bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

NEW RESULTS
MICROBIOLOGY

PanBGC: A Pangenome-inspired framework for comparative analysis of biosynthetic gene clusters

Davide Paccagnella [1 2], Caner Bagci [1 2 3], Athina Gavriilidou [1 2] and Nadine Ziemert [1 2 3 *]

[1] *Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine (IMIT)*

[2] *Institute for Bioinformatics and Medical Informatics (IBMI), University of Tuebingen, Tuebingen, Germany*

[3] *German Centre for Infection Research (DZIF), Tübingen, Germany*

[*] Corresponding author Nadine Ziemert Nadine.ziemert{at}uni-tuebingen.de

bioRxiv preprint DOI: https://doi.org/10.1101/2025.08.11.669102

Posted: August 11, 2025, Version 1

Abstract

Bacterial secondary metabolites are a major source of therapeutics and play key roles in microbial ecology. These compounds are encoded by biosynthetic gene clusters (BGCs), which show extensive genetic diversity across microbial genomes. While recent advances have enabled clustering of BGCs into gene cluster families (GCFs), there is still a lack of frameworks for systematically analysing their internal diversity at a population scale. Here, we introduce **PanBGC**, a pangenome-inspired framework that treats each GCF as a population of related BGCs. This enables classification of biosynthetic genes into core, accessory, and unique categories and provides openness metrics to quantify compositional diversity. Applied to over 250 000 BGCs from more than 35 000 genomes, PanBGC maps biosynthetic diversity of more than 80 000 GCFs. To facilitate exploration, we present **PanBGC-DB** (https://panbgc-db.cs.uni-tuebingen.de), an interactive web platform for comparative BGC analysis. PanBGC-DB offers gene- and domain-level visualizations, phylogenetic tools, openness metrics, and custom query integration. Together, PanBGC and PanBGC-DB provide a scalable framework for exploring biosynthetic gene clusters at population resolution and for contextualizing newly discovered BGCs within the global landscape of secondary metabolism.

Introduction

Microbial genomes are remarkably dynamic, shaped by an ongoing interplay of gene acquisition, loss, duplication, and rearrangement[1–4]. This evolutionary fluidity allows microorganisms to adapt to diverse ecological niches, develop resistance mechanisms, and expand their metabolic capacities[5–7]. As the volume of sequenced genomes continues to grow, comparative genomics has become a key approach for uncovering patterns of gene conservation and variation across related strains[8, 9].

The shift from analysing single genomes to comparing entire groups has given rise to powerful frameworks that help organize and interpret genomic diversity[10]. Among them, the pangenome model[11, 12] offers a structured view of how genes are distributed across populations, highlighting both shared and variable features that may underlie functional and ecological differences [13, 14].

The pangenome framework formalizes genomic diversity by organizing all genes found across a group of related organisms into three categories: core genes (present in all strains), accessory genes (shared by some but not all), and unique genes (found in only one strain)[9, 15–17]. This model captures both the conserved backbone of a species and its flexible genomic reservoir, providing a foundation for understanding how populations adapt, specialize, and diversify over time[13, 18, 19].

Beyond categorization, pangenomic analysis enables broader questions about genomic plasticity[19]. One such concept is pangenome openness, which quantifies how much new genetic material continues to be discovered as additional genomes are sampled[9, 18]. In open pangenomes, gene content continues to grow with each new genome, suggesting high rates of horizontal gene transfer and ecological versatility. Conversely, closed pangenomes saturate quickly, indicating more stable, conserved genetic repertoires.[17, 20] These metrics offer crucial insight into the evolutionary dynamics and adaptive strategies of microbial populations.

Over the past decade, this framework has been widely adopted in microbial genomics[21], helping to characterize evolutionary dynamics and adaptive potential in species ranging from pathogens to environmental isolates[8, 19, 22].

In recent years, similar approaches are now being extended to the study of secondary metabolism[25]. These specialized metabolites, which play key roles in microbial interactions and biotechnological applications, are encoded by biosynthetic gene clusters (BGCs)[26] (Fig.1a). BGCs exhibit complex evolutionary histories and can be considered evolutionary entities in their own right[27]. Recent studies of individual gene cluster families (GCFs), groups of BGCs that share similar biosynthetic architectures and typically produce structurally related

compounds, have shown that BGCs evolve through gene gain, loss, and rearrangement, mirroring the dynamics seen in microbial genomes[28–30].

Advances in tools such as BiG-SCAPE[24] and BiG-SLiCE[23] now make it possible to group BGCs into gene cluster families and study variation within them. This enables comparative analyses of modular organization and functional divergence[24]. Yet, while such studies have begun to uncover the evolutionary dynamics within individual GCFs[31], a global, population-level framework for analysing BGC diversity has been lacking.

Here, we introduce the PanBGC framework (**Fig. 1b**), which applies pangenome principles to biosynthetic gene clusters by treating each GCF as a population of related clusters. This enables the systematic classification of biosynthetic genes into core, accessory, and unique components, and allows us to quantify patterns of modularity and openness across thousands of families. Applied to over 80 000 GCFs derived from more than 250 000 BGCs, the present analysis reveals that biosynthetic innovation is primarily driven by combinatorial rearrangement of conserved gene sets.

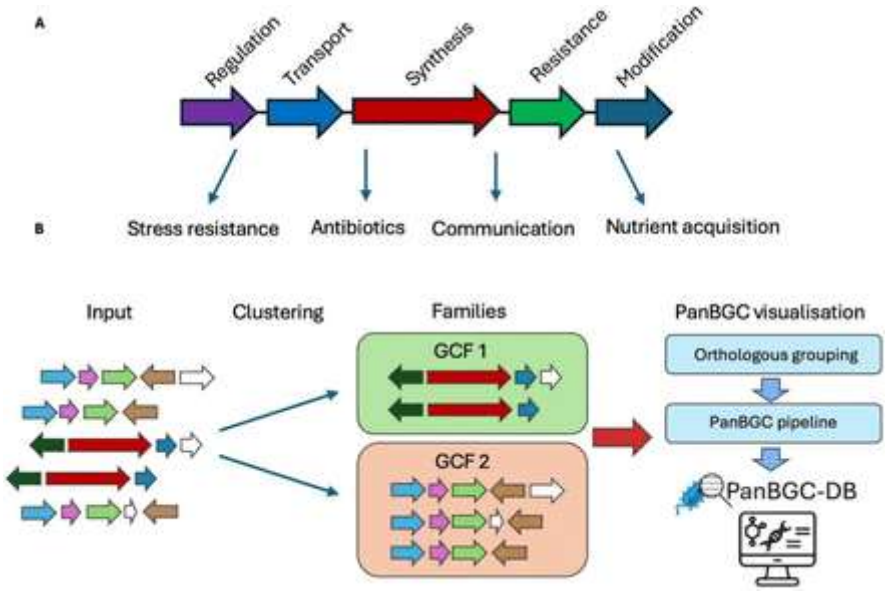

Figure 1:

Overview of biosynthetic gene cluster (BGC) function and PanBGC-DB workflow.
a Schematic representation of a typical BGC, highlighting the modular architecture with genes performing distinct biosynthetic functions. Each coloured arrow represents one gene in the BGC. The specialized metabolites produced by these BGCs serve diverse ecological functions in nature, like stress resistance, antibiotic production, intercellular communication, and nutrient acquisition. **b** Workflow of the PanBGC-DB pipeline. BGCs are first provided as input and clustered into gene cluster families (GCFs) based on sequence similarity and synteny by BiG-SLICE[**23**] and BiG-SCAPE[**24**]. Within each GCF, genes are grouped into orthologous groups to enable comparative analysis of core and accessory components. The resulting data are integrated into the PanBGC-DB platform for visualization and downstream exploration of BGC diversity and evolution.

