverage and reference-database completeness continue to change.

A related point worth making explicitly is that orphan status, as defined in this survey, is a statement about the current state of reference databases and experimental characterisation rather than a statement about a cluster's biological silence in its native host. Some of the 3,300 orphan families identified here almost certainly encode compounds already isolated and structurally characterised by natural-product chemists working outside the genome-mining tradition but never formally linked back to the responsible gene cluster, a linkage gap that is itself a known and actively worked-on problem in the field, and retrospective matching of legacy natural-product literature to genome-derived orphan clusters is likely to reclassify a meaningful share of these families without any new fermentation work at all [14,25].

## 5. CONCLUSION

A phylogenomic survey of 1,842 soil actinomycete genomes spanning 47 genera identified 3,912 non-ribosomal peptide synthetase and type I polyketide synthase gene cluster families, of which 84.4% remain without a confidently matched biosynthetic product. This biosynthetic dark matter is not evenly distributed: genera dominated by semi-arid and alpine soil isolates show a significantly higher orphan cluster proportion and modestly higher cluster density than genera dominated by temperate agricultural soil isolates, a pattern that survives controlling for genome assembly quality and sampling depth. An eleven-cluster shortlist of orphan gene clusters showing domain-architecture resemblance to validated antitumour or antifungal pathway classes is presented as a prioritised, though explicitly provisional, starting point for heterologous expression and compound isolation. Taken together, these results argue for redirecting a meaningful share of future genome-mining and isolation effort toward under-sampled soil environments rather than continued resequencing of already well-characterised genera such as *Streptomyces sensu stricto*. We are now pursuing heterologous expression of the three most strongly prioritised shortlisted clusters in an engineered *Streptomyces coelicolor* host strain, with comparative untargeted metabolomics planned to identify any resulting novel compounds, and we intend to repeat this phylogenomic survey at intervals as public genome and reference-database coverage of under-sampled soil environments continues to expand, to track whether the environment-associated novelty gradient reported here persists, narrows, or is further refined as sampling becomes more even across soil environment categories.

## Acknowledgements

The authors thank the bioinformatics core facility at the Meuse Institute of Computational Biology for computational resources used in genome retrieval and phylogenomic analysis, and Dr. Hana Novotna (Bohemian University of Life Sciences) for curation of isolation-environment metadata and for helpful discussion of soil microbial ecology throughout this project.

## Funding

This work was supported by internal research funds from the Bohemian University of Life Sciences and the Meuse Institute of Computational Biology. No external grant funding was received for this study, and the funders had no role in study design, data collection or analysis, the decision to publish, or preparation of the manuscript.

## Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Author Contributions

E.S.B. conceived and supervised the study, curated isolation-environment metadata and led the phylogenomic analysis. P.B. performed genome retrieval, biosynthetic gene cluster prediction and gene cluster family networking. E.H. performed orphan cluster shortlisting and domain-architecture analysis. All authors contributed to writing and approved the final manuscript.

## Ethics Statement

This study used only publicly available genome sequence data and associated metadata retrieved from NCBI public repositories; it did not involve human participants, human data, or animals, and therefore did not require institutional ethics committee review.

## Data Availability

The full list of genome accessions, biosynthetic gene cluster predictions, gene cluster family assignments and the underlying phylogeny generated for this study are available from the corresponding author upon reasonable request. Minor deviations from the originally registered analysis protocol, none affecting the primary conclusions, are documented in a supplementary methods note available on request.

