PanBGC: A pangenomic framework for systematic analysis of biosynthetic gene cluster diversity

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Davide Paccagnella

aus Bruneck/Italien

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	04.12.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Nadine Ziemert
2. Berichterstatter/-in:	PD Dr. Evi Stegmann

TABLE OF CONTENT

ACKNOWLEDGEMENTS

I would love to express my biggest thanks to:

My supervisor Nadine for giving me the opportunity to pursue my PhD in her group and for the invaluable support and excellent advice throughout my journey. Thank you, Nadine, for supporting me and guiding me every step of the way.

The Ziemertlab and all its current and former members for the wonderfully friendly atmosphere you created. Thank you for all the fun activities we did together and for making every day in the lab something to look forward to.

The collaborators and co-authors:

Caner for his constant help with running different bioinformatic tools and for always being there to support me with his expertise. Athina and Bita for their valuable contributions during the starting phase of my actual thesis topic and for helping to lay the foundation for this work. All the partners of the SECRETed project for forming such an interesting international collaboration and for the enriching experience of working together with such a vast variety of people with different backgrounds.

My funder, the SECRETed project funded by EU Horizon 2020. Thank you for the opportunity to work in this fantastic project and for making this research possible.

My family, which always believes in me, supports me and has an open ear for my problems. My parents for letting me know how proud they are and the immense support they gave me all my life. My sister for cheering me up and being there with a phone call when I needed it.

Franzi for everything we did since we met. For the support she gave me and for listening to my long endless complaint monolog when I was frustrated with my work. But most importantly, her ability to cheer me up in difficult times.

My Best friends Samu and Janes for being there since day one of my studies. For the long gaming nights, the long bib study days and all the memes we share in our group.

My lab family Alena, Jens and Dardan and our traditional breakfast before starting the day. Thank you for giving me a place where I could share problems, feelings and also some 10th floor gossip.

ABSTRACT

Microbial secondary metabolites are critical sources of antibiotics, anticancer agents, and other bioactive compounds, with over 60% of approved drugs derived from natural products. These compounds are encoded by biosynthetic gene clusters (BGCs), which contain the genetic instructions for producing complex chemical structures. While computational tools can identify and group BGCs into families based on similarity, current approaches lack systematic frameworks for analysing the internal genetic diversity that drives chemical innovation within these families. This limitation represents a significant gap in understanding how biosynthetic pathways evolve and generate the remarkable chemical diversity observed in microbial natural products.

In my PhD project, I developed PanBGC, a computational framework that adapts pangenomic principles to analyse biosynthetic gene cluster diversity. Just as pangenomics that analyses microbial genomic diversity within species populations, PanBGC treats gene cluster families as structured populations rather than isolated genomic islands. Applied to over 250,000 BGCs from more than 35,000 microbial genomes, representing over 80,000 gene cluster families, the PanBGC framework revealed patterns of genetic organization within biosynthetic families. The most significant finding was the identification of contrasting evolutionary dynamics: while most BGC families exhibit closed gene repertoires with limited acquisition of entirely novel genes, they simultaneously demonstrate high compositional plasticity in how existing genetic components of one family are combined within individual clusters. This finding indicates that biosynthetic innovation operates primarily through modular reorganization of evolutionarily validated genetic components rather than through continuous incorporation of novel genetic material.

Functional analysis revealed that core genes were predominantly associated with essential enzymatic activities required for basic metabolite biosynthesis, while accessory genes were associated with tailoring reactions and regulatory functions that contribute to structural diversification. To make these analyses accessible to the research community, PanBGC-DB was developed as an interactive web platform containing precomputed analyses of gene cluster families, enabling systematic investigations of BGC family diversity.

The framework addresses practical applications in natural product research. For genome mining, the systematic organization enables prioritization of clusters with unusual gene combinations that may represent novel chemical scaffolds. In synthetic biology, the identification of accessory genes that co-occur with specific core pathways provides evidence for functional compatibility, improving the success rate of biosynthetic engineering compared to traditional trial-and-error approaches.

The results of this PhD thesis establish a systematic framework for understanding diversification within BGC families. This work provides both fundamental insights into biosynthetic evolution and practical tools for natural product discovery, representing a shift toward population-level thinking in biosynthetic gene cluster analysis.

ZUSAMMENFASSUNG

Mikrobielle Sekundärmetabolite sind eine entscheidende Quelle für Antibiotika, Krebsmedikamente und andere bioaktive Verbindungen. Über 60% der zugelassenen Arzneimittel stammen von Naturstoffen ab. Diese Verbindungen werden von Biosynthese-Genclustern (BGCs) codiert, die die genetischen Anweisungen zur Herstellung komplexer chemischer Strukturen enthalten. Während bioinformatische Werkzeuge BGCs identifizieren und auf Basis ihrer Ähnlichkeit in Familien gruppieren können, fehlen den derzeitigen Ansätzen systematische Herangehensweise zur Analyse der inneren genetischen Diversität, die die chemische Innovation innerhalb dieser Familien antreibt. Diese Einschränkung stellt eine wesentliche Lücke im Verständnis dar, wie sich Biosynthesewege entwickeln und die bemerkenswerte chemische Vielfalt mikrobieller Naturstoffe erzeugt werden.

In meinem Promotionsprojekt habe ich PanBGC entwickelt, eine bioinformatische pipeline, die das pangenomische Prinzipien auf die Analyse der Diversität von Biosynthese-Genclustern anwendet. Ähnlich wie die Pangenomik, die die genetische Vielfalt mikrobieller Genome innerhalb von Spezies untersucht, behandelt PanBGC Genclusterfamilien als strukturierte Populationen statt als isolierte genomische Inseln. Angewendet auf über 250.000 BGCs aus mehr als 35.000 mikrobiellen Genomen, die über 80.000 Genclusterfamilien repräsentieren, offenbarte PanBGC Muster der genetischen Organisation innerhalb biosynthetischer Familien. Der bedeutendste Einblick war die Identifizierung kontrastierender evolutionärer Dynamiken: Während die meisten BGC-Familien geschlossene Gen-Repertoires mit begrenzter Aufnahme völlig neuer Gene aufweisen, zeigen sie gleichzeitig eine hohe Zusammensetzungsplastizität darin, wie bestehende genetische Komponenten einer Familie in einzelnen Clustern kombiniert werden. Dieser Einblick deutet darauf hin, dass biosynthetische Innovation primär durch modulare Reorganisation evolutionär bewährter genetischer Komponenten erfolgt und nicht durch die kontinuierliche Integration neuartiger genetischer Elemente.

Die funktionelle Analyse zeigte, dass Core-Gene vorwiegend mit essenziellen enzymatischen Aktivitäten verbunden sind, die für die grundlegende Metabolitenbiosynthese erforderlich sind, während akzessorische Gene mit Modifikationsreaktionen und regulatorischen Funktionen assoziiert waren, die zur strukturellen Diversifizierung beitragen. Um diese Analysen der Forschungsgemeinschaft zugänglich zu machen, wurde PanBGC-DB als interaktive Webplattform entwickelt, die vorkalkulierte Analysen von Genclusterfamilien enthält und systematische Untersuchungen der BGC-Familiendiversität ermöglicht.

Die PanBGC Herangehensweise bietet praktische Anwendungen in der Naturstoffforschung. Für das Genome Mining ermöglicht die systematische Organisation die Priorisierung von Clustern mit ungewöhnlichen Genkombinationen, die potenziell neuartige chemische Grundgerüste darstellen könnten. In der Synthetischen Biologie liefert die Identifizierung von akzessorischen Genen, die gemeinsam mit bestimmten Kernwegen auftreten, Hinweise auf funktionale Kompatibilität und verbessert damit die Erfolgsquote beim biosynthetischen Engineering im Vergleich zu herkömmlichen Trial-and-Error-Ansätzen.

Die Ergebnisse dieser Dissertation etablieren eine systematische pipeline zum Verständnis der Diversifizierung innerhalb von BGC-Familien. Diese Arbeit liefert sowohl grundlegende Erkenntnisse zur biosynthetischen Evolution als auch praktische Werkzeuge für die Naturstoffentdeckung und stellt einen Wechsel hin zu einem populationsbasierten Denken in der Analyse von Biosynthese-Genclustern dar.

ABBREVIATIONS

A	Adenylation (domains)
AI	Artificial Intelligence
ANI	Average Nucleotide Identity
antiSMASH	Antibiotics & Secondary Metabolite Analysis Shell
ARTS	Antibiotic Resistance Target Seeker
AT	Acyltransferase
BGC / BGCs	Biosynthetic Gene Cluster(s)
BiG-SCAPE	Biosynthetic Gene Similarity Clustering and Prospecting Engine
BiG-SLiCE	Biosynthetic Genes Super-Linear Clustering Engine
C	Condensation (domains)
CoA	Coenzyme A
E. coli	*Escherichia coli*
GCF / GCFs	Gene Cluster Family/Families
GNPS	Global Natural Products Social Molecular Networking
MIBiG	Minimum Information about a Biosynthetic Gene cluster
NRPS	Non-Ribosomal Peptide Synthetases
PanBGC-DB	Pan-Biosynthetic Gene Cluster Database
PKS	Polyketide Synthases
RiPP / RiPPs	Ribosomally Synthesized and Post-Translationally Modified Peptide(s)
S. aureus	*Staphylococcus aureus*
UV	Ultraviolet
ZOL	Zonal Orthologous Loci

THESIS – INTRODUCTION

THE IMPORTANCE OF SECONDARY METABOLITES FOR HUMANS

Microbial natural products produced by bacteria, fungi, and other microorganisms have fundamentally changed modern medicine and continue to serve as critical resources for human health[1–5]. The pharmaceutical industry has built its foundation on these compounds, with studies indicating that over 60% of approved drugs are derived from or inspired by natural products[1]. From the discovery of the first antibiotic[6] to today's more sophisticated therapeutic agents, microbial metabolites have provided us with life-saving medicines[7, 8]. Antibiotics such as penicillin, streptomycin, and erythromycin revolutionized the treatment of infectious diseases[9], while other microbial compounds have yielded effective therapies including anticancer agents like doxorubicin and bleomycin[10, 11], immunosuppressants such as cyclosporine and rapamycin[12], and cholesterol-lowering statins[13]. However, the rise of antibiotic-resistant pathogens is responsible for over 4,65 million deaths in 2019[14], creating an urgent need for new antimicrobials with different modes of action. The structural complexity and biological specificity of these compounds often cannot be easily replicated or created through synthetic chemistry[15–17], making natural product discovery an ongoing priority for pharmaceutical development. Recent calculations suggest that approximately only 3% of microbial natural products have been experimentally characterized, indicating a wide range of opportunities for future drug discovery[18].

Today, as we face continuing global challenges, some microbial natural products have emerged as important resources for addressing some of humanity’s most critical problems. Environmental applications of microbial natural products continue to grow as industries attempt to transition from oil-based chemicals to those from renewable sources. Biosurfactants from various bacterial species are a biodegradable substitute for synthetic surfactants with potential use in oil spill remediation, cosmetics, and industrial cleaning[19–22]. These compounds often possess more favourable chemical properties than their synthetic counterparts while being nontoxic and biodegradable[20, 22–24]. In agriculture, natural products provide ecological solutions for pest control and plant nutrition[25, 26]. Bacterial and fungal biopesticides, such as Bt-toxins and other antifungal compounds[5], provide effective crop protection with less environmental impact and toxicity compared to conventional pesticides. Growth promoters and biofertilizers of beneficial microorganisms also have the potential to optimize crop production while reducing the application of costly and ecologically harmful synthetic fertilizers[27].

Additionally, industrial biotechnology relies on enzymes and bioactive compounds that are produced by microorganisms[28, 29]. Applications range from food production, the textile sector to the manufacture of specific chemicals and pharmaceuticals[30, 31]. The growth of synthetic biology has seen these applications further extended through the ability to make compounds that are inspired from natural products, in designed microbial systems[32]. This

facilitates the manufacture of important chemicals that before only could be made through environmentally harmful synthetic approaches[33, 34].

BIOSYNTHETIC GENE CLUSTERS: THE RECIPES FOR A DIVERSE RANGE OF COMPOUNDS

The high chemical diversity of microbial secondary metabolites outlined in the previous sections is encoded in discrete genomic regions called biosynthetic gene clusters (BGCs)[35, 36]. These clusters represent the organizational schemes in microbial genomics, whereby genes that are responsible for the synthesis, modification, and regulation of a natural product are co-localized as compact genomic module[37]. This clustering strategy, first demonstrated in early studies on antibiotic biosynthesis in *Streptomyces* species, represents a profound evolutionary solution to the challenges of coordinating complex multi-step biosynthetic pathways and responding rapidly to ecological pressures[38, 39]. The evolutionary advantages of gene clustering become apparent when considering the dynamic nature of microbial environments and the need for rapid adaptation to changing ecological conditions[40] Secondary metabolite production often occurs in response to specific environmental triggers, competitive pressures, or resource limitations[41–44]. The clustering of biosynthetic genes enables coordinated regulation where entire pathways can be activated or silenced as functional units, allowing microorganisms to rapidly deploy chemical defenses, communication molecules, or resource acquisition tools when ecological conditions demand them[45]. This coordinated control is essential for producing complex molecules that often require multiple biosynthetic steps and precise regulation to compete effectively in dynamic microbial communities.

Perhaps most importantly for understanding BGC diversity, these clusters function as discrete evolutionary units that undergo selection, duplication, and recombination as coherent modules rather than as collections of individual genes[46]. The compact nature of BGCs makes them particularly suitable for horizontal gene transfer, allowing entire functional biosynthetic capabilities to be rapidly acquired by organisms facing similar ecological pressures[47, 48]. This horizontal mobility has proven crucial for the evolution and distribution of natural product biosynthesis, enabling the rapid dissemination of successful biosynthetic innovations across microbial communities when competitive advantages arise. The ecological drivers of BGC horizontal transfer include the selective advantage of acquiring new chemical capabilities for competition, communication, environmental adaptation, or exploitation of new ecological niches[49–52].

Furthermore, the modular organization within BGCs themselves facilitates evolutionary innovation under ecological pressure. The addition, loss, or modification of genes within clusters leads to chemically related compounds with distinct properties, allowing fine-tuning of bioactive molecules for specific ecological contexts[40, 53, 54]. For modular systems like Non-Ribosomal Peptide Synthetases (NRPS) and Polyketide Synthases (PKS), the rearrangement of biosynthetic modules can create new compounds that may provide advantages in different

competitive environments or against different target organisms[55, 56]. This modular flexibility represents a key mechanism through which BGCs can rapidly adapt to new ecological challenges while maintaining their core biosynthetic functions.

MAJOR CLASSES OF BIOSYNTHETIC GENE CLUSTERS

The chemical diversity of natural compounds is a reflection of the evolution of biosynthetic approaches, each defined by characteristic types of BGCs. Even though the products of these systems are extremely diverse, the underlying biosynthetic process can be categorized into a few major classes, each having its own genetic and biochemical features.

Non-Ribosomal Peptide Synthetase (NRPS) systems are an assembly line of multiple modules forming a mega-enzyme that constructs peptide compounds through a sequential series of condensations of amino acid substrates, including proteinogenic and non-proteinogenic amino acids which are not capable of being constructed into a protein by the ribosomal protein synthesis[57]. NRPS are responsible for the production of many of our most important antibiotics, including penicillin, vancomycin, and daptomycin, as well as immunosuppressants like cyclosporine. Due to the modular organization of NRPS clusters, it is possible to link gene structure to the chemical structure by predicting the amino acid sequence of the constructed compound[58]. This predictable relationship has made NRPS clusters particularly interesting to computational analysis and rational engineering approaches.

Polyketide Synthases (PKS) systems employ assembly-line logic similar to NRPS but utilize acyl-CoA and derivatives of it and uses condensation mechanisms to produce polyketide natural products[59]. Type I PKS systems use modular architectures where each module possesses the enzymatic domains for one round of chain extension and modification. Type II PKS systems are based on a different strategy using iterative enzymes that act repeatedly on elongating polyketide chains. Type III PKS systems possess a third structural form that produces shorter, often aromatic compounds[60–62]. Compounds produced by PKSs include many medically important compounds such as erythromycin, rapamycin, and the anticancer agent doxorubicin[63–65].

Ribosomally Synthesized and Post-Translationally Modified Peptides (RiPPs) represent a rapidly increasing class of natural products that begin as standard ribosomal peptides but are modified via various enzymatic reactions to create active molecules. Unlike NRPS systems, RiPPs utilize only proteinogenic amino acids and the existing cellular machinery for the initial peptide synthesis and later use specialized enzymes to introduce modifications such as cyclizations, cross-links, and non-canonical amino acid residues. The diversity of RiPP modifications have a wide range, from simple disulfide bridges to complex heterocyclic frameworks and new amino acid chemistry[66, 67]. To mention a few, lantibiotics like nisin, thiopeptides like thiostrepton, and lasso peptides are compounds produced by RiPPs systems[68, 69].

Terpene biosynthesis involves the cyclization and modification of isoprenoid precursors to create structurally complex molecules that range from simple monoterpenes to complex

polycyclic compounds. Terpene BGCs typically contain genes for terpene synthases that catalyze the formation of terpene skeletons, along with many modifying enzymes that introduce functional groups and structural complexity[70]. Microbial terpenes have found widespread applications in pharmaceuticals, with compounds such as artemisinin serving as a critical antimalarial drug[71], while others like β-carotene and astaxanthin are utilized as nutritional supplements and natural colorants in food and cosmetic industries[72].

In addition to the major modular systems, microorganisms use many other biosynthetic strategies to generate chemical diversity. These include saccharide natural products assembled through glycosyltransferases[73], alkaloid biosynthesis employing diverse strategies to generate nitrogen-containing heterocycles, and various other specialized pathways for producing unique chemical scaffolds[74]. One of the most intriguing aspects of natural product biosynthesis is the prevalent existence of hybrid systems that incorporate structural components from greater than one biosynthetic class. NRPS-PKS hybrids for example are able to yield compounds with peptide as well as polyketide components, while other mixtures yield additional chemical diversity[75].

This modular organization of biosynthetic genes enables not only experimental investigation but also computational analysis. As a result, tools for detection and specialized databases have been developed to identify, classify, and compare these genomic regions across the rapidly expanding amount of easily available sequenced microbial genomes.

ECOLOGICAL PRESSURES AS THE PRIMARY DRIVER OF BGC DIVERSITY

The architectural diversity observed in microbial natural products is not random but rather reflects millions of years of evolutionary selection driven by fundamental ecological challenges that microorganisms face in their natural environments. Understanding these ecological pressures is crucial to comprehend how and why such biosynthetic diversity evolved, and more importantly, how these pressures continue to drive diversification within families of related BGCs[76–78]. The ecological roles of natural products represent the primary selective forces that shape BGC evolution, creating the patterns of conservation and variation within biosynthetic families that form the central focus of this thesis.

In the microscopic world, natural products serve as molecular tools for survival and competitive advantage in resource-restricted environments and heavily competitive ecological niches[79]. These compounds are not merely secondary byproducts of metabolism but rather represent the outcome of millions of years of evolution toward fitness-enhancing characteristics[54]. Within microbial communities natural products serve as competition mediators, cooperation mediators, signallers, and adaptors to changing environmental conditions.

Chemical warfare is among the most notable ecological roles[80]. In densely populated environments such as soil or marine biofilms, microorganisms compete for limited space and resources[81]. The antibiotic-resistance arms race has stimulated fast innovation in microbial chemical warfare[82]. *Streptomyces* species, whose members have been shown to produce a

high quantity of secondary metabolites, each generate numerous different antimicrobial compounds, often with different modes of action[83]. Apart from direct antimicrobial activity, some natural products are more subtle competitive agents. Some will inhibit specific metabolism or interfere with essential cell operations without directly killing cells, creating competitive advantages without the severe selective pressure imposed by rapid killing. Others are analogs for signal molecules that can disrupt communication systems within competing organisms[84, 85].

Natural products of microbes also serve critical roles in cell-to-cell communication and community organization[86]. Quorum sensing molecules, including acyl-homoserine lactones, autoinducing peptides, and autoinducers, enable bacteria to coordinate population-wide behavior like biofilm formation, the expression of virulence factors, and sporulation[87, 88]. These signaling systems demonstrate how chemical communication can emerge as an organizing principle in microbial communities, enabling complex collective behaviors that enhance survival and competitiveness[89]. The complexity of microbial signaling also encompasses interspecies and even interkingdom interactions. 2,4-Diacetylphloroglucinol from *Pseudomonas* species, for instance, not only inhibits fungal competitors but also promotes plant growth, illustrating the polyfunctionality of some natural products[90, 91]. These molecules also serve as molecular mediators in complex ecological networks, influencing community composition and ecosystem function[92].

In nutrient-limited environments, microbial natural products are often part of the molecular machinery for resource acquisition. The best-studied example are siderophores. These compounds are highly selective iron-chelating molecules that enable survival in iron-limited environments. The structural diversity of siderophores is a reflection of the evolutionary arms race between iron acquisition and iron sequestration, with different classes of structures having specific advantages under specific environmental conditions[93].

Apart from nutrient acquisition, natural products have also a central function in stress response and adaptation. Pigments such as carotenoids and melanins are UV radiation and oxidative stress protectants[94], while compounds like ectoine and trehalose are produced to survive osmotic stress[95]. This sort of functional diversity demonstrates how natural product biosynthesis has evolved as a general strategy for environmental adaptation.

COMPUTATIONAL PIPELINE FOR BGC ANALYSIS: FROM DETECTION TO COMPARATIVE GENOMICS

The advancements in high-throughput genome sequencing have fundamentally reshaped how natural products are discovered and their biosynthesis is analyzed[96]. Whereas more traditional approaches relied on bioactivity-guided microbial isolation from cultivatable microbes, something that was inherently limited by cultivation biases and laboratory-expression conditions, the genomic era has revealed to us the vast, largely uncharted biosynthetic potential encoded in microbial genomes[97]. This new available data necessitated the development of computational algorithms and analytical tools with the ability to detect,

annotate, and align biosynthetic gene clusters from the increasing number of sequenced genomes.

The first step in the analysis pipeline is the detection of BGC in the sequenced genome. AntiSMASH (Antibiotics & Secondary Metabolite Analysis Shell)[98], which is considered the gold standard in BGC detection, was the first comprehensive, automated pipeline for BGC detection and annotation. AntiSMASH employs a multi-layered detection approach that entails the incorporation of signature gene identification, determination of domain architecture, and comparative genomics to predict BGC boundaries as well as classify clusters to known biosynthetic classes. The rule-based platform also uses databases of known biosynthetic domains with emphasis on the conserved catalytic domains that make up the large BGC classes such as NRPS, PKS, and RiPP systems. Complementing rule-based approaches, tools like DeepBGC introduced machine learning into BGC detection[99]. This machine learning approach is particularly efficient in discovering new BGC architectures that might be missed by rule-based approaches. Similarly, GECCO (Gene Cluster Prediction with Conditional Random Fields) employs ensemble methods combining multiple detection techniques to provide more comprehensive boundary predictions through probabilistic modeling of gene relationships[100].

The expanding number of computationally predicted BGCs has brought unprecedented data storage, organization, and access challenges. This led to the creation of different databases with curated datasets of BGCs. The gold standard for BGC curation is the MIBiG database (Minimum Information about a Biosynthetic Gene cluster) that provides manually curated, experimentally proven BGCs with clear cluster boundaries along with their characterized chemical products[101]. All entries in the MIBiG database include metadata covering biosynthetic pathways, chemical structures, biological activities, and the publication where the cluster was analyzed. This dataset is used for training computational tools and serves as reference standard for comparative BGC analysis. In addition to the MIBiG dataset the antiSMASH database provides BGCs of the entire collection of publicly released genomes and comprises over 250.000 predicted BGCs[102].

With the introduction of BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine) a systematic approach was enabled to organize BGCs in different groups, so called Gene Cluster Families (GCF), based on similarity[103]. Gene cluster families represent groups of BGCs that share similar genetic architecture and are likely to produce related chemical compounds. This clustering is crucial because it enables:

- Evolutionary insight: GCFs can reveal how biosynthetic pathways have evolved and spread across different organisms
- Functional prediction: BGCs within the same family often produce structurally related compounds, allowing prediction of chemical output from uncharacterized clusters
- Prioritization: Researchers can focus on novel GCFs that may represent new chemical scaffolds
- Comparative analysis: GCFs enable systematic comparison of how the same biosynthetic logic is implemented across different organisms and how diversity is generated

BiG-SCAPE constructs similarity networks by connecting BGCs based on their calculated distances, enabling cluster family identification. For larger-scale analyses, BiG-SLiCE (Biosynthetic Genes Super-Linear Clustering Engine) was developed to handle much larger BGCs datasets through the use of highly optimized clustering algorithms and dimensionality reductions[104].

In combination with advanced clustering methods, several tools have been created to support targeted comparative analyses and homolog identification. The cblaster tool enables the rapid identification of homologous gene clusters within large genomic datasets by combining sequence similarity analyses with synteny conservation analyses[105]. The strength of this tool lies in its ability to recognize biosynthetic gene cluster families based on gene content similarities while also taking into account gene order preservation. This makes it essential to find similar clusters and see how distributed a cluster is across different taxonomic groups. Furthermore, the IsaBGC tool performs comparative analyses of genomic islands, including biosynthetic gene clusters[106]. IsaBGC is especially focused on comparing BGCs through extensive sequence analysis, synteny analysis, and comparison of functional annotations. On the other hand, tools like Clinker provide visualization tools for the comparison of similar BGCs generating figures that highlight synteny relationships and structural diversity within BGC families[107].

Regarding systematic comparative studies of BGC families, the Zonal Orthologous Loci (ZOL) method represents a valuable tool for the detection of orthologous genes through the integrative application of sequence similarity clustering and synteny conservation[108]. As an orthologous gene group we understand a set of genes with high to identical functions. Unlike traditional methods for ortholog identification that ignore gene order, ZOL retains positional information, making it an especially useful approach for studies of gene conservation, loss, and rearrangement within phylogenetically related BGCs.

However, as the field advances and datasets continue to grow, new analytical opportunities emerge. While we can successfully identify and classify BGC families, we lack systematic frameworks for analyzing their internal diversity, as mentioned in previous chapters. This includes understanding which genes are conserved versus variable, quantifying compositional diversity, and assessing the evolutionary dynamics that drive biosynthetic innovation within families.

This analytical gap was especially evident, through collaborative research (manuscript 1 and 2) on caprazamycin biosynthesis and *Rhodococcus* lipopeptides, where I contributed comparative genomics analyses that mapped the distribution of specific BGCs across bacterial taxa. While these studies successfully achieved their objectives, they highlighted an intriguing analytical opportunity: the systematic analysis of diversity within BGC families. Understanding whether the thousands of related clusters we identified were compositionally identical or exhibited systematic variation could have provided crucial insights into biosynthetic evolution and functional specialization. For instance, compositional differences within the *liu* cluster family, which is responsible for the production of the antibiotic caprazamycin, might correlate with substrate specificity variations or ecological adaptations, while diversity patterns similar to the lipopeptide BGCs of *Rhodococcus* could reveal the genetic basis for structural

modifications that generate chemical novelty. Moreover, distinguishing universally conserved genes from variable accessory components could inform rational engineering approaches and guide the discovery of novel bioactive compounds by highlighting clusters with unusual gene combinations. Current BGC analysis tools excel at organizing clusters into families but lack frameworks for analyzing internal diversity within those families.

PANGENOMICS FOR UNDERSTANDING ECOLOGICAL EVOLUTION IN BGCS

The limitations of current BGC analysis approaches for understanding ecological diversification point toward a solution that has proven highly successful in microbial population genomics: the pangenome framework[109, 110]. Pangenomics provides the population-level analytical approach needed to understand how ecological pressures create patterns of genetic diversity within related groups of organisms, or in this case, related BGCs[111]. By adapting pangenomic concepts to BGC families, we can systematically analyze how ecological forces drive biosynthetic evolution and chemical diversification.

THE PANGENOME FRAMEWORK: A POPULATION-LEVEL MODEL FOR GENOMIC DIVERSITY

The pangenome concept emerged when it became clear that one genome could not possibly capture the entire genetic variation that exists within microbial species. By analyzing several genomes from the same species or closely related strains, Tettelin et al.[109] showed that bacterial populations have shared conserved genetic cores and variable accessory gene pools that allow adaptation and specialization. This observation challenged the traditional analysis of a single-reference genome and provided a population-centric model for understanding the genomic variation that exists in microbial organisms.

The main principle of the pangenome framework is to organize all the genes found across a group of related organisms into three functional categories. Core genes are found universally in all the genomes across one group and are mostly involved in functions that are necessary for basic cellular processes, such as metabolic pathways and organism survival. These genes form the conserved genetic core that defines the taxonomic group and maintains its most important biological characteristics. Accessory genes, on the other hand, occur in some but not all of the genomes in the population, often encoding functions that provide selective benefits in specific environmental settings, such as resistance to antibiotics, virulence, or specialized metabolism. Unique genes are found only in single genomes and can mark recent acquisitions, highly specialized functions, or evolutionary innovations that have not yet spread to the larger population[110, 112, 113].

Along with the above-described classification of genes, pangenomic studies provide quantitative analyses that facilitate more advanced understanding of evolutionary processes and adaptability[114, 115]. In the pangenomic framework this is quantified by the openness of

a population of closely related strains. The openness level within the pangenome, commonly quantified by Heaps' law γ-values, defines whether the population continues to gain new genetic content as more genomes are analyzed[116, 117]. Essentially, the level of openness determines if the addition of more relevant genomes to the group still reveals new genes or reaches a point of saturation. Open pangenomes ($\gamma > 0.6$) show an ongoing increase in the gene content with each subsequent genome, implying high horizontal gene transfer rates and ecological flexibility. Closed pangenomes ($\gamma < 0.3$) quickly reach a saturation level, indicating stable and conserved genetic arrangements with low rates of continued gene addition. Intermediate pangenomes ($0.3 \leq \gamma \leq 0.6$) show moderate levels of gene content increase.[117, 118]

The effectiveness of pangenomic approaches is supported by a vast range of novel studies involving different microbial systems. In the context of microbial pathogens, pangenomic analyses of *Escherichia coli* have revealed that this species contains a large reservoir of accessory genes, with every strain exhibiting only a part of the total genetic diversity, thus explaining the significant phenotypic differences between commensal and pathogenic isolates[119]. In a similar way, pangenomic analysis of *Salmonella* species has revealed accessory genes associated with host adaptation and virulence, providing essential insights into the genetic basis of host range and pathogenicity[120, 121]. In the realm of environmental microbiology, pangenomic analysis of marine *Prochlorococcus* populations has explained the maintenance of core metabolic functions across oceanic habitats, with accessory genes enabling adaptation to local variations in nutrient availability and environmental stresses[122]. Taken together, these studies highlight how the pangenome concept has enabled researchers to move beyond the single-genome paradigm, allowing a unified understanding of the genetic diversity and adaptive potential inherent within microbial populations.

AIM OF THE STUDY

To close the analytical gap that came apparent in the 2 projects mentioned in the previous chapter the central aim of this thesis is to create a framework to understand diversification of BGCs within gene cluster families by adapting the pangenome concept to BGC analysis. This approach will allow to identify core gene present in different gene cluster families which are most likely to produce the core structure of the compound, as well as accessory genes that are possibly used to introduce subtle modification to enable diversification. Additionally, by adapting, in multiple ways the openness metrics used in genomics, this research assesses whether BGC diversity is driven by the acquisition of more genes or by reshuffling existing genes within families.