To support broad access and interactive analysis, we present PanBGC-DB (https://panbgc-db.cs.uni-tuebingen.de), an open platform that enables global exploration of biosynthetic gene cluster families analysed using the PanBGC framework and integration of user-provided BGCs into this comparative landscape.

Methods

Biosynthetic gene cluster data were compiled from two publicly available sources. A total of 254 792 predicted BGCs were obtained from the antiSMASH-DB[32] (accessed date February 2025) by automated web crawling of all available JSON-formatted annotations. In addition, we included 2 635 experimentally validated BGCs from the MIBiG v4.0 database[33], using GenBank-format files. These datasets were used as the input for clustering the BGC in gene cluster families.

Clustering of Gene Clusters into GCFs

Biosynthetic gene clusters were clustered using a two-step strategy to define gene cluster families. First, all BGCs were processed using BiG-SLiCE[23] (v2.0.0; database release 2022-11-30) with default parameters and cut-off set to 0.7 to assign clusters to broadly defined gene cluster families based on global similarity. The computation was performed using 96 CPU cores. In the second step, each BiG-SLiCE family was subjected to fine-scale reclustering using BiG-SCAPE (v2.0-beta6)[24], which computes pairwise similarities based on domain architecture, composition and sequence similarity. BiG-SCAPE was executed with the following parameters: --input-mode recursive, --record-type region, --classify category, --gcf-cutoffs 0.4, --include-singletons, --hybrids-off, --no-trees, and --force-gbk, using the Pfam-A HMM database for domain annotation. This two-step approach allowed us to generate GCFs that reflect both broad biosynthetic relatedness and fine-grained architectural similarity.

Orthologous grouping of genes within Gene Cluster Families

For each GCF, genes from all associated BGCs were grouped into orthologous groups using ZOL v1.5.9[34], with the -r option to standardize locus tags. Functional annotations were assigned using ZOL's built-in annotation libraries. In addition to the ortholog assignments, we also extracted the consensus order, diversity, and average length of each orthologous group as computed by ZOL. The complete output was converted into a structured JSON format for downstream analysis and visualization.

Openness metrics calculation

To assess openness of the gene pool, a methodology of previous research[8, 35, 36], and was adapted for the PanBGC-DB. The code is available on the GitHub repo (https://github.com/ZiemertLab/PanBGC-DB/tree/master/Website_code/public/cluster_charts/scripts) to simulate gene accumulation across each GCF. For each family, 30 random BGC sampling permutations were generated to compute the cumulative number of genes as BGCs were incrementally added. The resulting gene accumulation curves were fitted to the Heaps' Law model $y = k * x^{\gamma}$ using three approaches:

1. Standard log-log linear regression, which applies linear regression on log-transformed gene and BGC counts.
2. Weighted regression, which emphasizes later sampling points by assigning increasing weights, improving fit when early data is noisy.
3. Non-linear optimization, which directly minimizes squared error using gradient descent without log transformation.

The best-fitting model was selected based on the highest R^2 value. The final γ value was used to quantify the openness of each GCF.

For the analysis based on unique gene composition in a BGC, the same simulation strategy and model-fitting approaches were applied. Instead of tracking the cumulative total of all genes, only the appearance of a BGC with a unique gene composition (not including gene order) at each sampling step was counted. The resulting curves capture the rate at which novel genes appear as more BGCs are added. As before, the Heaps' Law model was fit using all three approaches, and the model with the best R^2 was selected to report the final γ and k values.

To evaluate whether openness estimates based on unique gene composition differed significantly from those based on total gene count, we applied a Kruskal–Wallis rank-sum test on the corresponding γ values across all GCFs. This non-parametric test was used to assess differences in distributions without assuming normality.

Phylogenetic tree construction

Gene trees for each orthologous group were generated by ZOL. The BGC tree is a coalescent tree inferred by astral-pro3 version 1.19.3.5[37] using all OG trees of the respective GCF.

Website visualization

The interactive web platform was developed using JavaScript for both frontend and backend logic, with HTML and CSS for structure and styling. Visualizations were implemented using

different libraries. The input data is provided in JSON, Nexus, and CSV formats (Supplementary Tab. 1). Pre-processed data is dynamically loaded and rendered client-side. Code for the website is available in the GitHub repo (https://github.com/ZiemertLab/PanBGC-DB/tree/master/Website_code/).

Querying with user-provided BGCs

Users can upload a single BGC in GenBank format to identify the most similar GCF. For each GCF, a theoretical maximum BGC was constructed by merging all orthologous gene groups observed across its member clusters (https://github.com/ZiemertLab/PanBGC-DB/blob/master/Max_BGC.py). These representative BGCs were compiled into a searchable database using the makedb module from cblaster v1.3.0[**38**]. Upon upload, the user's BGC is queried against this database using the cblaster search function to determine the best-matching GCF (Supplementary Fig.3).

Visualization of user-generated GCFs

The Python pipeline used to process precomputed GCFs on the website was adapted to support user-provided data. Users can run this pipeline locally on one or more of their own GCFs to generate a structured JSON file compatible with the platform. By uploading this file, users can visualize their GCFs using the same interface and features as the preloaded dataset. The pipeline is available under https://panbgc-db.cs.uni-tuebingen.de/data/Scripts.zip.

Results

Building on the conceptual framework introduced before, we adapted the pangenome model to the analysis of BGCs. In this context, each gene cluster family is treated as a population-level unit analogous to a microbial species. This analogy is grounded in the fact that BGCs grouped into the same GCF share high architectural and functional similarity[**39**], often producing structurally related but still distinct compounds. Like individual genomes within a species, the BGCs within a GCF represent naturally occurring variants that reflect evolutionary diversification around a conserved biosynthetic theme[**29**]. This enables the application of core, accessory, and unique gene classification to BGCs, allowing us to investigate their diversity not as isolated cases but as structured populations with internal variability.

Construction and Clustering of the PanBGC-DB Dataset

To build a comprehensive dataset capturing bacterial BGC diversity, data from two major resources was compiled: the antiSMASH database[**40**], providing 254 792 BGCs from 35 726

bacterial genomes, and the MIBiG repository[33], contributing an additional 2 635 experimentally characterized BGCs.

To generate refined GCFs a two-stage clustering strategy was used. Initial clustering using BiG-SLiCE grouped 257 427 BGCs into 21 528 GCFs. Notably, 15 443 GCFs consisted of singletons (BGCs that did not cluster with any other BGC), representing ~6% of the total BGC dataset. To further refine family delineation each GCF was subsequently clustered using BiG-SCAPE v2.0, yielding a final set of 80 698 unique families. Of these, 58 700 GCFs were singletons, representing ~23% of the total BGC dataset. This indicates that the second clustering step identified additional fine-scale distinctions among loosely related BGCs (Fig.2a).

