Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative genome mining prioritizes low-reference-similarity biosynthetic gene-cluster families in a Streptomyces clavuligerus-enriched panel

Cem Boyraz

23050288@stu.omu.edu.tr

Ondokuz Mayis University Faculty of Agriculture: Ondokuz Mayis Universitesi Ziraat Fakultesi
https://orcid.org/0009-0005-7850-7754

Systematic Review

Keywords: biosynthetic gene clusters, comparative genomics, gene-cluster families, genome mining, natural products, Streptomyces clavuligerus

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10556669/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Streptomyces clavuligerus is an industrial source of clavulanic acid and other specialized metabolites, yet much of its predicted biosynthetic capacity remains uncharacterized. We applied a uniform genome-mining workflow to a purpose-built panel of 100 publicly available Streptomyces assemblies: 19 S. clavuligerus assemblies representing 16 strain labels, 17 established metabolite producers, and 64 additional species. antiSMASH 8.0.4 predicted 3,841 biosynthetic gene-cluster (BGC) regions. At a prespecified operational cutoff of ≤30% similarity to the best KnownClusterBlast/MIBiG match, 2,139 regions (55.7%) were classified as low-reference-similarity candidates, including 488 regions with no reported match. Within this curated panel, assembly-blocked analysis agreed: 79 of 96 non-tied assemblies contained more than 50% candidates (one-sided exact sign-test $P = 4.97 \times 10^{-11}$). The all-region fraction varied from 48.3% to 58.5% under ≤20% and ≤40% cutoffs, respectively, demonstrating threshold sensitivity. BiG-SCAPE 2.0.3 assigned the query regions to 2,070 gene-cluster families (GCFs), of which 1,960 lacked a MIBiG 3.1 reference member; 314 of these families were recurrent. All 19 S. clavuligerus assemblies contained KnownClusterBlast matches annotated as clavulanic acid/clavam, cephamycin C, holomycin, and tunicamycin clusters, and 51 MIBiG-unanchored GCFs were restricted to the focus group. BGC density remained higher than in the remainder of the curated panel after retaining one assembly per focus strain label. The resulting catalogue is a reproducible sequence-based prioritization resource, not evidence of chemical novelty, expression, or bioactivity.

Introduction

Members of the genus *Streptomyces* are among the most productive microbial sources of antibiotics and other specialized metabolites, with applications spanning medicine, agriculture, and industrial biotechnology (Barka et al. 2016; Genilloud 2017; Nett et al. 2009). Their filamentous development, large high-GC genomes, ecological interactions, and extensive regulatory networks support a chemically diverse secondary metabolism (van der Meij et al. 2017). The first complete genomes of the model actinomycete *Streptomyces coelicolor* A3(2) and the industrial producer *Streptomyces avermitilis* established a recurring genomic observation: the number of predicted biosynthetic loci substantially exceeds the number of metabolites known from routine cultivation (Bentley et al. 2002; Ōmura et al. 2001). This gap has made genome mining a central strategy for identifying cryptic or conditionally expressed biosynthetic pathways.

Biosynthetic gene clusters (BGCs) organize many specialized-metabolite pathways into physically linked sets of biosynthetic, tailoring, transport, resistance, and regulatory genes. Comparative analysis of these loci can support dereplication, reveal evolutionary relationships, and prioritize strains or pathways for experimental work (Medema and Fischbach 2015; Medema et al. 2021). Early actinomycete comparisons and global prokaryotic surveys showed that BGC diversity is extensive and unevenly distributed across taxa (Cimermancic et al. 2014; Doroghazi and Metcalf 2013; Doroghazi et al. 2014). Subsequent studies expanded this view across dozens to thousands of *Streptomyces* and actinomycete genomes (Belknap et al. 2020; Caicedo-Montoya et al. 2021; Chung et al. 2021; Gavriilidou et al. 2022;

Jørgensen et al. 2024; Nikolaidis et al. 2023; Otani et al. 2022). Scalable workflows such as BGCFlow now integrate pangenomics and BGC analysis, while a recent 2,371-genome study demonstrated that BGC-boundary uncertainty can materially alter estimates of family diversity (Mohite et al. 2025; Nuhamunada et al. 2024).

Modern genome-mining resources permit complementary comparisons at the region and family levels. antiSMASH predicts BGC regions and compares them with experimentally characterized clusters, whereas BiG-SCAPE groups related BGCs into gene-cluster families (GCFs) using domain content, sequence identity, and adjacency information (Blin et al. 2025; Draisma et al. 2026; Navarro-Muñoz et al. 2020). MIBiG supplies curated, experimentally supported reference clusters, and resources including BiG-FAM and BiG-SLiCE place newly predicted clusters within increasingly large sequence spaces (Kautsar et al. 2021a, b; Terlouw et al. 2023; Zdouc et al. 2025). These tools make large comparative screens reproducible, but their outputs require restrained interpretation. Low sequence similarity, absence of a reference member, or separation into a singleton GCF does not establish that a pathway is expressed or that its product is chemically or functionally novel (Medema et al. 2021; Mohite et al. 2025).

Streptomyces clavuligerus provides a compelling lineage in which to examine this distinction. The species was described as a beta-lactam producer and later identified as the source of clavulanic acid, a clinically important beta-lactamase inhibitor (Higgens and Kastner 1971; Reading and Cole 1977). Its characterized repertoire includes the cephamycin C pathway, the clavulanic acid and clavam loci, the dithiolopyrrolone holomycin cluster, and the tunicamycin pathway (Alexander and Jensen 1998; Jensen et al. 2000; Li and Walsh 2010; Tahlan et al. 2004; Wyszynski et al. 2010). The species also carries the approximately 1.8-Mb linear megaplasmid pSCL4, which is rich in specialized-metabolism loci and contributes to genome architecture (Gomez-Escribano et al. 2021; Medema et al. 2010). Genomic, regulatory, and strain-improvement studies have therefore focused heavily on clavulanic acid production (Liras and Martín 2021; Paradkar 2013; Song et al. 2010a, b), while more recent work has examined tunicamycin regulation and strain-level genomic diversity (Otur and Kurt-Kızıldoğan 2024; Ríos-Fernández et al. 2024).

The present work is not intended as the first or largest survey of *Streptomyces* biosynthetic diversity. In particular, Ríos-Fernández et al. (2024) compared 30 *S. clavuligerus* genomes, including their BGC content, and Mohite et al. (2025) provided a substantially larger genus-wide pangenome analysis. Closely related comparative studies have also documented BGC variation among selected *Streptomyces* lineages and have appeared within the scope of the present target journal (Passari et al. 2025; Vicente et al. 2018). Our narrower objective was to generate an *S. clavuligerus*-enriched, cross-genus benchmark using current antiSMASH and BiG-SCAPE releases, combine region-level reference similarity with GCF-level reference anchoring, quantify the sensitivity of the operational similarity cutoff, and release an auditable candidate catalogue. We tested whether low-reference-similarity regions were common across assemblies in this panel, whether the focus group retained characterized pathways alongside group-restricted MIBiG-unanchored GCFs, and whether BGC abundance and density varied with genome size

and sampling group. BGC-content clustering was used only to organize and visualize repertoire similarity.

Materials and methods

Study design and genome selection

All assemblies were obtained from the National Center for Biotechnology Information RefSeq/Assembly resources using NCBI Datasets command-line interface version 18.33.1 on 23 July 2026 (Goldfarb et al. 2025). No new sequence data or biological material were generated. The panel was intentionally stratified for comparative purposes and should not be interpreted as a random or exhaustive sample of the genus.