To make these population-level analyses accessible to the broader research community, this work developed PanBGC-DB (https://panbgc-db.cs.uni-tuebingen.de/), an interactive web platform enabling exploration of BGC family diversity patterns and custom pangenomic analyses. This research will provide the first systematic framework for understanding diversification within BGC families, with applications for both fundamental biosynthetic research and natural product discovery.

INFORMATION ON THE MANUSCRIPTS

MANUSCRIPT 1: ORIGIN OF THE 3-METHYLGLUTARYL MOIETY IN CAPRAZAMYCIN BIOSYNTHESIS

CONTENT

This research discovered two pathways that supply 3-methylglutaryl-CoA for caprazamycin biosynthesis: one from the caprazamycin gene cluster itself (involving *Cpz5*), and another hijacked from the host cell's leucine/isovalerate utilization pathway (Liu-pathway). This represents the first report of the intermediate 3-methylglutaconyl-CoA being repurposed from primary metabolism for natural product formation. The study also found that *Cpz20* and *Cpz25* function in a common route where both pathways converge, highlighting the interplay between primary and secondary metabolism. My contribution to this work was performing the cblaster analysis to investigate the distribution of *liu* clusters across Actinobacteria genomes.

STATUS

Published 05 November 2022, doi: https://doi.org/10.1186/s12934-022-01955-6
Microbial Cell Factories

MANUSCRIPT 2: INTEGRATED GENOME AND METABOLOME MINING UNVEILED STRUCTURE AND BIOSYNTHESIS OF NOVEL LIPOPEPTIDES FROM A DEEP-SEA *RHODOCOCCUS*

CONTENT

This research used integrated genome and metabolome mining to discover novel lipopeptides from a deep-sea bacterium *Rhodococcus sp. I2R*. The study identified two new families of bioactive compounds: rhodoheptins (20 cyclic lipopeptides) and rhodamides (33 glycolipopeptides), by linking NRPS gene clusters to their corresponding metabolite products using mass spectrometry-based molecular networking. These lipopeptides exhibited biosurfactant activity and showed moderate anticancer effects against melanoma cells, expanding the known biosurfactant repertoire from marine bacteria. The work demonstrates how combining genomics with metabolomics can reveal cryptic biosynthetic pathways and lead to the discovery of structurally diverse natural products with potential biotechnological applications. My contribution to this work was performing the cblaster analysis to investigate the distribution of rhodoheptin and rhodamide biosynthetic gene clusters across *Rhodococcus* species.

STATUS

Published 24 November 2024, doi: https://doi.org/10.1111/1751-7915.70011
Microbial Biotechnology

MANUSCRIPT 3: PANBGC-DB: ADAPTATION OF THE PANGENOMIC FRAMEWORK ON A BIOSYNTHETIC GENE CLUSTER LEVEL

CONTENT

This research introduces PanBGC, a pangenome-inspired framework that adapts the classic pangenomic approach to biosynthetic gene clusters (BGCs) by treating gene cluster families as structured populations rather than isolated genomic units. Applied to over 250,000 BGCs from more than 35,000 genomes, PanBGC systematically classifies biosynthetic genes into core, accessory, and unique categories and quantifies compositional diversity through openness metrics. The study reveals that BGC diversification is primarily driven by modular reorganization of existing gene sets inside a GCF rather than acquisition of entirely novel genes, with most gene cluster families showing closed gene repertoires but high compositional variability. To make this framework accessible, PanBGC-DB was developed as an interactive web platform that enables researchers to explore biosynthetic diversity at population resolution, visualize evolutionary relationships, and contextualize newly discovered BGCs within the global landscape of secondary metabolism.

STATUS

Preprinted on BioRxiv, August 2025, doi: https://doi.org/10.1101/2025.08.11.669102
Currently submitted to ISME communication, September 2025, under consideration

MANUSCRIPT 1

Title: Origin of the 3-methylglutaryl moiety in caprazamycin biosynthesis

Authors: Daniel Bär, Benjamin Konetschny, Andreas Kulik, Houchao Xu, Davide Paccagnella, Patrick Beller, Nadine Ziemert, Jeroen S. Dickschat, Bertolt Gust

DECLARATION ON THE CONTRIBUTION OF CO-AUTHORS TO THE MANUSCRIPT

Author	Author position	Scientific ideas %	Data generation %	Analysis & interpretation %	Paper writing %
D. Bär	First author	60	50	50	70
B. Konetschny	Co-author	-	15	5	-
Andreas Kulik	Co-author	-	10	5	-
Houchao Xu	Co-author	-	10	-	-
D. Paccagnella	Co-author	-	5	5	-
P. Beller	Co-author	-	10	-	-
N. Ziemert	Corresponding Author	5	-	5	5
J.S. Dickschat	Corresponding Author	10	-	10	10
Bertolt Gust	Corresponding Author	25	-	20	15
Title of paper	Origin of the 3-methylglutaryl moiety in caprazamycin biosynthesis				
Status in publication process	Published 05 November 2022, doi: https://doi.org/10.1186/s12934-022-01955-6 Microbial Cell Factories				

Microbial Cell Factories

RESEARCH

Open Access

Origin of the 3-methylglutaryl moiety in caprazamycin biosynthesis

Daniel Bär[1], Benjamin Konetschny[1], Andreas Kulik[2], Houchao Xu[3], Davide Paccagnella[4], Patrick Beller[1], Nadine Ziemert[4,5], Jeroen S. Dickschat[3] and Bertolt Gust[1*]

Abstract

Background: Caprazamycins are liponucleoside antibiotics showing bioactivity against Gram-positive bacteria including clinically relevant *Mycobacterium tuberculosis* by targeting the bacterial MraY-translocase. Their chemical structure contains a unique 3-methylglutaryl moiety which they only share with the closely related liposidomycins. Although the biosynthesis of caprazamycin is understood to some extent, the origin of 3-methylglutaryl-CoA for caprazamycin biosynthesis remains elusive.

Results: In this work, we demonstrate two pathways of the heterologous producer *Streptomyces coelicolor* M1154 capable of supplying 3-methylglutaryl-CoA: One is encoded by the caprazamycin gene cluster itself including the 3-hydroxy-3-methylglutaryl-CoA synthase Cpz5. The second pathway is part of primary metabolism of the host cell and encodes for the leucine/isovalerate utilization pathway (Liu-pathway). We could identify the *liu* cluster in *S. coelicolor* M1154 and gene deletions showed that the intermediate 3-methylglutaconyl-CoA is used for 3-methylglutaryl-CoA biosynthesis. This is the first report of this intermediate being hijacked for secondary metabolite biosynthesis. Furthermore, Cpz20 and Cpz25 from the caprazamycin gene cluster were found to be part of a common route after both individual pathways are merged together.

Conclusions: The unique 3-methylglutaryl moiety in caprazamycin originates both from the caprazamycin gene cluster and the leucine/isovalerate utilization pathway of the heterologous host. Our study enhanced the knowledge on the caprazamycin biosynthesis and points out the importance of primary metabolism of the host cell for biosynthesis of natural products.

Keywords: *Streptomyces coelicolor*, Caprazamycin, Leucine/isovalerate utilization pathway, 3-methylglutaryl-CoA, Primary metabolism

Background

Caprazamycins (CPZs) belong to the family of liponucleoside antibiotics and were first isolated from *Streptomyces sp.* MK730-62F2 [1]. Their chemical core structure is built up by (+)-caprazol which consists of 5′-glycyluridine, 5-amino-D-ribose and a permethylated diazepanone ring. β-hydroxylated fatty acids are attached to the diazepanone ring resulting in formation of the biosynthetic intermediates hydroxyacylcaprazols. Attached to the free hydroxyl group of the fatty acid is a 3-methylglutarate bound to a permethylated L-rhamnose, affording the final caprazamycins. Caprazamycins are classified by chain length and constitution of the fatty acid [2, 3] (Fig. 1).

Their bioactivity is based on inhibition of phospho-MurNAc-pentapeptide transferase (MraY), which transfers phospho-MurNAc-pentapeptide from UDP-MurNAc-pentapeptide onto undecaprenyl phosphate (C55-P) during bacterial cell wall

*Correspondence: bertolt.gust@uni-tuebingen.de

[1] Department of Pharmaceutical Biology, Eberhard-Karls University Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany
Full list of author information is available at the end of the article

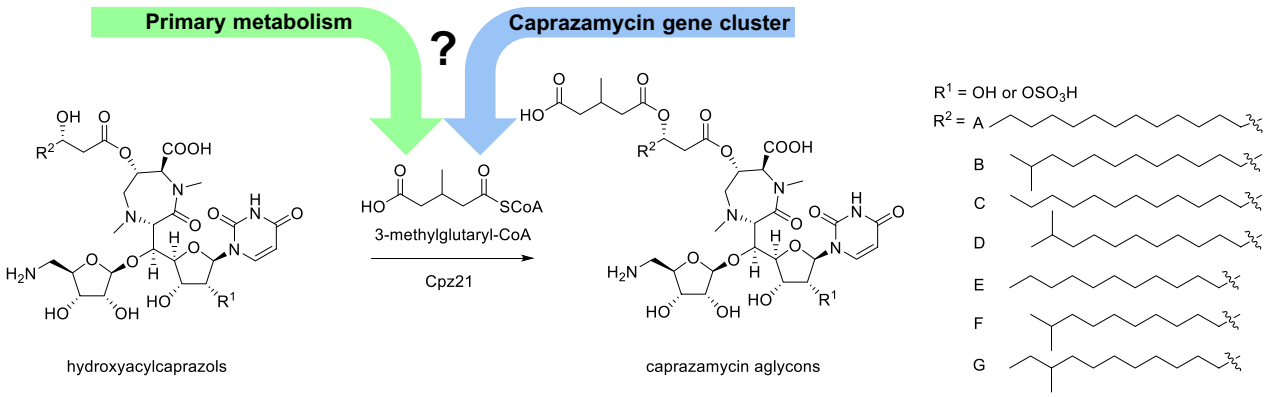


Fig. 1 Chemical structure of hydroxyacylcaprazols and caprazamycins aglycons. Attachment of 3-methylglutaryl-CoA is catalyzed by Cpz21

biosynthesis [4–6]. Caprazamycin B shows promising activity against Gram-positive bacteria including strains of the genus *Mycobacterium*, such as the clinically relevant pathogen *M. tuberculosis* [1, 7]. Although it was reported for liposidomycins, that the 3-methylglutaryl moiety increases inhibition of peptidoglycan synthesis, newly generated caprazamycin derivates were focused on truncating the chemical structure thus sacrificing this moiety [8]. In palmitoylcaprazol, the fatty acid residue was replaced by a simpler palmitoyl side chain resulting in similar potency against *M. smegmatis* and *M. tuberculosis* compared to caprazamycin B [9, 10]. Further studies showed that the key scaffold for antimicrobial activity consists of uridine, aminoribose, the diazepanone ring and the fatty acid chain, whereas the 3-methylglutaryl moiety and the rhamnose seemed to be dispensable [10, 11]. For CPZEN-45, replacing the fatty acid moiety by 4-butylanilide showed improved efficacy and lower toxicity compared to caprazamycins, but this compound shifted its target from MraY to WecA in *M. tuberculosis* [7, 12, 13].

The caprazamycin gene cluster was the first cluster of a MraY inhibitor to be identified and verified by heterologous expression in *S. coelicolor* M512. Later, the biosynthetic gene cluster of the closely related liposidomycins was reported [14, 15]. Genes required for biosynthesis of rhamnose are not located on the CPZ gene cluster but were identified elsewhere on the genome of *Streptomyces sp.* MK730-62F2. Since the heterologous host *S. coelicolor* is missing those genes, heterologous expression resulted in the formation of caprazamycin aglycons instead of caprazamycins [16]. Deletion of the carboxylesterase Cpz21 resulted in the formation of hydroxyacylcaprazols indicating Cpz21 to be responsible for the transfer of 3-methylglutaryl-CoA onto the β-hydroxyl group of hydroxyacylcaprazols [14].

Although the steps of caprazamycin biosynthesis are understood to some extent, no biosynthetic route leading to 3-methylglutaryl-CoA has been described so far [17]. Investigating the caprazamycin gene cluster, a starting point for 3-methylglutaryl-CoA biosynthesis could be catalyzed by the putative 3-hydroxy-3-methylglutaryl-CoA synthase Cpz5. This class of enzymes usually converts acetyl-CoA and acetoacetyl-CoA to 3-hydroxy-3-methylglutaryl-CoA [18–20]. A removal of the hydroxyl group would then lead to the desired 3-methylglutaryl-CoA.

Considering that precursors for natural products could also be provided by the host cell itself, we analyzed the primary metabolism of *S. coelicolor* M1154 for plausible intermediates leading to 3-methylglutaryl-CoA as well. A rich source of short-chain acyl-CoAs is the degradation of branched-chain amino acids (BCAA) leucine, valine and isoleucine [21]. The first step in degradation of all three branched-chain amino acids is transamination followed by oxidative decarboxylation to the corresponding acyl-CoA-thioesters carried out by a branched-chain α-keto acid dehydrogenase (BCDH) complex [22–24]. Next, this pathway diverges into three branches, one for each amino acid [21]. The leucine/isovalerate utilization pathway (Liu-pathway) of *P. aeruginosa* PAO1 is well described and it was also investigated in *P. putida* PpG2, *M. xanthus* DK1622 and *M. luteus* [25–32] (Additional file 1: Fig. S1). A comparative genomics study of the Liu-pathway regulation showed that this cluster is also widely distributed among protobacteria [33]. The catabolic pathway of leucine continues with isovaleryl-CoA being converted to 3-methylcrotonyl-CoA in a dehydrogenation step catalyzed by isovaleryl-CoA dehydrogenase LiuA. Enzymes facilitating this step have also been reported for *S. coelicolor* J802 encoded by *acdH* (*sco2779*) and for *S. avermitilis* ATCC 31272 encoded by *fadE4*

(*SAVERM_5275*) [34]. Next, a 3-methylcrotonyl-CoA carboxylase complex consisting of subunits α and β (LiuD and LiuB) generates 3-methylglutaconyl-CoA by transfer of acetyl-CoA onto 3-methylcrotonyl-CoA. The identification of a similar α-subunit was also reported for *S. toxytricini* [35]. 3-methylglutaconyl-CoA is further converted to 3-hydroxy-3-methylglutaryl-CoA by 3-methylglutaconyl-CoA hydratase LiuC and in a final step, the 3-hydroxy-3-methylglutaryl-CoA lyase LiuE generates acetyl-CoA and acetoacetate. Intriguingly, all intermediates of the Liu-pathway resembling 3-methylglutaryl-CoA in structure making them promising starting points for a biosynthetic route towards 3-methyglutaryl-CoA originating from the host cells primary metabolism.

In this work, we could show that two routes are leading to 3-methylglutaryl-CoA for caprazamycin biosynthesis. The first is encoded by the caprazamycin gene cluster and starts by the action of the 3-hydroxy-3-methylglutaryl-CoA synthase Cpz5. The second pathway derives from the catabolism of leucine as part of the primary metabolism of the host cell. Our study enhances the knowledge on the biosynthesis of liponucleoside antibiotics and lays the foundation for further studies on the alteration of the 3-methylglutaryl moiety towards new caprazamycin analogues.

Results

Deletion of the 3-hydroxy-3-methylglutaryl-CoA synthase Cpz5 does not abolish caprazamycin aglycon formation

The biosynthetic gene cluster of caprazamycins encodes for Cpz5, a putative 3-hydroxy-3-methylglutaryl-CoA synthase (Additional file 1: Fig. S2). This family of enzymes catalyze the Claisen-like condensation of acetyl-CoA and acetoacetyl-CoA to 3-hydroxy-3-methylglutaryl-CoA. So far, the role of Cpz5 during caprazamycin biosynthesis has not been investigated. Due to the structural similarity to 3-methylglutaryl-CoA, we predicted that 3-hydroxy-3-methylglutaryl-CoA generated by Cpz5 could be a promising starting point of a biosynthetic route towards 3-methylglutaryl-CoA entirely encoded by the caprazamycin gene cluster. To test this hypothesis, a deletion of *cpz5* was generated on cosmid cpzLK09, which contains the entire caprazamycin gene cluster, yielding cosmid cpzDB04. Heterologous expression of cpzLK09 in *S. coelicolor* leads to the accumulation of caprazamycin aglycons because the heterologous host is lacking the genes required to produce L-rhamnose, whereas a mutant not capable of providing 3-methylglutaryl-CoA should stop production at the stage of hydroxyacylcaprazols [16]. We generated three individual mutants either containing cpzLK09 resulting in *S. coelicolor* M1154/cpzLK09 (1)–(3) or harboring cpzDB04 resulting in *S. coelicolor* M1154/cpzDB04 (1)–(3). As expected, all three *S. coelicolor* M1154/cpzLK09 mutants were able to produce caprazamycin aglycons, showing signals with high intensities for caprazamycin aglycons A and B with *m/z* 958.5 at Rt of 14.9 min, caprazamycin aglycons C, D and G with *m/z* 944.5 at Rt of 14.3 min and caprazamycin aglycons E and F with *m/z* 930.5 at Rt of 13.8 min. Masses of hydroxyacylcaprazols A and B with *m/z* 830.5 expected at Rt of 14.0 min, hydroxyacylcaprazols C, D and G with *m/z* 816.5 expected at Rt of 13.4 min and hydroxyacylcaprazols E and F with *m/z* 802.5 expected at Rt of 12.9 min were not detected. In contrast to our hypothesis, a deletion of *cpz5* could not impair the formation of caprazamycin aglycons as they were still produced by all three gene deletion mutants (Additional file 1: Figs. S3–S5). Those findings indicated that *cpz5* is not exclusively responsible for 3-methylglutaryl-CoA formation.

Identification of the leucine/isovalerate utilization pathway as precursor supply for 3-methylglutaryl-CoA biosynthesis

Gene deletion could not confirm *cpz5* as the sole source of 3-methylglutaryl-CoA for caprazamycin biosynthesis. To discover other pathways for the generation of this intermediate, we investigated the primary metabolism of *S. coelicolor* M1154 in more detail. Since 3-methylglutaryl-CoA is not described to be part of a primary metabolism pathway so far, we decided to look for primary metabolism pathways processing plausible precursors of this compound, including 3-hydroxy-3-methylglutaryl-CoA, that can be readily transformed into 3-methylglutaryl-CoA. One promising lead were degradation pathways of amino acids, as they are ubiquitously distributed among bacteria and intermediates of their degradation processes are short-chain acyl-CoAs similar to 3-methylglutaryl-CoA. Utilization of leucine and isovalerate was described for *Pseudomonas aeruginosa* PAO1 in more detail [26–29]. In this strain, a gene cluster encoding for a reaction cascade called leucine/isovalerate utilization pathway was identified to process leucine and isovalerate to acetoacetate and acetyl-CoA. This gene cluster consists of six genes encoding for an isovaleryl-CoA dehydrogenase (*liuA*), two subunits of a 3-methylcrotonyl-CoA carboxylase (*liuB* and *liuD*), a 3-methylglutaconyl-CoA dehydratase (*liuC*), a 3-hydroxy-3-methylglutaryl-CoA lyase (*liuE*) and a transcriptional regulator (*liuR*). BLAST analysis of *S. coelicolor* M1154 revealed genes homologue to *liuA* (*sco2779*), *liuB* (*sco2776*), *liuD* (*sco2777*) and *liuE* (*sco2778*) (Additional file 1: Fig. S6). No homologue was found for *liuC* in this cluster though, raising the question if this gene is located elsewhere on the genome. Interestingly, besides the missing *liuC*, the *liu* homologues in *S. coelicolor* are ordered in a different genetic organization compared to *P. aeruginosa* PAO1. We found

the same genetic organization in caprazamycin wildtype producer *S. sp.* MK730-62F2 and it was also reported for *S. avermitilis* ATCC 31267 [36]. A cblaster analysis using the *liu* cluster from *S. coelicolor* M1154 including the putative regulator *sco2775* as query revealed, that only 4.5% of all *Actinobacteria* strains but more than 52% of *Streptomyces* strains listed in the Genome Taxonomy Database (GTDB) possess homologues of all five query genes. (Additional file 1: Fig. S7A). To investigate if the Liu-pathway is connected to caprazamycin biosynthesis, we deleted *sco2776* to *sco2779* in the heterologous host *S. coelicolor* M1154 resulting in *S. coelicolor* M1154Δ*sco2776-sco2779*. Successful deletion and absence of wildtype genes was verified by PCR and sequencing. Introduction of either cpzLK09 or cpzDB04 into this strain resulted in mutants *S. coelicolor* M1154Δ*sco2776-sco2779*/cpzLK09 (1)–(3) and *S. coelicolor* M1154Δ*sco2776-sco2779*/cpzDB04 (1)–(3). Successful cosmid integration was verified by PCR and sequencing. HPLC–MS analysis revealed that production of caprazamycin aglycons was abolished when both *cpz5* and *sco2776-sco2779* were deleted, whereas production of caprazamycin aglycons was still possible if only *sco2776-sco2779* were missing but *cpz5* was still intact (Additional file 1: Figs. S8, S9). This indicates that both pathways are generating 3-methylglutaryl-CoA independently. Formation of the correct product was confirmed by HPLC–MSMS (Additional file 1: Fig. S10). To find out which exact intermediate of the Liu-pathway is hijacked for caprazamycin biosynthesis, we next generated mutants with single deletions of *sco2776*, *sco2777*, *sco2778* or *sco2779* resulting in *S. coelicolor* M1154Δ*sco2776, S. coelicolor* M1154Δ*sco2777, S. coelicolor* M1154Δ*sco2778* and *S. coelicolor* M1154Δ*sco2779*. Successful deletion and absence of wildtype genes was again verified by PCR and sequencing. Introduction of cpzLK09 resulted in mutants *S. coelicolor* M1154Δ*sco2776*/cpzLK09 (1)–(3), *S. coelicolor* M1154Δ*sco2777*/cpzLK09 (1)–(3), *S. coelicolor* M1154Δ*sco2778*/cpzLK09 (1)–(3) and *S. coelicolor* M1154Δ*sco2779*/cpzLK09 (1)–(3) whereas introduction of cpzDB04 led to mutants *S. coelicolor* M1154Δ*sco2776*/cpzDB04 (1)-(3), *S. coelicolor* M1154Δ*sco2777*/cpzDB04 (1)–(3), *S. coelicolor* M1154Δ*sco2778*/cpzDB04 (1)–(3) and *S. coelicolor* M1154Δ*sco2779*/cpzDB04 (1)–(3). Successful cosmid integration was verified by PCR and sequencing. As expected, all mutants containing a complete caprazamycin gene cluster were still capable of producing caprazamycin aglycons. With *cpz5* deleted on the caprazamycin gene cluster, mutants with a missing *sco2776* or *sco2777* ceased caprazamycin aglycon production and accumulated hydroxyacylcaprazols instead, whereas a deletion of *sco2778* or *sco2779* had no impact on caprazamycin aglycon formation (Fig. 2 and Additional file 1: Figs. S11–S18).

Those results strongly suggest, that either 3-methylglutaconyl-CoA or 3-hydroxy-3-methylglutaryl-CoA is the common precursor for 3-methylglutaryl-CoA for both the gene cluster encoded pathway and the pathway originating from primary metabolism.

Two dehydrogenases are converting isovaleryl-CoA to 3-methylcrotonyl-CoA

Surprisingly, a single deletion of *sco2779* together with a *cpz5*-deletion did not abolish caprazamycin aglycon production, although we predicted this gene to encode for the enzyme catalyzing the first step in the Liu-pathway. To verify that 3-methylcrotonyl-CoA is truly a precursor in the biosynthesis of 3-methylglutaryl-CoA, we synthesized (1-^{13}C)-3-methylcrotonyl-SNAc. Feeding this compound to *S. coelicolor* M1154/cpzLK09 resulted in a change of isotope distribution towards heavier caprazamycin aglycons, indicating that (1-^{13}C)-3-methylglutaryl-SNAc was incorporated and therefore, 3-methylcrotonyl-CoA is utilized for caprazamycin biosynthesis. (Additional file 1: Fig. S19). Due to its location on the *liu*

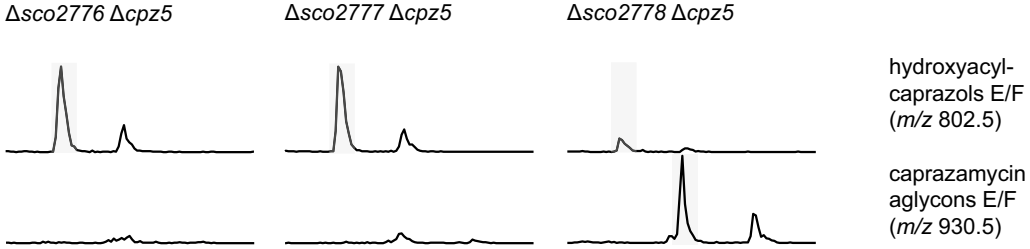


Fig. 2 Extracted ion chromatograms from extracts of *S. coelicolor* M1154 mutants with either *sco2776*, *sco2777* or *sco2778* deleted and harboring a caprazamycin gene cluster containing a *cpz5* deletion. Retention time from 12 to 16 min is shown. Captions indicate mutant genotypes. Highlighted are signals for hydroxyacylcaprazols E/F with *m/z* of 802.5 and caprazamycin aglycons E/F with *m/z* of 930.5 acquired in positive mode. Non-highlighted signals represent sulfated hydroxyacylcaprazols E/F or sulfated caprazamycin aglycons E/F respectively

operon, *sco2779* presumably encodes for an acyl-CoA dehydrogenase capable of converting isovaleryl-CoA to 3-methylcrotonyl-CoA, but our gene deletion experiment suggested that it is not the sole enzyme catalyzing this step. Thus, we searched for other genes annotated as acyl-CoA dehydrogenases on the genome of *S. coelicolor* A3(2) in the StrepDB database and found 38 candidates that could possibly complement *sco2779* [37]. A BLAST search revealed 23 homologues for this gene located on the *S. coelicolor* M1154 genome with identities ranging from about 20% to 40%. The most promising candidate turned out to be *sco2774*. This gene is located just next to the *liu* cluster and is orientated in the same direction as the cluster's putative regulator *sco2775* (Additional file 1: Fig. S6). To clarify the roles of *sco2774* and *sco2779*, we generated mutants with a single deletion of *sco2774* resulting in *S. coelicolor* M1154Δ*sco2774* and a mutant with a deletion of both *sco2774* and *sco2779* resulting in *S. coelicolor* M1154Δ*sco2774*Δ*sco2779*. Successful deletion and absence of wildtype genes was again verified by PCR and sequencing. After introduction of either cpzLK09 or cpzDB04, caprazamycin production of mutants *S. coelicolor* M1154Δ*sco2774*/cpzLK09 (1)–(3), *S. coelicolor* M1154Δ*sco2774*/cpzDB04 (1)–(3), *S. coelicolor* M1154Δ*sco2774*Δ*sco2779*/cpzLK09 (1)–(3) and *S. coelicolor* M1154Δ*sco2774*Δ*sco2779*/cpzDB04 (1)–(3) was analyzed by HPLC–MS. As expected, mutants containing the complete caprazamycin gene cluster were still capable of producing caprazamycin aglycons. A single deletion of *sco2774* together with a *cpz5* deletion did not impair caprazamycin aglycon production. However, if both *sco2774* and *sco2779* were deleted together with *cpz5*, hydroxyacylcaprazols accumulated instead (Fig. 3 and Additional file 1: Figs. S20–S23).

These findings strongly suggest that both acyl-CoA dehydrogenases are converting isovaleryl-CoA to 3-methylcrotonyl-CoA. To support this thesis, we performed a chemical complementation by adding 3-methylcrotonyl-SNAc to a culture of M1154Δ*sco2774*Δ*sco2779*/cpzDB04. HPLC–MS analysis revealed that caprazamycin aglycon biosynthesis could be restored by addition of 3-methylcrotonyl-SNAc, indicating a successful restoration of the Liu-pathway (Fig. 3 and Additional file 1: Fig. S24). Since our data suggests that *sco2774* is an extension of the *liu* cluster, we performed a cblaster analysis with this extended cluster as query sequence. Interestingly, 48.2% of *Streptomyces* strains still contained all six query genes compared to 52.2% of the query without *sco2774*, indicating that a *liu* cluster with a second acyl-CoA dehydrogenase is commonly distributed in *Streptomyces* (Additional file 1: Fig. S7B).

Involvement of the acyl-CoA synthase Cpz20 and the dehydrogenase Cpz25 in 3-methylglutaryl-CoA biosynthesis

So far, our results narrowed down precursor candidates for 3-methylglutaryl-CoA to either 3-methylglutaconyl-CoA synthesized by Sco2776 and Sco2778 or 3-hydroxy-3-methylglutaryl-CoA generated by Cpz5. A conversion of 3-methylglutaconyl-CoA to 3-methylglutaryl-CoA requires a reduction of the C–C double bond. The caprazamycin gene cluster encodes for Cpz25 which is a promising candidate for this reduction step. A BLAST analysis revealed that Cpz25 belongs to the medium chain reductase/dehydrogenase (MDR)/zinc-dependent alcohol dehydrogenase-like family of proteins (cd05188) and contains a conserved domain of enoyl-reductases from polyketide synthases (smart00829). Another enzyme that could be involved in 3-methylglutaryl-CoA biosynthesis is a putative acyl-CoA synthase encoded by *cpz20*. To investigate if *cpz20* and *cpz25* do in fact play a role in 3-methylglutaryl-CoA biosynthesis, both genes were deleted individually on the caprazamycin gene cluster, resulting in cosmids cpzDB05 and cpzDB06. Subsequent transfer into *S. coelicolor* M1154 resulted in mutants *S. coelicolor* M1154/cpzDB05 (1)–(3)

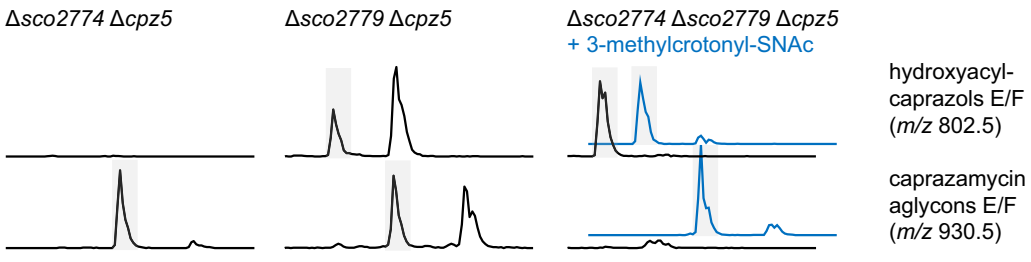


Fig. 3 Extracted ion chromatograms from extracts of *S. coelicolor* M1154 mutants with either *sco2774*, *sco2779* or both deleted and harboring a caprazamycin gene cluster containing a *cpz5* deletion. Retention time from 12 to 16 min is shown. Captions indicate mutant genotypes. Highlighted are signals for hydroxyacylcaprazols E/F with *m/z* of 802.5 and caprazamycin aglycons E/F with *m/z* of 930.5 acquired in positive mode. Non-highlighted signals represent sulfated hydroxyacylcaprazols E/F or sulfated caprazamycin aglycons E/F respectively. Extract of a culture supplemented with 3-methylcrotonyl-SNAc is shown in blue

and *S. coelicolor* M1154/cpzDB06 (1)-(3). Successful cosmid integration was verified by PCR and sequencing and caprazamycin production was analyzed by HPLC–MS. Neither *cpz20* nor *cpz25* deficient mutants were able to accumulate caprazamycin aglycons. However, hydroxyacylcaprazols were still produced, indicating that both *cpz20* and *cpz25* are indeed essential for supplying 3-methylglutaryl-CoA (Fig. 4 and Additional file 1: Figs. S25, S26).