Excluding singleton families, the average size of the refined GCFs was 9.4 BGCs (Supplementary Fig.1). These BGCs spanned 58 classes including nonribosomal peptide synthetases (NRPS), polyketide synthases (PKS), terpenes, ribosomally synthesized and post-translationally modified peptides (RiPPs), saccharides, others and diverse hybrid combinations thereof. Non-hybrids BGCs have the most GCFs, while hybrids of three or more different classes build the least GCFs (**Fig. 2b**)

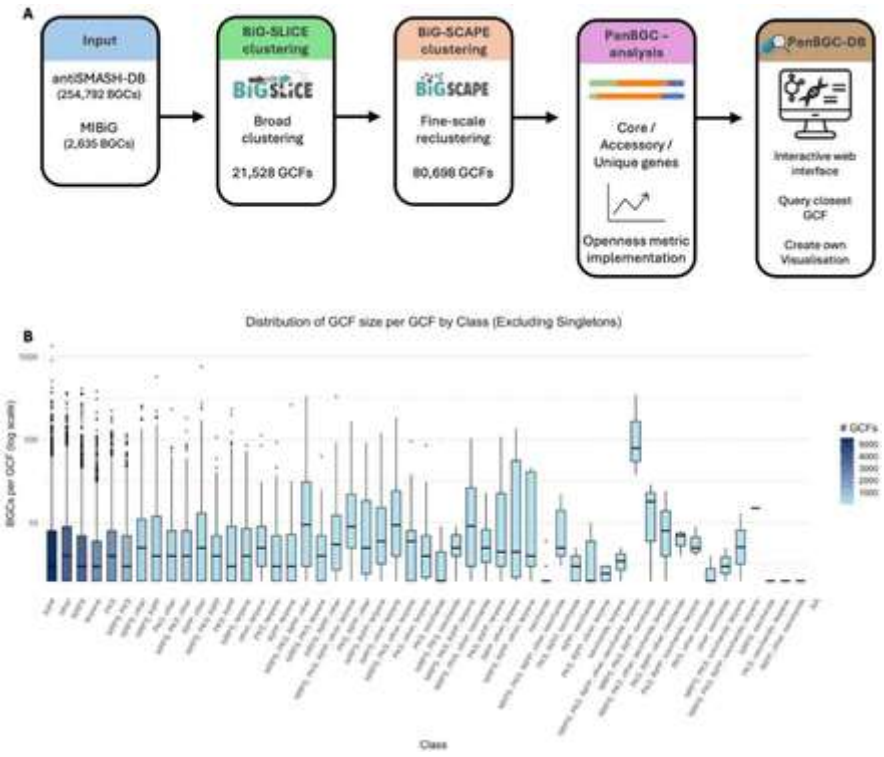

Figure 2:

PanBGC-DB pipeline overview and BGC family size distribution.
a BGCs from antiSMASH-DB and MIBiG are first clustered using BiG-SLiCE. Broad GCFs are further refined using BiG-SCAPE. The PanBGC analysis framework then identifies core, accessory, and unique genes within each family and implements openness metrics. The results are integrated into the PanBGC-DB platform, which provides an interactive web interface for querying, visualization, and user-uploaded comparisons. **b** Distribution of GCF sizes across classes, excluding singletons. Each boxplot shows the number of BGCs per GCF (log scale) for a given class, with colour gradient indicating the number of GCFs per class.

Core and Accessory Gene Classification within GCFs

Following standard pangenomic practice[8, 21, 41], all genes from BGCs belonging to the same GCF were clustered into orthologous gene groups (OGs). Each OG was subsequently classified as core, accessory, or unique according to its frequency across the clusters in that family. In bacterial pangenome studies, genes present in at least 95 % of genomes are typically considered core[12, 20, 21]. Because BGCs contain far fewer genes, applying a 95 % cutoff risks inflating the core fraction when only a handful of clusters differ and so loosing valuable information about GCF diversity. To avoid this bias, we adopted a more stringent criterion: an OG was deemed core only if it occurred in 100 % of BGCs within the family. OGs present in more than one but not all clusters were classified as accessory, whereas OGs exclusive to a single BGC were labelled unique. This stricter definition preserves meaningful distinctions between conserved and variable functions and provides a clearer view of genetic diversity within each GCF.

Functional Domain-Level Analysis

To explore trends in the functional roles of genes within BGCs, we analysed the Pfam domain annotations[42] associated with core, accessory, and unique orthologous groups across biosynthetic categories. Core genes were predominantly linked to essential enzymatic activities required for metabolite biosynthesis. For example, in NRPS clusters, condensation (C) domains were consistently classified as core due to their universal role in peptide bond formation[43]. Similarly, KS (ketosynthase) domains were frequently core in PKS systems[44]. In addition to biosynthetic enzymes, transporter-related domains were also commonly identified among core genes, reflecting the importance of compound export in BGC function.

Accessory genes displayed greater functional variability and were often associated with tailoring reactions or regulatory roles. By examining which accessory genes are recurrently present across BGCs within the same biosynthetic class, this analysis also highlights conserved auxiliary functions that may contribute to structural diversification or pathway regulation.

A complete overview of domain frequency distributions for core, accessory, and unique genes across all biosynthetic categories is available at https://panbgc-db.cs.uni-tuebingen.de/stats under the gene stats tab.

Compositional Insights and Boundary Considerations in Intra-Family BGC Comparisons

One of the key advantages of applying a pangenomic framework to BGCs is the ability to systematically compare gene composition within a gene cluster family and link these differences to chemical diversity. For instance, in a GCF containing experimentally validated Malleobactin A–D[45–48] and Ornibactin C4–C8 clusters[49], we identified clear substructuring based on accessory genes that correlate with distinct chemical features. Malleobactin-producing BGCs encode a formyltransferase absent from Ornibactin clusters, while the Ornibactin subset consistently features two acyltransferase genes not found in Malleobactin variants (**Fig. 3a**). These accessory elements are mutually exclusive and align with known structural modifications in their respective siderophores (**Fig. 3b-c**), highlighting the utility of the framework in pinpointing biosynthetic genes responsible for functional diversification. This comparative resolution also opens up possibilities for synthetic biology applications, where distinct accessory genes can be rationally combined to design novel hybrid clusters with tailored chemical output. Expanding from this example, systematic identification of gene-function relationships across thousands of GCFs paves the way for automated prediction of accessory gene functions and chemical modifications based on gene content, potentially accelerating the discovery and engineering of novel bioactive compounds.

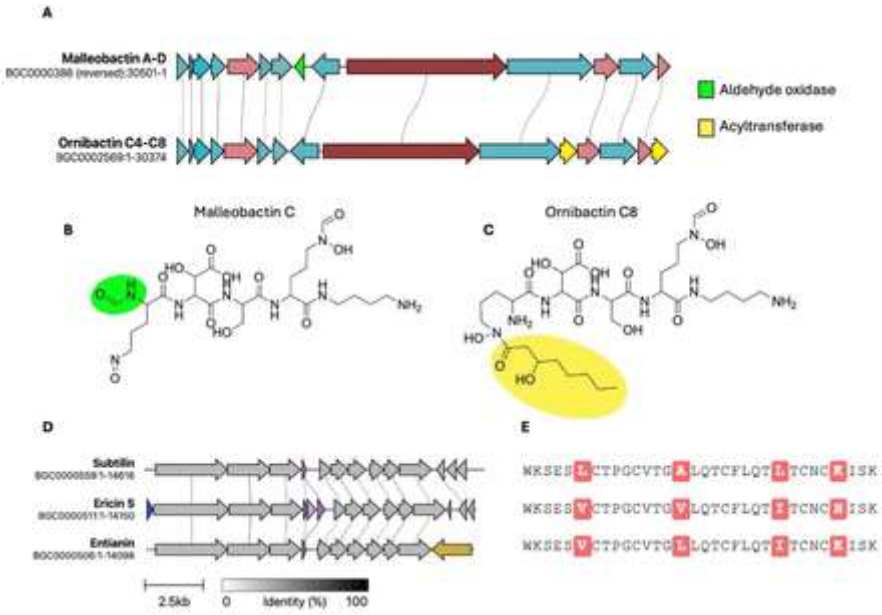

Figure 3:

Gene cluster variation and structural diversification in related siderophores.
a Comparative gene cluster alignment of the Malleobactin A–D and Ornibactin C4–C8 BGCs from the MIBiG database. Homologous genes are connected by grey lines. While both clusters share a conserved core biosynthetic architecture, they differ in accessory genes: the Malleobactin cluster encodes an aldehyde oxidase (green), whereas the Ornibactin cluster features additional acyltransferases (yellow). **b,c** Chemical structures of Malleobactin C and Ornibactin C8. Structural differences introduced by the respective accessory genes are highlighted with the respective colors. The remaining structure is shared by both compounds. **d** Compositional differences of the Subtilin, Ericin S and Entianin BGC. In blue a unique Ericin S transporter, in purple the duplicated structural genes and in brown the unique regulator of Entianin. **e** Core peptide alignment of Subtilin, Ericin S and Entianin.