The 100-assembly panel comprised three predefined groups. The focus group contained 19 *S. clavuligerus* assemblies representing 16 reported strain labels. The second group contained one assembly from each of 17 established metabolite-producing species. The third group contained 64 additional *Streptomyces* species chosen to broaden taxonomic and genome-size coverage. Selection prioritized RefSeq records, the “Complete Genome” assembly level, reference-genome status where available, and contiguity. The final panel contained 97 records labelled “Complete Genome,” two labelled “Chromosome,” and one labelled “Contig.” Assembly size ranged from 7.57 to 12.51 Mb, GC content from 69.5% to 72.5%, and contig count from 1 to 279. The full accession list, strain labels, selection group, assembly level, genome size, N50, contig count, and GC content are provided in Online Resource 1, Table S1. Assembly quality is important when comparing BGC calls, particularly near contig boundaries, and was therefore retained as an explicit limitation rather than collapsed into a claim that every record was complete (Studholme 2016).

BGC prediction and reference matching

Each genome FASTA file was analysed independently with antiSMASH version 8.0.4 in bacterial mode (Blin et al. 2025). Gene prediction was standardized with Prodigal, and the pinned antiSMASH database bundle supplied Pfam 35.0, ClusterBlast, KnownClusterBlast, Resfams, TIGRFAM, and MIBiG-derived comparison resources (Hyatt et al. 2010; Mistry et al. 2021). The command executed for each accession was:

```
antismash --cpus 4 --databases /db --genefinding-tool prodigal
    --output-dir <accession> --output-basename <accession>
    --minimal --cb-knownclusters --cc-mibig <accession>.fna
```

Runs were distributed across six workers with four CPUs per worker. All 100 jobs completed successfully. The antiSMASH JSON output was parsed to recover accession, record identifier, region coordinates, region length, predicted product, BGC category, protocluster count, and the top

KnownClusterBlast match. For matched regions, the parser retained the reference accession, description, cluster type, similarity percentage, BLAST score, core-gene hits, and synteny score. Regions without a reported KnownClusterBlast match were recorded as “no hit.” Because database resources evolve, the region-level result is described as the top match returned by the pinned antiSMASH 8.0.4 environment rather than as an immutable property of the sequence.

GCF clustering and MIBiG anchoring

The 3,841 antiSMASH region GenBank files were clustered with BiG-SCAPE version 2.0.3 together with 2,150 antiSMASH-processed MIBiG 3.1 reference clusters (Draisma et al. 2026; Terlouw et al. 2023). Pfam-A 35.0 profiles and the default BiG-SCAPE distance cutoff of 0.30 were used:

```
bigscape cluster -i /results --input-mode recursive
        -p /db/pfam/35.0/Pfam-A.hmm
        -m 3.1 -c 16 --label strepbgc -o bg_out
```

Assignments were read from the BiG-SCAPE SQLite database and clustering tables. Query regions placed in a family at the 0.30 cutoff were joined back to antiSMASH output by record identifier and region number. Query regions not assigned to a multi-member family were retained as individual singleton GCFs so that every query region had exactly one family identifier. A query GCF was termed “MIBiG-anchored” when it contained at least one MIBiG 3.1 reference member and “MIBiG-unanchored” otherwise. These labels are sequence-clustering outcomes, not evidence for or against metabolite production.

The region and family comparisons intentionally used two related but non-identical reference states: the MIBiG-derived resources shipped with the pinned antiSMASH 8.0.4 database bundle for KnownClusterBlast/ClusterCompare analyses, and the explicitly downloaded MIBiG 3.1 antiSMASH-formatted set required by the BiG-SCAPE run. This distinction is retained in the results and provenance files because MIBiG has since advanced to version 4.0 (Zdouc et al. 2025).

Operational candidate definition and prioritization

For each BGC region, a reference-distance score was calculated as:

reference-distance score = 100 - best KnownClusterBlast similarity (%)

Regions without a reported match received a score of 100. The prespecified primary candidate definition was a best-match similarity of ≤30%, equivalent to a reference-distance score of ≥70. This threshold was used only to prioritize low-reference-similarity sequences. It is not a validated predictor of chemical novelty. Sensitivity analyses repeated the classification at ≤20% and ≤40%, and a mutually exclusive four-

bin summary separated no-hit regions, matched regions at 1–30%, matched regions at 31–69%, and matched regions at 70–100%.

Candidate regions were ranked by a transparent composite score:

priority = reference-distance score
+ 10 × [GCF is MIBiG-unanchored]
+ 10 × [GCF contains at least two query BGCs]
+ min(number of represented genomes, 20)

This heuristic rewards low reference similarity, recurrence, and cross-genome conservation. It does not estimate biological activity or drug-development potential. Family-level rankings use the highest-scoring candidate region as the representative entry, with membership and group distribution retained for interpretation.

BGC-content structure

A genome-by-GCF presence/absence matrix was constructed from the query assignments. Pairwise Jaccard distances were calculated and subjected to average-linkage hierarchical clustering. Cutting the dendrogram at a maximum of eight clusters yielded five clusters with at least two genomes and three singleton clusters. The clusters are explicitly termed “BGC-content clusters”; they are not a species phylogeny and were not used to infer evolutionary ancestry. Per-genome candidate burden was calculated as the number of regions meeting the ≤30% criterion divided by the total number of predicted regions. Because the clusters were derived from the same GCF matrix used to describe biosynthetic repertoires, cluster labels were used for ordering and visualization only, without an inferential test of candidate burden among clusters.

Statistical analysis and robustness checks

Statistical analyses were performed with SciPy version 1.17.1, pandas version 2.3.3, and NumPy version 2.4.6 (Virtanen et al. 2020). Region totals were summarized descriptively because multiple regions from the same assembly are not independent observations. For each operational cutoff, the panel-conditional prevalence analysis was blocked by assembly: an assembly was scored positive when more than half of its regions met the candidate rule, negative when fewer than half met it, and tied when exactly half met it. Ties were excluded, and a one-sided exact sign test evaluated whether positive assemblies outnumbered negative assemblies under an equal-probability null. This test does not account for phylogenetic relatedness and is not a genus-wide prevalence estimate. Pearson correlation and linear regression quantified the linear relationship between genome size and BGC count; Spearman correlation was included as a rank-based sensitivity analysis. One-sided Mann-Whitney U tests compared BGC count and BGC density between the focus assemblies and the remaining panel because the directional

comparisons were specified before analysis. All other tests were two-sided. Exact or asymptotic *P* values are reported as generated by the stated procedures.

Four primary hypotheses were evaluated: assembly-blocked prevalence at the ≤30% cutoff, the genome-size/BGC-count association, and the focus-versus-remainder comparisons of BGC count and density. A Bonferroni-adjusted alpha of 0.0125 was used as a family-wise sensitivity criterion. Because four records represented the same *S. clavuligerus* strain label (ATCC 27064), a robustness analysis retained one focus assembly per strain label, prioritizing the RefSeq reference genome, then “Complete Genome” status, then N50. This yielded 16 focus assemblies and a 97-assembly unique-label panel. The prevalence sign test and focus-group comparisons were repeated on this reduced panel. Effect sizes and full statistics are reported in Online Resource 1, Tables S9, S11, and S15.

Reproducibility and data products

antiSMASH and BiG-SCAPE were run in pinned container images on cloud compute, and all commands, parsers, versions, accessions, derived tables, and figures needed to audit the reported summaries are included in Online Resource 1. The release follows established recommendations for transparent computational research and reusable data products (Grüning et al. 2018; Sandve et al. 2013; Stodden et al. 2016; Wilkinson et al. 2016). Public genome accessions permit reacquisition of the source assemblies; the release does not redistribute NCBI genome FASTA files.