Intriguingly, both *cpz20* and *cpz25* deletions were heterologously expressed in a *S. coelicolor* M1154 strain with *sco2774-sco2779* still intact. This strongly suggests that both enzymes are not specific for either one of the 3-methylglutaryl-CoA supply pathways. We predict that Cpz25 is the junction that merges both pathways into one common route by reduction of 3-methylglutaconyl-CoA and Cpz20 could be involved in an additional activation related step prior transfer onto hydroxyacylcaprazols by Cpz21.

Two pathways supply 3-methylglutaryl-CoA for caprazamycin biosynthesis

Based on the results obtained in this study, we are able to propose the first model of the origin of 3-methylglutaryl-CoA for caprazamycin biosynthesis (Scheme 1).

Our investigation identified two pathways involved in supplying this precursor. First, the degradation of branched-chain amino acids leucine and isovalerate as part of the hosts primary metabolism. Second, a pathway encoded on the caprazamycin gene cluster itself. Surprisingly, these two pathways are not fully independent from each other. The leucine/isovalerate degradation ends with generating acetyl-CoA and acetoacetate by action of the 3-hydroxy-3-methylglutaryl-CoA lyase Sco2779. The caprazamycin gene cluster encoded pathway starts with a similar reverse reaction catalyzed by the 3-hydroxy-3-methylglutaryl-CoA synthase Cpz5 utilizing acetoacetyl-CoA and acetyl-CoA continuing to work in the Liu-pathway's opposite direction until it reaches 3-methylglutaconyl-CoA. We identified this compound as the central intermediate where both pathways fuse and continue on a joint route. This route starts with a reduction step in which 3-methylglutaconyl-CoA is converted to the desired 3-methylglutaryl-CoA. Moreover, our results strongly suggest that Cpz20, a putative acyl-CoA synthase, is also required on this route because a single deletion of *cpz20* ceased caprazamycin aglycon production the same way as a *cpz25* deletion did. Finally, 3-methylglutaryl-CoA is transferred onto the hydroxyacylcaprazols by Cpz21 [14].

Discussion

The branched-chain amino acids degradation pathways as versatile suppliers of precursors

The catabolism of branched-chain amino acids is a rich source of precursors for metabolites and natural products. It starts with leucine, isoleucine and valine undergoing transamination and subsequent oxidative decarboxylation by the branched-chain dehydrogenase complex resulting in isovaleryl-CoA, 2-methylbutyryl-CoA and isobutyryl-CoA (Fig. 5).

Important products of these intermediates are branched-chain fatty acids [38]. Caprazamycins rely on all three of these intermediates by incorporating iso-even, iso-odd and anteiso fatty acids [2]. Our work showed that the 3-methylglutaryl moiety of caprazamycins is derived from another intermediate of leucine degradation, 3-methylglutaconyl-CoA, revealing an additional dependency of caprazamycin biosynthesis on this primary metabolism pathway. Members of A21978C, a complex of lipopeptide antibiotics including the clinically relevant daptomycin produced by *S. roseosporus*, and the tripropeptines produced by *Lysobacter sp.* BMK333-48F3 incorporate branched-chain fatty acids as well [39–41]. Iso-odd branched-chain fatty acids are further important for cell membrane fluidity and comprise about 75%

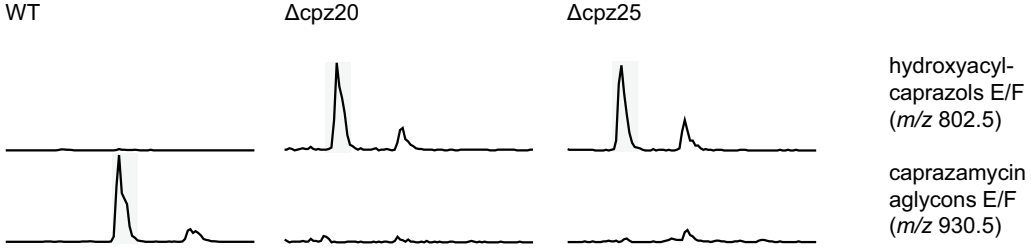
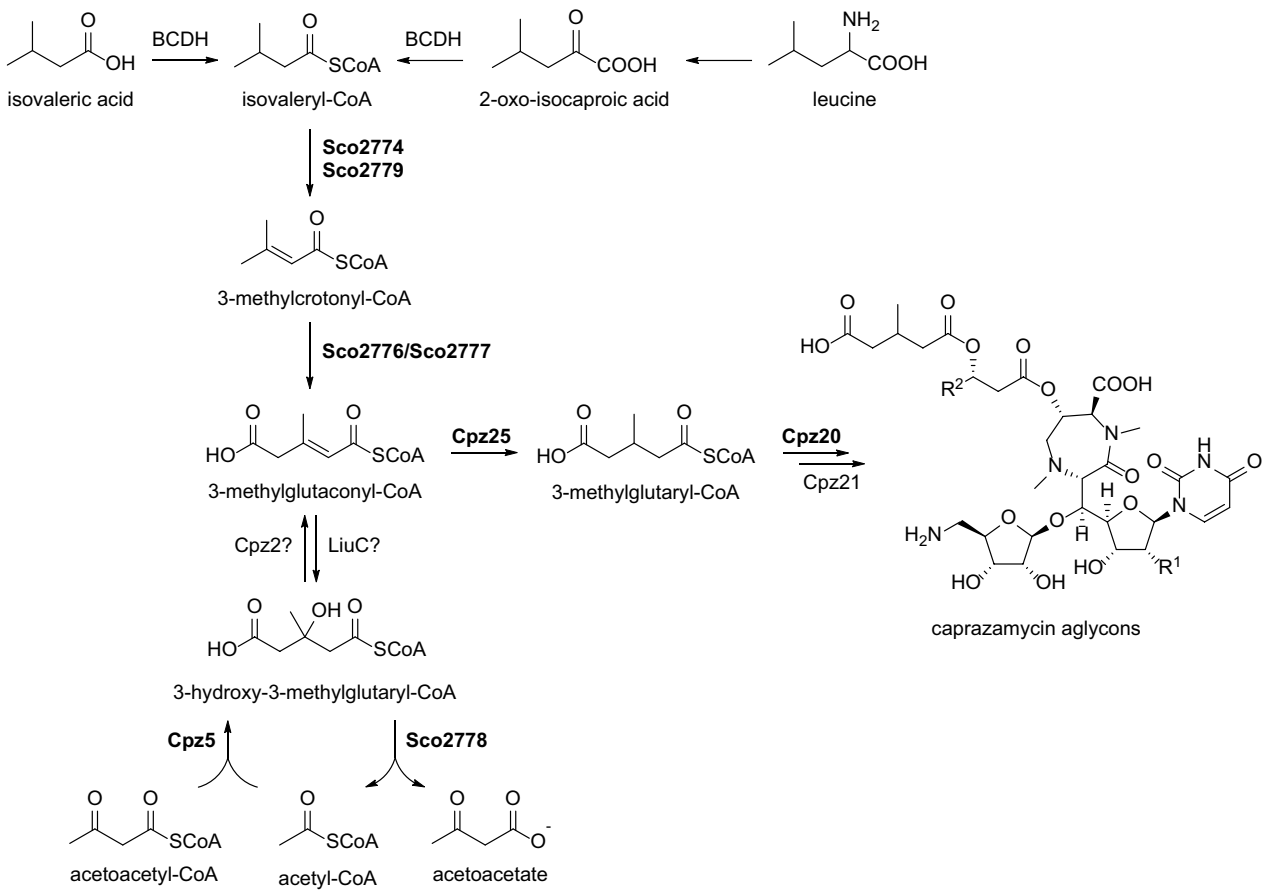


Fig. 4 Extracted ion chromatograms from extracts of *S. coelicolor* M1154 mutants harboring a caprazamycin gene cluster containing a deletion of either *cpz20* or *cpz25*. Retention time from 12 to 16 min is shown. Highlighted are signals for hydroxyacylcaprazols E/F with *m/z* of 802.5 and caprazamycin aglycons E/F with *m/z* of 930.5 acquired in positive mode. Non-highlighted signals represent sulfated hydroxyacylcaprazols E/F or sulfated caprazamycin aglycons E/F respectively



Scheme 1 Proposed biosynthesis model for the 3-methylglutaryl moiety in caprazamycin biosynthesis. In bold: Steps investigated in this study. BCDH: Branched-chain α-keto acid dehydrogenase complex. Sco2774: acyl-CoA dehydrogenase. Sco2779: isovaleryl-CoA dehydrogenase (LiuA). Sco2776: 3-methylcrotonyl-CoA carboxylase subunit β (LiuB). Sco2777: 3-methylcrotonyl-CoA carboxylase subunit α (LiuD). Sco2778: 3-hydroxy-3-methylglutaryl-CoA lyase (LiuE). Cpz5: 3-hydroxy-3-methylglutaryl-CoA synthase. Cpz20: acyl-CoA synthase. Cpz21: carboxylesterase. Cpz25: reductase

of all fatty acids in *Myxobacteria* [42]. *Stigmatella aurantiaca* DW4/3–1 utilizes isovaleryl-CoA as an unusual starter unit for biosynthesis of myxothiazol and aurafuron A and B [43–45]. The same starter unit is used for sulfangolides A-C and the closely related kulkenon isolated from different strains of *Sorangium cellulosum* [46]. Usage of the unusual extender unit 2-carboxy-3-hydroxy-5-methylhexanoyl-CoA derived from isovaleryl-CoA was reported for leupyrrin biosynthesis in *S. cellulosum* So ce690 [47]. An alternative pathway to isovaleryl-CoA was described in *M. xanthus* which reverses the Liu-pathway resembling the caprazamycin gene cluster encoded pathway for 3-methylglutaryl-CoA biosynthesis [48]. This alternative pathway contains a 3-hydroxy-3-methylglutaryl-CoA synthase (MvaS) similar to Cpz5 generating 3-hydroxy-3-methylglutaryl-CoA from acetoacetyl-CoA and acetyl-CoA [49]. Surprisingly, the next step in this cascade is catalyzed by the LiuC homologue of *M. xanthus* itself, indicating that LiuC is able to switch direction of catalysis depending on the metabolic state of the cell [50]. Since no LiuC homologue exists in the *liu* cluster of *S. coelicolor* M1154, we identified candidate genes homologue to LiuC from *M. xanthus* elsewhere on the genome that could fill the gap of the caprazamycin gene cluster encoded pathway. However, single gene inactivation by transposon insertion of those candidate genes (*sco1838*, *sco4384*, *sco4930* and *sco6732*) in a Liu-pathway deficient background could not impair caprazamycin aglycon production in a mutant harboring the complete caprazamycin gene cluster (Additional file 1: Figs. S27–S30). This raises the question whether the correct gene was amongst our candidates for inactivation, if more than one gene is responsible for the LiuC catalyzed reaction or if the LiuC homologue in *S. coelicolor* is not able to catalyze the reaction in both directions as it is the case in *M. xanthus*. Small adaption of an BCAA

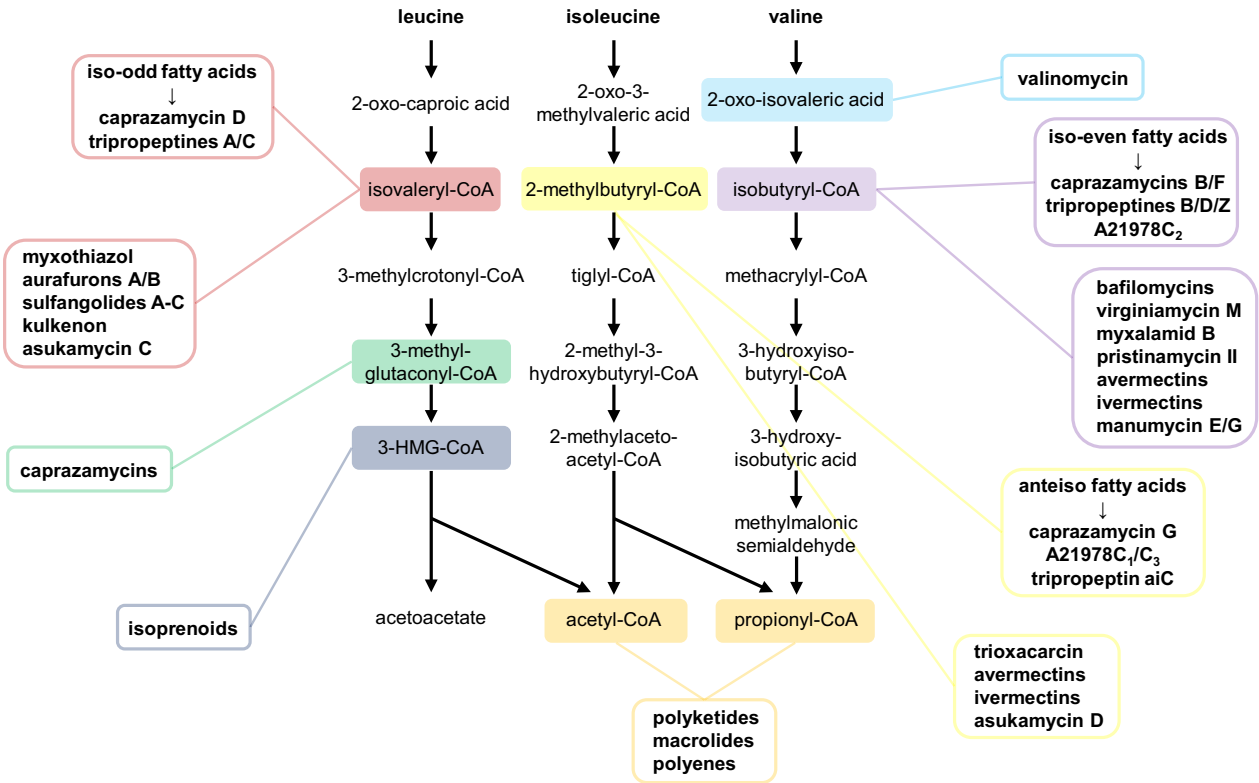


Fig. 5 Degradation pathways of branched-chain amino acids leucine, isoleucine and valine and natural products utilizing intermediates of those pathways for their biosynthesis

intermediate prior to incorporation is an observation we made for 3-methylglutaconyl-CoA converted by Cpz25 during this work. Another example is found in the biosynthesis of cyclic peptide valinomycin in *Streptomyces sp.* M10. This biosynthesis requires D-hydroxyisovaleric acid, which is afforded via reduction of 2-oxo-isovaleric acid, the first intermediate of valine degradation, by action of a D-hydroxyisovalerate dehydrogenase. In the same strain, production of the antifungal-active bafilomycins competes for 2-oxo-isovaleric acid by utilizing its BCDH product isobutyryl-CoA [51]. Isobutyryl-CoA was further suggested as the starter unit for the biosynthesis of virginiamycin M in *S. virginiae*, myxalamid B in *Stigmatella aurantiaca* Sg a15 and *Myxococcus xanthus* DK1622 and pristinamycin II in *S. pristinaespiralis* [52–54]. Trioxacarcin, first isolated from S. *bottropensis* DO-45, rather uses 2-methylbutyryl-CoA derived from isoleucine building an unusual spiro-epoxide structure which is believed to bind DNA as part of its mode of action [55, 56]. Avermectins and their hydrogenated derivates ivermectins, antiparasitic compounds isolated from *S. avermitilis*, carrying either a 2-methylbutyryl or isobutyryl moiety, relying their production on the degradation of isoleucine or valine [22, 57, 58]. Members of the manumycin family are distinguished by the utilized PKS starter unit including all three: isovaleryl-CoA, 2-methylbutyryl-CoA and isobutyryl-CoA [59, 60].

The end products of BCAA catabolism, propionyl-CoA and acetyl-CoA, are versatile starter units for PKS derived natural products. Those include therapeutically important antimicrobial polyketides such as erythromycin and antifungal polyenes such as nystatin [61–63]. Although acetyl-CoA can be obtained from several sources including glycolysis, studies in *S. coelicolor* demonstrated that branched-chain amino acid degradation is an important pathway for acetyl-CoA supply for actinorhodin biosynthesis [24]. However, overexpression of the BCDH cluster aiming to elevate methylmalonyl-CoA levels increased pikromycin production only by 1.3-fold, whereas overexpression of methylmalonyl-CoA mutase achieved a 1.7-fold increase, pointing out a possible limitation of methylmalonyl-CoA supply by this pathway [64].

Despite the numerous natural products derived from either BCDH complex intermediates or the end products of BCAA degradation, surprisingly little is known about other BCAA degradation intermediates being utilized for secondary metabolite biosynthesis. The mevalonate pathway leading to the large group of isoprenoids relies

on 3-hydroxy-3-methylglutaryl-CoA [18]. However, the mevalonate pathway is equipped with a 3-hydroxy-3-methylglutaryl-CoA synthase similar to Cpz5 generating 3-hydroxy-3-methylglutaryl-CoA and is therefore not dependent on BCAA degradation. In this work, we report for the first time that the leucine/isovalerate degradation pathway intermediate 3-methylglutaconyl-CoA is being utilized as a precursor for natural product biosynthesis.

The caprazamycin gene cluster provides 3-methylglutaryl-CoA from acetyl-CoA and acetoacetyl-CoA

The 3-methylglutaryl in caprazamycin is a unique moiety only found in the closely related liposidomycins, muraminomicins, A-84830A and A-90289A [65, 66]. The gene cluster of liposidomycin was identified in *S. sp.* SN-1061 M. Since liposidomycins share the same core structure as caprazamycins and only differ in the fatty acid chain composition and absence of the rhamnosyl moiety, both clusters share highly similar genetic organization and homology [14, 15]. The same biosynthetic machinery leading to 3-methylglutaryl-CoA we discovered in this work could also be assembled by the liposidomycin gene cluster encoding for the 3-hydroxy-3-methylglutaryl-CoA synthase LpmA (81/89% identity/similarity to Cpz5), the acyl-CoA synthase LpmR (88/91% identity/similarity to Cpz20) and the acyl dehydrogenase LpmW (91/95% identity/similarity to Cpz25). The biosynthetic gene cluster of A-90289 in *S. sp.* SANK 60,405 encodes for LipC (82/88% identity/similarity to Cpz5), LipQ (88/92% identity/similarity to Cpz20) and LipV (90/95% identity/similarity to Cpz25) [67]. Although the muraminomicin gene cluster encodes for Mra13 (87/90% identity/similarity to Cpz20) and Mra8 (88/93% identity/similarity to Cpz25), it lacks a homologue of *cpz5*, raising the question if this biosynthesis relies exclusively on primary metabolism for 3-methylgutaryl-CoA precursor supply [68].

One step in the caprazamycin gene cluster encoded pathway is dehydration of 3-hydroxy-3-methylglutaryl-CoA generating 3-methylglutaconyl-CoA. As already described, we could not identify a gene from the heterologous host *S. coelicolor* M1154 responsible this reaction. The caprazamycin gene cluster itself encodes with the putative dehydratase *cpz2* for another candidate suitable for catalyzing this reaction [14]. However, a single deletion of *cpz2* did not interfere with production of caprazamycin aglycons in a Liu-pathway deficient background (Additional file 1: Fig. S31). Furthermore, analyzing the gene clusters of liposidomycin, A-90289A and muraminomicins, no homologues of *cpz2* could be identified [15, 67, 68]. It remains unclear if *cpz2* is participating in the generation of 3-methylglutaconyl-CoA along with other enzymes from the primary metabolism or if other enzymes are responsible for this dehydration step.

So far, no other natural products than the liposidomycin family of nucleoside antibiotics are known to contain a 3-methylglutaryl moiety. We could show that generating this moiety is fully dependent on genes located on the caprazamycin gene cluster. However, a degradation pathway for 4-methylbenzoyl-CoA utilized under anaerobic conditions was reported for *Magnetospirillum sp.* pMbN1. An intermediate of this degradation process is indeed 3-methylgluaryl-CoA which is converted to 3-methylglutaconyl-CoA followed by the same steps as shown for the Liu-pathway, revealing that 3-methylglutaryl-CoA can also be part of the cell's metabolic pathway [69]. Finding new pathways involving 3-methylglutaryl-CoA and investigating the strains containing these could be a promising starting point for the discovery of novel natural products containing this unusual moiety.

Material and methods

Bacterial strains and culture conditions

Escherichia coli DH5α (Thermo Fisher Scientific) was used as general cloning host (Additional file 1: Table S1). *E. coli* BW25113/pIJ790 was used for Red/ET-mediated recombination and *E. coli* BT340 facilitated FLP-recombination [70–72]. *E. coli* ET12567 was used for triparental conjugation [73]. *E. coli* DH5α was generally cultivated in LB-medium or on LB-agar plates at 37° C. *E. coli* BW25113/pIJ790 and *E. coli* BT340 were cultivated as described before [70]. *Streptomyces coelicolor* M1154 was used for heterologous expression of the caprazamycin gene cluster [74]. *Streptomyces* strains were generally cultivated on mannitol soya flour (MS) agar plates supplemented with $MgCl_2$ (1 mg/ml) or in tryptone soy broth (TSB) at 30° C [75]. Apramycin (50 µg/ml), carbenicillin (100 µg/ml), chloramphenicol (25 µg/ml), kanamycin (50 µg/ml), nalidixic acid (25 µg/ml) and tetracycline (5 µg/ml) were added if required.

DNA isolation and manipulation

Isolation and manipulation of DNA was carried out according to standard procedures described for *E. coli* and *Streptomyces* [75, 76].

Generation of *liu* deletion mutants

An apramycin resistance cassette was amplified from plasmid pIJ773 using primer pairs *liuA_F* and *liuA_R, liuB_F* and *liuB_R, liuD_F* and *liuD_R, liuE_F* and *liuE_R* or *sco2774_773_FW* and *sco2774_773_RV* (Additional file 1: Table S2). The purified cassettes were introduced into electro-competent *E. coli* ET12567/pIJ790/StC105 to replace the corresponding gene by Red/ET-mediated

recombination, respectively. The resulting cosmids were verified by restriction digest and PCR using primer pairs *LiuA_verify_1154_neu_F* and *LiuA_verify_1154_neu_R*, *LiuB_verify_1154_neu2_F* and *LiuB_verify_1154_neu2_R*, *LiuD_verify_1154_neu_F* and *LiuD_verify_1154_neu_R*, *LiuE_verify_1154_neu_F* and *LiuE_verify_1154_neu_R* or *sco2774_verify_FW* and *sco2774_verify_RV*. FLP-mediated excision of the resistance cassette was achieved by introducing the cosmids into *E. coli* BT340. Obtained cosmids were verified by restriction digest, PCR and sequencing of the resulting PCR products using the same primer pairs as above. A tetracycline-*oriT*-cassette was amplified from plasmid pIJ787 using primer pairs *bla-oriT_cassette_787* and *bla-tet_cassette_787* [77]. The purified cassette was introduced into *E. coli* ET12567/pIJ790 containing the desired *liu* deletion cosmid followed by Red/ET-mediated recombination. The resulting cosmids were verified by restriction digest, PCR and sequencing using primer pairs *bla_tet_oriT_v4_F* and *bla_tet_oriT_v4_R*. Final cosmids were introduced into *E. coli* ET12567 followed by triparental conjugation into *S. coelicolor* M1154 using *E. coli* ET12567/pR9406. Dilution series of single exconjugants were cultivated on MS-agar plates supplemented with nalidixic acid (25 µg/ml) and $MgCl_2$ (1 mg/ml) for several rounds until complete loss of the kanamycin resistance, indicating a successful double crossover event. Deletion of the desired gene was verified by PCR and sequencing using genomic DNA as template and primer pairs as above. Absence of wildtype gene was tested by PCR using primer pairs *LiuA_in-out_1154_F* and *LiuA_in-out_1154_R*, *LiuB_in-out_1154_F* and *LiuB_in-out_1154_R*, *LiuD_in-out_1154_neu_F* and *LiuD_in-out_1154_neu_R*, *LiuE_in-out_1154_neu_F* and *LiuE_in-out_1154_neu_R* or *sco2774_in-out_FW* and *sco2774_in-out_RV*.

Generation of mutants with transposon insertions

Cosmids containing transposon insertions were obtained from Paul Dyson (Swansea University, UK) and introduced into *E. coli* ET12567 followed by triparental conjugation into *S. coelicolor M1154Δsco2776-2779* using *E. coli* ET12567/pR9406 [78]. Dilution series of single exconjugants were cultivated on MS-agar plates containing apramycin (50 µg/ml) nalidixic acid (25 µg/ml) and $MgCl_2$ (1 mg/ml) for several rounds until complete loss of the kanamycin resistance, indicating a successful double crossover event. Inactivation of the desired gene was verified by PCR and sequencing using genomic DNA as template and primer pairs *SCO1838_tra_ver_FW* and *SCO1838_tra_ver_RV*, *SCO4384_tra_ver_FW* and *SCO4384_tra_ver_RV*, *SCO4930_tra_ver_FW* and *SCO4930_tra_ver_RV* or *SCO6732_tra_ver_FW* and *SCO6732_tra_ver_RV*. Absence of wildtype gene was tested by PCR using primer pairs *SCO1838_tra_ver_FW* and *1838_gene_RV*, *SCO4384_tra_ver_FW* and *4384_gene_RV*, *SCO4930_tra_ver_FW* and *4930_gene_RV* or *SCO6732_tra_ver_FW* and *6732_gene_RV*.

Generation of *cpz* deletion mutants

An apramycin resistance cassette was amplified from plasmid pIJ773 using primer pairs *cpz2_773_FW* and *cpz2_773_RV*, *cpz5_773_F* and *cpz5_773_R*, *cpz20_773_F* and *cpz20_773_R* or *cpz25_773_F* and *cpz25_773_R*, respectively. The purified cassettes were introduced into electro-competent *E. coli* ET12567/pIJ790/cpzLK09 to replace the correspondent gene by Red/ET-mediated recombination. The resulting cosmids were verified by restriction digest and PCR using primer pairs *cpz2_verify_FW* and *cpz2_verify_RV*, *cpz5_verifyKO_F* and *cpz5_verifyKO_R*, *cpz20_verifyKO_F* and *cpz20_verifyKO_R* or *cpz25_verifyKO_F* and *cpz25_verifyKO_R*. FLP-mediated excision of the resistance cassette was achieved by introducing the cosmids into *E. coli* BT340. Resulting cosmids were verified by restriction digest, PCR and sequencing using the same primer pairs as above. Final cosmids were introduced into *E. coli* ET12567 followed by triparental conjugation into *S. coelicolor* M1154 mutants using *E. coli* ET12567/pR9406. Dilution series of single exconjugants were cultivated on MS-agar plates supplemented with nalidixic acid (25 µg/ml), kanamycin (50 µg/ml) and $MgCl_2$ (1 mg/ml) for several rounds. Successful introduction of the caprazamycin gene cluster was verified by PCR and sequencing using genomic DNA as template and primer pairs as above.

Production and extraction of secondary metabolites

For production of caprazamycin aglycons or hydroxyacylcaprazols, 10–20 µl of *Streptomyces* spores were cultivated in 2 ml of TSB for 2 days at 30° C and 200 rpm. 100 µl of the preculture was transferred into 3 ml of P-medium (10 g/L soytone, 10 g/L soluble starch, 20 g/L D-maltose) supplemented with 150 µl trace elements solution (40 mg/L $ZnCl_2$, 200 mg/L $FeCl_3 \times 6$ H_2O, 10 mg/L $CuCl_2 \times 2$ H_2O, 10 mg/L $MnCl_2 \times 4$ H_2O, 10 mg/L $Na_2B_4O_6 \times 10$ H_2O and 10 mg/L $(NH_4)_6Mo_7O_{24} \times 4$ H_2O) in a 24-square deepwell plate and cultivated for another 5–7 days at 30° C and 200 rpm [79]. 3-methylcrotonyl-SNAc or 1-^{13}C-3-methylcrotonyl-SNAc were added to a final concentration of 0.8 mM, if required. To extract secondary metabolites, 1 ml of culture supernatant was adjusted to pH of 4 with 1 M HCl, 1 ml of n-butanol was added, mixed vigorously followed by centrifugation (13,000 rpm, 4° C, 15 min). The

n-butanol phase was separated, evaporated and the residue resuspended in 500 µl methanol.

Synthesis of 3-methylcrotonyl-SNAc (1) and (1-^{13}C)-3-methylcrotonyl-SNAc (5)

To a solution of 3,3-dimethylacrylic acid (0.50 g, 5.0 mmol) in CH_2Cl_2 (20 mL) was added *N*,*N*′-dicyclohexylcarbodiimide (1.14 g, 5.5 mmol; Additional file 1: Fig. S32). Then the reaction mixture was cooled to 0 °C, followed by the addition of 4-dimethylaminopyridine (0.12 g, 1.0 mmol) and *N*-acetylcysteamine (0.48 g, 4.0 mmol). After stirring at room temperature for 2 h, the reaction was quenched by adding aq. HCl (1 m, 20 mL) and then extracted with EtOAc (3 × 20 mL). The combined organic phases were dried with $MgSO_4$ and concentrated under reduced pressure. The residue was purified by column chromatography on silica gel (EtOAc: cyclohexane = 3: 1) to give 3-methylcrotonyl-SNAc (**1**) as white solid.

3-Methylcrotonyl-SNAc (1) TLC (EtOAc: cyclohexane = 3: 1): R_f = 0.20; yield: 0.58 g (2.9 mmol, 72%); ^{1}H-NMR (500 MHz, $CDCl_3$): δ = 5.91 (m, 1H), 5.88 (br s, 1H), 3.35 (dt, $^3J_{H,H}$ = 6.5, 5.7 Hz, 2H), 2.94 (dd $^3J_{H,H}$ = 6.8, 5.8 Hz, 2H), 2.06 (d, $^4J_{H,H}$ = 1.2 Hz, 3H), 1.86 (s, 3H), 1.79 (d, $^4J_{H,H}$ = 1.3 Hz, 3H) ppm; ^{13}C-NMR (126 MHz, $CDCl_3$): δ = 189.6 (C), 170.5 (C), 155.1 (C), 123.1 (CH), 40.1 (CH_2), 28.5 (CH_2), 27.4 (CH_3), 23.4 (CH_3), 21.4 (CH_3) ppm.