However, interpreting such compositional differences also requires caution due to limitations inherent in automated boundary predictions. As the BGC definitions in our dataset rely on antiSMASH outputs, the predicted cluster boundaries are often influenced by synteny and genomic context.[32] This can lead to the inclusion of conserved flanking regions that are not functionally related to the BGC, particularly in closely related species where genomic neighbourhoods are similar. This challenge is exemplified by the GCF containing the enterobactin pathway. The MIBiG entry for enterobactin (BGC0002685) and amonabactin P 750 (BGC0001502) define compact and experimentally validated clusters[50, 51], yet related BGCs detected from antiSMASH extend considerably beyond this boundary, capturing additional genes on both ends (**Fig.4**). It is unclear whether these genes represent novel accessory functions or unrelated genomic content.

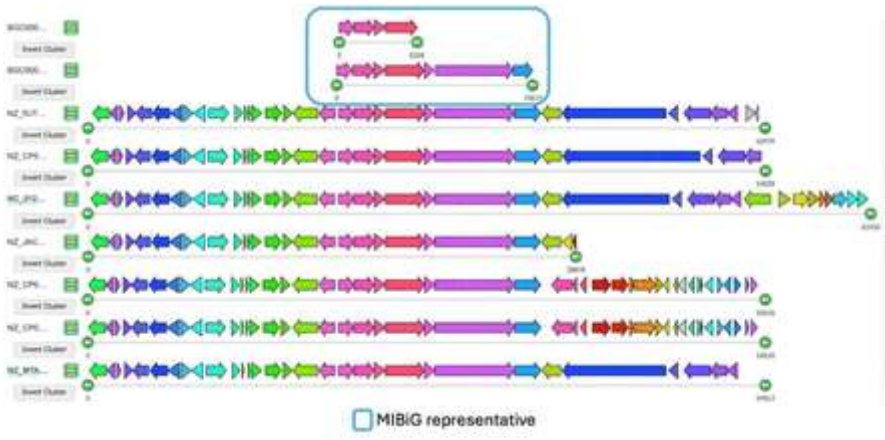

Figure 4:

Gene cluster diversity within the Enterobactin and amonabactin P 750 family.
Comparative visualization of BGCs belonging to the Enterobactin gene cluster families. Each row represents a single BGC, with genes colored by orthologous group and arrows indicating gene orientation. MIBiG reference cluster is highlighted with a blue box.

Despite this limitation, the comparative approach offers a strategy to refine cluster boundaries. When antiSMASH-predicted clusters can be directly compared with a curated MIBiG entry, shared core regions can be confidently delineated. Accessory genes that are consistently absent in MIBiG but variably present in antiSMASH predictions may be deprioritized for functional interpretation. In this way, the pangenome model not only supports the discovery of biosynthetic variability but also provides a framework for improving cluster annotation fidelity through comparative context.

Gene Composition Openness Reveals Modular Innovation in BGCs

To quantify the degree of compositional variability within gene cluster families, we adapted openness metrics from microbial pangenome analysis using Heaps' law γ-values [**8**, **12**]. These metrics assess whether the gene content within a family is relatively saturated (closed) or continues to expand with the inclusion of new members (open), reflecting either genetic stability or ongoing diversification. Unlike species-level pangenomes, where hundreds to thousands of genomes are typically analysed, most GCFs consist of only a few BGCs. To address this constraint, we implemented modified curve fitting strategies tailored to smaller sample sizes and restricted openness analysis to GCFs with at least three BGCs, as reliable γ-value estimation was not feasible for smaller families.

To capture different dimensions of diversity, openness was defined in three distinct forms. First, gene-based openness quantifies the increase in the total number of orthologous groups (OGs) with each additional BGC, reflecting expansion of the overall gene repertoire (**Fig. 5a**). Second, composition-based openness measures how consistently OGs are reused across BGCs

in a GCF, indicating variability in how subsets of the PanBGC are deployed. This captures the rate at which novel gene combinations (distinct sets of orthologous groups) appear with each additional BGC, regardless of gene order (**Fig. 5b**). Third, we also considered the rate at which entirely novel genes, those not previously observed in a family, appear with the addition of new BGCs. However, our main focus remained on gene- and composition-based openness, which together capture both repertoire size and modular flexibility.

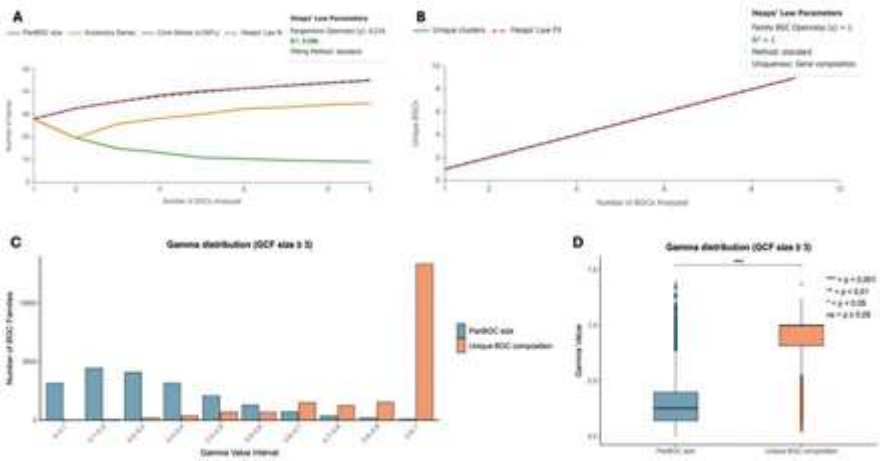

Figure 5:

Openness metrics adapted for biosynthetic gene clusters.
a Example Heaps' Law curve showing the accumulation of total (blue curve), core (green curve), and accessory genes (orange curve) as additional BGCs are sampled from a representative GCF. The red dotted line represents the fitted curve. **b** Heaps' Law modeling of BGC uniqueness using gene composition data. The red dotted line represents the fitted curve. **c** Histogram of gamma values for BGC families with ≥3 members, calculated using either PanBGC size (blue) or unique gene composition (orange). **d** Boxplot comparing gamma distributions between the PanBGC size and unique BGC composition metrics for GCFs with ≥3 BGCs n =14 716.

Among the 80 698 GCFs in the final dataset, 14 716 families contained at least three BGCs and were retained for openness calculations. Using gene-based openness (i.e., increase in PanBGC size with each added BGC), the average γ-value across all biosynthetic categories was 0.286. According to established thresholds (γ < 0.3: closed; 0.3-0.6: intermediate; >0.6: open)[**8**, **36**], this indicates that most GCFs are closed, with relatively stable gene repertoires. This suggests a limited influx of novel genes as more BGCs belonging to the same GCF are sampled. In contrast, openness based on gene composition diversity within BGCs (i.e., how consistently subsets of PanBGC genes appear across clusters) showed an average γ-value of 0.841, indicating a high degree of structural variability (**Fig. 5c**). These findings highlight that gene composition reshuffling, rather than acquisition of new genes, is the dominant driver of diversity within GCFs. This suggests that in natural systems, BGC diversity emerges primarily through modular reorganization of existing genes rather than through frequent incorporation of entirely novel genes, reinforcing the idea that structural variability is a key evolutionary mechanism in BGC innovation. A Kruskal-Wallis test comparing γ-values from PanBGC size-

based and composition-based openness confirmed a statistically significant difference ($p < 0.001$) between the two distributions (Fig.5d).