During manuscript preparation, OpenAI Codex was used to assist with language editing, literature organization, derived consistency checks, and document formatting. It was not used to generate or alter primary sequence data or the original antiSMASH and BiG-SCAPE outputs. All analyses, citations, interpretations, and final wording must be reviewed and approved by the human authors, who retain full responsibility for the submitted work.

Results

Assembly panel and predicted biosynthetic landscape

The purpose-built panel contained 19 focus, 17 known-producer, and 64 genus-diversity assemblies. antiSMASH detected 3,841 BGC regions, corresponding to a median of 38 per assembly (range 26–58). BGC count showed a modest positive linear association with assembly size (Pearson $r = 0.341$, $P = 5.2 \times 10^{-4}$; slope 2.08 BGC/Mb; Fig. 1a). The rank-based sensitivity analysis was weaker and did not reach conventional significance (Spearman $\rho = 0.178$, $P = 0.0759$), indicating that the size-count relationship should be viewed as a modest linear trend rather than a robust monotonic rule across the deliberately stratified panel.

[Fig. 1 near here]

Reference-similarity classes and threshold sensitivity

At the prespecified ≤30% cutoff, 2,139 of 3,841 regions (55.7%) were classified as low-reference-similarity candidates (Fig. 1b). This set comprised 488 no-hit regions (12.7% of all regions) and 1,651 matched regions with similarities between 1% and 30% (43.0%). A further 466 regions (12.1%) had intermediate similarity of 31–69%, whereas 1,236 (32.2%) had similarity of at least 70%. These categories are mutually exclusive and sum to the total set (Table 2). Within the curated panel, 79 of 96 non-tied assemblies contained more than 50% candidates, compared with 17 containing fewer than 50%; four assemblies were tied (one-sided exact sign-test $P = 4.97 \times 10^{-11}$).

The candidate fraction was sensitive to the operational threshold. At ≤20% similarity, 1,854 regions (48.3%) qualified and 33 of 93 non-tied assemblies exceeded a 50% candidate fraction. At ≤40%, 2,246 regions (58.5%) qualified and 91 of 99 non-tied assemblies exceeded 50% (Fig. 4b). In the 97-assembly unique-label panel, the primary ≤30% result remained: 76 of 93 non-tied assemblies exceeded 50% candidates (four ties; one-sided exact sign-test $P = 2.22 \times 10^{-19}$). This sensitivity supports use of the score for graded prioritization rather than a binary claim of confirmed novelty.

Most query GCFs lacked a MIBiG reference member

At the BiG-SCAPE 0.30 cutoff, the 3,841 query regions were represented by 2,070 query GCFs: 371 multi-member families and 1,699 singleton families. Of these, 110 GCFs contained at least one MIBiG 3.1 reference member and 1,960 (94.7%) were MIBiG-unanchored (Fig. 2). The unanchored set comprised 314 recurrent multi-member families and 1,646 singletons. Across all query GCFs, 59 occurred in at least 10 genomes; 47 of those widely distributed families were MIBiG-unanchored. Because singletons dominate the family count, the 94.7% value is interpreted as a clustering result that is sensitive to BGC boundaries and the selected distance cutoff, not as the proportion of chemically novel pathways.

[Fig. 2 near here]

The focus group retained hallmark pathways and group-restricted GCFs

KnownClusterBlast returned matches to the clavulanic acid/clavam, cephamycin C, holomycin, and tunicamycin reference clusters in all 19 *S. clavuligerus* assemblies. Thirty-six GCFs were present in at least 80% of the focus assemblies. Fifty-four query GCFs occurred only in the focus group, and 51 of these were MIBiG-unanchored. Several highly ranked families were conserved across all 19 focus assemblies, including terpene, non-ribosomal peptide synthetase, lanthipeptide, and butyrolactone-associated predictions (Table 1). Conversely, the highest-ranked broadly distributed family, GCF71, was a predicted non-ribosomal siderophore family observed in 36 non-focus genomes, showing that the prioritization also recovers cross-genus candidates.

The focus assemblies contained more predicted regions than the rest of the panel (median 40 versus 37; Mann-Whitney $P = 2.3 \times 10^{-3}$). Their median BGC density was also higher (5.14 versus 3.64 BGC/Mb; $P = 3.0 \times 10^{-11}$; Fig. 4a). The unique-strain-label sensitivity analysis retained the same direction: medians were 39 versus 37 regions ($P = 0.0121$; Cliff's $\delta = 0.358$) and 5.05 versus 3.64 BGC/Mb ($P = 6.38 \times 10^{-10}$; Cliff's $\delta = 0.965$). The density contrast was therefore not explained by repeated ATCC 27064 assemblies.

BGC-content organization provided a descriptive repertoire map

The genome-by-GCF display grouped assemblies with similar predicted BGC repertoires and placed the focus assemblies in a visually coherent portion of the ordering (Fig. 3). Candidate fractions were shown beside the category profiles to support assembly-level inspection, while Fig. 4b summarized their sensitivity to the operational cutoff. No significance test was applied across the derived BGC-content clusters because those labels arise from the same GCF representation being described. The ordering is therefore a descriptive map of repertoire similarity, not an independent grouping variable or evidence of phylogenetic signal.

[Fig. 3 and Fig. 4 near here]

Prioritized catalogue

The complete 2,139-region candidate catalogue records accession, genomic position, predicted category and product, top reference match, similarity metrics, GCF assignment, family recurrence, genome spread, and priority score (Online Resource 1, Table S8). A family-level table contains 1,349 MIBiG-unanchored GCFs represented by at least one region meeting the ≤30% criterion (Online Resource 1, Table S6). The highest priority score was 140. GCF71 reached this score through complete reference distance and distribution across 36 genomes, whereas GCF312 combined complete reference distance with recurrence in 21 BGCs across 20 genomes, including all 19 focus assemblies. The 51 focus-restricted, MIBiG-unanchored GCFs are provided separately to support lineage-focused follow-up (Online Resource 1, Table S7).

Discussion

This study provides a current, reproducible comparison of predicted BGCs in an *S. clavuligerus*-enriched panel and makes three contributions. First, it separates region-level similarity from family-level reference anchoring rather than treating either measure as direct evidence of chemical novelty. Second, it quantifies how the candidate fraction changes under alternative similarity cutoffs. Third, it releases both region- and GCF-level rankings with explicit provenance, allowing investigators to choose candidates according to recurrence, lineage distribution, predicted class, or reference distance.

The observed abundance of low-reference-similarity BGCs is consistent with earlier comparative and global analyses showing that experimentally characterized clusters occupy only a fraction of bacterial biosynthetic sequence space (Cimermancic et al. 2014; Doroghazi et al. 2014; Gavriilidou et al. 2022). However, direct percentage comparisons among studies are not appropriate without accounting for panel composition, assembly quality, antiSMASH version, database state, BGC boundaries, and clustering parameters. For example, Mohite et al. (2025) analysed 2,371 *Streptomyces* genomes and showed that boundary refinement can substantially consolidate apparent GCF diversity. The present panel is much smaller, is deliberately enriched for *S. clavuligerus*, and includes multiple assemblies from one reported strain label. Its 2,070 GCFs should therefore be understood as operational families under one reproducible workflow, not a definitive estimate of genus-wide biosynthetic diversity.