Ethyl (1-^{13}C)bromoacetate (2.00 g, 11.9 mmol) and triethyl phosphite (1.98 g, 11.9 mmol) were added to a reaction flask. The reaction mixture was refluxed at 130 °C for 8 h to obtain triethyl (1-^{13}C)phosphonoacetate (**2**) which was used in the following reaction without further purification. A solution of NaH (60% in mineral oil, 0.18 g, 4.4 mmol) in THF (12 mL) was cooled to 0 °C and **2** (1.00 g, 4.4 mmol) was then added dropwise. After stirring at 0 °C for 15 min, acetone (0.26 g, 4.4 mmol) was added. Stirring of the reaction mixture was continued at 0 °C for 30 min, followed by stirring at room temperature for 7 h. The reaction was quenched with sat. aq. NH_4Cl solution (20 mL) and then extracted with Et_2O (3 × 20 mL). The organic layers were combined, dried with $MgSO_4$ and concentrated under reduced pressure. The residue was purified by column chromatography on silica gel (petrol ether: Et_2O = 20: 1) to give ethyl (1-^{13}C)-3,3-dimethylacrylate (**3**) as colorless oil.

Ethyl (1-^{13}C)-3,3-dimethylacrylate (3) TLC (petrol ether: Et_2O = 10: 1): R_f = 0.41; yield: 0.45 g (3.5 mmol, 79%); ^{1}H-NMR (500 MHz, $CDCl_3$): δ = 5.57 (m, 1H), 4.04 (qd, $^3J_{H,H}$ = 7.1, $^3J_{C,H}$ = 3.0 Hz, 2H), 2.06 (dd, $^4J_{H,H}$ = 1.2 Hz, $^3J_{C,H}$ = 1.2 Hz, 3H), 1.79 (d, $^4J_{H,H}$ = 1.3 Hz, 3H), 1.17 (t, $^3J_{H,H}$ = 7.1 Hz, 3H) ppm. ^{13}C-NMR (126 MHz, $CDCl_3$): δ = 166.8 (^{13}C), 156.4 (d, $^2J_{C,C}$ = 2.2 Hz, C), 116.2 (d, $^1J_{C,C}$ = 75.7 Hz, CH), 59.5 (d, $^2J_{C,C}$ = 2.3 Hz, CH_2), 27.4 (d, $^3J_{C,C}$ = 7.6 Hz, CH_3), 20.2 (d, $^3J_{C,C}$ = 1.5 Hz, CH_3), 14.4 (d, $^3J_{C,C}$ = 2.2 Hz, CH_3) ppm.

To an aq. KOH solution (0.5 n, 5 mL) was added **3** (39 mg, 0.3 mmol). The reaction mixture was refluxed at 100 °C overnight until the oil layer disappeared. Then aq. HCl solution (1 n, 10 mL) was added and the reaction mixture was extracted with Et_2O (3 × 10 mL). The organic layers were combined, washed with brine (10 mL), dried with $MgSO_4$ and then concentrated under reduced pressure to give (1-^{13}C)-3,3-dimethylacrylic acid (**4**) which was used in the following reaction without further purification. Compound **4** and *N*-acetylcysteamine (35 mg, 0.3 mmol) were added to CH_2Cl_2 (3 mL). The solution was then cooled to 0 °C, followed by the addition of 1-ethyl-3-(3-dimethylaminopropyl)carbodiimid (58 mg, 0.3 mmol) and 4-dimethylaminopyridine (7 mg, 0.06 mmol). After stirring at room temperature overnight, the reaction solution was diluted with EtOAc (20 mL), washed with sat. aq. NH_4Cl solution, sat. aq. $NaHCO_3$ solution and brine, dried with $MgSO_4$, and then concentrated under reduced pressure. The residue was purified by column chromatography on silica gel (EtOAc) to give (1-^{13}C)-3-methylcrotonyl-SNAc (**5**) as white solid.

(1-^{13}C)-3-Methylcrotonyl-SNAc (5) Yield: 45 mg (0.22 mmol, 74%); ^{1}H-NMR (500 MHz, $CDCl_3$): δ = 5.91 (m, 1H), 5.83 (br s, 1H), 3.35 (dt, $^3J_{H,H}$ = 6.6, 5.7 Hz, 2H), 2.95 (ddd, $^3J_{H,H}$ = 7.1, 5.8 Hz, $^3J_{C,H}$ = 4.7 Hz, 2H), 2.06 (dd, $^4J_{H,H}$ = 1.2 Hz, $^3J_{C,H}$ = 1.2 Hz, 3H), 1.86 (s, 3H), 1.79 (d, $^4J_{H,H}$ = 1.3 Hz, 3H) ppm; ^{13}C-NMR (126 MHz, $CDCl_3$): δ = 189.6 (^{13}C), 170.4 (C), 155.1 (C), 123.1 (d, $^1J_{C,C}$ = 63.5 Hz, CH), 40.2 (CH_2), 28.5 (CH_2), 27.4 (d, $^3J_{C,C}$ = 7.3 Hz, CH_3), 23.4 (CH_3), 21.4 (d, $^3J_{C,C}$ = 1.5 Hz, CH_3) ppm.

HPLC–ESI–MS and MS^2

Analysis of extracts was performed on an HPLC-LC/MSD Ultra Trap System XCT 6330 (Agilent Technologies) equipped with a Luna Omega Polar C18 (5 µm, 150 × 2.1 mm; Phenomenex) column. Elution was performed with 0.1% formic acid (solvent A) and acetonitrile containing 0.06% formic acid (solvent B) at a flow rate of 0.4 ml/min using the following gradient: 0 to 100% solvent B over 20 min followed by an isocratic step of 100% B for 3 min. UV spectra were recorded between 230 and 600 nm by diode array detector. Mass spectrometry was performed using positive electrospray ionization (electrospray voltage 3.5 kV, heated capillary temperature 350° C, acquired mass range from 100 to 2200 m/*z*).

In-silico analysis of liu cluster distribution in *Actinobacteria*

A database containing all 31.598 assembled *Actinobacteria* genomes listed in the Genome Taxonomy Database (GTDB; last checked 4th August 2022) was created [80]. The cblaster search module (default settings) was used to analyze genomes from this database using query sequences ranging from *sco2775-sco2779* or *sco2774-sco2779* from the genome of *S. coelicolor* M1154 [81].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12934-022-01955-6.

Additional file 1: Figure S1. Degradation of leucine and isovalerate via the Liu-pathway as described for *P. aeruginosa* PAO1 [29]. **Figure S2.** Genetic organization of the caprazamycin biosynthetic gene cluster. Genes putatively involved in colour coding according to McErlean et al. [17]. **Figure S3.** Extracted ion chromatograms of *S. coelicolor* M1154 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z of* 816.5 and hydroxyacylcaprazols A/B with *m/z* of 830.5. **Figure S4.** Extracted ion chromatograms of *S. coelicolor* M1154/cpzLK09 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z* of 816.5 and hydroxyacylcaprazols A/B with *m/z* of 830.5. **Figure S5.** Extracted ion chromatograms of *S. coelicolor* M1154/cpzDB04 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z* of 816.5 and hydroxyacylcaprazols A/B with *m/z* of 830.5. **Figure S6.** Genetic organization of clusters encoding for Liu-pathway from *P. aeruginosa* PAO1, caprazamycin wildtype producer *S. sp.* MK730-62F2 and the heterologous caprazamycin producer *S. coelicolor* M1154. Additional genes in *Streptomyces* strains are shown transparent. Table shows genes from *P. aeruginosa* PAO1 and their proposed function along with homologue genes found in *S. sp.* MK730-62F2 and *S. coelicolor* M1154. Values in brackets indicate % identities/similarities (n.s. no significant similarities). **Figure S7.** Distribution of *liu* clusters in *Actinobacteria* and *Streptomyces*. **A**: Cblaster detected 4837 similar clusters containing at least three genes homolog to the query sequence *sco2775-sco2779* from *S. coelicolor* M1154. We discovered that 1.431 out of 31.598 (4.5%) *Actinobacteria* assemblies and 1.344 out of 2.574 (52.2%) *Streptomyces* assemblies and listed in the GTDB contain a *liu* cluster with homologues of all five query genes. **B**: Cblaster detected 5009 similar clusters containing at least three genes homolog to the query sequence *sco2774-sco2779* from *S. coelicolor* M1154. We discovered that 1.246 out of 31.598 (3.9%) *Actinobacteria* assemblies and 1.241 out of 2.574 (48.2%) *Streptomyces* assemblies and listed in the GTDB contain a *liu* cluster with homologues of all six query genes. A homologue of *sco2774* could only be found in 1.376 out of 5009 (27,5%) of the *Actinobacteria* clusters detected by cblaster, whereas the majority of detected *Streptomyces* cluster contain this gene with 1.343 out of 1.652 sequences (81,3%). **Figure S8.** Extracted ion chromatograms of *S. coelicolor* M1154Δ*sco2776-sco2779*/cpzLK09 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z of* 816.5 and hydroxyacylcaprazols A/B with *m/z* of 830.5. **Figure S9.** Extracted ion chromatograms of *S. coelicolor* M1154Δ*sco2776-sco2779*/cpzDB04 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z of* 816.5 and hydroxyacylcaprazols A/B with *m/z* of 830.5. **Figure S10.** Total ion chromatograms of *S. coelicolor* M1154Δ*sco2776-sco2779*/cpzDB04 and *S. coelicolor* M1154Δ*sco2776-sco2779*/cpzLK09. MS^2-fragmentation patterns of peak 1 (hydoxyacylcaprazol E, R_t 12.9 min) and peak 2 (caprazamycin aglycon E, R_t 13.8 min) are shown together with corresponding fragmentation schemes. **Figure S11.** Extracted ion chromatograms of *S. coelicolor* M1154Δ*sco2779*/cpzLK09 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z of* 816.5 and hydroxyacylcaprazols A/B with *m/z* of 830.5. **Figure S12.** Extracted ion chromatograms of *S. coelicolor* M1154Δ*sco2779*/cpzDB04 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z of* 816.5 and hydroxyacylcaprazols A/B with *m/z* of 830.5. **Figure S13.** Extracted ion chromatograms of *S. coelicolor* M1154Δ*sco2776*/cpzLK09 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z of* 816.5 and hydroxyacylcaprazols A/B with *m/z* of 830.5. **Figure S14.** Extracted ion chromatograms of *S. coelicolor* M1154Δ*sco2776*/cpzDB04 (three individual mutants). Masses are shown for caprazamycin aglycons E/F with *m/z* of 930.5, caprazamycin aglycons C/D/G with *m/z* of 944.5, caprazamycin aglycons A/B with *m/z* of 958.5 and the respective hydroxyacylcaprazols E/F with *m/z* of 802.5, hydroxyacylcaprazols C/D/G with *m/z of* 816.5 and hydroxyacylcaprazols A/B with *m/z*

Acknowledgements

We thank Prof. Paul Dyson from Swansea University (UK) for providing *S. coelicolor* cosmid StC105 and cosmids containing transposon insertions. We thank David Figurski for providing plasmid pR9406.

Author contributions

DB, NZ, JSD and BG deigned the research. DB generated gene deletion mutants of *liu* homologues in *Streptomyces coelicolor* M1154 and *cpz2*, *cpz5*, *cpz20* and *cpz25* gene deletions on the caprazamycin gene cluster and generated strains with combined mutations thereof. DB cultivated, fed and extracted mutants and interpreted HPLC–MS and MS^2 results. BK and PB generated and analyzed transposon mutants of *liuC* homologues in *Streptomyces coelicolor* M1154. AK performed HPLC–MS and MS^2 analysis. HX synthesized 3-methylcrotonyl-SNAc and (1-^{13}C)-3-methylcrotonyl-SNAc. DP performed *in-silico* analysis of *liu* cluster distribution in *Actinobacteria*. DB, NZ, JSD and BG analyzed the data and wrote the manuscript. BG supervised the project. All the authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) TRR 261, project-ID No. 398967434-TRR 261. N.Z. thanks the DZIF TTU09.716 for funding. D.P. and N.Z. thank H2020-FNR-11–2020: SECRETED for funding.

Availability of data and materials

The data generated and/or analyzed during this study is included in this article and Additional file 1. The constructed mutant strains, cosmids and plasmids are available at the Department of Pharmaceutical Biology of the University of Tübingen, Tübingen, Germany.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Pharmaceutical Biology, Eberhard-Karls University Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany. [2]Department of Microbial Bioactive Compounds, Interfaculty Institute of Microbiology and Infection Medicine, Eberhard-Karls University Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany. [3]Kekulé-Institute for Organic Chemistry and Biochemistry, University of Bonn, Gerhard-Domagk-Straße 1, 53121 Bonn, Germany. [4]Interfaculty Institute of Microbiology and Infection Medicine, Institute for Bioinformatics and Medical Informatics, Eberhard-Karls University Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany. [5]German Center for Infection Research (DZIF), Partner Site Tübingen, Tübingen, Germany.

Received: 15 August 2022 Accepted: 13 October 2022
Published online: 05 November 2022

References

1. Igarashi M, Nakagawa N, Doi N, Hattori S, Naganawa H, Hamada M. Caprazamycin B, a novel anti-tuberculosis antibiotic, from Streptomyces sp. J Antibiot (Tokyo). 2003;56(6):580–3.
2. Igarashi M, Takahashi Y, Shitara T, Nakamura H, Naganawa H, Miyake T, et al. Caprazamycins, novel lipo-nucleoside antibiotics, from Streptomyces sp. II. Structure elucidation of caprazamycins. J Antibiot (Tokyo). 2005;58(5):327–37.
3. Siebenberg S, Kaysser L, Wemakor E, Heide L, Gust B, Kammerer B. Identification and structural elucidation of new caprazamycins from Streptomyces sp. MK730–62F2 by liquid chromatography/electrospray ionization tandem mass spectrometry. Rapid Commun Mass Spectrom. 2011;25(4):495–502.
4. Ikeda M, Wachi M, Jung HK, Ishino F, Matsuhashi M. The Escherichia coli mraY gene encoding UDP-N-acetylmuramoyl-pentapeptide: undecaprenyl-phosphate phospho-N-acetylmuramoyl-pentapeptide transferase. J Bacteriol. 1991;173(3):1021–6.
5. Brandish PE, Kimura KI, Inukai M, Southgate R, Lonsdale JT, Bugg TD. Modes of action of tunicamycin, liposidomycin B, and mureidomycin A: inhibition of phospho-N-acetylmuramyl-pentapeptide translocase from *Escherichia coli*. Antimicrob Agents Chemother. 1996;40(7):1640–4.
6. Dini C, Collette P, Drochon N, Guillot JC, Lemoine G, Mauvais P, et al. Synthesis of the nucleoside moiety of liposidomycins: elucidation of the pharmacophore of this family of MraY inhibitors. Bioorg Med Chem Lett. 2000;10(16):1839–43.
7. Ishizaki Y, Hayashi C, Inoue K, Igarashi M, Takahashi Y, Pujari V, et al. Inhibition of the first step in synthesis of the mycobacterial cell wall core, catalyzed by the GlcNAc-1-phosphate transferase WecA, by the novel caprazamycin derivative CPZEN-45. J Biol Chem. 2013;288(42):30309–19.
8. Kimura K, Ikeda Y, Kagami S, Yoshihama M, Suzuki K, Osada H, et al. Selective inhibition of the bacterial peptidoglycan biosynthesis by the new types of liposidomycins. J Antibiot (Tokyo). 1998;51(12):1099–104.
9. Hirano S, Ichikawa S, Matsuda A. Design and synthesis of diketopiperazine and acyclic analogs related to the caprazamycins and liposidomycins as potential antibacterial agents. Bioorg Med Chem. 2008;16(1):428–36.
10. Hirano S, Ichikawa S, Matsuda A. Structure–activity relationship of truncated analogs of caprazamycins as potential anti-tuberculosis agents. Bioorg Med Chem. 2008;16(9):5123–33.
11. Hirano S, Ichikawa S, Matsuda A. Development of a highly beta-selective ribosylation reaction without using neighboring group participation: total synthesis of (+)-caprazol, a core structure of caprazamycins. J Org Chem. 2007;72(26):9936–46.
12. Takahashi Y, Igarashi M, Miyake T, Soutome H, Ishikawa K, Komatsuki Y, et al. Novel semisynthetic antibiotics from caprazamycins A–G: caprazene derivatives and their antibacterial activity. J Antibiot (Tokyo). 2013;66(3):171–8.
13. Ishizaki Y, Takahashi Y, Kimura T, Inoue M, Hayashi C, Igarashi M. Synthesis and biological activity of analogs of CPZEN-45, a novel antituberculosis drug. J Antibiot (Tokyo). 2019;72(12):970–80.
14. Kaysser L, Lutsch L, Siebenberg S, Wemakor E, Kammerer B, Gust B. Identification and manipulation of the caprazamycin gene cluster lead to new simplified liponucleoside antibiotics and give insights into the biosynthetic pathway. J Biol Chem. 2009;284(22):14987–96.
15. Kaysser L, Siebenberg S, Kammerer B, Gust B. Analysis of the liposidomycin gene cluster leads to the identification of new caprazamycin derivatives. ChemBioChem. 2010;11(2):191–6.
16. Kaysser L, Wemakor E, Siebenberg S, Salas JA, Sohng JK, Kammerer B, et al. Formation and attachment of the deoxysugar moiety and assembly of the gene cluster for caprazamycin biosynthesis. Appl Environ Microbiol. 2010;76(12):4008–18.
17. McErlean M, Liu X, Cui Z, Gust B, Van Lanen SG. Identification and characterization of enzymes involved in the biosynthesis of pyrimidine nucleoside antibiotics. Nat Prod Rep. 2021;38(7):1362–407.
18. Dairi T. Studies on biosynthetic genes and enzymes of isoprenoids produced by actinomycetes. J Antibiot (Tokyo). 2005;58(4):227–43.
19. Bock T, Kasten J, Muller R, Blankenfeldt W. Crystal structure of the HMG-CoA Synthase MvaS from the Gram-Negative Bacterium *Myxococcus xanthus*. ChemBioChem. 2016;17(13):1257–62.
20. Campobasso N, Patel M, Wilding IE, Kallender H, Rosenberg M, Gwynn MN. *Staphylococcus aureus* 3-hydroxy-3-methylglutaryl-CoA synthase: crystal structure and mechanism. J Biol Chem. 2004;279(43):44883–8.
21. Massey LK, Sokatch JR, Conrad RS. Branched-chain amino acid catabolism in bacteria. Bacteriol Rev. 1976;40(1):42–54.
22. Denoya CD, Fedechko RW, Hafner EW, McArthur HA, Morgenstern MR, Skinner DD, et al. A second branched-chain alpha-keto acid dehydrogenase gene cluster (bkdFGH) from *Streptomyces avermitilis*: its relationship to avermectin biosynthesis and the construction of a bkdF mutant suitable for the production of novel antiparasitic avermectins. J Bacteriol. 1995;177(12):3504–11.
23. Sprusansky O, Stirrett K, Skinner D, Denoya C, Westpheling J. The bkdR gene of Streptomyces coelicolor is required for morphogenesis and antibiotic production and encodes a transcriptional regulator of a branched-chain amino acid dehydrogenase complex. J Bacteriol. 2005;187(2):664–71.
24. Stirrett K, Denoya C, Westpheling J. Branched-chain amino acid catabolism provides precursors for the type II polyketide antibiotic, actinorhodin, via pathways that are nutrient dependent. J Ind Microbiol Biotechnol. 2009;36(1):129–37.
25. Hoschle B, Gnau V, Jendrossek D. Methylcrotonyl-CoA and geranyl-CoA carboxylases are involved in leucine/isovalerate utilization (Liu) and acyclic terpene utilization (Atu), and are encoded by liuB/liuD and atuC/atuF in *Pseudomonas aeruginosa*. Microbiology (Reading). 2005;151(Pt 11):3649–56.
26. Forster-Fromme K, Hoschle B, Mack C, Bott M, Armbruster W, Jendrossek D. Identification of genes and proteins necessary for catabolism of acyclic terpenes and leucine/isovalerate in *Pseudomonas aeruginosa*. Appl Environ Microbiol. 2006;72(7):4819–28.
27. Forster-Fromme K, Jendrossek D. Biochemical characterization of isovaleryl-CoA dehydrogenase (LiuA) of *Pseudomonas aeruginosa* and the importance of liu genes fora functional catabolic pathway of methyl-branched compounds. FEMS Microbiol Lett. 2008;286(1):78–84.
28. Chavez-Aviles M, Diaz-Perez AL, Reyes-de la Cruz H, Campos-Garcia J. The *Pseudomonas aeruginosa* liuE gene encodes the 3-hydroxy-3-methylglutaryl coenzyme A lyase, involved in leucine and acyclic terpene catabolism. FEMS Microbiol Lett. 2009;296(1):117–23.
29. Chavez-Aviles M, Diaz-Perez AL, Campos-Garcia J. The bifunctional role of LiuE from *Pseudomonas aeruginosa*, displays additionally HlHG-CoA lyase enzymatic activity. Mol Biol Rep. 2010;37(4):1787–91.
30. Massey LK, Conrad RS, Sokatch JR. Regulation of leucine catabolism in *Pseudomonas putida*. J Bacteriol. 1974;118(1):112–20.
31. Bode HB, Ring MW, Schwar G, Altmeyer MO, Kegler C, Jose IR, et al. Identification of additional players in the alternative biosynthesis pathway to isovaleryl-CoA in the myxobacterium *Myxococcus xanthus*. ChemBioChem. 2009;10(1):128–40.
32. Surger MJ, Angelov A, Stier P, Ubelacker M, Liebl W. Impact of branched-chain amino acid catabolism on fatty acid and alkene biosynthesis in *Micrococcus luteus*. Front Microbiol. 2018;9:374.

33. Kazakov AE, Rodionov DA, Alm E, Arkin AP, Dubchak I, Gelfand MS. Comparative genomics of regulation of fatty acid and branched-chain amino acid utilization in proteobacteria. J Bacteriol. 2009;191(1):52–64.
34. Zhang YX, Denoya CD, Skinner DD, Fedechko RW, McArthur HAI, Morgenstern MR, et al. Genes encoding acyl-CoA dehydrogenase (AcdH) homologues from *Streptomyces coelicolor* and *Streptomyces avermitilis* provide insights into the metabolism of small branched-chain fatty acids and macrolide antibiotic production. Microbiology (Reading). 1999;145(Pt 9):2323–34.
35. Demirev AV, Lee JS, Sedai BR, Ivanov IG, Nam DH. Identification and characterization of acetyl-CoA carboxylase gene cluster in *Streptomyces toxytricini*. J Microbiol. 2009;47(4):473–8.
36. Lyu M, Cheng Y, Han X, Wen Y, Song Y, Li J, et al. AccR, a TetR family transcriptional repressor, coordinates short-chain acyl coenzyme a homeostasis in *Streptomyces avermitilis*. Appl Environ Microbiol. 2020;86(12): e00508.
37. Fernandez-Martinez LT, Del Sol R, Evans MC, Fielding S, Herron PR, Chandra G, et al. A transposon insertion single-gene knockout library and new ordered cosmid library for the model organism *Streptomyces coelicolor* A3(2). Antonie Van Leeuwenhoek. 2011;99(3):515–22.
38. Kaneda T. Iso- and anteiso-fatty acids in bacteria: biosynthesis, function, and taxonomic significance. Microbiol Rev. 1991;55(2):288–302.
39. Debono M, Barnhart M, Carrell CB, Hoffmann JA, Occolowitz JL, Abbott BJ, et al. A21978C, a complex of new acidic peptide antibiotics: isolation, chemistry, and mass spectral structure elucidation. J Antibiot (Tokyo). 1987;40(6):761–77.
40. Hashizume H, Hirosawa S, Sawa R, Muraoka Y, Ikeda D, Naganawa H, et al. Tripropeptins, novel antimicrobial agents produced by Lysobacter sp. J Antibiot (Tokyo). 2004;57(1):52–8.
41. Hashizume H, Igarashi M, Sawa R, Adachi H, Nishimura Y, Akamatsu Y. A new type of tripropeptin with anteiso-branched chain fatty acid from Lysobacter sp. BMK333–48F3. J Antibiot (Tokyo). 2008;61(9):577–82.
42. Díaz-Pérez AL, Díaz-Pérez C, Campos-García J. Bacterial l-leucine catabolism as a source of secondary metabolites. Rev Environ Sci Biotechnol. 2015;15(1):1–29.
43. Trowitzsch-Kienast W, Wray V, Gerth K, Reichenbach H, Höfle G. Antibiotika aus Gleitenden Bakterien, XXVIII1) Biosynthese des Myxothiazols in *Myxococcus fulvus* Mx f16. Liebigs Ann Chem. 1986;1986(1):93–8.
44. Silakowski B, Schairer HU, Ehret H, Kunze B, Weinig S, Nordsiek G, et al. New lessons for combinatorial biosynthesis from myxobacteria. The myxothiazol biosynthetic gene cluster of *Stigmatella aurantiaca* DW4/3–1. J Biol Chem. 1999;274(52):37391–9.
45. Frank B, Wenzel SC, Bode HB, Scharfe M, Blocker H, Muller R. From genetic diversity to metabolic unity: studies on the biosynthesis of aurafurones and aurafuron-like structures in myxobacteria and streptomycetes. J Mol Biol. 2007;374(1):24–38.
46. Zander W, Irschik H, Augustiniak H, Herrmann M, Jansen R, Steinmetz H, et al. Sulfangolids, macrolide sulfate esters from *Sorangium cellulosum*. Chemistry. 2012;18(20):6264–71.
47. Kopp M, Irschik H, Gemperlein K, Buntin K, Meiser P, Weissman KJ, et al. Insights into the complex biosynthesis of the leupyrrins in *Sorangium cellulosum* So ce690. Mol Biosyst. 2011;7(5):1549–63.
48. Mahmud T, Bode HB, Silakowski B, Kroppenstedt RM, Xu M, Nordhoff S, et al. A novel biosynthetic pathway providing precursors for fatty acid biosynthesis and secondary metabolite formation in myxobacteria. J Biol Chem. 2002;277(36):32768–74.
49. Bode HB, Ring MW, Schwar G, Kroppenstedt RM, Kaiser D, Muller R. 3-Hydroxy-3-methylglutaryl-coenzyme A (CoA) synthase is involved in biosynthesis of isovaleryl-CoA in the myxobacterium *Myxococcus xanthus* during fruiting body formation. J Bacteriol. 2006;188(18):6524–8.
50. Li Y, Luxenburger E, Muller R. An alternative isovaleryl CoA biosynthetic pathway involving a previously unknown 3-methylglutaconyl CoA decarboxylase. Angew Chem Int Ed Engl. 2013;52(4):1304–8.
51. Lee DW, Ng BG, Kim BS. Increased valinomycin production in mutants of Streptomyces sp. M10 defective in bafilomycin biosynthesis and branched-chain alpha-keto acid dehydrogenase complex expression. J Ind Microbiol Biotechnol. 2015;42(11):1507–17.
52. Pulsawat N, Kitani S, Kinoshita H, Lee CK, Nihira T. Identification of the bkdAB gene cluster, a plausible source of the starter-unit for virginiamycin M production in *Streptomyces virginiae*. Arch Microbiol. 2007;187(6):459–66.
53. Bode HB, Meiser P, Klefisch T, Cortina NS, Krug D, Gohring A, et al. Mutasynthesis-derived myxalamids and origin of the isobutyryl-CoA starter unit of myxalamid B. ChemBioChem. 2007;8(17):2139–44.
54. Mast Y, Weber T, Golz M, Ort-Winklbauer R, Gondran A, Wohlleben W, et al. Characterization of the 'pristinamycin supercluster' of *Streptomyces pristinaespiralis*. Microb Biotechnol. 2011;4(2):192–206.
55. Tamaoki T, Shirahata K, Iida T, Tomita F. Trioxacarcins, novel antitumor antibiotics. II. Isolation, physico-chemical properties and mode of action. J Antibiot (Tokyo). 1981;34(12):1525–30.
56. Zhang M, Hou XF, Qi LH, Yin Y, Li Q, Pan HX, et al. Biosynthesis of trioxacarcin revealing a different starter unit and complex tailoring steps for type II polyketide synthase. Chem Sci. 2015;6(6):3440–7.
57. Hafner EW, Holley BW, Holdom KS, Lee SE, Wax RG, Beck D, et al. Branched-chain fatty acid requirement for avermectin production by a mutant of *Streptomyces avermitilis* lacking branched-chain 2-oxo acid dehydrogenase activity. J Antibiot (Tokyo). 1991;44(3):349–56.
58. Ikeda H, Omura S. Avermectin biosynthesis. Chem Rev. 1997;97(7):2591–610.
59. Moore BS, Hertweck C. Biosynthesis and attachment of novel bacterial polyketide synthase starter units. Nat Prod Rep. 2002;19(1):70–99.
60. Pospisil S, Petrickova K, Sedmera P, Halada P, Olsovska J, Petricek M. Effect of starter unit availability on the spectrum of manumycin-type metabolites produced by *Streptomyces nodosus* sp. asukaensis. J Appl Microbiol. 2011;111(5):1116–28.
61. Weissman KJ, Bycroft M, Staunton J, Leadlay PF. Origin of starter units for erythromycin biosynthesis. Biochemistry. 1998;37(31):11012–7.
62. Park SR, Han AR, Ban YH, Yoo YJ, Kim EJ, Yoon YJ. Genetic engineering of macrolide biosynthesis: past advances, current state, and future prospects. Appl Microbiol Biotechnol. 2010;85(5):1227–39.
63. Kong D, Lee MJ, Lin S, Kim ES. Biosynthesis and pathway engineering of antifungal polyene macrolides in actinomycetes. J Ind Microbiol Biotechnol. 2013;40(6):529–43.
64. Yi JS, Kim M, Kim EJ, Kim BG. Production of pikromycin using branched chain amino acid catabolism in *Streptomyces venezuelae* ATCC 15439. J Ind Microbiol Biotechnol. 2018;45(5):293–303.
65. Ichikawa S, Yamaguchi M, Matsuda A. Antibacterial nucleoside natural products inhibiting phospho-MurNAc-pentapeptide translocase; chemistry and structure-activity relationship. Curr Med Chem. 2015;22(34):3951–79.
66. Kimura KI. Liposidomycin, the first reported nucleoside antibiotic inhibitor of peptidoglycan biosynthesis translocase I: the discovery of liposidomycin and related compounds with a perspective on their application to new antibiotics. J Antibiot (Tokyo). 2019;72(12):877–89.
67. Funabashi M, Baba S, Nonaka K, Hosobuchi M, Fujita Y, Shibata T, et al. The biosynthesis of liposidomycin-like A-90289 antibiotics featuring a new type of sulfotransferase. ChemBioChem. 2010;11(2):184–90.
68. Chi X, Baba S, Tibrewal N, Funabashi M, Nonaka K, Van Lanen SG. The muraminomicin biosynthetic gene cluster and enzymatic formation of the 2-deoxyaminoribosyl appendage. Medchemcomm. 2013;4(1):239–43.
69. Lahme S, Eberlein C, Jarling R, Kube M, Boll M, Wilkes H, et al. Anaerobic degradation of 4-methylbenzoate via a specific 4-methylbenzoyl-CoA pathway. Environ Microbiol. 2012;14(5):1118–32.
70. Gust B, Challis GL, Fowler K, Kieser T, Chater KF. PCR-targeted Streptomyces gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. Proc Natl Acad Sci U S A. 2003;100(4):1541–6.
71. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci U S A. 2000;97(12):6640–5.
72. Cherepanov PP, Wackernagel W. Gene disruption in *Escherichia coli*: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. Gene. 1995;158(1):9–14.
73. MacNeil DJ, Gewain KM, Ruby CL, Dezeny G, Gibbons PH, MacNeil T. Analysis of *Streptomyces avermitilis* genes required for avermectin biosynthesis utilizing a novel integration vector. Gene. 1992;111(1):61–8.
74. Gomez-Escribano JP, Bibb MJ. Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. Microb Biotechnol. 2011;4(2):207–15.
75. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA. Practical Streptomyces genetics. Norwich: John Innes Foundation; 2000.