The *Bacillus* lanthipeptide family (family 415_FAM_00315) exemplifies this pattern. The clusters belonging to this family encode for subtilin, ericin S and entianin, which are antimicrobial peptides with potential applications as natural food preservatives and medicine[**52**]. While all three clusters share conserved core genes for lanthionine ring formation they exhibit compositional differences: the ericin cluster contains duplicated structural genes separated by an inserted *lanC* fragment, while the entianin cluster features different regulatory genes compared to the standard subtilin architecture (**Fig. 3d-e**)[**53–55**]. This family shows low gene-based openness ($\gamma = 0.195$) reflecting limited novel gene acquisition, yet a relative high composition-based openness ($\gamma = 0.641$) indicating modular rearrangement of existing genes.

To ensure that openness values were not biased by differences in family size, we also assessed correlations between γ-values and the number of BGCs per GCF (Supplementary Fig.2). No significant association was detected, confirming that openness metrics robustly reflect compositional dynamics independent of cluster count.

Evolutionary Dynamics Assessed Through Tanglegram Analysis

To explore the evolutionary relationships among gene clusters within each family, we applied tanglegram-based comparisons between gene trees and BGC trees. Phylogenetic trees represent evolutionary relationships, with closely related sequences grouped on nearby branches. In our analysis, gene trees show the evolutionary history of individual orthologous groups (how each gene family evolved), while BGC trees represent the overall evolutionary relationships between entire gene clusters (how the complete BGCs are related to each other). For all GCFs with at least three BGCs, individual gene trees were constructed for each orthologous group and aligned against the BGC tree. These tanglegrams visually represent the congruence between gene-level and cluster-level relationships, enabling the identification of structural similarities or differences across BGCs in a family.

In some GCFs, gene trees closely mirrored the coalescent BGC tree, indicating structural consistency and shared evolutionary trajectories. In contrast, tanglegrams with extensive crossover lines suggest high plasticity of the GCF with possible horizontal gene transfers. Families with many crossover lines suggest higher evolutionary plasticity and potential for gene module exchange or recombination, while minimal crossovers indicate structurally conserved BGCs.

Web-Based Visualization and User Tools for Exploring BGC Diversity

To make this conceptual framework accessible and interpretable, we developed PanBGC-DB (https://panbgc-db.cs.uni-tuebingen.de/), an interactive web platform for exploring biosynthetic gene cluster families (**Fig. 6 a-h**). The website allows users to browse thousands of precomputed GCFs (**Fig. 6a**) and interactively visualize their internal diversity. By presenting the results of the pangenome adaptation in an intuitive, visual format, PanBGC-DB provides a practical entry point into the population-level analysis of BGCs.

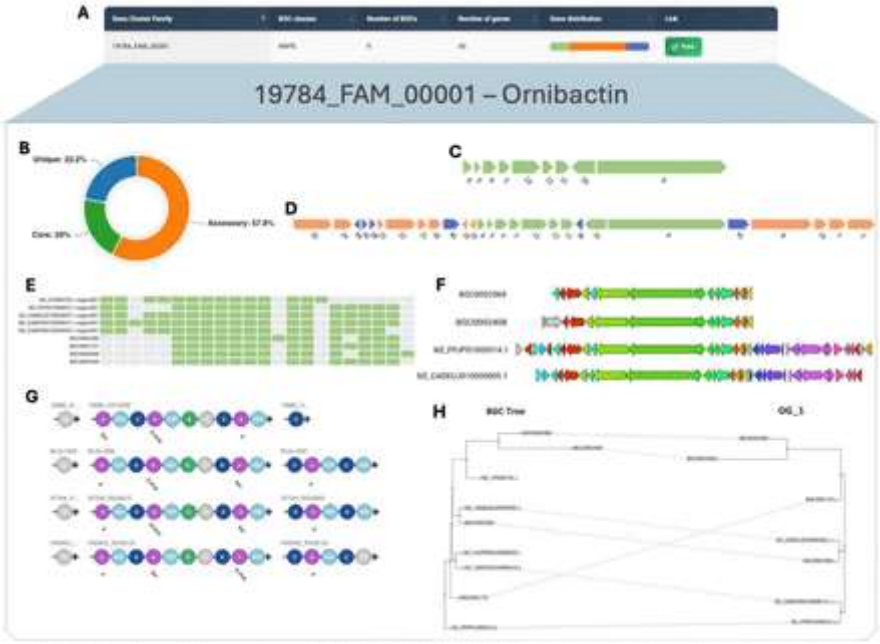

Figure 6:

Multiple screenshots from the PanBGC-DB web interface for the Ornibactin and Ornibactin GCF
a Overview table showing metadata of gene cluster families, including class, number of BGCs, total genes and summary of gene distribution (core / accessory / unique). **b** Donut plot illustrating the proportion of core, accessory, and unique genes in this GCF. **c** Visual representation of the core BGC. **d** Visual representation of the maximum BGC (PanBGC). **e** Gene presence-absence heatmap across all BGCs within the family. **f** Gene cluster comparison plot showing structural conservation across BGCs. **g** Domain architecture viewer showing module organization in biosynthetic genes across the family. **h** Interactive tanglegram linking the phylogenetic tree of BGCs (left) with that of a selected orthologous group (right).

Each GCF page offers a suite of interactive modules enabling in-depth exploration of its composition and structure. Users can adjust the core gene threshold dynamically (e.g., from 100% to lower cutoffs), which updates the classification of core, accessory, and unique genes across the cluster family (**Fig. 6b**). Based on this threshold, the platform reconstructs a core

BGC (comprising only genes that meet the cutoff) (**Fig. 6c**) and a maximum BGC (all genes observed in any cluster) (**Fig. 6d**), both displayed in consensus gene order. These representations allow for immediate insight into conserved biosynthetic cores versus variable extensions.

To evaluate diversity within a GCF, the platform provides openness visualizations, including curves for both the increase in PanBGC size and compositional diversity as more BGCs are added (**Fig. 5 a-b**). These charts help interpret whether a family is genomically saturated (closed) or still expanding (open), capturing both the total repertoire and the variability in gene usage.

Further modules include a presence-absence heatmap showing the distribution of orthologous groups across BGCs in a family (**Fig. 6e**), and a clinker[**56**] inspired synteny view that groups structurally identical BGCs and aligns them for side-by-side comparison (**Fig. 6f**). For modular BGCs such as NRPS and PKS, domain-level annotations are visualized, allowing users to assess differences in enzymatic architecture and modular composition across clusters (**Fig. 6g**).

To explore evolutionary dynamics, the platform features tanglegram visualizations, comparing individual gene trees to the coalescent BGC tree (**Fig. 6h**). These allow users to assess evolutionary congruence or structural rearrangements within each GCF. A high number of crossover lines between trees suggests evolutionary plasticity or potential horizontal gene transfer, while low crossover indicates conserved gene arrangements.

Beyond internal visualization, PanBGC-DB also includes two tools that extend its utility to custom datasets and external queries:

1. Custom BGC Visualization Pipeline

Users can download a Python-based pipeline that processes user-provided BGCs into PanBGC-style visualizations. With a single command, the pipeline generates interactive displays including core/accessory gene maps, heatmaps, domain alignments, and presence-absence matrices. This allows researchers to analyse their own gene clusters within the same conceptual framework used in the public database.

2. Query Interface for Cluster Matching

A built-in search function enables users to upload a BGC of interest and identify the closest matching GCFs in the PanBGC-DB reference dataset using the cblaster tool. The query returns the most similar families based on gene content similarity, allowing users to contextualize their BGC within broader patterns of diversity and conservation (Supplementary Fig.3).

Together, these features make PanBGC-DB not only a static repository of precomputed clusters, but also a dynamic environment for hypothesis generation, comparative analysis, and integration of user-generated data within a population-level framework of BGC diversity.