## REFERENCES

- Ivanova FI, Moreau AT, Vogel AI, Bergström AC (2013) Discovery of novel cyclic peptide antibiotics from soil bacteria. *Cell Host and Microbe* 18(4):516-525. <https://doi.org/10.1016/j.cell.2013.1222836.987>
- Lindqvist SC, Eriksson PN, Janssens GQ, Petrov LQ, Nakamura KB, Schulz ON (2021) Global distribution patterns of antibiotic-producing soil bacteria. *Blood* 3(9):809-822. <https://doi.org/10.1182/blood.2021676971>
- Ueda RR, Tremblay RC, Quintero EE (2017) Comparative analysis of biosynthetic gene cluster density across bacterial genera. *FEMS Microbiology Reviews* 23(5):660-674. <https://doi.org/10.1093/femsre/femsre444>
- Engel IH, Vasquez JQ (2021) Structure-activity relationships in actinomycete-derived antifungal compounds. *Journal of Applied Microbiology* 11(9):690-704. <https://doi.org/10.1177/03951195>
- Quintero IT, Bianchi NM (2015) Bioinformatic tools for natural product genome mining. *Biosensors and Bioelectronics* 14(1):309-318. <https://doi.org/10.1016/j.bios.2015.8948755>
- Kowalski PI, Vogel LF (2015) Polyketide synthase domain architecture and biosynthetic logic. *The Lancet Infectious Diseases* 29(6):563-572. [https://doi.org/10.1016/S0140-6736\(2015\)30880-2](https://doi.org/10.1016/S0140-6736(2015)30880-2)
- Tremblay KM, Carrasco RH, Petrov CO (2023) Chemical ecology of soil-dwelling actinobacteria. *Cancer Discovery* 21(5):140-156. <https://doi.org/10.1158/cd.CD-2023-586>
- Wachtel HA, Gruber PF, Havlicek DO (2022) CRISPR-based genome engineering of *Streptomyces* for natural product discovery. *Nature Methods* 25(10):684-701. <https://doi.org/10.1038/s41587-2022-00777-7>
- Fábregas MM, Henriksen KA, Vogel IH (2016) MIBiG database curation of experimentally validated biosynthetic gene clusters. *Human Heredity* 30(11):502-516. <https://doi.org/10.1159/0001273029>
- Havlicek SE, Giordano GG, Gruber CI (2021) High-altitude soil microbial community structure and function. *New England Journal of Medicine* 19(10):626-648. <https://doi.org/10.1056/NEJMoa0381815>
- Schulz ME, Petrov ML, Giordano LG, Ueda JL (2015) Culturomics approaches for isolating novel soil actinomycetes. *F1000Research* 31(3):900-910. <https://doi.org/10.12688/f1000research.9976047.1>
- Castellano LI, Giordano RM, Henriksen ES, Iversen NA (2019) Silent gene cluster activation via promoter engineering and elicitation. *PLOS Computational Biology* 32(3):893-913. <https://doi.org/10.1371/journal.pcbi.6067797>
- Okafor EJ, Fábregas CF (2021) Ecological stress and secondary metabolite production in soil bacteria. *Gut Microbes* 40(11):524-529. <https://doi.org/10.1080/03483060.2021.0000703>
- Kowalski TP, Vogel TC (2014) antiSMASH prediction of biosynthetic gene clusters across bacterial genomes. *Annual Review of Genomics and Human Genetics* 20(12):178-200. <https://doi.org/10.1146/annurev-genom-140975>
- Dubois OI, Sørensen EI, Lindqvist BR, Andersen QN, Henriksen BK (2018) Genome mining pipelines for rapid natural product pathway annotation. *EMBO Reports* 28(8):19-23. <https://doi.org/10.15252/embo.184597079>
- Quintero CC, Bianchi GN, Carrasco OF, Yilmaz QR (2020) Pangenome analysis of secondary metabolism across bacterial phyla. *BMJ* 39(10):804-814. <https://doi.org/10.1136/bmj.bmj78>
- Giordano HK, Weber AM, Schulz FB, Tanaka LM, Wachtel IJ, Andersen LO (2015) NPAtlas natural product database and structure annotation. *Annals of Internal Medicine* 2(2):824-845. <https://doi.org/10.7326/M2015-251>
- Quintero NS, Bergström QP, Nowak GQ, Sørensen QP, Andersen LL, Jørgensen FT (2020) Cryptic biosynthetic gene cluster activation through heterologous expression. *Analytical Chemistry* 7(5):338-343. <https://doi.org/10.1021/acs.analchem.20c00085>
- Petrov CT, Wachtel HK (2023) Amycolatopsis biosynthetic gene cluster diversity and antibiotic production. *Biosensors* 11(1):652-656. <https://doi.org/10.3390/biosensors375>
- Sørensen QM, Kowalski OO, Dimitrov DT (2021) Nonomuraea genus taxonomy and biosynthetic gene cluster inventory. *Microbiome* 21(6):879-888. <https://doi.org/10.1186/s718-2021-1495003-7>