76. Sambrook J, Russell DW. Molecular cloning. A laboratory manual. New York: Cold Spring Harbor Laboratory Press; 2001.
77. Gust B, Chandra G, Jakimowicz D, Yuqing T, Bruton CJ, Chater KF. Lambda red-mediated genetic manipulation of antibiotic-producing Streptomyces. Adv Appl Microbiol. 2004;54:107–28.
78. Bishop A, Fielding S, Dyson P, Herron P. Systematic insertional mutagenesis of a streptomycete genome: a link between osmoadaptation and antibiotic production. Genome Res. 2004;14(5):893–900.
79. Siebenberg S, Bapat PM, Lantz AE, Gust B, Heide L. Reducing the variability of antibiotic production in Streptomyces by cultivation in 24-square deepwell plates. J Biosci Bioeng. 2010;109(3):230–4.
80. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 2022;50(D1):D785–94.
81. Gilchrist CLM, Booth TJ, van Wersch B, van Grieken L, Medema MH, Chooi Y-H, et al. cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. Bioinform Adv. 2021;1(1):vbab016.

Publisher's Note

MANUSCRIPT 2

Title: Integrated genome and metabolome mining unveiled structure and biosynthesis of novel lipopeptides from a deep-sea *Rhodococcus*

Authors: Constanza Ragozzino, Fortunato Palma Esposito, Carmine Buonocore, Pietro Tedesco, Daniela Coppola, Davide Paccagnella, Nadine Ziemert, Gerardo Della Sala, Donatella de Pascale

DECLARATION ON THE CONTRIBUTION OF CO-AUTHORS TO THE MANUSCRIPT

Author	Author position	Scientific ideas %	Data generation %	Analysis & interpretation %	Paper writing %
C. Ragozzino	First Author	50	60	50	40
F.P. Esposito	Co-Author	10	30	10	5
C. Buonocore	Co-Author	-	-	5	5
P. Tedesco	Co-Author	-	-	-	5
D. Coppola	Co-Author	-	-	-	5
D. Paccagnella	Co-Author	-	-	5	5
N. Ziemert	Corresponding author	5	-	5	10
G. Della Sala	Corresponding author	20	10	15	15
D. Pascale	Corresponding author	15	-	10	10
Title of paper	Integrated genome and metabolome mining unveiled structure and biosynthesis of novel lipopeptides from a deep-sea *Rhodococcus*				

Received: 22 May 2024 | Accepted: 23 August 2024

DOI: 10.1111/1751-7915.70011

MICROBIAL BIOTECHNOLOGY Open Access

RESEARCH ARTICLE

Integrated genome and metabolome mining unveiled structure and biosynthesis of novel lipopeptides from a deep-sea *Rhodococcus*

Costanza Ragozzino[1,2] | **Fortunato Palma Esposito**[1] | **Carmine Buonocore**[1] | **Pietro Tedesco**[1] | **Daniela Coppola**[1] | **Davide Paccagnella**[3] | **Nadine Ziemert**[3,4] | **Gerardo Della Sala**[1] | **Donatella de de Pascale**[1]

[1]Department of Ecosustainable Marine Biotechnology, Stazione Zoologica Anton Dohrn, Giardini del Molosiglio, Naples, Italy

[2]Department of Chemical, Biological, Pharmaceutical and Environmental Sciences, University of Messina, Messina, Italy

[3]Interfaculty Institute of Microbiology and Infection Medicine Tuebingen, Microbiology/Biotechnology, University of Tuebingen, Tuebingen, Germany

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

Correspondence
Gerardo Della Sala and Donatella de Pascale, Department of Ecosustainable Marine Biotechnology, Stazione Zoologica Anton Dohrn, Giardini del Molosiglio, Naples, Italy.
Email: gerardo.dellasala@szn.it and donatella.depascale@szn.it

Funding information
H2020 Societal Challenges, Grant/Award Number: 101000794

Abstract

Microbial biosurfactants have garnered significant interest from industry due to their lower toxicity, biodegradability, activity at lower concentrations and higher resistance compared to synthetic surfactants. The deep-sea *Rhodococcus* sp. I2R has been identified as a producer of glycolipid biosurfactants, specifically succinoyl trehalolipids, which exhibit antiviral activity. However, genome mining of this bacterium has revealed a still unexplored repertoire of biosurfactants. The microbial genome was found to host five non-ribosomal peptide synthetase (NRPS) gene clusters containing starter condensation domains that direct lipopeptide biosynthesis. Genomics and mass spectrometry (MS)-based metabolomics enabled the linking of two NRPS gene clusters to the corresponding lipopeptide families, leading to the identification of 20 new cyclolipopeptides, designated as rhodoheptins, and 33 new glycolipopeptides, designated as rhodamides. An integrated *in silico* gene cluster and high-resolution MS/MS data analysis allowed us to elucidate the planar structure, inference of stereochemistry and reconstruction of the biosynthesis of rhodoheptins and rhodamides. Rhodoheptins are cyclic heptapeptides where the N-terminus is bonded to a β-hydroxy fatty acid forming a macrolactone ring with the C-terminal amino acid residue. Rhodamides are linear 14-mer glycolipopeptides with a serine- and alanine-rich peptide backbone, featuring a distinctive pattern of acetylation, glycosylation and succinylation. These molecules exhibited biosurfactant activity in the oil-spreading assay and showed moderate antiproliferative effects against human A375 melanoma cells.

INTRODUCTION

Surfactants are amphiphilic molecules with hydrophilic heads and hydrophobic tails, conferring these compounds the natural tendency to form self-aggregates and lower water surface tension. Due to their various applications, such as detergents in household and industrial cleaning products, or as vehicles for active ingredients in pharmaceutical formulations, surfactants are extremely important from an industrial perspective.

Costanza Ragozzino and Fortunato Palma Esposito shared first co-authorship.

Gerardo Della Sala and Donatella de Pascale shared last co-authorship.

Currently, exhaustible fossil fuels are the starting material for surfactant production (Nagtode et al., 2023). Therefore, in response to increasing industrial demand, finding alternative and renewable sources for surfactants presents both economic and ecological imperatives. Microbial biosurfactants have raised significant attention as being less toxic, biodegradable, active at lower concentrations and more resistant to extreme conditions as compared to synthetic ones (Jahan et al., 2020). Moreover, biosurfactants are often endowed with pharmacological properties, such as antimicrobial, antitumor, antiviral and anti-inflammatory activities (Buonocore et al., 2023; Giugliano et al., 2023; Ceresa et al., 2023; Pilz et al., 2023; Subramaniam et al., 2020).

Bacteria usually secrete biosurfactants to emulsify hydrophobic substrates and improve nutrient availability in harsh conditions. Moreover, biosurfactants exert antibacterial functions to gain an advantage over competitors and mediate physiological processes such as *quorum sensing*, biofilm formation, cell attachment/dissociation to surfaces and swarming motility (Dias & Nitschke, 2023).

Among microbial surfactants, lipopeptides (LPs) are low-molecular-weight metabolites, which have emerged as promising anticancer and antimicrobial agents. For instance, pseudofactin II triggers apoptosis in melanoma A375 cells as a consequence of plasma membrane permeabilization (Janek et al., 2013). Iturin A prompts apoptotic cell death in breast cancer via inhibition of the Akt pathway (Dey et al., 2015). Surfactin enhances a ROS/JNK-mediated mitochondrial/caspase pathway to kill MCF-7 cancer cells, but also disrupts cell membrane and protein synthesis in pathogenic bacteria (Cao et al., 2010; Chen et al., 2022). Remarkably, the LP daptomycin is a membrane-active antibiotic currently used in clinical settings (Huang, 2020).

Biosynthetically, LPs are assembled primarily through the sequential condensation of proteinogenic and non-proteinogenic amino acids by large multimodular enzymes, known as non-ribosomal peptide synthetases (NRPSs). Only recently were LPs of ribosomal origin, such as lipolanthines, discovered (Wiebach et al., 2018).

Typically, each NRPS module contains a condensation (C) domain, an adenylation (A) domain, and a peptidyl carrier protein (PCP) domain, adding a single amino acid to the peptidyl backbone. The specific fatty acyl chains anchored to the peptide core in LPs form hydrophobic tails that strongly influence LP bioactivity. Fatty acyl chains are incorporated into the peptidyl backbone during the first step of LP biosynthesis through the *N*-acylation of the α-amino group of the first amino acid, a process known as lipoinitiation.

Bacteria use different lipoinitiation strategies to build up LPs (Figure 1) (Chooi & Tang, 2010).

Within the biosynthetic machinery of puwainaphycins (Mareš et al., 2014), a hybrid polyketide synthase and

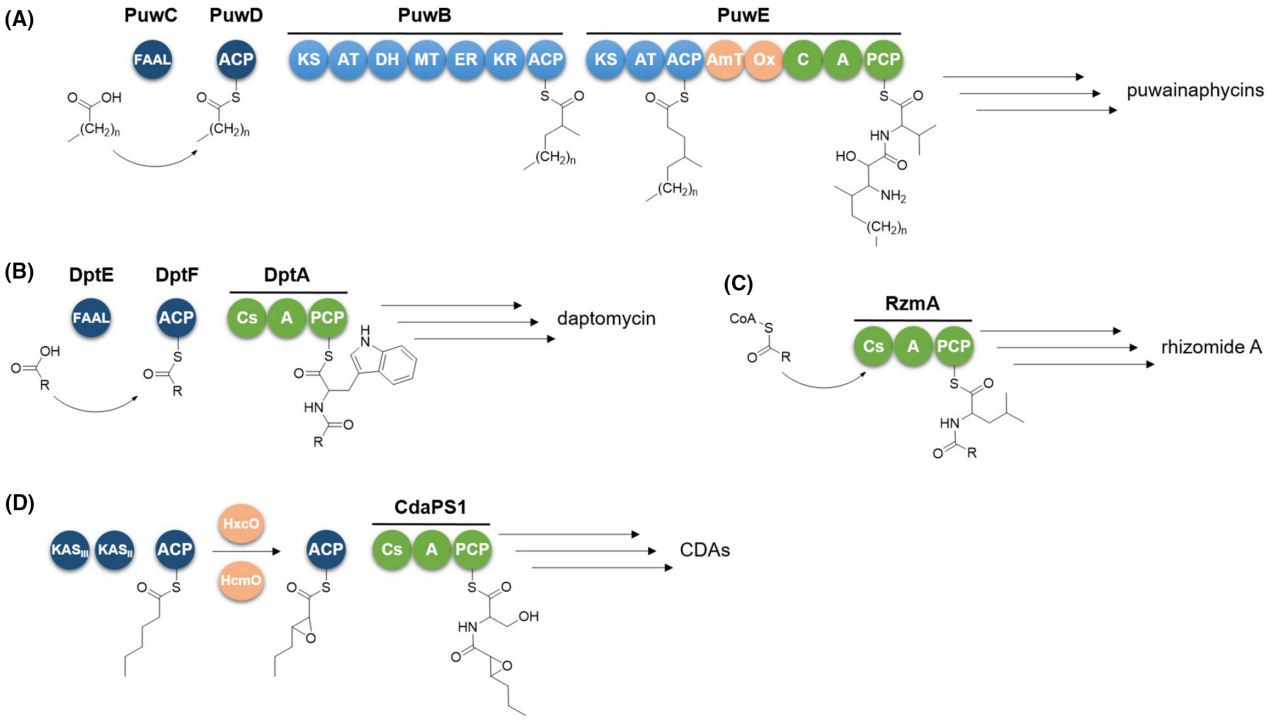


FIGURE 1 Lipoinitiation strategies employed in the biosynthesis of puwainaphycins (A), daptomycin (B), rhizomide A (C), and calcium-dependent antibiotics (CDAs) (D). FAAL, fatty acyl-AMP ligase; ACP, acyl carrier protein; KS, ketosynthase; AT, acyltransferase; DH, dehydratase; MT, methyltransferase; ER, enoylreductase; KR, ketoreductase; AmT, aminotransferase; Ox, monooxygenase; C, condensation domain; A, adenylation domain; PCP, peptidyl carrier protein; Cs, starter condensation domain; KAS_{II}, beta-ketoacyl-acyl carrier protein synthase II; KAS_{III}, beta-ketoacyl-acyl carrier protein synthase III; HxcO and HcmO, FAD-dependent oxidases.

non-ribosomal synthetase (PKS/NRPS) system incorporates the lipid chain (Figure 1A). The fatty acyl-AMP ligase (FAAL) PuwC selects and activates the fatty acid, which is then transferred to the acyl carrier protein (ACP) PuwD. The fatty acid is subsequently elongated, aminated and hydroxylated by two adjacent PKS modules (PuwB and PuwE). Finally, the modified lipid chain is incorporated by the first NRPS module, priming the synthesis of the peptide backbone.

In most cases, lipoinitiation is directed by the starter condensation domain (Cs) located in the initial NRPS module, which catalyses formation of an amide bond between the fatty acyl chain and the first amino acid. Cs domains can be promptly detected through bioinformatics, as being phylogenetically distinguishable from other C-domain subtypes (Rausch et al., 2007). In ramoplanin and daptomycin biosynthesis (Hoertz et al., 2012; Wittmann et al., 2008), FAAL and ACP directly activate and transfer the starter fatty acyl unit to the Cs of the initial NRPS module (Figure 1B). In case of the calcium-dependent antibiotics (CDAs), a dedicated *fab* operon within the biosynthetic gene cluster (BGC) synthesizes 2,3-epoxyhexanoic acid, which is tethered to an ACP and transferred to the Cs domain to be channelled into the NRPS assembly line (Hojati et al., 2002) (Figure 1D). The fatty acyl unit can be also delivered to the Cs domain as a free acyl-CoA derivative, without FAAL and ACP-mediation, as in the case of SrfAA surfactin, RzmA rhizomide and HolA holrhizin synthetases (Kraas et al., 2010; Zhong et al., 2021) (Figure 1C).

Following the One Strain Many Compounds (OSMAC) approach (Pan et al., 2019) with 22 different culture media, the marine isolate *Rhodococcus* sp. I2R (actinomycetota) was identified as a producer of more than 30 novel glycolipid biosurfactants, specifically succinoyl trehalolipids, which exhibit potent antiviral effects against herpes simplex virus and human coronaviruses, likely through a detergent-like mechanism (Palma Esposito et al., 2021). However, genome mining of *Rhodococcus* sp. I2R (from now on *R*. I2R) highlighted a number of cryptic BGCs, some of which are predicted to encode new surface-active natural products. Given the potential of this strain, we have taken an interest in further exploring its biosurfactant repertoire using a metabologenomic approach.

Herein, we report on the structure and biosynthesis of two LP families from *R*. I2R. Mining the bacterial genome for Cs domain-containing NRPS revealed four putative LP biosynthetic operons. Linking genomics with mass spectrometry (MS)-based metabolomics enabled us to connect two NRPS pathways to the corresponding LP families, leading to the identification of 20 new cyclic lipoheptapeptides, designated as rhodoheptins, and 33 new glycolipopeptides, designated as rhodamides. The integrated *in silico* genome analysis and comprehensive investigation of high-resolution MS/MS data from individual LPs proved valuable in elucidating the planar structure, inferring stereochemistry and reconstructing the biosynthetic route of rhodoheptins and rhodamides, as they are produced in quantities too small for isolation and characterization by NMR. This study represents a step forward in exploring the biotechnological potential of the *Rhodococcus* genus.

RESULTS AND DISCUSSION

Identification of non-ribosomal peptide lipopeptide gene clusters

The draft genome of *R*. I2R (accession: JAHUTG000000000) consists of 72 contigs, with an overall size of 5.29 Mb and GC content of 64.01%. A preliminary antiSMASH analysis (Blin et al., 2023) revealed several BGCs split across multiple contigs (Palma Esposito et al., 2021). To improve assembly continuity, the 72 contigs were joined into 36 scaffolds using the multi-draft-based scaffolder MEDUSA (http://150.217.159.17/medusa, accessed on 07/12/2023) (Bosi et al., 2015). Aiming to identify genes encoding the biosynthesis of LPs, *de novo* assembled scaffolds were mined for the presence of NRPS modules containing a Cs domain. Five novel intact multimodular NRPS systems with Cs domains were detected using antiSMASH. Four NRPS were predicted to catalyse LP biosynthesis, one of them being a hydroxamate siderophore synthetase. The last one was predicted to assemble a hydroxamate-catechol hybrid siderophore, where the Cs domain is expected to incorporate a 2,3-dihydroxybenzoic acid moiety. As modular NRPS biosynthesis usually follows the collinearity rule, and functions of catalytic domains within each module can be fairly inferred *in silico* (Della Sala et al., 2020), the putative structures and/or chemical substructures of the encoded LPs could be predicted from gene cluster sequences. Mass spectrometry data from *R*. I2R extracts were analysed for metabolites matching the chemical features predicted. This 'feature-based matching approach' enabled us to link two of the five NRPS gene clusters, designated as *rhp* and *rmd*, to their corresponding LP families, namely rhodoheptins and rhodamides respectively. These LPs were detected via molecular networking in the *R*. I2R extract and structurally characterized using liquid chromatography coupled with high-resolution tandem mass spectrometry (LC-HRMS^2) (see Section 2.4). Rhodoheptins were shown to be cyclic lipoheptapeptides while rhodamides were identified as linear glycolipopeptides.

Analysis of the *rhp* biosynthetic gene cluster

The putative *rhp* gene is 26,760 bp long and encodes for a heptamodular NRPS, which is expected

to assemble a lipoheptapeptide molecule (Figure 2, Table S1). AntiSMASH analysis of *rhp* allowed for the rough prediction of the peptide backbone of rhodoheptins by integrating the A domain specificity conferring code identified by Stachelhaus et al. and the NRPyS library (Blin et al., 2023) (Table 1). Rhodoheptin biosynthesis starts with the loading module which selects and loads an L-Leu residue undergoing *N*-acylation catalysed by the Cs. A putative N-acyl transferase (orf5, Table S1) is located upstream of the NRPS gene and presumably recruits and loads the acyl group to the loading module. Based on the subsequent MS-based structure elucidation of rhodoheptins (Section 2.5), the Cs domain recognizes saturated and monounsaturated β-hydroxy fatty acids ranging from C16 to C24. The *N*-acylated Leu residue is subsequently transferred to the first elongation module featuring an $^{L}C_{L}$-A-PCP domain architecture and catalysing the condensation with an L-Ser residue ($^{L}C_{L}$ is a C domain catalysing condensation of two L amino acids). The following elongation module ($^{L}C_{L}$-A-PCP-E) extends the N-acylated leucylseryl intermediate with a D-Leu or *allo*-Ile residue, generated by the epimerization domain (E) from the corresponding L-epimer. Accordingly, module 4 starts with a $^{D}C_{L}$ forming a peptide bond between the D-Leu/*allo*-Ile and the fourth unknown L-monomer. The growing peptide chain is forwarded sequentially to the following three extension modules, each adding one amino acid, predicted to be threonine, valine, and X. Specificity-pocket structures predicted for the Rhp_A4 and Rhp_A7 domains did not resolve to a single building block, likely due to either a relaxed substrate selectivity or an unprecedented binding site signature. The antiSMASH algorithm refined the prediction of Rhp_A4 selectivity for branched aliphatic amino acids, while it resulted in ambiguous matches for Rhp_A7. In light of these findings, Rhp_A4 was hypothesized to select L-Leu/Ile, consistent with the peptide sequence deduced by MS/MS analysis of rhodoheptins (Section 2.5). On the other hand, structural analysis of the rhodoheptin variants demonstrated that Rhp_A7 possesses substrate promiscuity, being able to activate L-Leu/Ile, L-Phe and L-MetO (Section 2.5). Module 5 displays the same organization as module 3 ($^{L}C_{L}$-A-PCP-E),

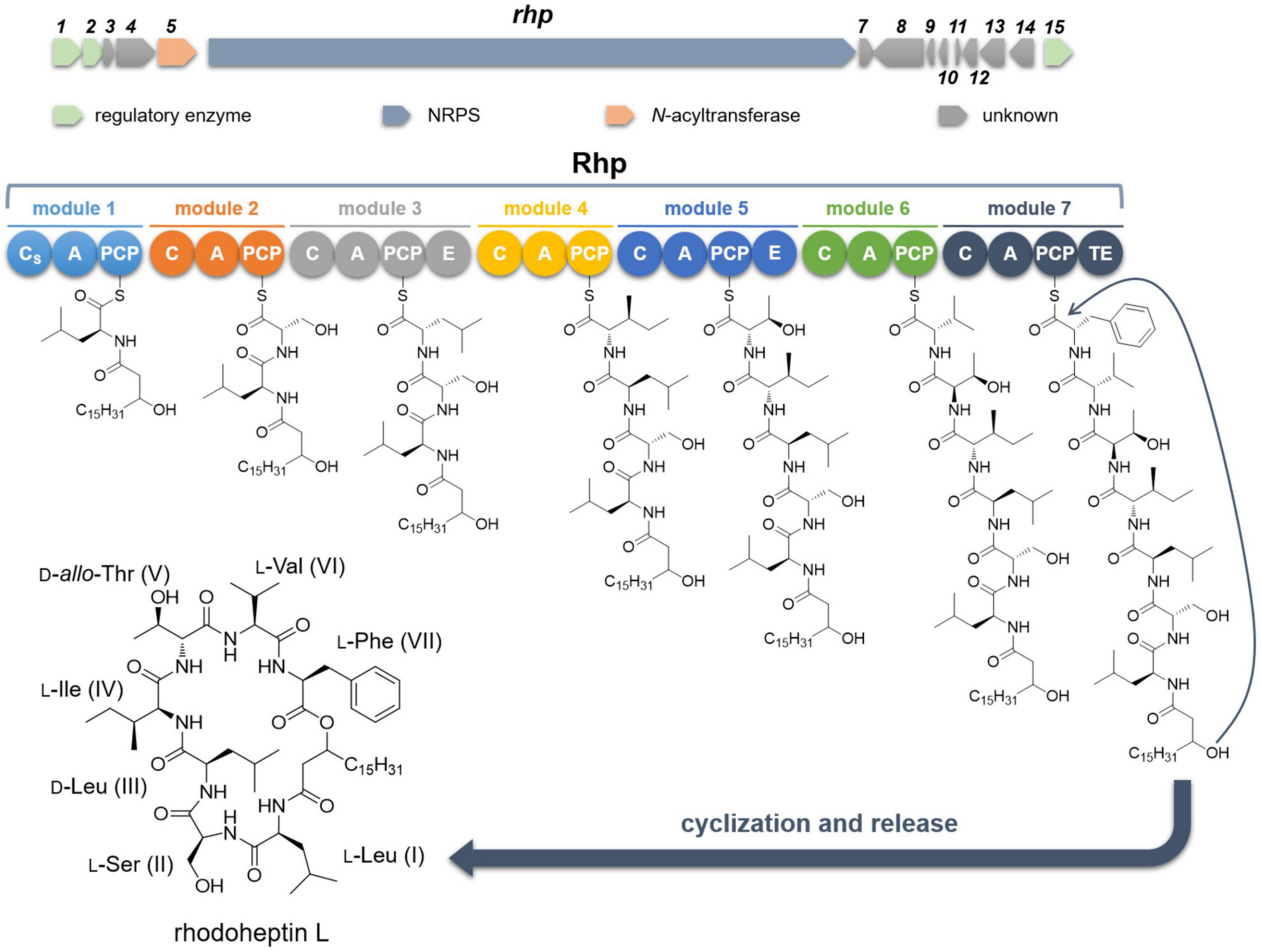


FIGURE 2 Putative biosynthesis of rhodoheptins. C, condensation domain; A, adenylation domain; PCP, peptidyl-carrier protein; E, epimerase; Cs, starter condensation domain; TE, thioesterase.

TABLE 1 Binding pocket signatures of adenylation domains from the Rhp synthetase.

		Binding pocket signatures (position)[a]									
	AntiSMASH prediction	**235**	**236**	**239**	**278**	**299**	**301**	**322**	**330**	**331**	**517**
Rhp_A1	Leu	D	A	L	F	V	G	A	V	F	K
Rhp_A2	Ser	D	V	W	H	F	S	L	V	D	K
Rhp_A3	Leu/Ile	D	A	L	F	A	G	A	I	F	K
Rhp_A4	Unknown	D	A	L	F	V	G	A	V	F	K
Rhp_A5	Thr	D	F	W	N	I	G	M	V	H	K
Rhp_A6	Val	D	A	L	F	V	G	G	I	M	K
Rhp_A7	Unknown	D	A	Y	F	A	G	G	I	Q	K

[a]As reported by Della Sala et al. (2020).

accounting for the incorporation of D-Thr at position 5 of the peptide moiety. As expected, module 6 has a $^{D}C_{L}$-A-PCP domain architecture and is involved in the condensation between D-Thr and L-Val. Finally, the last module terminates with a canonical thioesterase/cyclase domain, which may catalyse the LP release either through hydrolysis or cyclization. In a BlastP search against thioesterases (TEs) from the ThYme database (https://thyme.engr.unr.edu/v2.0/, accessed on 20/12/2023), TEs belonging to the TE16 family were identified as the closest homologues to RhP_TE (E value = $1e\text{-}^{98} \leq \times \leq 2e\text{-}^{38}$ for the first 500 hits). TE16 enzymes have α/β-hydrolase folds and in several PKSs and NRPs may use a hydroxyl group from the substrate chain for lactonization, thus yielding macrocyclic products (Cantu et al., 2010). This is consistent with the structure of rhodoheptins, where the β-hydroxyl group of the N-terminal fatty acid forms an ester bond with the last amino acid incorporated during biosynthesis (Figure 2), similar to the process observed in surfactin biosynthesis (Kraas et al., 2010).

Analysis of the *rmd* biosynthetic gene cluster

The putative *rmd* operon is 47,562 bp long and consists of 14 NRPS modules (Figure 3, Table S2). In addition, the putative biosynthetic genes, *rmd*E and *rmd*F, encode for two distinct acyl-CoA synthetases, which are predicted to cooperate with the NRPS machinery to assemble rhodamides. RmdE and RmdF showed 50% and 33% identity, respectively, with YngI (ABS74201.1) and VrtB (ADI24927.1) fatty acyl-CoA ligases in a BlastP search against the MIBiG database (Terlouw et al., 2023). Therefore, they are expected to catalyse formation of CoA-activated fatty acids to fuel the *rmd* biosynthetic route. NRPS lipopeptide synthetase assembly lines very often require availability of acyl-CoA thioesters to be initiated, as in the case of surfactin biosynthesis, where four different fatty acyl-CoA ligases—LcfA, YhfL, YhfT, and YngI—are involved in providing the activated pool of fatty acid starter units (Kraas et al., 2010). Based on the subsequent MS-based structural characterization of rhodamides (Section 2.6), RmdE and RmdF catalyse thioester formation with CoA from saturated and monounsaturated C16, C20, C22, and C24 fatty acids. The RmdA_Cs domain catalyses the transfer of CoA-activated fatty acids from its donor site to the PCP-bound amino acid, namely glycine. The *N*-acylated glycyl intermediate is elongated by modules 2–14 in an assembly-line fashion as amino acid residues are added sequentially to yield the LP backbone. Although antiSMASH analysis revealed the modular architecture of the *rmd* gene cluster, as well as the domain organization of each NRPS module, substrate selectivity of the A-domain-binding pocket could be predicted only for 6 out of 14 modules, as indicated in Table 2. RmdA_A3, RmdA_A6 and RmdB_A9 activate and select serine residues while RmdA_A7, RmdB_A10 and RmdD_A13 show selectivity for Thr, Ala, and Leu, respectively. Building block prediction based on the adenylation domain specificity was consistent with the structure of rhodamides, except for RmdA_A3. This is not surprising considering that according to the model proposed by Rausch et al. (2005), the 8 Å signature sequence extracted from the RmdA_A3 active site shares only 65% identity with Ser-selective A domains. Notably, modules 5 and 12 of the rhodamide synthetase apparently lack a functional A domain. Prior examples reported the complementation of adenylation activity of NRPS modules lacking their own A domain with upstream/downstream A domains from other modules (Du et al., 2000; Felnagle et al., 2007; Magarvey et al., 2006; Thomas et al., 2003; Zhang et al., 2020). According to this biosynthetic model, which specifies serine specificity, RmdA_A6 and/or RmdB_A9 should deliver Ser to modules 5 and 12 to match the chemical structures of rhodamides. However, it cannot be excluded that the two A-less modules may result from contig/scaffold misassembly caused by either low-depth sequencing data or the presence of highly similar A-domain sequences within the BGC.

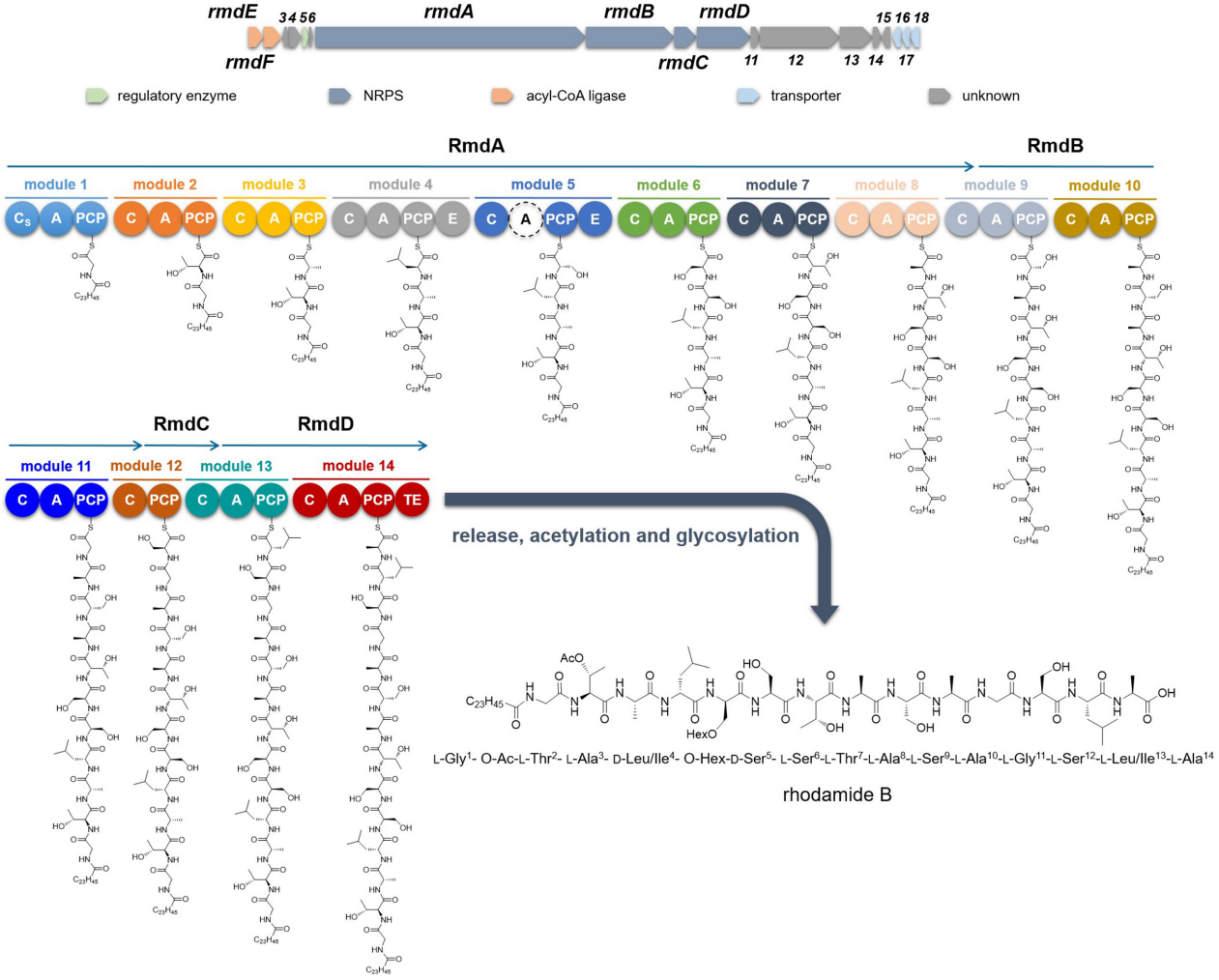


FIGURE 3 Putative biosynthesis of rhodamides. C, condensation domain; A, adenylation domain; PCP, peptidyl carrier protein; E, epimerase; Cs, starter condensation domain; TE, thioesterase; Hex, hexosyl; Ac, acetyl. The A domain labelled with the dashed circle is a putative incomplete or inactive A domain.