Discussion

In this study, we introduce a conceptual shift in the analysis of biosynthetic gene clusters by adapting the pangenome framework to the level of gene cluster families. Rather than treating BGCs as isolated genomic islands, we organize them into structured populations of related clusters, enabling comparative analyses grounded in evolutionary principles. This approach positions each GCF analogously to a microbial species in classical pangenomics[12, 16], allowing us to systematically partition biosynthetic diversity into core, accessory, and unique components [12, 57]. By doing so, we provide a scalable framework for interpreting the modular architecture of secondary metabolism across large genomic datasets and open the door to population-level reasoning in a domain traditionally dominated by individual-case studies. To translate this conceptual shift into an accessible resource, we developed PanBGC-DB, an interactive web platform that enables users to explore population-level diversity of BGCs across thousands of gene cluster families.

While other resources exist for organizing biosynthetic gene clusters at scale [23, 24], including the widely used BiG-FAM database [39], these platforms have primarily focused on high-level classification and dereplication of BGCs across global datasets. Workflow tools like BGCflow similarly examine BGC distribution across pangenomes, treating entire clusters as discrete genomic units[58]. In contrast, PanBGC-DB is tailored toward the in-depth analysis of variation within gene cluster families. By adopting a population-genomic framework and applying a two-step clustering strategy, PanBGC-DB generates more granular GCFs, allowing closely related BGCs to be studied as coherent evolutionary units. This finer resolution is not intended to replace broader classification schemes, but rather to support the complementary goal of revealing how modular rearrangement, gene loss, and duplication shape the diversity within biosynthetic lineages. In doing so, PanBGC-DB extends the comparative scope of BGC analysis beyond mere grouping, towards understanding the internal dynamics that drive natural product diversification.

The openness metrics for BGC gene pool and BGC composition introduced here provide a valuable framework for interpreting the evolutionary and functional potential of biosynthetic gene clusters. Our analysis revealed that while most GCFs exhibit limited acquisition of entirely novel genes, indicating a relatively closed gene content, the combinations in which these genes appear across BGCs of the same GCF are highly variable. This compositional fluidity suggests that the primary driver of BGC diversification is not the continual integration of new biosynthetic genes, but rather the modular reshuffling of a conserved gene set. This

modularity is exemplified by the *Bacillus* lanthipeptide family, where conserved ring-forming enzymes are coupled with variable resistance, transport, and regulatory modules. Such flexibility allows organisms to fine-tune metabolic pathways, generate structurally distinct metabolites, and adapt to shifting ecological contexts, all without expanding their gene pool. [53, 59–61] These findings position gene composition plasticity as a central mechanism of biosynthetic innovation and underscore the value of viewing GCFs as evolutionary populations rather than isolated units. Thus, modular reshuffling allows organisms to repurpose existing biosynthetic elements into new configurations, enabling the generation of diverse metabolites without the need to acquire entirely novel genes. This strategy not only supports metabolic adaptability across ecological niches[60] but may also facilitate the emergence of novel functions by reassembling familiar parts in previously untested ways.

The ability to dissect biosynthetic gene clusters at the population level opens new opportunities for natural product discovery and design. By clearly distinguishing conserved core genes from variable accessory components within GCFs, PanBGC-DB helps researchers identify families with modular flexibility, which often correlates with chemical novelty[62]. This makes it possible to prioritize gene cluster families that exhibit unexplored biosynthetic potential, particularly those with unusual combinations of biosynthetic domains or accessory genes. At the same time, the structured view of naturally (co-) occurring genes[63] provides a valuable basis for synthetic biology, offering a blueprint for reconstructing or modifying pathways using genes present. By using gene combinations that are already observed together in nature and present in the same gene cluster family, synthetic biology can draw on pathways that are more likely to be functionally compatible. [29, 64–66] A concrete example of this design-guiding potential is illustrated by the Ornibactin/Malleobactin gene cluster family. Using PanBGC-DB, we identified accessory gene differences, specifically two acyltransferases in Ornibactin C4–C8 and a formyltransferase in Malleobactin A–D, that correlate with known structural variations between the compounds [67, 68]. These structural differences translate into functional divergence. Ornibactin C8 exhibits strong siderophore activity, while most malleobactins require concentrations exceeding 400 µM for minimal iron-chelating function. Additionally, malleobactin and Ornibactin display a different role during infection. Ornibactins are essential virulence factors for *B. cenocepacia* pathogenesis, whereas malleobactins are dispensable for *B. pseudomallei* virulence. The accessory gene modifications therefore appear to have specialized ornibactin for virulence-associated iron acquisition, while malleobactins may have evolved broader alternative biological functions. Based on this observation, we propose the rational construction of a hybrid cluster incorporating both functionalities, potentially generating a novel metabolite that combines the potent iron-chelating capacity of

ornibactins with the alternative biological activities of malleobactins.[69] This demonstrates how PanBGC-DB can move beyond passive exploration to actively guide the design of new biosynthetic pathways, grounded in observed natural configurations and evolutionary compatibility. This increases the chances that genes will integrate successfully into engineered systems, both structurally and biochemically. In this way, PanBGC-DB provides a practical and biologically grounded resource for guiding pathway engineering with greater precision and confidence.

While PanBGC-DB provides a scalable framework for analysing biosynthetic gene clusters at the population level, limitations of the platform should be acknowledged. One of the key trade-offs in our approach lies in the GCF construction. By using a two-step clustering strategy to ensure that only closely related BGCs are grouped together, we enhance the resolution of within-family comparisons and make diversity patterns more interpretable. However, this increased specificity may come at the cost of excluding more distantly related clusters that, while functionally relevant, fall outside the defined similarity thresholds. As a result, broader biosynthetic relationships might be fragmented across multiple families, potentially limiting comparative insights at higher levels of divergence. In addition, the BGCs used in this study were sourced from the antiSMASH-DB, where cluster boundaries are inferred based on the position of core biosynthetic genes and domain architecture, but are not experimentally validated. As such, inaccuracies in boundary prediction may lead to the inclusion of non-functional genes or the omission of relevant tailoring enzymes, which could dilute or distort the inferred gene content of a family. Moreover, the high number of singleton clusters observed in our analysis may reflect the uniqueness of these biosynthetic systems but could also be artificially inflated by the predominance of cultivated strains in the antiSMASH database, which may not fully represent the true diversity of microbial communities and could therefore lead to an incomplete or skewed assessment of BGC diversity patterns. However, the growing interest in metagenomics and the increasing incorporation of metagenomic-derived BGCs into databases should help mitigate this bias in future. Finally, orthologous gene groups in PanBGC-DB are determined using ZOL[34]. ZOL clusters genes based on sequence similarity and positional conservation across gene clusters, which is highly scalable but may group structurally similar yet functionally divergent genes into the same OG. This can blur subtle functional differences, potentially reducing the resolution of accessory versus core gene identification. However, functionally precise orthology inference is an open problem[70], and ZOL is one of the tools performing very well with cluster genes.

Despite these limitations, PanBGC-DB represents a significant advance in the systematic exploration of biosynthetic diversity. By reframing gene cluster families as structured populations, the platform provides a powerful conceptual and analytical foundation for understanding how secondary metabolism evolves, diversifies, and adapts. Its integration of scalable clustering, gene-level pangenomic metrics, and interactive visualization tools makes the database a unique and accessible resource for both hypothesis-driven research and data exploration. As new genomes and metagenomes continue to be sequenced, and as tools for BGC boundary refinement and gene function prediction advance, the precision and utility of PanBGC-DB will continue to grow. The structured, population-level representation of BGCs provided by PanBGC also creates new opportunities for machine learning applications. For example, core/accessory gene classification enables standardized feature extraction for models predicting metabolite structure, function, or ecological role. Openness scores and domain architectures provide rich quantitative descriptors for prioritizing BGCs with biosynthetic novelty, while curated GCFs define empirically co-occurring gene sets that are valuable for training generative models. We anticipate that PanBGC will serve as a foundational resource for both experimental and computational advances in secondary metabolism.