The threshold-sensitivity analysis is especially important. The 55.7% primary all-region candidate fraction falls to 48.3% at the stricter ≤20% cutoff. Within this panel, assembly-blocked analysis supported a majority of candidates at the prespecified cutoff and remained positive after retaining one focus assembly per strain label, but it did not do so at ≤20%. These results limit any majority-of-regions interpretation to the stated operational rule and curated panel and reinforce the terminology used here: low-reference-similarity candidate. Reference similarity is useful for triage, but sequence divergence may reflect boundary differences, domain rearrangement, incomplete assemblies, or database coverage rather than a new molecular scaffold. Conversely, a high similarity score does not guarantee identical regulation, tailoring, or product profiles. The family-level MIBiG anchor and recurrence terms add context, but experimental evidence remains essential.

Within the focus group, recovery of KnownClusterBlast matches annotated as clavulanic acid/clavam, cephamycin C, holomycin, and tunicamycin provides a reference-matching check consistent with established *S. clavuligerus* biology (Alexander and Jensen 1998; Jensen et al. 2000; Li and Walsh 2010; Wyszynski et al. 2010). These matches do not by themselves establish locus completeness or function. The higher BGC density within this panel also agrees with the known contribution of pSCL4 and with prior genome-based estimates of this species' specialized-metabolite capacity (Liras and Martín 2021; Medema et al. 2010; Ríos-Fernández et al. 2024). Importantly, the density contrast persisted after retaining one assembly per strain label. Nevertheless, the focus-exclusive GCFs cannot yet be called species-specific: the cross-genus comparison panel contains only 81 non-focus assemblies and is not an exhaustive taxonomic screen.

The current work complements, rather than replaces, the 30-genome *S. clavuligerus* analysis of Ríos-Fernández et al. (2024). That study examined a larger intraspecific set and integrated pangenome, structural, and production-related questions. Our contribution is the use of newer BGC/GCF software in a cross-genus comparison, dual reference anchoring, threshold sensitivity, and a candidate-oriented release. Similar recent comparative work in *Streptomyces* has demonstrated the value of integrating stress-resistance, regulatory, and ecological features with BGC mining (Augustijn et al. 2025; Passari et al. 2025; Salamzade and Kalan 2025). Adding such orthogonal evidence would be a logical next step for the present catalogue.

The BGC-content ordering is useful as a descriptive map, but it must not be confused with organismal phylogeny or used as independent evidence that candidate burden is structured. BGC repertoires are shaped by vertical inheritance, loss, duplication, recombination, and horizontal transfer, and their evolutionary histories may differ from those of core genomes (Challis and Hopwood 2003; Chevrette et al. 2020; Ziemert et al. 2014). A future analysis should infer a core-gene phylogeny, quantify phylogenetic signal, and model BGC burden while accounting for shared ancestry. The present analysis supports visualization and content-based stratification only.

The most actionable candidates are recurrent MIBiG-unanchored families with interpretable biosynthetic classes. GCF312, for example, is a RiPP-like prediction represented in 20 genomes, whereas several focus-conserved candidates are annotated as lanthipeptide, terpene, NRPS, or butyrolactone-associated loci. Their next evaluation should include boundary refinement, domain architecture, synteny, regulatory context, self-resistance or transporter genes, and comparison with larger family databases. Transcriptomics can test expression under varied conditions, while metabolomics and molecular networking can connect candidate loci to chemical features (Nothias et al. 2020; van der Hooft et al. 2020). Heterologous expression, targeted gene disruption, and purified-product characterization would ultimately be required to establish structure and activity.

Several limitations remain. The panel is selected rather than population representative; three records are not labelled Complete Genome; fragmented contigs can split BGCs; and repeated assemblies of one strain label reduce independence in assembly-level comparisons. The unique-label analysis addresses the last issue but not all strain relatedness. KnownClusterBlast similarity is an algorithm- and database-dependent summary. The antiSMASH and BiG-SCAPE analyses use different MIBiG reference states, and the BiG-SCAPE distance cutoff was not varied. Online Resource 1 supplies accessions, commands, derived tables, clustering exports, and analysis code, but does not redistribute NCBI FASTA files, raw antiSMASH JSON/GenBank output, or the BiG-SCAPE SQLite database; exact regeneration therefore requires reacquisition and rerunning the pinned workflow. The priority score is intentionally simple and contains user-defined weights. No transcriptomic, metabolomic, biochemical, or antimicrobial assay was performed. These constraints are why the release uses candidate language throughout.

Despite these limitations, the study offers a transparent starting point for experimental triage. The catalogue can be filtered without rerunning the complete workflow, while the accession list and scripts allow regeneration as databases improve. This combination of restrained interpretation and reusable outputs is valuable in natural-product discovery, where the principal bottleneck is increasingly the selection of tractable, information-rich targets rather than the mere detection of additional BGCs.

Conclusions

A uniform antiSMASH 8/BiG-SCAPE 2 workflow identified extensive low-reference-similarity and MIBiG-unanchored biosynthetic sequence diversity in a purpose-built panel of 100 *Streptomyces* assemblies. The primary ≤30% threshold classified 2,139 regions as candidates, but the candidate fraction was

threshold dependent, and the GCF count was dominated by singletons. *Streptomyces clavuligerus* assemblies retained matches to expected characterized pathways, had higher BGC density than the remainder of this curated panel in the unique-label sensitivity analysis, and contained 51 focus-restricted MIBiG-unanchored GCFs. The released catalogue is best viewed as a ranked hypothesis set for boundary refinement, comparative genomics, expression analysis, metabolomics, and experimental validation rather than as evidence of new compounds.

Declarations

Acknowledgements

[AUTHOR INPUT REQUIRED: identify any contributors who do not meet authorship criteria, or state “Not applicable.”]

Funding

The authors declare that no external funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

[ALL AUTHORS MUST CONFIRM BEFORE SUBMISSION.] The intended declaration is: “The authors have no relevant financial or non-financial interests to disclose.”

Author contributions

[AUTHOR INPUT REQUIRED. Replace with a CRediT-style statement naming every author and describing conceptualization, methodology, software, validation, formal analysis, data curation, writing, visualization, supervision, and project administration as applicable.]

Data availability

All source assemblies are publicly available through NCBI RefSeq/Assembly under the accessions listed in Online Resource 1, Table S1. Derived region- and family-level tables, sensitivity analyses, figure source data, run-status records, software versions, and analysis scripts are supplied with the submission as Online Resource 1. No new nucleotide or amino-acid sequences were generated.

Code availability

The scripts used to run and parse antiSMASH and BiG-SCAPE, together with the scripts used to regenerate the submission-level summary tables, are included in Online Resource 1.