TABLE 2 Binding pocket signatures of adenylation domains from the Rmd synthetase.

		Binding pocket signatures (position)[a]									
	AntiSMASH prediction	235	236	239	278	299	301	322	330	331	517
RmdA_A1	Unknown	D	V	W	D	F	I	L	V	S	K
RmdA_A2	Unknown	D	P	L	I	S	G	A	I	V	K
RmdA_A3	Ser	D	V	W	S	L	A	L	V	H	K
RmdA_A4	Unknown	D	P	L	I	S	G	G	I	I	K
RmdA_A6	Ser	D	V	W	H	F	S	L	V	D	K
RmdA_A7	Thr	D	F	W	N	V	G	M	V	H	K
RmdA_A8	Unknown	D	L	W	H	L	I	L	V	S	K
RmdB_A9	Ser	D	V	W	H	F	S	L	V	D	K
RmdB_A10	Ala	D	V	W	S	Q	A	L	V	H	K
RmdB_A11	Unknown	D	V	W	D	F	I	L	V	S	K
RmdD_A13	Leu	D	A	L	F	V	G	A	V	V	K
RmdD_A14	Unknown	D	L	W	H	L	I	V	V	S	K

[a]As reported by Della Sala et al. (2020).

Two canonical epimerization domains have been detected in modules 4 and 5, indicating conversion of L-Ser to the D-configuration at positions 4 and 5 of the peptide backbone. The last NRPS module contains a terminal thioesterase domain belonging to the TE16 family mediating the LP release through hydrolysis, thus yielding a linear product as revealed by MS/MS analysis of rhodamide congeners (Section 2.6).

To achieve the end-product, the LP backbone undergoes (a) side-chain glycosylation at Ser-5 and/or Thr-7 and/or Ser-12 and (b) side-chain acetylation at Thr/Dab-2 and/or Thr-4. Since suitable tailoring enzymes could not be bioinformatically detected within the relevant modules, *trans*-acting glycosyl- and acetyl-transferases are likely to cooperate with the Rhd synthetase to yield the final product. Two putative glycosyltransferases (WP_230557595.1 and WP_230555296.1) belonging to the glycosyltransferase family A are located adjacent to the Rmd NRPS and could O-glycosylate the rhodamide peptide backbone. On the other hand, no acetyl-transferases were identified within the *rmd* gene cluster. More broadly, it cannot be excluded that glycosyl- and acetyl-transferases located at different loci in the genome of *R*. I2R may catalyse these tailoring steps in rhodamide biosynthesis, as reported for the glycosylation of macrolactin and bacillaene (Qin et al., 2014). In addition, the glycosyl moiety in most rhodamide variants undergoes O-succinylation. *R*. I2R has been already reported to assemble succinic saccharide esters, and a probably responsible acyltransferase (WP_230555977) has been identified in the bacterial genome (Palma Esposito et al., 2021).

Tandem mass spectrometry molecular networking analysis of *R*. I2R culture

Our previous OSMAC screening (22 culture conditions) failed to trigger LP production in *R*. I2R (Palma Esposito et al., 2021). In this work, we tried to enhance LP biosynthesis in ASG medium by hexadecane and iron supplementation, as previously described (de Oliveira Schmidt et al., 2021; Peng et al., 2008). After cultivation, biomass of *R*. I2R and exhausted culture broth were harvested separately and extracted using MeOH and AcOEt, respectively. Then, crude extracts were combined and fractionated by C18 reversed-phase (RP) chromatography, thus yielding five fractions F1-F5, eluted using different mixtures of H_2O and MeOH (100% H_2O, 50% MeOH, 90% MeOH, 100% MeOH, and 100% MeOH supplemented with 0.1% TFA).

To map a metabolic profile of *R*. I2R, each fraction was analysed by untargeted LC-HRMS^2. MS^2 data from all fractions were used to generate a global molecular network using the feature-based molecular networking (FBMN) tool (Nothias et al., 2020), which enables clustering of similar molecules based on MS/MS fragmentation similarity (Figure S1). Molecules are subsequently visualized in cytoscape as nodes, connected by edges, with the thickness of each edge corresponding to the similarity of the MS/MS spectra. Linking the molecular networking analysis with the manual detection of specific substructures in MS^2 data led to the identification of two large molecular clusters, which matched the predictions of the new *rhp* and *rmd* BGCs, respectively. The rhodoheptins and rhodamides were shown to be mainly located in the F5 fraction eluted with 100% MeOH plus 0.1% TFA. Therefore, following the workflow described above, a molecular network of the F5 fraction was created to assess its chemical composition (Figure 4). Besides the presence of rhodoheptins (red nodes) and rhodamides (orange nodes), F5 also contained a minor molecular group composed of succinoyl trehalolipids (blue nodes).

To search for LPs encoded by the *rhp* and *rmd* BGCs, LC-HRMS^2 data were imported into the MZmine platform. The data were mined for precursor ions that generated fragment ions and/or neutral losses diagnostic of chemical substructures inferred by genome mining. Specifically, rhodoheptins were identified by looking for precursor ions showing neutral losses of 313.2002 amu ($C_{15}H_{27}N_3O_4$) and/or 200.1161 amu ($C_9H_{16}N_2O_3$) corresponding to the Leu-Ser-Ile/Leu and Thr-Val motifs, respectively, consistent with the prediction of the *rhp* BGC (Table 1, Figure S2). Additionally, rhodamides were detected by searching for precursor ions that generated the fragment ions at *m/z* 159.0764 ($C_6H_{11}N_2O_3^+$) and 189.0870 ($C_7H_{13}N_2O_4^+$) corresponding to the Ser-Ala and Thr-Ser motifs, respectively, as inferred from the prediction of the *rmd* gene cluster (Table 2, Figure S3).

Structure elucidation of rhodoheptins by mass spectrometry

For structural characterization of rhodoheptins, F5 was analysed by LC-HRMS^2 in the positive ion detection mode on a Q Exactive Focus Orbitrap mass spectrometer. Precursor ions included in the rhodoheptin network (red nodes, Figure 4) were selected to be fragmented in the higher-energy collision dissociation cell (HCD). The high-resolution (HR) masses of the pseudo-molecular ions $[M+H]^+$ indicated the molecular formulas reported in Table 3 (mass accuracy ≤1.4 ppm).

Mass fragmentation spectra of rhodoheptins suggested a cyclic LP structure (Figure 5), as being characterized by several *b*- and *y*-type ion series arising from random ring opening at different amide bonds and at the ester bond. Particularly, most rhodoheptin variants first undergo α-ε bond cleavages, thus yielding five different linear fragments (Figure 6 and Figures S5–S23). Then, sequential N- and C-terminal cleavage of these linear fragments resulted in five series of ions,

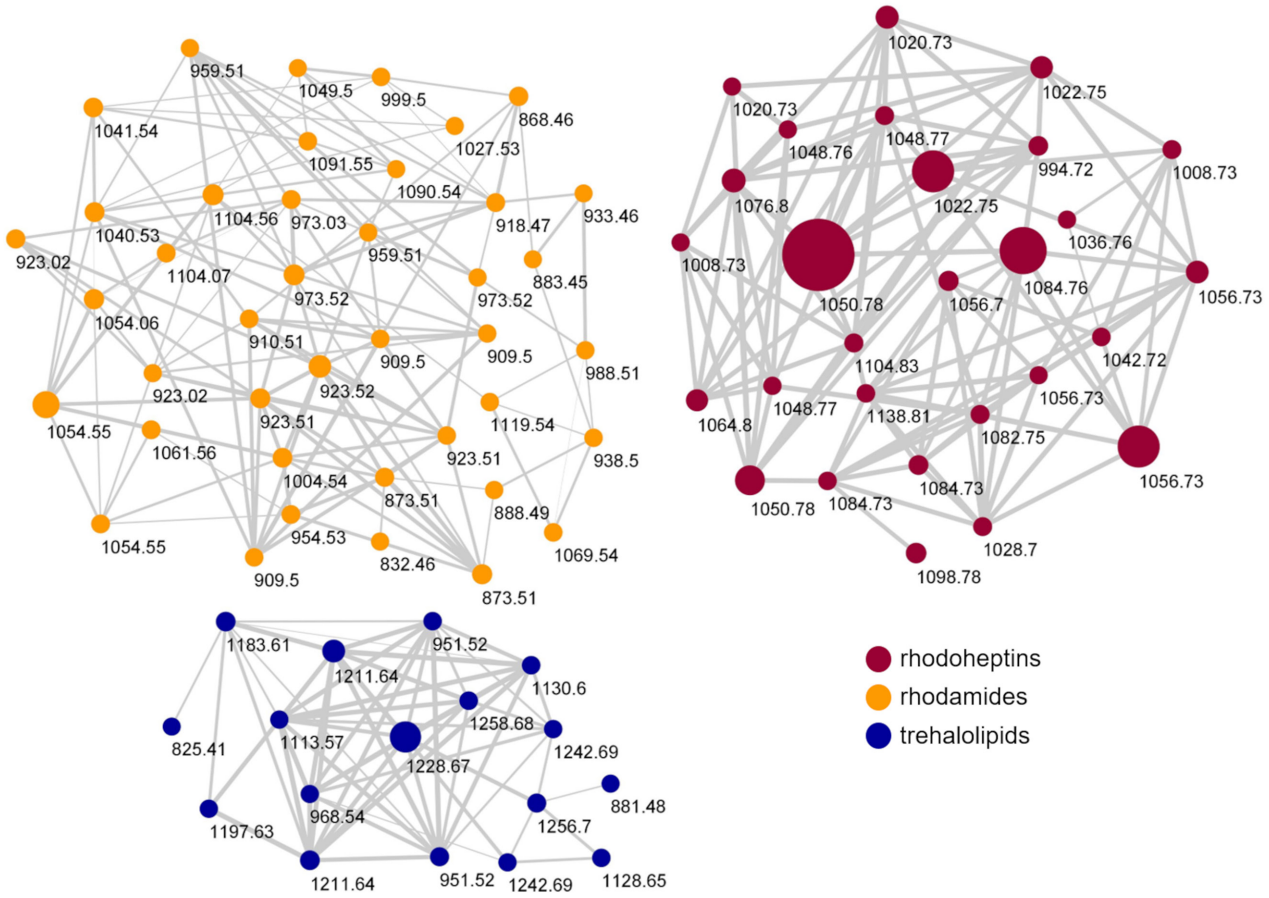


FIGURE 4 Molecular network of the F5 fraction obtained by reversed-phase chromatography of the organic extract of *Rhodococcus* sp. I2R. Nodes are coloured according to biosurfactant classes and their size is related to metabolite amounts. Edge thickness reflects cosine scores which reveal the MS/MS spectra similarity.

which allowed for the assignment of the amino acid sequence of rhodoheptins. In addition, the linear fragment β (Figure 6) displays neutral loss of a β-hydroxy fatty acyl moiety as ketene, which is then linked to Leu^1 through an amide bond. This is consistent with the genome mining prediction results, thereby indicating an LP structure. The β-hydroxy function of the fatty acyl group forms an ester bond with the carboxylic group of aa-7, as inferred from the generation of (a) the y_6 and y_5 ions during fragmentation of the δ linear fragment and (b) the y_5, y_4 and $y_4^{\#}$ ions during fragmentation of the ε linear fragment (Figure 6). As a result, the structure of rhodoheptins was determined as indicated in Table 3. Rhodoheptins have the common tetrapeptide motif Leu^1-Ser^2-Leu^3-Ile^4, while structural diversification results from amino acid substitutions in positions 5, 6, and 7 and/or fatty acyl degree of unsaturation and length. Thus, either Thr or Ser (the latter only in rhodoheptins C (**3**) and G (**7**)) is incorporated into position 5, whereas Val or Ile (the latter only in rhodoheptin N (**14**)) may occupy position 6. However, the highest variation is due to exchange of amino acids in position 7 where Leu/Ile, Phe or MetO may be present. Incorporation of MetO in rhodoheptins K (**11**) and Q (**17**) is corroborated by the presence of fragment ions arising from the neutral loss of methanesulfenic acid (CH_3SOH, 63.9983 Da), which is indicative of the sulfoxide group in side-chain of MetO (Figures S14 and S20). Saturated and monounsaturated β-hydroxy fatty acids ranging from C16 to C24 are the fatty-acyl moieties embedded in the peptide ring of rhodoheptins.

Finally, the structures of 21 rhodoheptins have been characterized, 20 of them being new compounds. Based on the similarity of the MS/MS spectrum, only one congener, that is, rhodoheptin R (**18**), was already reported from *Rhodococcus equi* and *Rhodococcus opacus* (Frankfater et al., 2020), which were shown to produce a family of cyclic and linear LPs closely related to rhodoheptins. By analogy with the structure of rhodoheptin R (**18**) and the other peptidolipid variants elucidated by MS, NMR spectroscopy and GC/MS by Frankfater and co-workers (Frankfater et al., 2020), we tentatively distinguished between the isomeric Ile and Leu residues at positions 1, 3, 4, and 6 in rhodoheptins, as indicated in Table 3. Notably, Leu^1 also correlates with prediction of the adenylation domain selectivity (Table 1). However, it was not possible to recognize by analogy the Ile/Leu^7 moieties in rhodoheptins A-E

TABLE 3 Structures of rhodoheptins A-U (**1–21**) from *Rhodococcus* sp. I2R.

Compound	$[M+H]^+$	m/z	R_t (min)	R (FA)[a]	Aa-5	Aa-6	Aa-7
Rhodoheptin A (**1**)	$C_{52}H_{96}N_7O_{11}$	994.7152	22.3	$C_{13}H_{27}$ (C16:0;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin B (**2**)	$C_{53}H_{98}N_7O_{11}$	1008.7305	26.0	$C_{14}H_{29}$ (C17:0;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin C (**3**)	$C_{53}H_{98}N_7O_{11}$	1008.7311	27.3	$C_{15}H_{31}$ (C18:0;O)	D-Ser	L-Val	L-Leu/Ile
Rhodoheptin D (**4**)	$C_{54}H_{98}N_7O_{11}$	1020.7308	23.2	$C_{15}H_{29}$ (C18:1;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin E (**5**)	$C_{54}H_{100}N_7O_{11}$	1022.7470	30.1	$C_{15}H_{31}$ (C18:0;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin F (**6**)	$C_{55}H_{94}N_7O_{11}$	1028.7001	21.4	$C_{13}H_{27}$ (C16:0;O)	D-*allo*-Thr	L-Val	L-Phe
Rhodoheptin G (**7**)	$C_{56}H_{96}N_7O_{11}$	1042.7154	26.6	$C_{15}H_{31}$ (C18:0;O)	D-Ser	L-Val	L-Phe
Rhodoheptin H (**8**)	$C_{56}H_{102}N_7O_{11}$	1048.7637	29.8	$C_{17}H_{33}$ (C20:1;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin I (**9**)	$C_{56}H_{104}N_7O_{11}$	1050.7787	19.1	$C_{17}H_{35}$ (C20:0;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin J (**10**)	$C_{57}H_{96}N_7O_{11}$	1054.7154	22.3	$C_{15}H_{29}$ (C18:1;O)	D-*allo*-Thr	L-Val	L-Phe
Rhodoheptin K (**11**)	$C_{53}H_{98}N_7O_{12}S$	1056.6981	21.8	$C_{15}H_{31}$ (C18:0;O)	D-*allo*-Thr	L-Val	L-MetO
Rhodoheptin L (**12**)	$C_{57}H_{98}N_7O_{11}$	1056.7317	28.8	$C_{15}H_{31}$ (C18:0;O)	D-*allo*-Thr	L-Val	L-Phe
Rhodoheptin M (**13**)	$C_{57}H_{106}N_7O_{11}$	1064.7940	22.1	$C_{18}H_{37}$ (C21:0;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin N (**14**)	$C_{58}H_{100}N_7O_{11}$	1070.7470	30.4	$C_{15}H_{31}$ (C18:0;O)	D-*allo*-Thr	L-Ile	L-Phe
Rhodoheptin O (**15**)	$C_{58}H_{106}N7O_{11}$	1076.7948	14.7	$C_{19}H_{37}$ (C22:1;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin P (**16**)	$C_{59}H_{100}N_7O_{11}$	1082.7472	28.6	$C_{17}H_{33}$ (C20:1;O)	D-*allo*-Thr	L-Val	L-Phe
Rhodoheptin Q (**17**)	$C_{55}H_{102}N_7O_{12}S$	1084.7290	25.8	$C_{17}H_{35}$ (C20:0;O)	D-*allo*-Thr	L-Val	L-MetO
Rhodoheptin R (**18**)[b]	$C_{59}H_{102}N_7O_{11}$	1084.7633	15.3	$C_{17}H_{35}$ (C20:0;O)	D-*allo*-Thr	L-Val	L-Phe
Rhodoheptin S (**19**)	$C_{60}H_{104}N_7O_{11}$	1098.7781	18.9	$C_{18}H_{37}$ (C21:0;O)	D-*allo*-Thr	L-Val	L-Phe
Rhodoheptin T (**20**)	$C_{60}H_{110}N_7O_{11}$	1104.8262	34.8	$C_{21}H_{41}$ (C24:1;O)	D-*allo*-Thr	L-Val	L-Leu/Ile
Rhodoheptin U (**21**)	$C_{63}H_{108}N_7O_{11}$	1138.8105	32.3	$C_{21}H_{41}$ (C24:1;O)	D-*allo*-Thr	L-Val	L-Phe

Abbreviations: aa, amino acid; FA, fatty acid; MetO, methionine sulfoxide; R_t, retention time.

[a]Fatty acids have been reported in parenthesis using the LIPID MAPS shorthand notation (Liebisch et al., 2020). Fatty acyl chains are indicated as C:N;O, where C is the number of carbon atoms, N is the number of double bond equivalents and O is the number of additional oxygen atoms linked to the hydrocarbon chain.

[b]Frankfater et al. (2020).

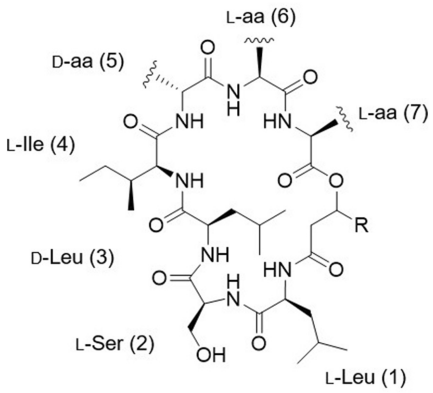


FIGURE 5 General structure of rhodoheptins.

(**1**–**5**), H-I (**8**–**9**), M (**13**), O (**15**) and T (**20**), as such a residue is not present in any peptidolipid variants from *R. equi* and *R. opacus*, where the last amino acid is always Phe.

Rhodoheptins are composed of mixed D- and L-amino acids. The absolute configuration of the seven amino acids was assigned through bioinformatic prediction. As the rhodoheptin synthetase includes two epimerization domains within modules 3 and 5 (Figure 2), respectively, the presence of D-Leu^3 and D-*allo*-Thr^5/D-Ser^5 was inferred, while the remaining amino acids were assumed to possess the L configuration. Nevertheless, the absolute configuration of the β-carbon of the β-hydroxy fatty acids in rhodoheptins remains unsolved, as not predictable by gene cluster analysis.

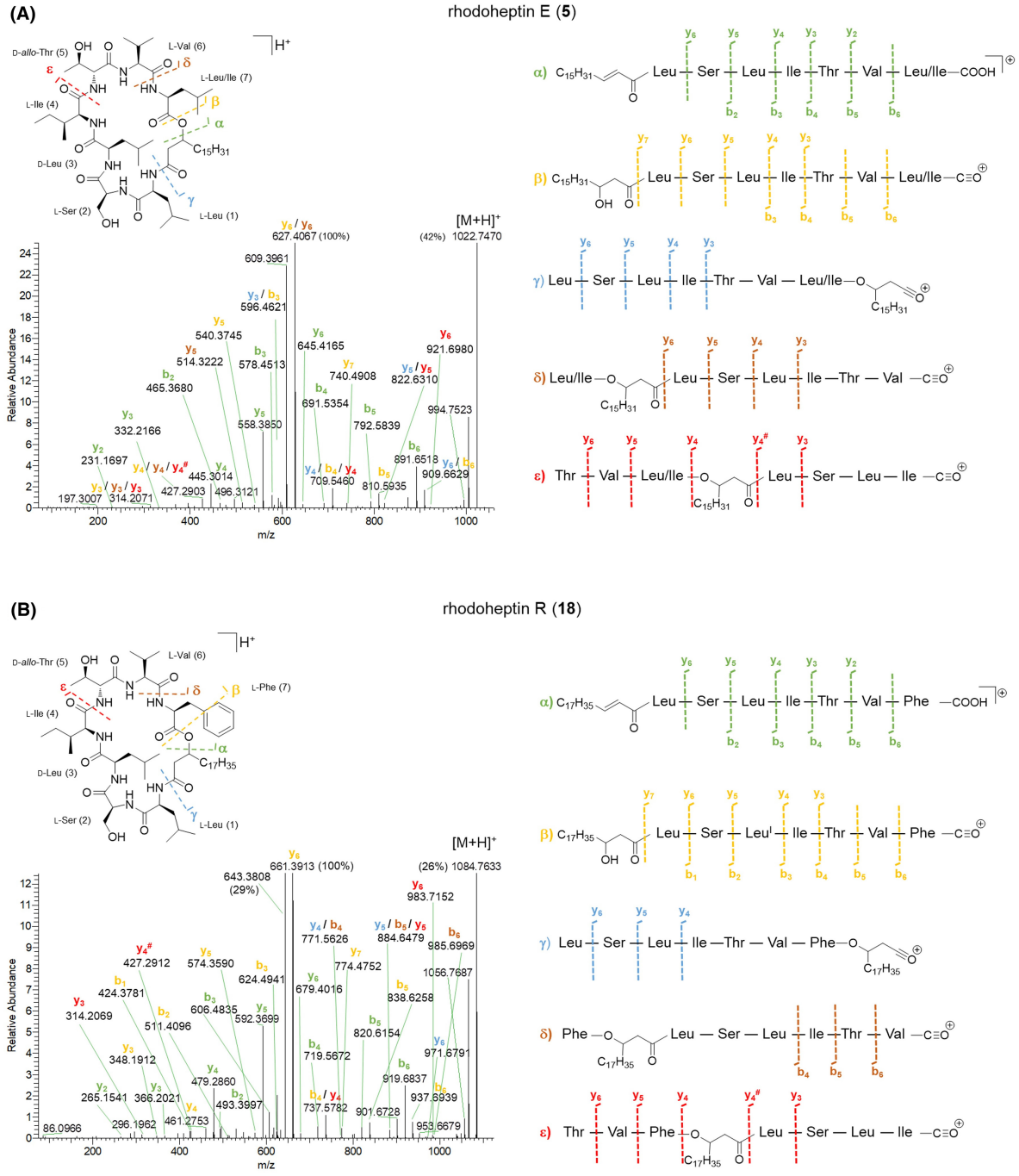


FIGURE 6 HR-MS^2 spectra of the [M + H]$^+$ pseudo-molecular ions of rhodoheptin E (**5**) (A) and rhodoheptin R (**18**) (B) and their proposed fragmentation pathways.

Structure elucidation of rhodamides by mass spectrometry

After reversed-phase liquid chromatography, MS/MS spectra of [M + 2H]$^{2+}$ ions of rhodamides (orange nodes, Figure 4) were acquired at lower and higher normalized collision energies (NCE), that is, 15 and 30, to obtain both larger and smaller fragment ions for *de novo* sequencing of the peptide backbone. The high-resolution ESI mass spectrum of [M + 2H]$^{2+}$ ions of rhodamides was indicative of the molecular formulas reported in Table 4 (mass accuracy ≤3 ppm). Rhodamides were shown to be linear glycosylated and acetylated peptidolipids consisting of a tetradecapeptide linked to a saturated or monounsaturated fatty acid via an amide bond with the N-terminal amino acid residue (Figure 7).

The product ion spectra of [M + 2H]$^{2+}$ ions of rhodamides at NCE 15 displayed y_{12}-y_2 and b_{12}-b_1 ions together with *b*-type internal fragments arising from fragmentation of the y_{10} ion, the latter being

TABLE 4 Structures of rhodamides A-Z and A1-D1 (**22**–**54**) from *Rhodococcus* sp. I2R.

Compound	$[M+2H]^{2+}$	m/z	R_t (min)	R_1 (FA)[a]	R_2	R_3	R_4	L-aa-2	D-aa-4
Rhodamide A (**22**)	$C_{76}H_{132}N_{14}O_{31}$	868.4610	5.9	$C_{15}H_{31}$ (C16:0)	Succ-Hex	H	H	O-Ac-Thr	Leu/Ile
Rhodamide B (**23**)	$C_{80}H_{142}N_{14}O_{28}$	873.5066	20.3	$C_{23}H_{45}$ (C24:1)	Hex	H	H	O-Ac-Thr	Leu/Ile
Isorhodamide B (**24**)	$C_{80}H_{142}N_{14}O_{28}$	873.5066	21.6	$C_{23}H_{45}$ (C24:1)	H	Hex	H	O-Ac-Thr	Leu/Ile
Rhodamide C (**25**)	$C_{76}H_{130}N_{14}O_{33}$	883.4483	5.2	$C_{15}H_{31}$ (C16:0)	Succ-Hex	H	H	O-Ac-Thr	O-Ac-Thr
Rhodamide D (**26**)	$C_{80}H_{140}N_{14}O_{30}$	888.4929	18.4	$C_{23}H_{45}$ (C24:1)	Hex	H	H	O-Ac-Thr	O-Ac-Thr
Rhodamide E (**27**)	$C_{82}H_{142}N_{14}O_{31}$	909.4992	14.4	$C_{21}H_{41}$ (C22:1)	Succ-Hex	H	H	O-Ac-Thr	Leu/Ile
Isorhodamide E (**28**)	$C_{82}H_{142}N_{14}O_{31}$	909.4992	15.2	$C_{21}H_{41}$ (C22:1)	H	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide F (**29**)	$C_{82}H_{144}N_{14}O_{31}$	910.5064	21.0	$C_{21}H_{43}$ (C22:0)	Succ-Hex	H	H	O-Ac-Thr	Leu/Ile
Rhodamide G (**30**)	$C_{80}H_{136}N_{14}O_{34}$	918.4683	6.1	$C_{15}H_{31}$ (C16:0)	Disucc-Hex	H	H	O-Ac-Thr	Leu/Ile
Rhodamide H (**31**)	$C_{84}H_{147}N_{15}O_{30}$	923.0219	18.2	$C_{23}H_{45}$ (C24:1)	Succ-Hex	H	H	N-γ-Ac-Dab	Leu/Ile
Rhodamide I (**32**)	$C_{84}H_{146}N_{14}O_{31}$	923.5143	20.0	$C_{23}H_{45}$ (C24:1)	Succ-Hex	H	H	O-Ac-Thr	Leu/Ile
Isorhodamide I (**33**)	$C_{84}H_{146}N_{14}O_{31}$	923.5143	21.4	$C_{23}H_{45}$ (C24:1)	H	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide J (**34**)	$C_{80}H_{134}N_{14}O_{36}$	933.4566	5.4	$C_{15}H_{31}$ (C16:0)	Disucc-Hex	H	H	O-Ac-Thr	O-Ac-Thr
Rhodamide K (**35**)	$C_{84}H_{144}N_{14}O_{33}$	938.5007	18.4	$C_{23}H_{45}$ (C24:1)	Succ-Hex	H	H	O-Ac-Thr	O-Ac-Thr
Rhodamide L (**36**)	$C_{86}H_{152}N_{14}O_{33}$	954.5326	18.6	$C_{23}H_{45}$ (C24:1)	Hex	Hex	H	O-Ac-Thr	Leu/Ile
Rhodamide M (**37**)	$C_{86}H_{146}N_{14}O_{34}$	959.5073	14.4	$C_{21}H_{41}$ (C22:1)	Disucc-Hex	H	H	O-Ac-Thr	Leu/Ile
Rhodamide N (**38**)	$C_{88}H_{151}N_{15}O_{33}$	973.0306	17.9	$C_{23}H_{45}$ (C24:1)	Disucc-Hex	H	H	N-γ-Ac-Dab	Leu/Ile
Rhodamide O (**39**)	$C_{88}H_{150}N_{14}O_{34}$	973.5222	19.7	$C_{23}H_{45}$ (C24:1)	Disucc-Hex	H	H	O-Ac-Thr	Leu/Ile
Rhodamide P (**40**)	$C_{88}H_{148}N_{14}O_{36}$	988.5094	18.2	$C_{23}H_{45}$ (C24:1)	H	Disucc-Hex	H	O-Ac-Thr	O-Ac-Thr
Rhodamide Q (**41**)	$C_{86}H_{146}N_{14}O_{39}$	999.4961	5.2	$C_{15}H_{31}$ (C16:0)	Succ-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide R (**42**)	$C_{90}H_{156}N_{14}O_{36}$	1004.5404	18.4	$C_{23}H_{45}$ (C24:1)	Succ-Hex	Hex	H	O-Ac-Thr	Leu/Ile
Rhodamide S (**43**)	$C_{90}H_{154}N_{14}O_{39}$	1027.5263	13.1	$C_{19}H_{39}$ (C20:0)	Succ-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide T (**44**)	$C_{92}H_{156}N_{14}O_{39}$	1040.5341	13.1	$C_{21}H_{41}$ (C22:1)	Succ-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide U (**45**)	$C_{92}H_{158}N_{14}O_{39}$	1041.5411	19.2	$C_{21}H_{43}$ (C22:0)	Succ-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide V (**46**)	$C_{90}H_{150}N_{14}O_{42}$	1049.5042	5.4	$C_{15}H_{31}$ (C16:0)	Disucc-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide W (**47**)	$C_{94}H_{161}N_{15}O_{38}$	1054.0575	16.5	$C_{23}H_{45}$ (C24:1)	Succ-Hex	H	Succ-Hex	N-γ-Ac-Dab	Leu/Ile
Rhodamide X (**48**)	$C_{94}H_{160}N_{14}O_{39}$	1054.5492	18.3	$C_{23}H_{45}$ (C24:1)	Succ-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide Y (**49**)	$C_{94}H_{158}N_{14}O_{41}$	1069.5369	16.9	$C_{23}H_{45}$ (C24:1)	Succ-Hex	H	Succ-Hex	O-Ac-Thr	O-Ac-Thr
Rhodamide Z (**50**)	$C_{96}H_{160}N_{14}O_{42}$	1090.5419	13.2	$C_{21}H_{41}$ (C22:1)	Disucc-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide A1 (**51**)	$C_{96}H_{162}N_{14}O_{42}$	1091.5485	18.8	$C_{21}H_{43}$ (C22:0)	Disucc-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide B1 (**52**)	$C_{98}H_{165}N_{15}O_{41}$	1104.0656	16.3	$C_{23}H_{45}$ (C24:1)	Disucc-Hex	H	Succ-Hex	N-γ-Ac-Dab	Leu/Ile
Rhodamide C1 (**53**)	$C_{98}H_{164}N_{14}O_{42}$	1104.5573	18.0	$C_{23}H_{45}$ (C24:1)	Disucc-Hex	H	Succ-Hex	O-Ac-Thr	Leu/Ile
Rhodamide D1 (**54**)	$C_{98}H_{162}N_{14}O_{44}$	1119.5451	16.8	$C_{23}H_{45}$ (C24:1)	Disucc-Hex	H	Succ-Hex	O-Ac-Thr	O-Ac-Thr

Abbreviations: aa, amino acid; Dab, diaminobutyric acid; Disucc, disuccinyl; FA, fatty acid; Hex, hexosyl; R_t, retention time; Succ, succinoyl.