Data availability

PanBGC-DB is freely available at https://panbgc-db.cs.uni-tuebingen.de and can be accessed using any web browser with JavaScript support. The full source code for the website, as well as all scripts and pipelines used for clustering, orthologous grouping, and openness calculations, are available at https://github.com/ZiemertLab/PanBGC-DB.

Supporting information

Supplementary Figures and Table | [supplements/669102_file08.pdf]

Author contributions

D.P. and N.Z. wrote the main manuscript and designed the research; D.P. build the pipeline and created the Web interface; C.B. conducted the GCF creation; A.G. validated the tools used for orthologous clustering.

Competing interests

The authors declare no competing interests.

Acknowledgments

D.P. and N.Z. were supported by H2020-FNR-11-2020: SECRETED (grant agreement: 101000794). N.Z. was supported by the German Center for Infection Research (TTU09.717); AG was supported by the Deutsche Forschungsgemeinschaft (DFG; Project ID # 398967434-TRR 261). CB was supported by the German Center for Infection Research (TTU09.716); The authors thank the Cluster of Excellence: EXC 2124: Controlling Microbes to Fight Infection (CMFI, project ID 390838134) for the structural support. We thank the Interfaculty Institute for Biomedical Informatics (IBMI) at the University of Tübingen for providing the computational resources.

FUNDER INFORMATION DECLARED

H2020-FNR-11-2020: SECRETED, 101000794

German Center for Infection Research, https://ror.org/028s4q594, TTU09.716

Deutsche Forschungsgemeinschaft, https://ror.org/018mejw64, 398967434-TRR 261

Footnotes

-
- https://github.com/ZiemertLab/PanBGC-DB
- https://panbgc-db.cs.uni-tuebingen.de/

References

1.Bolotin E, Hershberg R. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. *Genome Biol Evol* 2015;**7**:2173–2187. doi:10.1093/GBE/EVV135

2.Puigbò P et al. Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. *BMC Med* 2014;**12**:1–19. doi:10.1186/S12915-014-0066-4/FIGURES/11

3.Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics 2015 16:8* 2015;**16**:472–482. doi:10.1038/nrg3962

4.Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS Genet* 2011;**7**:e1001284. doi:10.1371/JOURNAL.PGEN.1001284

5.Rosconi F et al. A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nature Microbiology 2022 7:10* 2022;**7**:1580–1592. doi:10.1038/s41564-022-01208-7

6.Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 2005;**3**:679–687. doi:10.1038/NRMICRO1204

7.Power JJ et al. Adaptive evolution of hybrid bacteria by horizontal gene transfer. *Proc Natl Acad Sci U S A* 2021;**118**:e2007873118. doi:10.1073/PNAS.2007873118/SUPPL_FILE/PNAS.2007873118.SD05.XL SX

8.Hyun JC, Monk JM, Palsson BO. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics* 2022;**23**:1–18. doi:10.1186/S12864-021-08223-8/FIGURES/7

9.Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. *Trends in Genetics* 2009;**25**:107–110. doi:10.1016/j.tig.2008.12.004

10.Guimarães LC et al. Inside the Pan-genome - Methods and Software Overview. *Curr Genomics* 2015;**16**:245. doi:10.2174/1389202916666150423002311

11.Tettelin H et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *Proc Natl Acad Sci U S A* 2005;**102**:13950–13955. doi:10.1073/PNAS.0506758102

12.Medini D, et al. The microbial pan-genome. *Curr Opin Genet Dev* 2005;**15**:589–594. doi:10.1016/J.GDE.2005.09.006

13.Conrad RE et al. Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. *ISME J* 2022;**16**:1222–1234. doi:10.1038/S41396-021-01149-9

14.Donati C et al. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol* 2010;**11**:1–19. doi:10.1186/GB-2010-11-10-R107/FIGURES/11

15.Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports 2016 6:1* 2016;**6**:1–10. doi:10.1038/srep24373

16.Tettelin H et al. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008;**11**:472–477. doi:10.1016/J.MIB.2008.09.006

17.Costa SS et al. First Steps in the Analysis of Prokaryotic Pan-Genomes. *Bioinform Biol Insights* 2020;**14**:1177932220938064. doi:10.1177/1177932220938064

18.Terra LA et al. Pangenome analysis indicates evolutionary origins and genetic diversity: emphasis on the role of nodulation in symbiotic Bradyrhizobium. *Front Plant Sci* 2025;**16**:1539151. doi:10.3389/FPLS.2025.1539151/BIBTEX

19.Brockhurst MA, et al. The Ecology and Evolution of Pangenomes. *Current Biology* 2019;**29**:R1094–R1103. doi:10.1016/J.CUB.2019.08.012/ASSET/6E841833-BF0B-497D-AE60-4EB4414C1218/MAIN.ASSETS/GR2.JPG

20.Blaustein RA et al. Pangenomic Approach To Understanding Microbial Adaptations within a Model Built Environment, the International Space Station, *Relative to Human Hosts and Soil. mSystems* 2019;**4**. doi:10.1128/MSYSTEMS.00281-18/SUPPL_FILE/SYS00119231OST4.XLSX

21.Rouli L et al. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect* 2015;**7**:72. doi:10.1016/J.NMNI.2015.06.005

22.Vernikos G et al. Ten years of pan-genome analyses. *Curr Opin Microbiol* 2015;**23**:148–154. doi:10.1016/J.MIB.2014.11.016

23.Kautsar SA et al. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience* 2021;**10**:giaa154. doi:10.1093/GIGASCIENCE/GIAA154

24.Navarro-Muñoz JC et al. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 2019;**16**:60. doi:10.1038/S41589-019-0400-9

25.Mohite OS et al. Pangenome analysis of Enterobacteria reveals richness of secondary metabolite gene clusters and their associated gene sets. *Synth Syst Biotechnol* 2022;**7**:900–910. doi:10.1016/J.SYNBIO.2022.04.011

26.Cimermancic P et al. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell* 2014;**158**:412–421. doi:10.1016/J.CELL.2014.06.034

27.Chevrette MG et al. Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat Prod Rep* 2020;**37**:566–599. doi:10.1039/C9NP00048H

28.Ziemert N et al. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc Natl Acad Sci U S A* 2014;**111**:E1130–E1139. doi:10.1073/PNAS.1324161111/SUPPL_FILE/PNAS.201324161SI.PDF

29.Medema MH et al. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. *PLoS Comput Biol* 2014;**10**:e1004016. doi:10.1371/JOURNAL.PCBI.1004016

30.Hansen MH et al. Resurrecting ancestral antibiotics: unveiling the origins of modern lipid II targeting glycopeptides. *Nature Communications 2023 14:1* 2023;**14**:1–16. doi:10.1038/s41467-023-43451-4

31.Gavriilidou A et al. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nature Microbiology 2022 7:5* 2022;**7**:726–735. doi:10.1038/s41564-022-01110-2

32.Blin K et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Res* 2025;**53**:W32–W38. doi:10.1093/NAR/GKAF334

33.Zdouc MM et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res* 2025;**53**:D678–D690. doi:10.1093/NAR/GKAE1115

34.Salamzade R et al. zol and fai: large-scale targeted detection and evolutionary investigation of gene clusters. *Nucleic Acids Res* 2025;**53**:45. doi:10.1093/NAR/GKAF045

35.Sun B et al. PanKB: An interactive microbial pangenome knowledgebase for research, biotechnological innovation, and knowledge mining. *Nucleic Acids Res* 2025;**53**:D806– D818. doi:10.1093/NAR/GKAE1042