Ethics approval

Not applicable. This computational study analysed publicly available microbial genome assemblies and did not involve humans, animals, or newly collected biological material.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Alexander DC, Jensen SE (1998) Investigation of the Streptomyces clavuligerus Cephamycin C Gene Cluster and Its Regulation by the CcaR Protein. Journal of Bacteriology 180(16):4068-4079. https://doi.org/10.1128/jb.180.16.4068-4079.1998
2. Augustijn HE, van Nassauw D, Cernat S, Reitz ZL, van Wezel GP, Medema MH (2025) Regulatory Genes as Beacons for Discovery and Prioritization of Biosynthetic Gene Clusters in Streptomyces. Biochemistry 64(13):2877-2885. https://doi.org/10.1021/acs.biochem.4c00711
3. Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Klenk HP, Clément C, Ouhdouch Y, van Wezel GP (2016) Taxonomy, Physiology, and Natural Products of Actinobacteria. Microbiology and Molecular Biology Reviews 80(1):1-43. https://doi.org/10.1128/mmbr.00019-15
4. Belknap KC, Park CJ, Barth BM, Andam CP (2020) Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. Scientific Reports 10(1):2003. https://doi.org/10.1038/s41598-020-58904-9
5. Bentley SD, Chater KF, Cerdeño-Tárraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA et al (2002) Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature 417(6885):141-147. https://doi.org/10.1038/417141a
6. Blin K, Shaw S, Vader L, Szenei J, Reitz ZL, Augustijn HE, Cediel-Becerra JDD, de Crécy-Lagard V et al (2025) antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. Nucleic Acids Research 53(W1):W32-W38. https://doi.org/10.1093/nar/gkaf334

7. Caicedo-Montoya C, Manzo-Ruiz M, Ríos-Estepa R (2021) Pan-Genome of the Genus Streptomyces and Prioritization of Biosynthetic Gene Clusters With Potential to Produce Antibiotic Compounds. Frontiers in Microbiology 12:677558. https://doi.org/10.3389/fmicb.2021.677558
8. Challis GL, Hopwood DA (2003) Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by Streptomyces species. Proceedings of the National Academy of Sciences 100(suppl_2):14555-14561. https://doi.org/10.1073/pnas.1934677100
9. Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A, Ramos-Aboites HE, Hoskisson PA, Barona-Gómez F (2020) Evolutionary dynamics of natural product biosynthesis in bacteria. Natural Product Reports 37(4):566-599. https://doi.org/10.1039/c9np00048h
10. Chung YH, Kim H, Ji CH, Je HW, Lee D, Shim SH, Joo HS, Kang HS (2021) Comparative Genomics Reveals a Remarkable Biosynthetic Potential of the Streptomyces Phylogenetic Lineage Associated with Rugose-Ornamented Spores. mSystems 6:e00489-21. https://doi.org/10.1128/msystems.00489-21
11. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA et al (2014) Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. Cell 158(2):412-421. https://doi.org/10.1016/j.cell.2014.06.034
12. Doroghazi JR, Metcalf WW (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. BMC Genomics 14(1):611. https://doi.org/10.1186/1471-2164-14-611
13. Doroghazi JR, Albright JC, Goering AW, Ju KS, Haines RR, Tchalukov KA, Labeda DP, Kelleher NL, Metcalf WW (2014) A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nature Chemical Biology 10(11):963-968. https://doi.org/10.1038/nchembio.1659
14. Draisma A, Loureiro C, Louwen NLL, Kautsar SA, Navarro-Muñoz JC, Doering DT, Mouncey NJ, Medema MH (2026) BiG-SCAPE 2.0 and BiG-SLiCE 2.0: scalable, accurate and interactive sequence clustering of metabolic gene clusters. Nature Communications 17(1):2000. https://doi.org/10.1038/s41467-026-68733-5
15. Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, Ziemert N (2022) Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nature Microbiology 7(5):726-735. https://doi.org/10.1038/s41564-022-01110-2
16. Genilloud O (2017) Actinomycetes: still a source of novel antibiotics. Natural Product Reports 34(10):1203-1232. https://doi.org/10.1039/c7np00026j
17. Goldfarb T, Kodali VK, Pujar S, Brover V, Robbertse B, Farrell CM, Oh DH, Astashyn A et al (2025) NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Research 53(D1):D243-D257. https://doi.org/10.1093/nar/gkae1038
18. Gomez-Escribano JP, Algora Gallardo L, Bozhüyük KAJ, Kendrew SG, Huckle BD, Crowhurst NA, Bibb MJ, Collis AJ et al (2021) Genome editing reveals that pSCL4 is required for chromosome linearity in Streptomyces clavuligerus. Microbial Genomics 7:000669. https://doi.org/10.1099/mgen.0.000669

19. Grüning B, Chilton J, Köster J, Dale R, Soranzo N, van den Beek M, Goecks J, Backofen R, Nekrutenko A, Taylor J (2018) Practical Computational Reproducibility in the Life Sciences. Cell Systems 6(6):631-635. https://doi.org/10.1016/j.cels.2018.03.014
20. Higgens CE, Kastner RE (1971) Streptomyces clavuligerus sp. nov., a beta-lactam antibiotic producer. International Journal of Systematic Bacteriology 21(4):326-331. https://doi.org/10.1099/00207713-21-4-326
21. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11(1):119. https://doi.org/10.1186/1471-2105-11-119
22. Jensen SE, Elder KJ, Aidoo KA, Paradkar AS (2000) Enzymes Catalyzing the Early Steps of Clavulanic Acid Biosynthesis Are Encoded by Two Sets of Paralogous Genes in Streptomyces clavuligerus. Antimicrobial Agents and Chemotherapy 44(3):720-726. https://doi.org/10.1128/aac.44.3.720-726.2000
23. Jørgensen TS, Mohite OS, Sterndorff EB, Alvarez-Arevalo M, Blin K, Booth TJ, Charusanti P, Faurdal D et al (2024) A treasure trove of 1034 actinomycete genomes. Nucleic Acids Research 52(13):7487-7503. https://doi.org/10.1093/nar/gkae523
24. Kautsar SA, Blin K, Shaw S, Weber T, Medema MH (2021) BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Research 49(D1):D490-D497. https://doi.org/10.1093/nar/gkaa812
25. Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH (2021) BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience 10(1):giaa154. https://doi.org/10.1093/gigascience/giaa154
26. Li B, Walsh CT (2010) Identification of the gene cluster for the dithiolopyrrolone antibiotic holomycin in Streptomyces clavuligerus. Proceedings of the National Academy of Sciences 107(46):19731-19735. https://doi.org/10.1073/pnas.1014140107
27. Liras P, Martín JF (2021) Streptomyces clavuligerus: The Omics Era. Journal of Industrial Microbiology and Biotechnology 48(9-10):kuab072. https://doi.org/10.1093/jimb/kuab072
28. Medema MH, Trefzer A, Kovalchuk A, van den Berg M, Müller U, Heijne W, Wu L, Alam MT et al (2010) The Sequence of a 1.8-Mb Bacterial Linear Plasmid Reveals a Rich Evolutionary Reservoir of Secondary Metabolic Pathways. Genome Biology and Evolution 2:212-224. https://doi.org/10.1093/gbe/evq013
29. Medema MH, Fischbach MA (2015) Computational approaches to natural product discovery. Nature Chemical Biology 11(9):639-648. https://doi.org/10.1038/nchembio.1884
30. Medema MH, de Rond T, Moore BS (2021) Mining genomes to illuminate the specialized chemistry of life. Nature Reviews Genetics 22(9):553-571. https://doi.org/10.1038/s41576-021-00363-7
31. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L et al (2021) Pfam: The protein families database in 2021. Nucleic Acids Research 49(D1):D412-D419. https://doi.org/10.1093/nar/gkaa913