[a]Fatty acids have been reported in parenthesis using the LIPID MAPS shorthand notation (Liebisch et al., 2020). Fatty acyl chains are indicated as C:N;O, where C is the number of carbon atoms, N is the number of double bond equivalents and O is the number of additional oxygen atoms linked to the hydrocarbon chain.

useful to identify the last two amino acids as Ile/Leu[13] and Ala[14] (Figures 8 and Figure S24–S26). The observed fragmentation pattern at NCE 15 allowed for the assignment of the peptide sequence as shown in Table 4, however, it was necessary to increase collision energy (NCE 30) to identify the structure of the b_1 fragment ion, which features in all rhodamide isoforms a Gly residue loss to yield the acylium ion corresponding to the fatty acid starter unit (Figure 8B). With this regard, rhodamides were shown to possess C16:0, C20:0 C22:0, C22:1 and C24:1 fatty acyl chains (Table 4).

As a result, the peptide backbone of rhodamides was determined as Gly[1]-aa[2]-Ala[3]-aa[4]-Ser[5]-Ser[6]-Thr[7]-Ala[8]-Ser[9]-Ala[10]-Gly[11]-Ser[12]-Leu/Ile[13]-Ala[14], where the different congeners contain (a) O-acetylthreonine (O-Ac-Thr) or *N*-γ-acetyl-2,4-diaminobutyric acid (*N*-γ-Ac-Dab) and (b) Leu/Ile or O-Ac-Thr at the variable amino acid positions 2 and 4, respectively. Particularly, the presence of O-Ac-Thr or *N*-γ-Ac-Dab at positions 2 and/or 4 was inferred from the loss of the acetyl group as ketene (C_2H_2O, 42.0106 Da) from b_2 and/or b_4 fragment ions, followed by the loss of a Thr or Dab residue (Figures 8, and Figure S24–S26). An in-depth investigation of the MS/MS spectra led to the identification and position assignment of the substitution patterns of rhodamides as detailed in Table 4. Rhodamides bear up to two pendant hexosyl groups, which may undergo or not mono- or di-succinylation. Indeed, the MS/MS spectra of rhodamides displayed low-abundance fragment ions at *m/z* 163.0601 ($C_6H_{11}O_5^+$), *m/z* 145.0495 ($C_6H_9O_4^+$) and/or *m/z* 263.0761 ($C_{10}H_{15}O_8^+$) and/or *m/z* 363.0922 ($C_{14}H_{19}O_{11}^+$), as well as the corresponding neutral losses, which were indicative of the dehydrated forms of the hexoside (Hex), succinoyl-hexoside (Succ-Hex) and disuccinoyl-hexoside (Disucc-Hex) moieties (Figures 8B and Figure S24B–S26B). Indeed, fragment ions at *m/z* 263.0761 and *m/z* 363.0922 undergo further fragmentation resulting in the neutral loss of one molecule of succinic acid ($C_4H_6O_4$, 118.0260 amu) and sequential losses of succinic acid and succinic anhydride ($C_4H_4O_3$, 100.0160 amu), respectively, thereby suggesting one or two succinate units ester linked to the hexosyl group. The product ion spectra of rhodamides bearing the Hex/Succ-Hex/Disucc-Hex on Ser[5] side-chain hydroxy function exhibit glycosylated b_{12}-b_5 ion fragments (besides the corresponding non-glycosylated counterparts) but only non-glycosylated b_4-b_1 ions, thus allowing for the unambiguous assignment of the substituent position (Figure S24). Also, rhodamide variants with a pendant

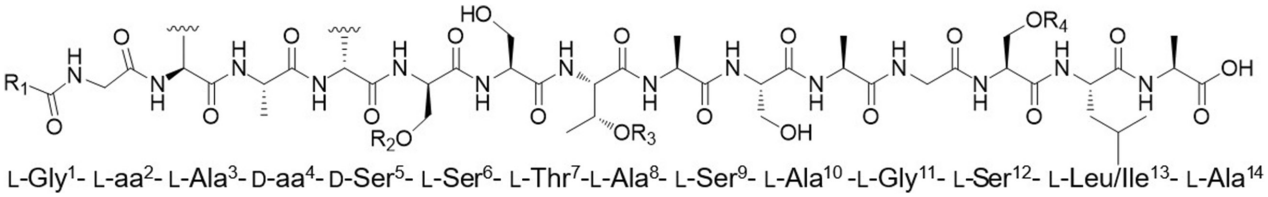


FIGURE 7 General structure of rhodamides.

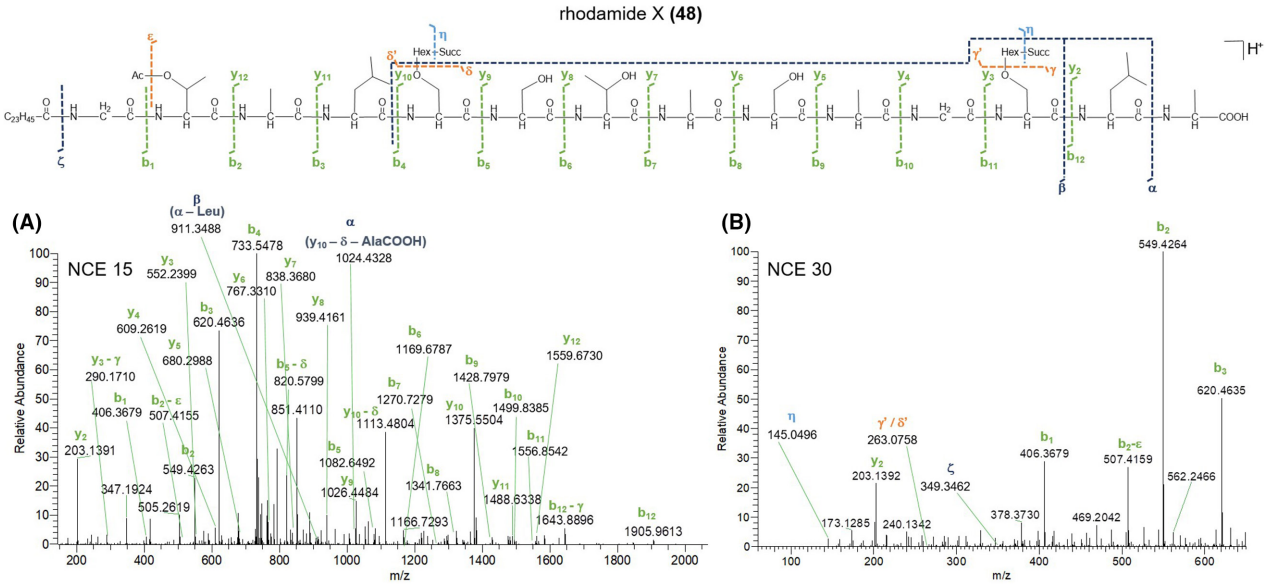


FIGURE 8 HR-MS^2 spectra of the $[M+2H]^{2+}$ pseudo-molecular ion of rhodamide X (**48**) acquired at NCE 15 (A) and 30 (B) and its proposed fragmentation pathway. Succ, succinoyl; Hex, hexosyl; Ac, acetyl.

glycosyl moiety on Thr^7 or Ser^{12} side-chain hydroxy group display glycosylated y_{12}-y_8 or y_{12}-y_3 ions (together with the corresponding non-glycosylated fragments) but only the non-glycosylated y_7-y_2 or y_2 ions (Figure S25 and S26). The simultaneous presence of glycosylated y_{12}-y_8/y_{12}-y_3 and b_{12}-b_5 ion series in MS/MS spectra was indicative of rhodamide congeners featuring substituents on both the Thr^7/Ser^{12} and Ser^5 side-chains (Figure 8).

The genomics-driven absolute configuration assignment of rhodamides indicated the presence of D-Leu/Ile[4], D-*allo*-O-Ac-Thr^4 and D-Ser^5, as the rhodamide synthetase features two epimerization domains within modules 4 and 5, respectively (Figure 3). Consequently, the remaining amino acids were assumed to possess the L configuration.

Finally, the structures of 33 glycolipopeptides have been reported here for the first time. Interestingly, a group of cell wall glycolipopeptides related to rhodamides has been isolated from *R. erytropolis* 134 and differs from rhodamides in the amino acid composition, the substitution pattern and the N-terminal fatty acid moiety, including mycolic acids besides normal fatty acids (Koronelli, 1988).

In addition, while glycolipopeptides from *R. erytropolis* 134 and peptidolipids from *R. equi* and *R. opacus* have been described as cell wall components, rhodamides and rhodoheptins have been detected both in the cell pellet and the supernatant of *R.* I2R. Therefore, it can be argued that the bacterium may secrete these biosurfactants to emulsify and intake hexadecane from the growth medium (see Section 4.2), as widely reported for *Corynebacteriaceae* (Koronelli, 1988).

Bioactivity evaluation of rhodoheptins and rhodamides

The F5 fraction (including rhodoheptins, rhodamides, and trehalolipids) exhibited biosurfactant activity in the oil-spreading assay, causing oil displacement and forming a clear zone in the oil layer (Figure S4) as for other surface active agents of microbial origin (Giugliano et al., 2023; Buonocore et al., 2020). Therefore, the identification of rhodoheptins and rhodamides further expands the biosurfactant repertoire of *R.* I2R, previously recognized solely as a source of glycolipid biosurfactants (Palma Esposito et al., 2021). An increasing body of evidence suggests that many biosurfactants may interfere with different steps of cancer development and bacterial infections (Wang et al., 2024). Consequently, we assessed potential antimicrobial and antiproliferative effects of LPs from *R.* I2R.

F5 did not show any activity in antibacterial assays against *Staphylococcus aureus* ATCC 6538, *Listeria monocytogenes* MB 677, *E. coli* ATCC 10536, and *Pseudomonas aeruginosa* PA01.

Moreover, F5 and a purified rhodoheptin-enriched fraction were evaluated for antiproliferative activity against human A375 melanoma cells using the MTT assay (Figure 9). The purified rhodoheptin-enriched fraction was obtained following the procedure described in Section 4.2 and its composition was assessed using LC-$HRMS^2$ (Figure S27 and Table S4), thus confirming the presence of all the rhodoheptin variants listed in Table 3. F5 induced approximately 50% cancer cell death after a 48-hour treatment at the highest concentration (500 μg/mL). In contrast, the purified rhodoheptin-enriched fraction elicited weaker yet significant growth inhibition (33%) in A375 cells at the same concentration. The synergestic effects of rhodoheptins, trehalolipids and presumably rhodamides may explain the increased reduction in cell viability, considering that trehalolipids from *R.* I2R have been already shown to induce cytotoxicity against PC3 prostate cancer cells (Palma Esposito et al., 2021).

Rhodoheptin and rhodamide-related lipopeptides from other *Rhodococcus* strains

As *R. equi* and *R. opacus* are able to synthesize cyclic and linear LPs structurally related to rhodoheptins,

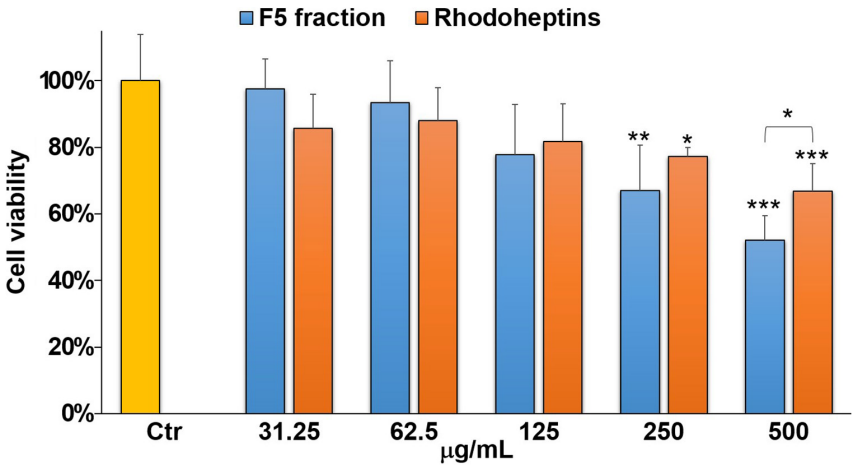


FIGURE 9 Cell viability of A375 cells exposed to 0.25% DMSO vehicle (Ctr) and different concentrations of F5 and rhodoheptin mixture (31.25, 62.5, 125, 250 or 500 μg/mL). Data have been normalized to the relevant control (Ctr) and are presented as mean ± SD; $n = 4$. Where not indicated, statistical significances are referred to as control (Ctr). *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

genomes of both species (NCBI accession numbers NZ_CP027793.1 and NZ_CP080954.1) were mined *in silico* for the relevant BGCs. Unfortunately, the genome of *R. erytropolis* 134 is not publicly available and, therefore, it was not possible to look for the BGC encoding for the biosynthesis of the rhodamide-related glycolipopeptides described by Koronelli. (1988). Nevertheless, an antiSMASH analysis of several *R. erytropolis* strains (CCM2595, JCM 2895, X5, CERE8 and D310) disclosed the presence of multimodular NRPS putatively involved in the assemblage of serine- and threonine-containing LPs resembling rhodamide architecture.

As expected, *R. equi* and *R. opacus* were found to host the heptamodular NRPS gene cluster, specifically WP_106851421.1 and WP_025433610.1, respectively, featuring the same organization as *rhp* and showing collinearity with the peptidolipid family described by Frankfater and co-workers (Figure S28). These findings further support the link between the BGCs and the encoded molecules, thus shedding light on a biosynthetic route conserved across different *Rhodococcus* strains. Basically, rhodoheptins from *R.* I2R differ from peptidolipid variants of *R. equi* and *R. opacus* as a result of (a) the relaxed substrate selectivity of the adenylation domain present in the last module, which allows for the accommodation of Leu/Ile and MetO, beyond Phe, and (b) the fatty acyl moiety specificity of the Cs domain, which selects C16 and C17, beyond C18-C24 β-hydroxy fatty acids.

In the most general terms, the LP profile of *R.* I2R is more similar to that of *R. equi* as being exclusively characterized by cyclic LPs. Notably, detection of a putative esterase-encoding gene (WP_005249251.1) flanking the NRPS cluster solely in *R. opacus* could explain why this strain is able to produce also linear LPs, likely arising from the corresponding cyclic variants after hydrolytic cleavage of the lactone ring. As for the rhodoheptin synthetase, the heptamodular NRPS from *R. equi* contains two E domains within the modules 3 and 5, thus unveiling the conserved D absolute configuration of aa-3 and -5 (Figure S28). On the other hand, the heptamodular NRPS from *R. opacus* bears two E domains within modules 1 and 6, thus accounting for a different stereochemistry of the entire LP family in this strain (Figure S28).

To determine if the rhodoheptin BGC is widespread in *Rhodococcus* species, a cblaster analysis (Gilchrist et al., 2021) was performed. All genomic data from *Rhodococcus* available on NCBI were used. We detected a total of 2224 clusters distributed across 1936 genomic scaffolds originating from 856 different organisms. Among these, 322 unique species were identified to have at least 60% of the genes present in the query cluster (Figure S29). For the rhodamide BGC, cblaster detected 2499 clusters across 2150 genomic scaffolds from 860 organisms within the NCBI *Rhodococcus* data. Of these, 478 unique species were found to contain at least 60% of the genes present in the query cluster (Figure S30). This indicates a wide distribution of the rhodoheptin and rhodamide biosynthetic pathways across a diverse set of *Rhodococcus* species.

CONCLUSIONS

The genome and metabolome analysis of *R.* I2R enabled the structure and biosynthetic pathway discovery of two LP families with biosurfactant and moderate antiproliferative activity. These families would have remained cryptic if a genomic bottom-up approach had not been applied. Indeed, genome mining guided the selection of tailored growth media to unlock LP biosynthesis and supported detection and stereo-structural elucidation of rhodoheptins and rhodamides as well. So far, few studies report about LPs from *Rhodococcus* spp. (Chiba et al., 1999; Habib et al., 2020; Peng et al., 2008), as the most prolific strains belong to *Bacillus* and *Paenibacillus* species, *Pseudomonas* spp., Cyanobacteria and Actinomycetota, mainly represented by the *Streptomyces* genus (Carolin et al., 2021; Cock & Cheesman, 2023; Li et al., 2021; Zhang et al., 2023). Therefore, our results represent a step forward in the investigation of the underexplored *Rhodococcus* genus as a biotechnological resource. Nevertheless, as widely documented for biosurfactants (Hashemi et al., 2024), the low microbial productivity remained a bottleneck which hampered isolation and NMR structural characterization of rhodoheptins and rhodamides. Notably, *R.* I2R was shown to possess multiple biosynthetic routes for biosurfactant production, thus making this bacterium a factory of diverse surface-active natural products, i.e., saccharide succinic esters (Palma Esposito et al., 2021), LPs and glycolipopeptides. As *R.* I2R has been isolated from deep-sea sediments, the bacterium has likely maintained this synergy as a result of the environmental pressure (e.g. light and nutrient shortage, high pressure, and salinity), which might have imposed *R.* I2R to synthesize molecules such as biosurfactants, for complex carbon source degradation, nutrient retrieval and chemical defence. Due to the novelty of the chemical structures and the availability of the BGCs of rhodoheptins and rhodamides, our future work aims at the large-scale production of these potentially industrially relevant biosurfactants through heterologous expression.

EXPERIMENTAL PROCEDURES

Bioinformatics

The draft genome of *R.* I2R (accession: JAHUTG 000000000) consisting of 72 contigs was assembled into 36 scaffolds employing the software MeDuSa (Multi-Draft

based Scaffolder) (Bosi et al., 2015) and selecting the genomes of *Rhodococcus* sp. P1Y (NZ_CP032762.1) and *Rhodococcus* sp. BP-261 (NZ_JABUCO000000000.1) as references to guide contigs ordering and orientating, as previously described (Palma Esposito et al., 2023). BGCs encoding the biosynthesis of LPs were identified by using the genome mining tool antiSMASH 7.0 (Blin et al., 2023). The rhodoheptin and rhodamide gene clusters were deposited in GenBank under accession numbers PP729154 and PP729155, respectively. Cblaster (Gilchrist, et al. 2021) was run using the search module (default settings), against the NCBI Database (limited to *Rhodococcus* genomes). To visualize the similarity of the clusters, clinker was used.

Cultivation, extraction and fractionation

The marine bacterium *R*. I2R was isolated on agar plates containing mineral salt medium (MSM) supplemented with 1.0 mM of phenanthrene from deep-sea sediments collected in the Southern Tyrrhenian Sea, as previously reported (Palma Esposito et al., 2021). For LP production, *R*. I2R was grown at 28°C under shaking (200 rpm) for 6 days in 200 mL of ASG medium (casamino acids 3 g/L; $MgSO_4 \times 7\ H_2O$ 12 g/L; KCl 0.75 g/L; NaCl 15 g/L; $CaCl_2$ 3 g/L; NH_4Cl 1 g/L; $NaHCO_3$ 0.17 g/L) supplemented with 0.5% (v/v) hexadecane and 0.5 μM $FeCl_3$. Hexadecane and iron were added as potential hydrophobic and hydrophilic inducers, respectively, to stimulate LP synthesis (de Oliveira Schmidt et al., 2021; Peng et al., 2008; Solyanikova & Golovleva, 2019). After the 6 days of growth, biomass of *R*. I2R and exhausted broth were harvested separately and extracted with MeOH (3 × 100 mL) and AcOEt (2 × 200 mL) respectively. The combined MeOH and AcOEt extracts (1.734 g) were purified by reversed-phase chromatography using a C18 SPE column eluted with different mixtures of H_2O and MeOH to obtain five fractions: F1, 100% H_2O (88.9 mg); F2, 50% MeOH (32.7 mg); F3, 90% MeOH (51,8 mg); F4, 100% MeOH (20.4 mg); F5, 100% MeOH supplemented with 0.1% TFA (6.5 mg). To obtain the rhodoheptin-enriched fraction, crude bacterial extract was separated over a C18 SPE column using the above-mentioned gradient elution and finally flushed with $CHCl_3$, to collect an additional fraction (0.5 mg) containing only residual rhodoheptins previously retained by the stationary phase.

Liquid chromatography–high-resolution tandem mass spectrometry (LC-HRMS^2)

The five SPE fractions were suspended in CH_3OH at a concentration of 1 mg/mL and analysed by untargeted LC-HRMS^2 on a Thermo Scientific Q Exactive Focus Orbitrap mass spectrometer coupled to a Thermo Ultimate 3000 HPLC System equipped with an Hypersil C18 column (100 × 4.6 mm, 3 μm). The RP18 column was maintained at 25°C and eluted at 400 μL/min with H_2O and CH_3CN, both supplemented with 0.1% formic acid, setting the following gradient program: 70% CH_3CN 10 min (equilibration), 70% → 75% CH_3CN over 3 min, 75% → 85% CH_3CN over 5 min, 85% → 100% CH_3CN over 22 min, 100% CH_3CN 11 min. MS spectra were recorded in the positive ion detection mode and HESI source parameters were set as follows: a sheath gas flow rate of 32 units N_2, an auxiliary gas flow rate of 15 units N_2, a spray voltage of 4.8 kV, a capillary temperature of 285°C, an S-lens RF level of 55 and an auxiliary gas heater temperature of 150°C. MS survey scans (800–1400 m/z) were acquired at a resolution of 70,000 and an AGC target of $1e^6$. HRMS^2 spectra were acquired in the DDA mode at a resolution of 70,000 and an AGC target of $5e^4$, setting three MS^2 events after each full MS scan. HRMS^2 scans were obtained with HCD fragmentation, using an isolation width of 2.0 m/z, a normalized collision energy of 15 and 30 units and an automated injection time.

Feature-based molecular networking (FBMN)

To chart a metabolic overview of *R*. I2R, the molecular network of the F1-F5 fractions from *R*. I2R was generated with the FBMN workflow (Nothias et al., 2020; Teta et al., 2016) on GNPS (https://gnps.ucsd.edu), as previously reported (Teta et al., 2021). MS raw data were initially processed by MZmine2 (Pluskal et al., 2010) using the parameters reported in the Supplementary Material (Table S3). Briefly, after chromatogram building and deconvolution, peaks from the SPE fractions and blank sample (i.e. methanol used for fraction suspension) were aligned to subtract the background spectrum and eliminate interference. In addition, [M + Na–H], [M + K–H], [M + Mg − 2H], [M + NH_3], [M-Na + NH_4], [M + 1, ^{13}C] adducts and peaks without associated MS^2 spectra were filtered out. The results were exported as mgf file to GNPS for FBMN analysis. For FBMN analysis, the precursor ion mass tolerance was set to 0.02 Da and the MS^2 fragment ion tolerance to 0.05 Da. A molecular network was then created where edges were filtered to have a cosine score above 0.7 and more than three matched peaks. The molecular network (Figure S1) was visualized using the Cytoscape software (Shannon et al., 2003) and can be publicly accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1a8c0d5b8c0a42aaaa1a8b5520f5764c. Following the same workflow as described above, a molecular network of the F5 fraction containing rhodoheptins and rhodamides was created (Figure 4) and is available at the following link https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4a962a758d304461a855f935f0295109.

Oil-spreading assay

The oil-spreading assay was performed following the protocol previously reported (Giugliano et al., 2023). In brief, 60 μL of exhaust motor oil was added to 25 mL of distillate water in a Petri dish (Ø 9 cm) to create a thin oil layer. Then, F1-F5 fractions were dissolved in DMSO at 50 mg/mL concentration and 1 μL of each fraction was poured into the centre of the oil layer. The presence of biosurfactants was indicated by oil displacement and formation of a clear zone in the oil layer.

Cell viability assay

Cell viability was assessed by the MTT assay as previously described (Villanova et al., 2022). A375 melanoma cells were cultured in RPMI medium supplemented with 10% fetal bovine serum, penicillin–streptomycin (100 U/mL) and 2 mM L-glutamine, at 37°C under a 5% CO_2 humidified atmosphere. A375 cells were seeded at a cell density of 6000 cells/well in a final volume of 100 μL per well. Approximately 24 h after seeding, medium was removed and cells were treated with medium containing DMSO vehicle, F5 fraction and rhodoheptin mixture at the indicated concentrations for 48 h. Samples were dissolved in DMSO to prepare 200 mg/mL stock solutions and then diluted in RPMI medium for antiproliferative assays. In all experiments, the final concentration of DMSO did not exceed 0.25%, as it was not toxic to A375 cancer cells. After 48 h treatment, 0.5 mg/mL MTT was added to each well. Then, cells were incubated for 3 h at 37°C in a 5% CO_2 humidified atmosphere. After incubation, medium was discarded and formazan salts generated by living cells were dissolved in 100 μL isopropanol at room temperature for 60 min under shaking. To assess cell viability, sample absorbance was measured at 570 nm using an Infinite M1000Pro (TECAN, Männedorf, Switzerland) plate reader.

Data from MTT assay are reported as the mean ± standard deviation (SD) of four independent experiments. Data statistical analysis was performed by using the GraphPad Prism Software version 5. One-way analysis of variance (ANOVA) and Student's *t*-test were used to compare means between groups. Dunnett test was applied as a post hoc test in ANOVA for multiple comparisons with the control group. Differences were considered statistically significant if $p < 0.05$.

Antibacterial assays

Antibacterial activity was evaluated following the broth microdilution method, as previously reported (Giugliano et al., 2023). Briefly, after reaching the log-phase, *Staphylococcus aureus* ATCC 6538, *E. coli* ATCC 10536, and *Pseudomonas aeruginosa* PA01 were diluted in MH (Mueller Hinton) medium and dispensed into a microtitre plate to have a titre of 5×10^4 CFU per well in a final volume of 200 μL. Unlike the other pathogens, *Listeria monocytogenes* MB 677 was diluted in TSB medium supplemented with 0.6% yeast extract. Pathogens were treated with 2% DMSO vehicle (control) and the F5 fraction at different concentrations (1.0–0.500–0.250–0.125–0.062 mg/mL) for 24 h at 37°C. Antibacterial activity was assessed by comparing the optical density (OD) values for the treated strain with the control. OD was measured at 600 nm by using a Tecan plate reader (Tecan, Männedorf, Switzerland).

AUTHOR CONTRIBUTIONS

Costanza Ragozzino: Data curation; methodology; investigation; validation. **Fortunato Palma Esposito:** Investigation; methodology; data curation; writing – review and editing; conceptualization. **Carmine Buonocore:** Investigation; writing – review and editing; visualization. **Pietro Tedesco:** Visualization; writing – review and editing. **Daniela Coppola:** Visualization; writing – review and editing. **Davide Paccagnella:** Investigation; formal analysis; data curation; writing – review and editing. **Nadine Ziemert:** Writing – review and editing; visualization; validation; formal analysis; supervision. **Gerardo Della Sala:** Conceptualization; investigation; methodology; supervision; writing – review and editing; writing – original draft; validation; formal analysis; data curation. **Donatella de Pascale:** Conceptualization; funding acquisition; writing – review and editing; visualization; resources; project administration; supervision.

ACKNOWLEDGEMENTS

This research was funded by H2020-FNR-11-2020: SECRETED—Grant agreement: 101000794.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The molecular networks and mass specrometry data can be publicly accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1a8c0d5b8c0a42aaaa1a8b5520f5764c and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4a962a758d304461a855f935f0295109. The rhodoheptin and rhodamide gene clusters were deposited in GenBank under accession numbers PP729154 and PP729155, respectively.