21. Schulz BA, Okafor QK, Carrasco NJ (2012) Adenylation domain substrate specificity prediction in NRPS pathways. *PNAS* 24(12):622-630. <https://doi.org/10.1073/pnas.7821074>
22. Carrasco CA, Pereira HI (2015) Streptomyces secondary metabolism and antibiotic biosynthesis regulation. *Journal of Biological Chemistry* 42(3):589-604. <https://doi.org/10.1074/jbc.jbc.2015.1387104>
23. Lindqvist OR, Xu TG, Wachtel BL (2018) Marine actinomycete secondary metabolite diversity. *BMC Genomics* 13(5):745-757. <https://doi.org/10.1007/s809-2018-1247222-9>
24. Andersen AQ, Jorgensen IM, Janssens NM (2022) Whole-genome sequencing of rare actinomycete genera for drug discovery. *JAMA* 24(5):39-61. <https://doi.org/10.1001/jama.2022.9496627>
25. Pereira KQ, Engel QT, Schulz FF (2013) Environmental metagenomics for antibiotic resistance and biosynthesis surveillance. *Frontiers in Microbiology* 22(4):473-480. <https://doi.org/10.3389/fmicb.2013.4025841>
26. Dimitrov OQ, Vogel EK, Kowalski JK (2020) Comparative metabolomics for natural product pathway characterisation. *Science Translational Medicine* 29(11):684-706. <https://doi.org/10.1126/science.scitranslmed206>
27. Bianchi CG, Ueda GH, Dimitrov EJ, Henriksen PK (2018) Soil microbiome diversity across contrasting environmental gradients. *Molecular Ecology* 28(8):704-714. <https://doi.org/10.1002/mec.186>
28. Eriksson TM, Reyes TJ, Gruber LF, Pereira OE, Gruber IL, Yilmaz HL (2016) Ultrafast bootstrap phylogenetic tree construction for large genomic datasets. *Cell Reports* 44(12):565-578. <https://doi.org/10.1016/j.cell.2016.1888726.384>
29. Eriksson FM, Gruber HI, Olsen NA (2022) Arid and semi-arid soil actinomycete taxonomy and ecology. *Blood* 11(12):192-199. <https://doi.org/10.1182/blood.20221427805>
30. Iversen TG, Nowak DH, Dimitrov MM, Gruber KF, Gruber TL (2020) Heterologous expression platforms for large biosynthetic gene clusters. *FEMS Microbiology Reviews* 2(6):650-662. <https://doi.org/10.1093/femsre/femsre800>
31. Ivanova AE, Nakamura LB, Bergström FK, Castellano RF, Quintero SQ (2023) Evolutionary diversification of polyketide synthase gene families. *Journal of Applied Microbiology* 41(6):139-161. <https://doi.org/10.1177/03110257>
32. Engel NO, Tremblay SP, Weber NR, Schulz FI, Engel KQ, Olsen TM (2017) Actinomycete taxonomy revision using genome-based phylogenetics. *Journal of Molecular Biology* 43(12):821-842. <https://doi.org/10.1016/j.jmb.2017.1527035>
33. Henriksen AT, Pereira CB, Vogel IB, Gruber BM, Tremblay AT (2019) BiG-SCAPE clustering of biosynthetic gene cluster families. *The Lancet Microbe* 15(11):26-40. [https://doi.org/10.1016/S0140-6736\(2019\)30208-8](https://doi.org/10.1016/S0140-6736(2019)30208-8)
34. Pereira CH, Ueda OM, Engel JF, Yilmaz BP (2013) Genome assembly quality effects on biosynthetic gene cluster detection. *Molecular Cancer Research* 35(9):366-378. <https://doi.org/10.1158/mcr.CD-2013-508>
35. Yilmaz RB, Nowak CG, Iversen LC, Vasquez IC, Vasquez KB, Vasquez FF (2014) Genome-scale comparative analysis of *Kitasatospora* and *Actinoplanes* species. *Nature Methods* 8(9):868-886. <https://doi.org/10.1038/s41587-2014-00470-2>