32. Mohite OS, Jørgensen TS, Booth TJ, Charusanti P, Phaneuf PV, Weber T, Palsson BO (2025) Pangenome mining of the Streptomyces genus redefines species' biosynthetic potential. Genome Biology 26(1):9. https://doi.org/10.1186/s13059-024-03471-9
33. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M et al (2020) A computational framework to explore large-scale biosynthetic diversity. Nature Chemical Biology 16(1):60-68. https://doi.org/10.1038/s41589-019-0400-9
34. Nett M, Ikeda H, Moore BS (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. Natural Product Reports 26(11):1362. https://doi.org/10.1039/b817069j
35. Nikolaidis M, Hesketh A, Frangou N, Mossialos D, Van de Peer Y, Oliver SG, Amoutzias GD (2023) A panoramic view of the genomic landscape of the genus Streptomyces. Microbial Genomics 9:001028. https://doi.org/10.1099/mgen.0.001028
36. Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, Protsyuk I, Ernst M et al (2020) Feature-based molecular networking in the GNPS analysis environment. Nature Methods 17(9):905-908. https://doi.org/10.1038/s41592-020-0933-6
37. Nuhamunada M, Mohite OS, Phaneuf PV, Palsson BO, Weber T (2024) BGCFlow: systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. Nucleic Acids Research 52(10):5478-5495. https://doi.org/10.1093/nar/gkae314
38. Ōmura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, Takahashi Y, Horikawa H et al (2001) Genome sequence of an industrial microorganism Streptomyces avermitilis : Deducing the ability of producing secondary metabolites. Proceedings of the National Academy of Sciences 98(21):12215-12220. https://doi.org/10.1073/pnas.211433198
39. Otani H, Udwary DW, Mouncey NJ (2022) Comparative and pangenomic analysis of the genus Streptomyces. Scientific Reports 12(1):18909. https://doi.org/10.1038/s41598-022-21731-1
40. Otur ÇD, Kurt-Kızıldoğan AH (2024) Global regulator AdpA directly binds to tunicamycin gene cluster and negatively regulates tunicamycin biosynthesis in Streptomyces clavuligerus. World Journal of Microbiology and Biotechnology 40(11):360. https://doi.org/10.1007/s11274-024-04160-2
41. Paradkar A (2013) Clavulanic acid production by Streptomyces clavuligerus: biogenesis, regulation and strain improvement. The Journal of Antibiotics 66(7):411-420. https://doi.org/10.1038/ja.2013.26
42. Passari AK, Caicedo-Montoya C, Manzo-Ruiz M, Gomez-Roman MP, Ruiz-Villafán B, Rivero-Cruz JF, Singh BP, Rodríguez-Sanoja R, Sanchez S (2025) Comparative genomics of the Streptomyces genus: insights into multi-stress-resistant genes for bioremediation. World Journal of Microbiology and Biotechnology 41(12):492. https://doi.org/10.1007/s11274-025-04703-1
43. Reading C, Cole M (1977) Clavulanic Acid: a Beta-Lactamase-Inhibiting Beta-Lactam from Streptomyces clavuligerus. Antimicrobial Agents and Chemotherapy 11(5):852-857. https://doi.org/10.1128/aac.11.5.852
44. Ríos-Fernández P, Caicedo-Montoya C, Ríos-Estepa R (2024) Genomic Diversity of Streptomyces clavuligerus: Implications for Clavulanic Acid Biosynthesis and Industrial Hyperproduction.

International Journal of Molecular Sciences 25(20):10992. https://doi.org/10.3390/ijms252010992

45. Salamzade R, Kalan LR (2025) Context matters: assessing the impacts of genomic background and ecology on microbial biosynthetic gene cluster evolution. mSystems 10(3):e01538-24. https://doi.org/10.1128/msystems.01538-24
46. Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology 9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285
47. Song JY, Jensen SE, Lee KJ (2010) Clavulanic acid biosynthesis and genetic manipulation for its overproduction. Applied Microbiology and Biotechnology 88(3):659-669. https://doi.org/10.1007/s00253-010-2801-2
48. Song JY, Jeong H, Yu DS, Fischbach MA, Park HS, Kim JJ, Seo JS, Jensen SE et al (2010) Draft Genome Sequence of Streptomyces clavuligerus NRRL 3585, a Producer of Diverse Secondary Metabolites. Journal of Bacteriology 192(23):6317-6318. https://doi.org/10.1128/jb.00859-10
49. Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, Heroux MA, Ioannidis JPA, Taufer M (2016) Enhancing reproducibility for computational methods. Science 354(6317):1240-1241. https://doi.org/10.1126/science.aah6168
50. Studholme DJ (2016) Genome Update. Let the consumer beware: Streptomyces genome sequence quality. Microbial Biotechnology 9(1):3-7. https://doi.org/10.1111/1751-7915.12344
51. Tahlan K, Park HU, Wong A, Beatty PH, Jensen SE (2004) Two Sets of Paralogous Genes Encode the Enzymes Involved in the Early Stages of Clavulanic Acid and Clavam Metabolite Biosynthesis in Streptomyces clavuligerus. Antimicrobial Agents and Chemotherapy 48(3):930-939. https://doi.org/10.1128/aac.48.3.930-939.2004
52. Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, Lee S, Meijer D et al (2023) MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Research 51(D1):D603-D610. https://doi.org/10.1093/nar/gkac1049
53. van der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH (2020) Linking genomics and metabolomics to chart specialized metabolic diversity. Chemical Society Reviews 49(11):3297-3314. https://doi.org/10.1039/d0cs00162g
54. van der Meij A, Worsley SF, Hutchings MI, van Wezel GP (2017) Chemical ecology of antibiotic production by actinomycetes. FEMS Microbiology Reviews 41(3):392-416. https://doi.org/10.1093/femsre/fux005
55. Vicente CM, Thibessard A, Lorenzi JN, Benhadj M, Hôtel L, Gacemi-Kirane D, Lespinet O, Leblond P, Aigle B (2018) Comparative Genomics among Closely Related Streptomyces Strains Revealed Specialized Metabolite Biosynthetic Gene Cluster Diversity. Antibiotics 7(4):86. https://doi.org/10.3390/antibiotics7040086
56. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17(3):261-272. https://doi.org/10.1038/s41592-019-0686-2

57. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1):160018. https://doi.org/10.1038/sdata.2016.18
58. Wyszynski FJ, Hesketh AR, Bibb MJ, Davis BG (2010) Dissecting tunicamycin biosynthesis by genome mining: cloning and heterologous expression of a minimal gene cluster. Chemical Science 1(5):581. https://doi.org/10.1039/c0sc00325e
59. Zdouc MM, Blin K, Louwen NLL, Navarro J, Loureiro C, Bader CD, Bailey CB, Barra L et al (2025) MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Research 53(D1):D678-D690. https://doi.org/10.1093/nar/gkae1115
60. Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. Proceedings of the National Academy of Sciences 111:E1130-E1139. https://doi.org/10.1073/pnas.1324161111

Tables

Table 1 Highest-ranked MIBiG-unanchored candidate GCFs. “Group distribution” gives the number of represented genomes in the focus/known-producer/genus-diversity groups. The representative reference-distance score is the maximum among candidate members. Rankings are sequence-based priorities and do not demonstrate product novelty or activity

GCF	Predicted class/product	BGC regions	Genomes	Group distribution	Reference-distance score	Priority
GCF71	Other / non-ribosomal siderophore	36	36	0 / 7 / 29	100	140
GCF312	RiPP / RiPP-like	21	20	19 / 0 / 1	100	140
GCF32	Terpene / terpene	19	19	19 / 0 / 0	100	139
GCF33	Terpene / terpene	19	19	19 / 0 / 0	100	139
GCF79	Other / butyrolactone	19	19	19 / 0 / 0	100	139
GCF155	NRPS / NRPS	19	19	19 / 0 / 0	100	139
GCF311	RiPP / class-I lanthipeptide	19	19	19 / 0 / 0	100	139
GCF316	RiPP / class-I lanthipeptide	19	19	19 / 0 / 0	100	139
GCF22	Terpene / terpene precursor	22	18	18 / 0 / 0	100	138
GCF10	Terpene / terpene precursor	20	20	19 / 1 / 0	96	136