36.Rajput A et al. Pangenome analysis reveals the genetic basis for taxonomic classification of the Lactobacillaceae family. *Food Microbiol* 2023;**115**:104334. doi:10.1016/J.FM.2023.104334

37.Zhang C, Mirarab S. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. doi:10.1093/bioinformatics/btac620

38.Gilchrist CLM et al. cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. *Bioinformatics Advances* 2021;**1**. doi:10.1093/BIOADV/VBAB016

39.Kautsar SA et al. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res* 2021;**49**:D490–D497. doi:10.1093/NAR/GKAA812

40.Blin K et al. The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more. *Nucleic Acids Res* 2024;**52**:D586–D589. doi:10.1093/NAR/GKAD984

41.Mohite OS et al. Pangenome analysis of Enterobacteria reveals richness of secondary metabolite gene clusters and their associated gene sets. *Synth Syst Biotechnol* 2022;**7**:900–910. doi:10.1016/J.SYNBIO.2022.04.011

42.Paysan-Lafosse T et al. The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res* 2025;**53**:D523–D534. doi:10.1093/NAR/GKAE997

43.Süssmuth RD, Mainz A. Nonribosomal Peptide Synthesis—Principles and Prospects. *Angewandte Chemie International Edition* 2017;**56**:3770–3821. doi:10.1002/ANIE.201609079

44.Nivina A et al. Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem Rev* 2019;**119**:12524–12547. doi:10.1021/ACS.CHEMREV.9B00525

45.Franke J, Ishida K, Hertweck C. Plasticity of the Malleobactin Pathway and Its Impact on Siderophore Action in Human Pathogenic Bacteria. *Chemistry – A European Journal* 2015;**21**:8010–8014. doi:10.1002/CHEM.201500757

46.Franke J et al. Nitro versus hydroxamate in siderophores of pathogenic bacteria: effect of missing hydroxylamine protection in malleobactin biosynthesis. *Angew Chem Int Ed Engl* 2013;**52**:8271–8275. doi:10.1002/ANIE.201303196

47.Franke J, Ishida K, Hertweck C. Evolution of siderophore pathways in human pathogenic bacteria. *J Am Chem Soc* 2014;**136**:5599–5602. doi:10.1021/JA501597W

48.Alice AF et al. Genetic and transcriptional analysis of the siderophore malleobactin biosynthesis and transport genes in the human pathogen Burkholderia pseudomallei K96243. *J Bacteriol* 2006;**188**:1551–1566. doi:10.1128/JB.188.4.1551-1566.2006

49.Agnoli K et al. The ornibactin biosynthesis and transport genes of Burkholderia cenocepacia are regulated by an extracytoplasmic function σ factor which is a part of the fur regulon. *J Bacteriol* 2006;**188**:3631–3644. doi:10.1128/JB.188.10.3631-3644.2006/SUPPL_FILE/SUPPLEMENTARY_TABLE_1.DOC

50.Curson ARJ et al. Identification of genes for dimethyl sulfide production in bacteria in the gut of Atlantic Herring (Clupea harengus). *ISME J* 2010;**4**:144–146. doi:10.1038/ISMEJ.2009.93

51.Esmaeel Q et al. Nonribosomal peptide synthetase with a unique iterative-alternative-optional mechanism catalyzes amonabactin synthesis in Aeromonas. *Appl Microbiol Biotechnol* 2016;**100**:8453–8463. doi:10.1007/S00253-016-7773-4

52.Wang X et al. Nisin: harnessing nature's preservative for the future of food safety and beyond. *Crit Rev Food Sci Nutr* 2025. doi:10.1080/10408398.2025.2517822

53.Stein T et al. Two different lantibiotic-like peptides originate from the ericin gene cluster of Bacillus subtilis A1/3. *J Bacteriol* 2002;**184**:1703–1711. doi:10.1128/JB.184.6.1703-1711.2002/ASSET/71664147-8734-4B32-B6EF-5A1B258D42BB/ASSETS/GRAPHIC/JB0621355006.JPEG

54.Fuchs SW et al. Entianin, a novel subtilin-like lantibiotic from Bacillus subtilis subsp. spizizenii DSM 15029T with high antimicrobial activity. *Appl Environ Microbiol* 2011;**77**:1698–1707. doi:10.1128/AEM.01962-10

55.Bochmann SM et al. Synthesis and Succinylation of Subtilin-Like Lantibiotics Are Strongly Influenced by Glucose and Transition State Regulator AbrB. *Appl Environ Microbiol* 2015;**81**:614. doi:10.1128/AEM.02579-14

56.Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* 2021;**37**:2473–2475. doi:10.1093/BIOINFORMATICS/BTAB007

57.Page AJ et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;**31**:3691–3693. doi:10.1093/BIOINFORMATICS/BTV421

58.Nuhamunada M et al. BGCFlow: systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. *Nucleic Acids Res* 2024;**52**:5478–5495. doi:10.1093/NAR/GKAE314

59.Chen J, Kuipers OP. Analysis of Cross-Functionality within LanBTC Synthetase Complexes from Different Bacterial Sources with Respect to Production of Fully Modified Lanthipeptides. *Appl Environ Microbiol* 2022;**88**. doi:10.1128/AEM.01618-21/SUPPL_FILE/AEM.01618-21-S0001.PDF

60.Repka LM et al. Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chem Rev* 2017;**117**:5457–5520. doi:10.1021/ACS.CHEMREV.6B00591/ASSET/IMAGES/MEDIUM/CR-2016-00591E_0039.GIF

61.Khosa S, Lagedroste M, Smits SHJ. Protein defense systems against the lantibiotic nisin: Function of the immunity protein NisI and the resistance protein NSR. *Front Microbiol* 2016;**7**:187154. doi:10.3389/FMICB.2016.00504/BIBTEX

62.Medema MH et al. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nature Reviews Microbiology 2010 9:2* 2010;**9**:131–137. doi:10.1038/nrmicro2478

63.Del Carratore F et al. Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. *Communications Biology 2019 2:1* 2019;**2**:1–10. doi:10.1038/s42003-019-0333-6

64.Smanski MJ et al. Synthetic biology to access and expand nature's chemical diversity. *Nature Reviews Microbiology 2016 14:3* 2016;**14**:135–149. doi:10.1038/nrmicro.2015.24

65.Alam K et al. Synthetic biology-inspired strategies and tools for engineering of microbial natural product biosynthetic pathways. *Biotechnol Adv* 2021;**49**:107759. doi:10.1016/J.BIOTECHADV.2021.107759

66.Baltz RH. Combinatorial biosynthesis of cyclic lipopeptide antibiotics: A model for synthetic biology to accelerate the evolution of secondary metabolite biosynthetic pathways. *ACS Synth Biol* 2014;**3**:748–758. doi:10.1021/SB3000673/ASSET/IMAGES/MEDIUM/SB-2012-000673_0004.GIF

67.Vences-Guzmán MÁ et al. Discovery of a bifunctional acyltransferase responsible for ornithine lipid synthesis in Serratia proteamaculans. *Environ Microbiol* 2015;**17**:1487–1496. doi:10.1111/1462-2920.12562/SUPPINFO

68.Franke J et al. Nitro versus Hydroxamate in Siderophores of Pathogenic Bacteria: Effect of Missing Hydroxylamine Protection in Malleobactin Biosynthesis. *Angewandte Chemie International Edition* 2013;**52**:8271–8275. doi:10.1002/ANIE.201303196

69.Franke J, Ishida K, Hertweck C. Plasticity of the malleobactin pathway and its impact on siderophore action in human pathogenic bacteria. *Chemistry* 2015;**21**:8010–8014. doi:10.1002/CHEM.201500757

70.Glover N et al. Advances and Applications in the Quest for Orthologs. *Mol Biol Evol* 2019;**36**:2157–2164. doi:10.1093/MOLBEV/MSZ150