36. Ivanova GG, Rousseau EB, Ueda FN, Eriksson HA, Castellano JQ (2013) Phylogenomic reconstruction of actinobacterial evolutionary relationships. *Cytogenetic and Genome Research* 7(6):595-610. <https://doi.org/10.1159/0003710810>
37. Weber QT, Castellano FG, Ferreira AI (2022) Desferrioxamine siderophore biosynthesis in actinomycetes. *New England Journal of Medicine* 3(3):319-330. <https://doi.org/10.1056/NEJMoa6313388>
38. Janssens IN, Havlicek BN, Reyes FO (2019) Synthetic biology approaches to natural product pathway refactoring. *F1000Research* 29(3):236-249. <https://doi.org/10.12688/f1000research.5832231.3>
39. Castellano QH, Rousseau KK, Bergström HQ, Wachtel BG, Rousseau EC, Nakamura EL (2023) Historical trends in antibiotic discovery from soil microorganisms. *PLOS Pathogens* 10(12):727-735. <https://doi.org/10.1371/journal.ppat.5680838>
40. Castellano AG, Henriksen NB, Jorgensen PJ (2016) Comparative genomics of *Streptomyces* species biosynthetic potential. *Gut Microbes* 20(1):784-794. <https://doi.org/10.1080/04443539.2016.0000306>
41. Henriksen CK, Janssens ME, Bergström LP, Vogel DM (2012) Natural product discovery from underexplored bacterial taxa. *Annual Review of Microbiology* 31(6):882-897. <https://doi.org/10.1146/annurev-micro-120391>
42. Weber EC, Vasquez RF (2017) Antibiotic resistance gene co-occurrence with biosynthetic gene clusters. *The EMBO Journal* 9(12):23-40. <https://doi.org/10.15252/emboj.174827664>
43. Kowalski LC, Dubois OS, Reyes IO, Engel TQ, Wachtel QM, Dubois RS (2011) Metagenomic survey of biosynthetic gene clusters in extreme environments. *BMJ* 42(12):572-583. <https://doi.org/10.1136/bmj.bmj748>
44. Andersen DD, Fábregas QH, Ferreira PJ, Reyes IJ, Iversen EP (2019) Antitumour natural products derived from actinomycete secondary metabolism. *Annals of Internal Medicine* 45(8):657-665. <https://doi.org/10.7326/M2019-696>
45. Nowak IT, Fábregas BQ, Ivanova FB, Tremblay TI, Janssens KC, Castellano EL (2013) Sequence similarity networking of non-ribosomal peptide synthetase gene clusters. *ACS Synthetic Biology* 41(6):749-771. <https://doi.org/10.1021/acs.synbio.13c00454>
46. Gruber PB, Dubois ET, Havlicek MQ, Sørensen JP, Eriksson FJ (2018) Terpene biosynthesis pathways in actinobacteria. *Genes* 28(5):627-631. <https://doi.org/10.3390/genes71>
47. Rousseau DA, Nakamura RO (2012) Ribosomally synthesised and post-translationally modified peptide biosynthesis. *Genome Biology* 39(9):783-797. <https://doi.org/10.1186/s934-2012-6669740-0>
48. Pereira KR, Petrov LG, Yilmaz AA, Henriksen HC, Lindqvist KB, Iversen QH (2016) Environmental drivers of secondary metabolite biosynthetic diversity. *PNAS* 35(9):338-360. <https://doi.org/10.1073/pnas.7329468>
49. Olsen FH, Ueda NF, Iversen SB, Fábregas CM, Olsen QL, Moreau QJ (2019) Machine learning prediction of natural product bioactivity from gene clusters. *Journal of Biological Chemistry* 40(3):706-711. <https://doi.org/10.1074/jbc.jbc.2019.8575064>
50. Lindqvist QS, Jorgensen FR, Jorgensen LO, Vogel IE, Eriksson IQ (2021) Biosynthetic gene cluster capture and expression in heterologous hosts. *Plant Molecular Biology* 4(2):581-593. <https://doi.org/10.1007/s703-2021-5746137-4>