Table 2 Mutually exclusive KnownClusterBlast reference-similarity classes and candidate-threshold sensitivity. No-hit regions are included in candidate counts at every threshold

Analysis	Region count	Fraction
No reported MIBiG/KnownClusterBlast hit	488	12.7%
Matched at 1–30% similarity	1,651	43.0%
Matched at 31–69% similarity	466	12.1%
Matched at 70–100% similarity	1,236	32.2%
Candidate at ≤20% similarity, including no-hit regions	1,854	48.3%
Candidate at ≤30% similarity, including no-hit regions	2,139	55.7%
Candidate at ≤40% similarity, including no-hit regions	2,246	58.5%

Figures

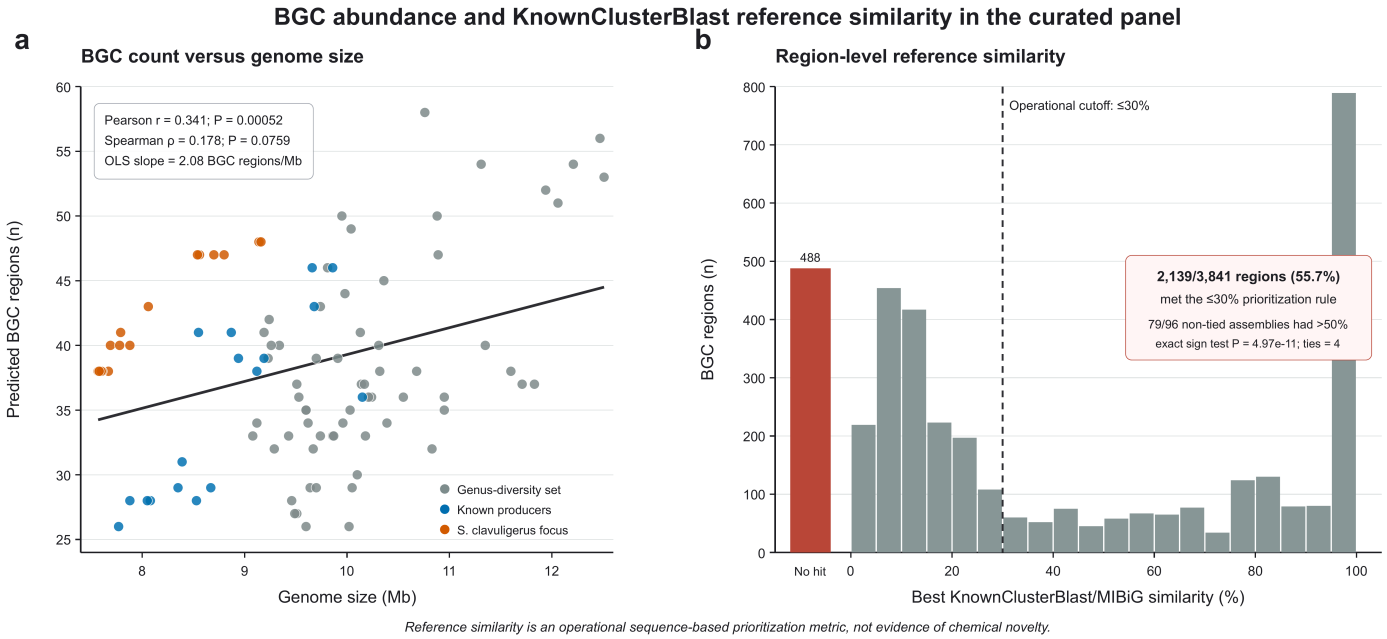


Figure 1

BGC abundance and reference similarity across the 100-assembly panel. a Predicted BGC-region count plotted against assembly size; colours denote the three predefined sampling groups and the line shows the ordinary least-squares fit. b Distribution of the best KnownClusterBlast/MIBiG similarity score. The red bar denotes 488 regions without a reported hit, and the dashed line marks the prespecified ≤30% candidate threshold. The annotation distinguishes the descriptive all-region proportion from the assembly-blocked exact sign test.

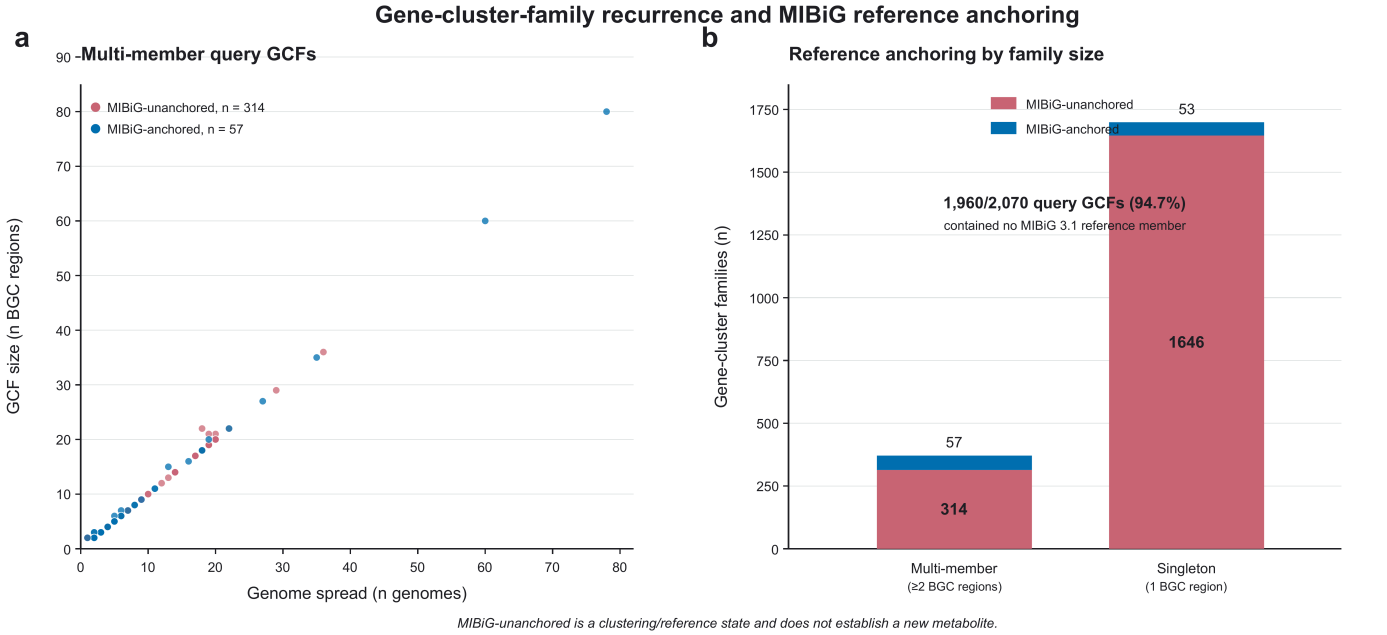


Figure 2

Query GCF recurrence and MIBiG anchoring at the BiG-SCAPE 0.30 cutoff. a Family size versus the number of represented genomes for multi-member MIBiG-anchored and MIBiG-unanchored query GCFs. b Counts of multi-member and singleton query GCFs partitioned by MIBiG anchoring. Family calls are dependent on predicted BGC boundaries, reference-set version, and clustering parameters.

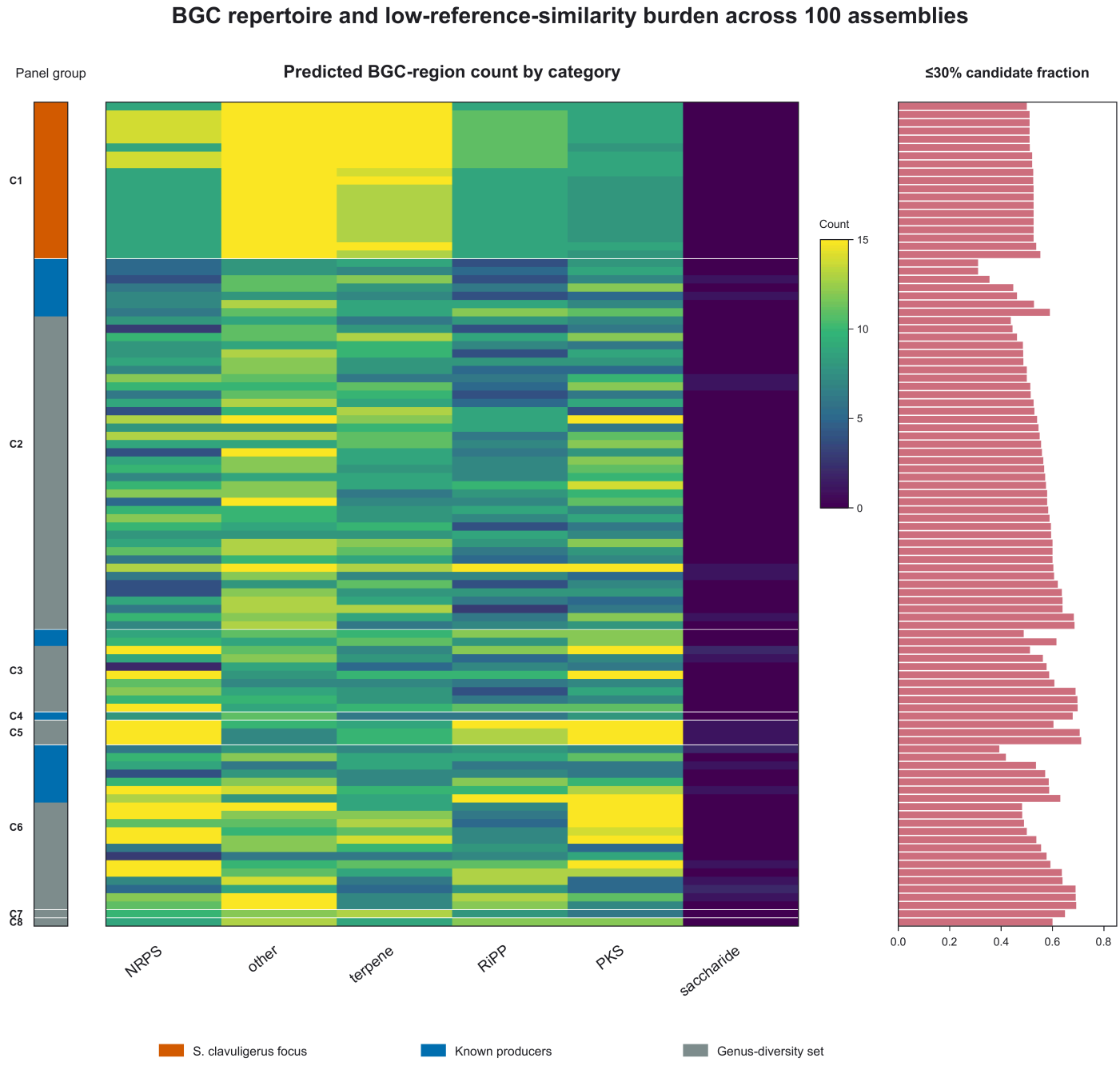


Figure 3

BGC-category profiles and low-reference-similarity burden across the 100 assemblies. Rows are ordered by average-linkage clustering of Jaccard distances calculated from the genome-by-GCF presence/absence matrix. The left annotation indicates sampling group, the central heat map shows per-

genome BGC counts by broad category, and the right bars show the fraction of regions meeting the ≤30% criterion. The ordering represents BGC-content similarity and is not a phylogeny.

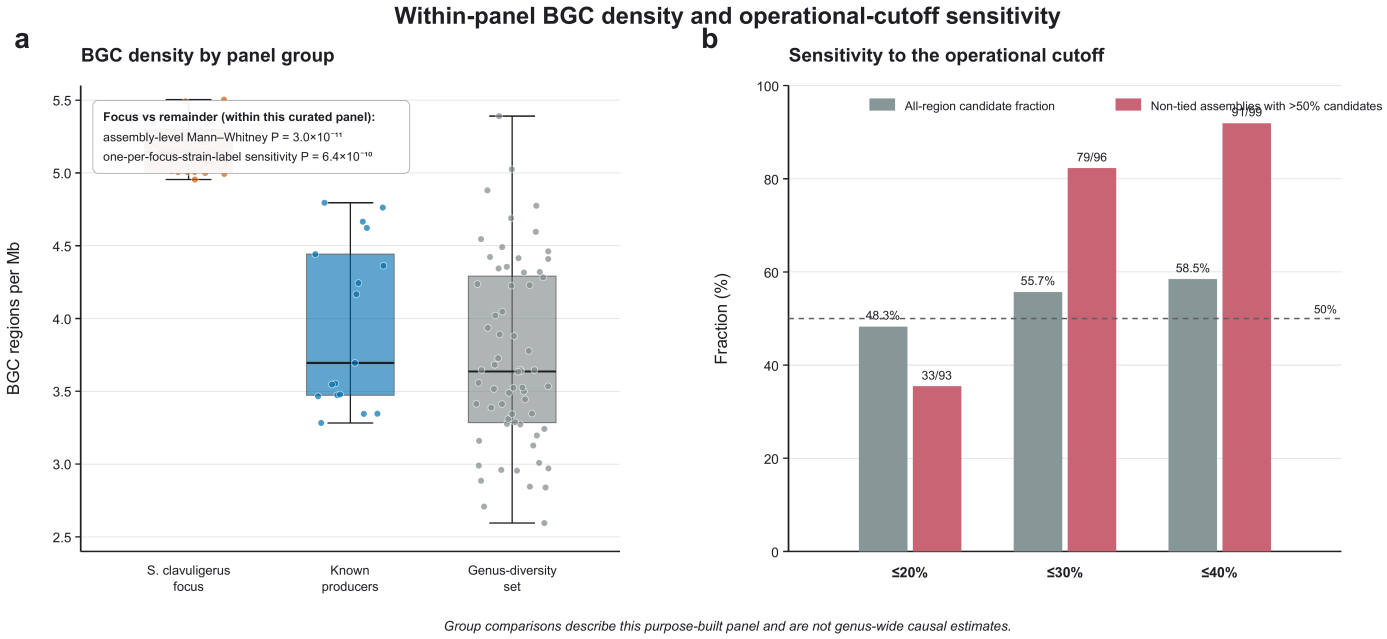


Figure 4

Within-panel BGC density and operational-cutoff sensitivity. a BGC-region density in the focus, known-producer, and genus-diversity sampling groups; points denote assemblies, while inferential annotations compare the focus group with the remainder of this curated panel. b All-region candidate fractions and the proportion of non-tied assemblies containing more than 50% candidates at ≤20%, ≤30%, and ≤40% KnownClusterBlast-similarity cutoffs. Fractions above the rose bars give positive assemblies over all non-tied assemblies; the dashed line marks 50%. Group comparisons are within-panel descriptions and are not genus-wide causal estimates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1StrepBGCMinesupportingdataandcode.zip