ORCID

Carmine Buonocore https://orcid.org/0000-0002-8327-954X

Gerardo Della Sala https://orcid.org/0000-0001-8957-3259

REFERENCES

Blin, K., Shaw, S., Augustijn, H.E., Reitz, Z.L., Biermann, F., Alanjary, M. et al. (2023) antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research*, 51(W1), W46–W50. Available from: https://doi.org/10.1093/nar/gkad344

Bosi, E., Donati, B., Galardini, M., Brunetti, S., Sagot, M.F., Lió, P. et al. (2015) MeDuSa: a multi-draft based scaffolder. *Bioinformatics (Oxford, England)*, 31, 2443–2451. Available from: https://doi.org/10.1093/bioinformatics/btv171

Buonocore, C., Giugliano, R., Della Sala, G., Palma Esposito, F., Tedesco, P., Folliero, V. et al. (2023) Evaluation of antimicrobial properties and potential applications of *pseudomonas gessardii* M15 Rhamnolipids towards multiresistant *Staphylococcus aureus*. *Pharmaceutics*, 15(2), 700. Available from: https://doi.org/10.3390/pharmaceutics15020700

Buonocore, C., Tedesco, P., Vitale, G.A., Esposito, F.P., Giugliano, R., Monti, M.C. et al. (2020) Characterization of a new mixture of mono-Rhamnolipids produced by *pseudomonas gessardii* isolated from Edmonson point (Antarctica). *Marine Drugs*, 18(5), 269. Available from: https://doi.org/10.3390/md18050269

Cantu, D.C., Chen, Y. & Reilly, P.J. (2010) Thioesterases: a new perspective based on their primary and tertiary structures. *Protein Science: A Publication of the Protein Society*, 19(7), 1281–1295. Available from: https://doi.org/10.1002/pro.417

Cao, X.H., Wang, A.H., Wang, C.L., Mao, D.Z., Lu, M.F., Cui, Y.Q. et al. (2010) Surfactin induces apoptosis in human breast cancer MCF-7 cells through a ROS/JNK-mediated mitochondrial/caspase pathway. *Chemico-Biological Interactions*, 183(3), 357–362. Available from: https://doi.org/10.1016/j.cbi.2009.11.027

Carolin, C.F., Kumar, P.S. & Ngueagni, P.T. (2021) A review on new aspects of lipopeptide biosurfactant: types, production, properties and its application in the bioremediation process. *Journal of Hazardous Materials*, 407, 124827. Available from: https://doi.org/10.1016/j.jhazmat.2020.124827

Ceresa, C., Fracchia, L., Sansotera, A.C., De Rienzo, M.A.D. & Banat, I.M. (2023) Harnessing the potential of biosurfactants for biomedical and pharmaceutical applications. *Pharmaceutics*, 15(8), 2156. Available from: https://doi.org/10.3390/pharmaceutics15082156

Chen, X., Lu, Y., Shan, M., Zhao, H., Lu, Z. & Lu, Y. (2022) A mini-review: mechanism of antimicrobial action and application of surfactin. *World Journal of Microbiology and Biotechnology*, 38(8), 143. Available from: https://doi.org/10.1007/s11274-022-03323-3

Chiba, H., Agematu, H., Sakai, K., Dobashi, K. & Yoshioka, T. (1999) Rhodopeptins, novel cyclic tetrapeptides with antifungal activities from Rhodococcus sp. III. Synthetic study of rhodopeptins. *The Journal of Antibiotics*, 52(8), 710–720. Available from: https://doi.org/10.7164/antibiotics.52.710

Chooi, Y.H. & Tang, Y. (2010) Adding the lipo to lipopeptides: do more with less. *Chemistry and Biology*, 17(8), 791–793. Available from: https://doi.org/10.1016/j.chembiol.2010.08.001

Cock, I.E. & Cheesman, M.J. (2023) A review of the antimicrobial properties of cyanobacterial natural products. *Molecules (Basel, Switzerland)*, 28(20), 7127. Available from: https://doi.org/10.3390/molecules28207127

de Oliveira Schmidt, V.K., de Souza Carvalho, J., de Oliveira, D. & de Andrade, C.J. (2021) Biosurfactant inducers for enhanced production of surfactin and rhamnolipids: an overview. *World Journal of Microbiology and Biotechnology*, 37(2), 21. Available from: https://doi.org/10.1007/s11274-020-02970-8

Della Sala, G., Mangoni, A., Costantino, V. & Teta, R. (2020) Identification of the biosynthetic gene cluster of Thermoactinoamides and discovery of new congeners by integrated genome mining and MS-based molecular networking. *Frontiers in Chemistry*, 8, 397. Available from: https://doi.org/10.3389/fchem.2020.00397

Dey, G., Bharti, R., Dhanarajan, G., Das, S., Dey, K.K., Kumar, B.N. et al. (2015) Marine lipopeptide Iturin a inhibits Akt mediated GSK3β and FoxO3a signaling and triggers apoptosis in breast cancer. *Scientific Reports*, 5, 10316. Available from: https://doi.org/10.1038/srep10316

Dias, M.A.M. & Nitschke, M. (2023) Bacterial-derived surfactants: an update on general aspects and forthcoming applications. *Brazilian Journal of Microbiology*, 54(1), 103–123. Available from: https://doi.org/10.1007/s42770-023-00905-7

Du, L., Sánchez, C., Chen, M., Edwards, D.J. & Shen, B. (2000) The biosynthetic gene cluster for the antitumor drug bleomycin from Streptomyces verticillus ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. *Chemistry and Biology*, 7(8), 623–642. Available from: https://doi.org/10.1016/s1074-5521(00)00011-9

Esposito, F.P., Vecchiato, V., Buonocore, C., Tedesco, P., Noble, B., Basnett, P. et al. (2023) Enhanced production of biobased, biodegradable, poly(3-hydroxybutyrate) using an unexplored marine bacterium Pseudohalocynthiibacter aestuariivivens, isolated from highly polluted coastal environment. *Bioresource Technology*, 368, 128287. Available from: https://doi.org/10.1016/j.biortech.2022.128287

Felnagle, E.A., Rondon, M.R., Berti, A.D., Crosby, H.A. & Thomas, M.G. (2007) Identification of the biosynthetic gene cluster and an additional gene for resistance to the antituberculosis drug capreomycin. *Applied and Environmental Microbiology*, 73(13), 4162–4170. Available from: https://doi.org/10.1128/AEM.00485-07

Frankfater, C., Henson, W.R., Juenger-Leif, A., Foston, M., Moon, T.S., Turk, J. et al. (2020) Structural determination of a new Peptidolipid family from *Rhodococcus opacus* and the pathogen *Rhodococcus equi* by multiple stage mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 31(3), 611–623. Available from: https://doi.org/10.1021/jasms.9b00059

Gilchrist, C.L.M., Booth, T.J., van Wersch, B., van Grieken, L., Medema, M.H. & Chooi, Y.H. (2021) Cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. *Bioinformatics Advances*, 1, vbab016. Available from: https://doi.org/10.1093/bioadv/vbab016

Giugliano, R., Della Sala, G., Buonocore, C., Zannella, C., Tedesco, P., Palma Esposito, F. et al. (2023) New Imidazolium alkaloids with broad Spectrum of action from the marine bacterium *Shewanella aquimarina*. *Pharmaceutics*, 15(8), 2139. Available from: https://doi.org/10.3390/pharmaceutics15082139

Habib, S., Ahmad, S.A., Wan Johari, W.L., Abd Shukor, M.Y., Alias, S.A., Smykla, J. et al. (2020) Production of Lipopeptide biosurfactant by a hydrocarbon-degrading Antarctic *Rhodococcus*. *International Journal of Molecular Sciences*, 21(17), 6138. Available from: https://doi.org/10.3390/ijms21176138

Hashemi, S.Z., Fooladi, J., Vahidinasab, M., Hubel, P., Pfannstiel, J., Pillai, E. et al. (2024) Toward effects of hydrophobicity on biosurfactant production by *Bacillus subtilis* isolates from crude-oil-exposed environments. *Applied Microbiology*, 4, 215–236. Available from: https://doi.org/10.3390/applmicrobiol4010015

Hoertz, A.J., Hamburger, J.B., Gooden, D.M., Bednar, M.M. & McCafferty, D.G. (2012) Studies on the biosynthesis of the lipodepsipeptide antibiotic Ramoplanin A2. *Bioorganic and Medicinal Chemistry*, 20(2), 859–865. Available from: https://doi.org/10.1016/j.bmc.2011.11.062

Hojati, Z., Milne, C., Harvey, B., Gordon, L., Borg, M., Flett, F. et al. (2002) Structure, biosynthetic origin, and

engineered biosynthesis of calcium-dependent antibiotics from Streptomyces coelicolor. *Chemistry and Biology*, 9(11), 1175–1187. Available from: https://doi.org/10.1016/s1074-5521(02)00252-1

Huang, H.W. (2020) DAPTOMYCIN, its membrane-active mechanism vs. that of other antimicrobial peptides. *Biochimica et Biophysica Acta, Biomembranes*, 1862(10), 183395. Available from: https://doi.org/10.1016/j.bbamem.2020.183395

Jahan, R., Bodratti, A.M., Tsianou, M. & Alexandridis, P. (2020) Biosurfactants, natural alternatives to synthetic surfactants: physicochemical properties and applications. *Advances in Colloid and Interface Science*, 275, 102061. Available from: https://doi.org/10.1016/j.cis.2019.102061

Janek, T., Krasowska, A., Radwańska, A. & Łukaszewicz, M. (2013) Lipopeptide biosurfactant pseudofactin II induced apoptosis of melanoma a 375 cells by specific interaction with the plasma membrane. *PLoS One*, 8(3), e57991. Available from: https://doi.org/10.1371/journal.pone.0057991

Koronelli, T.V. (1988) Investigation of the lipids of saprophytic mycobacteria in the U.S.S.R. *Journal of Chromatography*, 440, 479–486. Available from: https://doi.org/10.1016/s0021-9673(00)94553-9

Kraas, F.I., Helmetag, V., Wittmann, M., Strieker, M. & Marahiel, M.A. (2010) Functional dissection of surfactin synthetase initiation module reveals insights into the mechanism of lipoinitiation. *Chemistry and Biology*, 17(8), 872–880. Available from: https://doi.org/10.1016/j.chembiol.2010.06.015

Li, J.Y., Wang, L., Liu, Y.F., Zhou, L., Gang, H.Z., Liu, J.F. et al. (2021) Microbial Lipopeptide-producing strains and their metabolic roles under anaerobic conditions. *Microorganisms*, 9(10), 2030. Available from: https://doi.org/10.3390/microorganisms9102030

Liebisch, G., Fahy, E., Aoki, J., Dennis, E.A., Durand, T., Ejsing, C.S. et al. (2020) Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures. *Journal of Lipid Research*, 61(12), 1539–1555. Available from: https://doi.org/10.1194/jlr.S120001025

Magarvey, N.A., Haltli, B., He, M., Greenstein, M. & Hucul, J.A. (2006) Biosynthetic pathway for mannopeptimycins, lipoglycopeptide antibiotics active against drug-resistant gram-positive pathogens. *Antimicrobial Agents and Chemotherapy*, 50(6), 2167–2177. Available from: https://doi.org/10.1128/AAC.01545-05

Mareš, J., Hájek, J., Urajová, P., Kopecký, J. & Hrouzek, P. (2014) A hybrid non-ribosomal peptide/polyketide synthetase containing fatty-acyl ligase (FAAL) synthesizes the β-amino fatty acid lipopeptides puwainaphycins in the cyanobacterium Cylindrospermum alatosporum. *PLoS One*, 9(11), e111904. Available from: https://doi.org/10.1371/journal.pone.0111904

Nagtode, V.S., Cardoza, C., Yasin, H.K.A., Mali, S.N., Tambe, S.M., Roy, P. et al. (2023) Green surfactants (biosurfactants): a petroleum-free substitute for sustainability-comparison, applications, market, and future prospects. *ACS Omega*, 8(13), 11674–11699. Available from: https://doi.org/10.1021/acsomega.3c00591

Nothias, L.F., Petras, D., Schmid, R., Dührkop, K., Rainer, J., Sarvepalli, A. et al. (2020) Feature-based molecular networking in the GNPS analysis environment. *Nature Methods*, 17(9), 905–908. Available from: https://doi.org/10.1038/s41592-020-0933-6

Palma Esposito, F., Giugliano, R., Della Sala, G., Vitale, G.A., Buonocore, C., Ausuri, J. et al. (2021) Combining OSMAC approach and untargeted metabolomics for the identification of new glycolipids with potent antiviral activity produced by a marine *Rhodococcus*. *International Journal of Molecular Sciences*, 22(16), 9055. Available from: https://doi.org/10.3390/ijms22169055

Pan, R., Bai, X., Chen, J., Zhang, H. & Wang, H. (2019) Exploring structural diversity of microbe secondary metabolites using OSMAC strategy: a literature review. *Frontiers in Microbiology*, 10, 294. Available from: https://doi.org/10.3389/fmicb.2019.00294

Peng, F., Wang, Y., Sun, F., Liu, Z., Lai, Q. & Shao, Z. (2008) A novel lipopeptide produced by a Pacific Ocean deep-sea bacterium, Rhodococcus sp. TW53. *Journal of Applied Microbiology*, 105(3), 698–705. Available from: https://doi.org/10.1111/j.1365-2672.2008.03816.x

Pilz, M., Cavelius, P., Qoura, F., Awad, D. & Brück, T. (2023) Lipopeptides development in cosmetics and pharmaceutical applications: a comprehensive review. *Biotechnology Advances*, 67, 108210. Available from: https://doi.org/10.1016/j.biotechadv.2023.108210

Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11, 395. Available from: https://doi.org/10.1186/1471-2105-11-395

Qin, W., Liu, Y., Ren, P., Zhang, J., Li, H., Tian, L. et al. (2014) Uncovering a glycosyltransferase provides insights into the glycosylation step during macrolactin and bacillaene biosynthesis. *Chembiochem: A European Journal of Chemical Biology*, 15(18), 2747–2753. Available from: https://doi.org/10.1002/cbic.201402384

Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D.H. (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evolutionary Biology*, 7, 78. Available from: https://doi.org/10.1186/1471-2148-7-78

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Research*, 33(18), 5799–5808. Available from: https://doi.org/10.1093/nar/gki885

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504. Available from: https://doi.org/10.1101/gr.1239303

Solyanikova, I.P. & Golovleva, L.A. (2019) Hexadecane and hexadecane-degrading bacteria: mechanisms of interaction. *Microbiology*, 88, 15–26. Available from: https://doi.org/10.1134/S0026261718060152

Subramaniam, M.D., Venkatesan, D., Iyer, M., Subbarayan, S., Govindasami, V., Roy, A. et al. (2020) Biosurfactants and anti-inflammatory activity: a potential new approach towards COVID-19. *Current Opinion in Environmental Science & Health*, 17, 72–81. Available from: https://doi.org/10.1016/j.coesh.2020.09.002

Terlouw, B.R., Blin, K., Navarro-Muñoz, J.C., Avalon, N.E., Chevrette, M.G., Egbert, S. et al. (2023) MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research*, 51(D1), D603–D610. Available from: https://doi.org/10.1093/nar/gkac1049

Teta, R., Della Sala, G., Mangoni, A., Lega, M. & Costantino, V. (2016) Tracing cyanobacterial blooms to assess the impact of wastewaters discharges on coastal areas and lakes. *International Journal of Sustainable Development and Planning*, 11, 804–811. Available from: https://doi.org/10.2495/SDP-V11-N5-804-811

Teta, R., Sala, G.D., Esposito, G., Stornaiuolo, M., Scarpato, S., Casazza, M. et al. (2021) Monitoring cyanobacterial blooms during the COVID-19 pandemic in Campania, Italy: the case of Lake Avernus. *Toxins*, 13(7), 471. Available from: https://doi.org/10.3390/toxins13070471

Thomas, M.G., Chan, Y.A. & Ozanick, S.G. (2003) Deciphering tuberactinomycin biosynthesis: isolation, sequencing, and annotation of the viomycin biosynthetic gene cluster. *Antimicrobial Agents and Chemotherapy*, 47(9), 2823–2830. Available from: https://doi.org/10.1128/AAC.47.9.2823-2830.2003

Villanova, V., Galasso, C., Vitale, G.A., Della Sala, G., Engelbrektsson, J., Strömberg, N. et al. (2022) Mixotrophy in a local strain of Nannochloropsis granulata for renewable high-value biomass production on the west coast of Sweden. *Marine Drugs*, 20(7), 424. Available from: https://doi.org/10.3390/md20070424

Wang, X., An, J., Cao, T., Guo, M. & Han, F. (2024) Application of biosurfactants in medical sciences. *Molecules (Basel, Switzerland)*, 29(11), 2606. Available from: https://doi.org/10.3390/molecules29112606

Wiebach, V., Mainz, A., Siegert, M.J., Jungmann, N.A., Lesquame, G., Tirat, S. et al. (2018) The anti-staphylococcal lipolanthines are ribosomally synthesized lipopeptides. *Nature Chemical Biology*, 14(7), 652–654. Available from: https://doi.org/10.1038/s41589-018-0068-6

Wittmann, M., Linne, U., Pohlmann, V. & Marahiel, M.A. (2008) Role of DptE and DptF in the lipidation reaction of daptomycin. *The FEBS Journal*, 275(21), 5343–5354. Available from: https://doi.org/10.1111/j.1742-4658.2008.06664.x

Zhang, J.J., Tang, X., Huan, T., Ross, A.C. & Moore, B.S. (2020) Pass-back chain extension expands multimodular assembly line biosynthesis. *Nature Chemical Biology*, 16(1), 42–49. Available from: https://doi.org/10.1038/s41589-019-0385-4

Zhang, S., Chen, Y., Zhu, J., Lu, Q., Cryle, M.J., Zhang, Y. et al. (2023) Structural diversity, biosynthesis, and biological functions of lipopeptides from *Streptomyces*. *Natural Product Reports*, 40(3), 557–594. Available from: https://doi.org/10.1039/d2np00044j

Zhong, L., Diao, X., Zhang, N., Li, F., Zhou, H., Chen, H. et al. (2021) Engineering and elucidation of the lipoinitiation process in nonribosomal peptide biosynthesis. *Nature Communications*, 12(1), 296. Available from: https://doi.org/10.1038/s41467-020-20548-8

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ragozzino, C., Palma Esposito, F., Buonocore, C., Tedesco, P., Coppola, D., Paccagnella, D. et al. (2024) Integrated genome and metabolome mining unveiled structure and biosynthesis of novel lipopeptides from a deep-sea *Rhodococcus*. *Microbial Biotechnology*, 17, e70011. Available from: https://doi.org/10.1111/1751-7915.70011

MANUSCRIPT 3

Title: PanBGC: A pangenome-inspired framework for comparative analysis of biosynthetic gene clusters.

Authors: Davide Paccagnella, Caner Bagci, Athina Gavrillidou, Nadine Ziemert

DECLARATION ON THE CONTRIBUTION OF CO-AUTHORS TO THE MANUSCRIPT

Author	Author position	Scientific ideas %	Data generation %	Analysis & interpretation %	Paper writing %
D. Paccagnella	First Author	60	70	60	70
C. Bagci	Co-Author	5	20	10	5
A. Gavrillidou	Co-Author	5	10	10	5
N. Ziemert	Corresponding author	30	-	20	20
Title of paper	PanBGC: A pangenome-inspired framework for comparative analysis of biosynthetic gene clusters.				
Status in publication process	Preprint on biorxiv.org/content/10.1101/2025.08.11.669102v1 Submitted and currently under revision in ISME communication (08/09/2025)				

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].

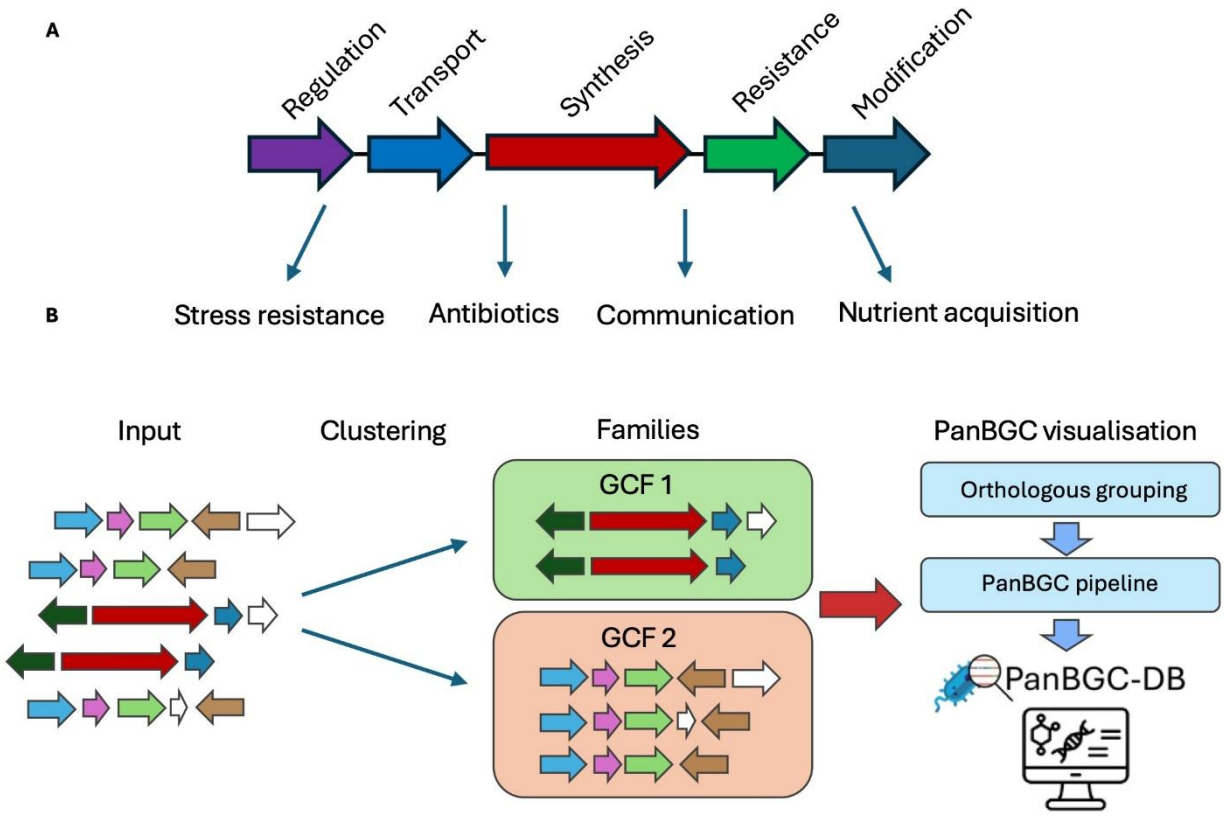


Figure 1: Overview of biosynthetic gene cluster (BGC) function and PanBGC-DB workflow. **a** Schematic representation of a typical BGC, highlighting the modular architecture with genes performing distinct biosynthetic functions. Each coloured arrow represents one gene in the BGC. The specialized metabolites 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.

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.

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)

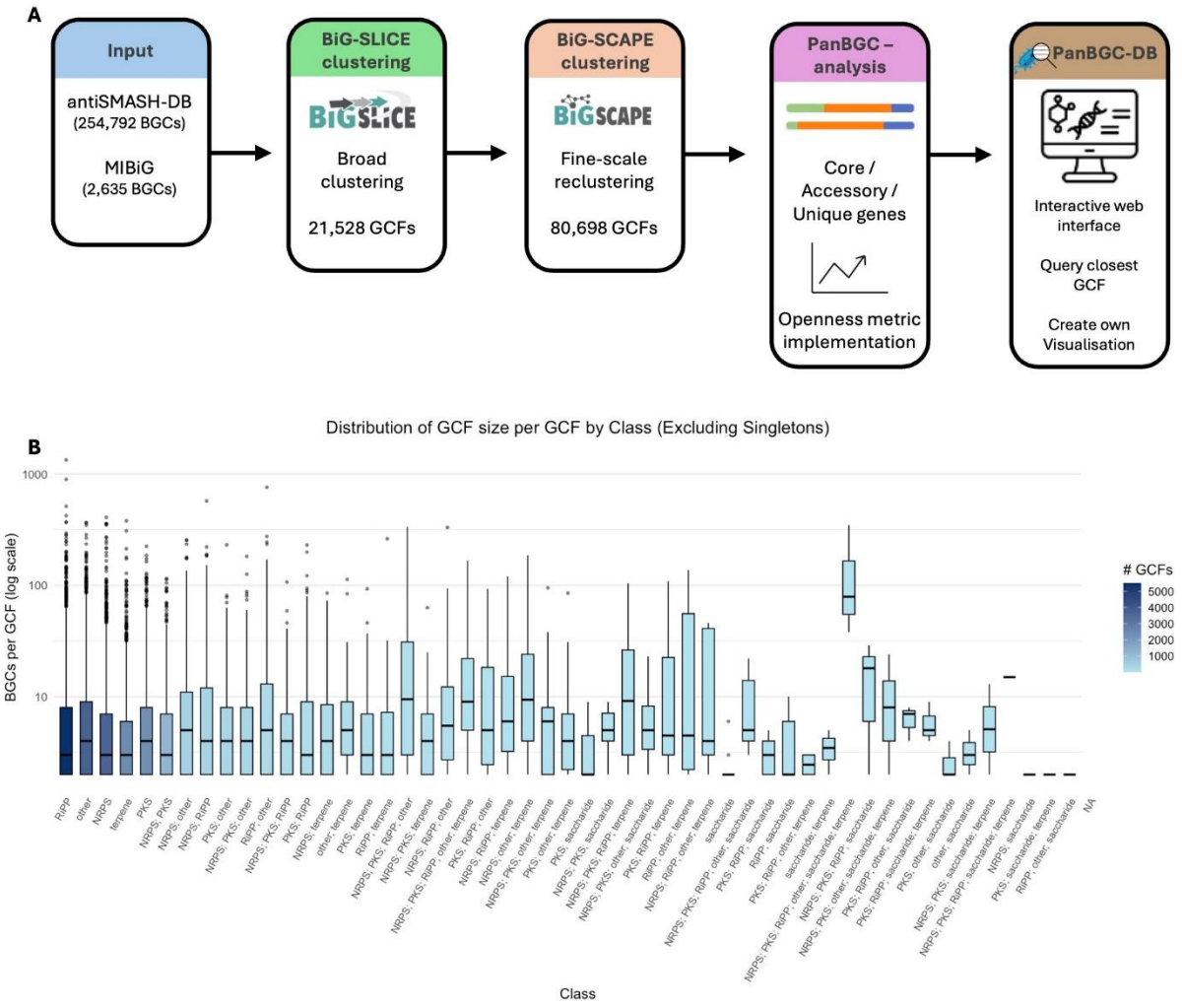


Figure 2: PanBGC-DB pipeline overview and BGC family size distribution.
a BGCs from antiSMASH-DB and MIBiG are first clustered using BiG-SLiCE. Broad GCFs are further refined using BiG-SCAPE. The PanBGC analysis framework then identifies core, accessory, and unique genes within each family and implements 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

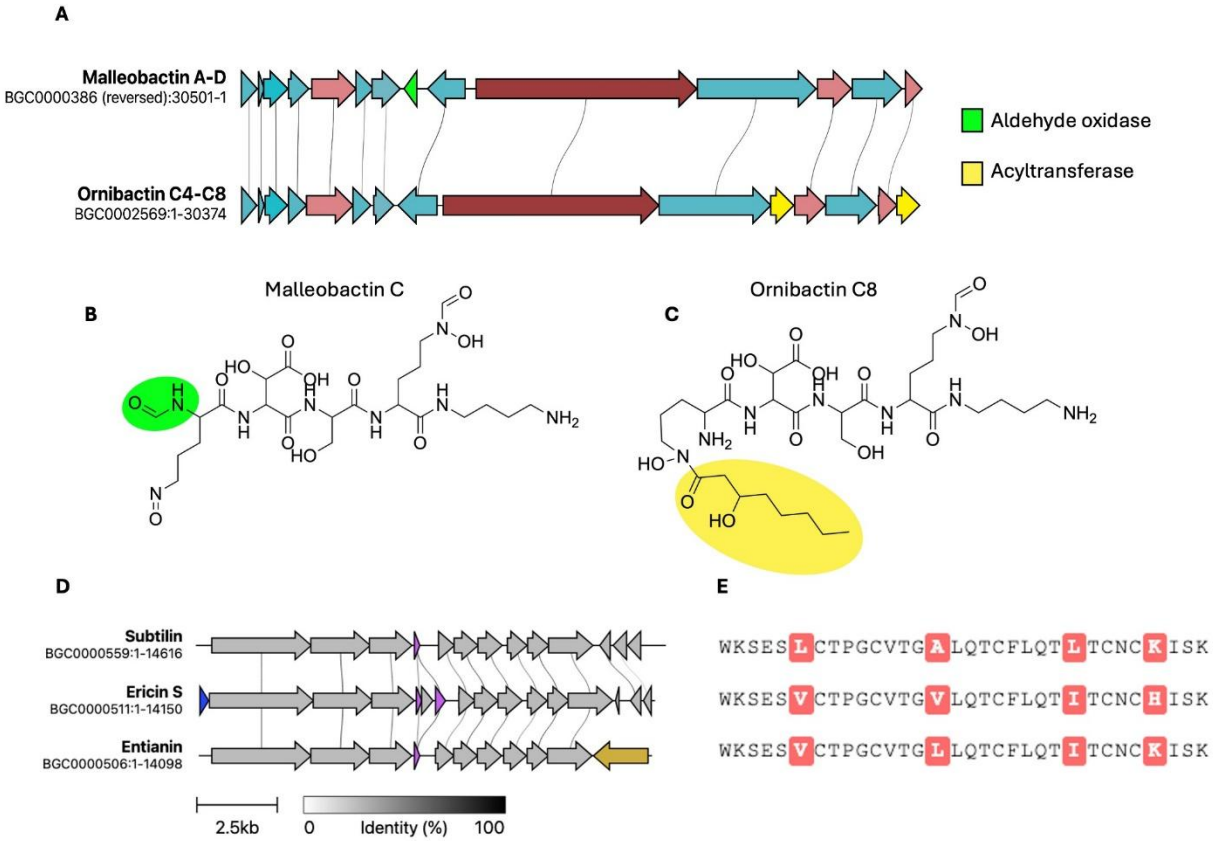


Figure 3: Gene cluster variation and structural diversification in related siderophores.
a Comparative gene cluster alignment of the Malleobactin A–D and Ornibactin C4–C8 BGCs from the MIBiG database. Homologous genes are connected by grey lines. While both clusters share a conserved core biosynthetic 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.

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.

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.

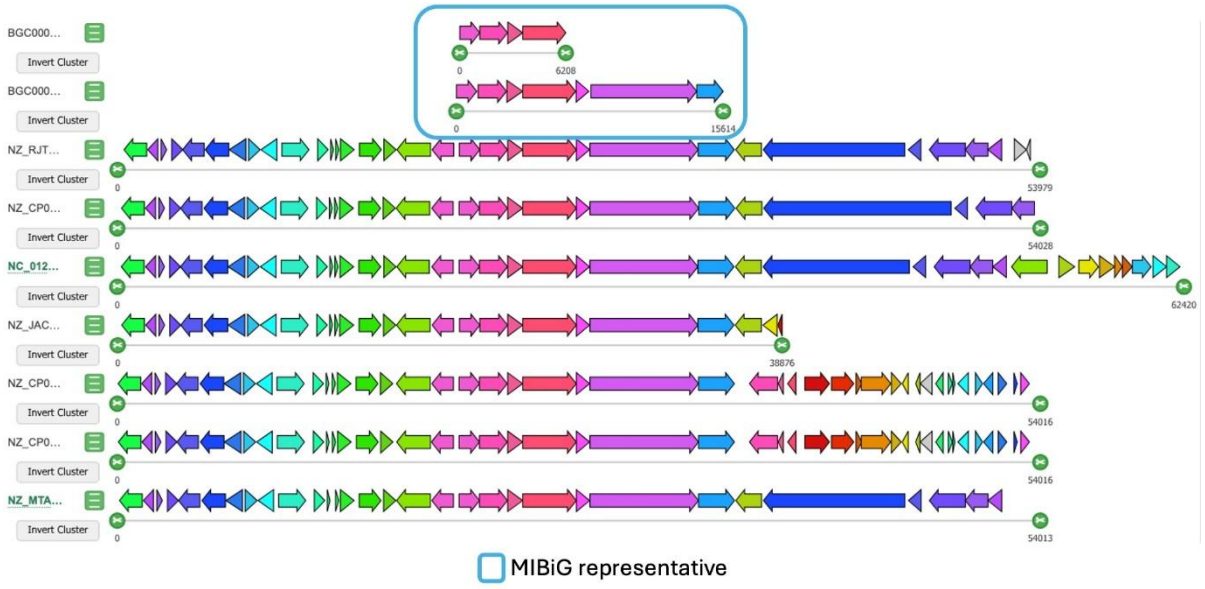


Figure 4: Gene cluster diversity within the Enterobactin and amonabactin P 750 family.
Comparative visualization of BGCs belonging to the Enterobactin gene cluster families. Each row represents a single BGC, with genes colored by orthologous group and arrows indicating gene orientation. MIBiG reference 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.

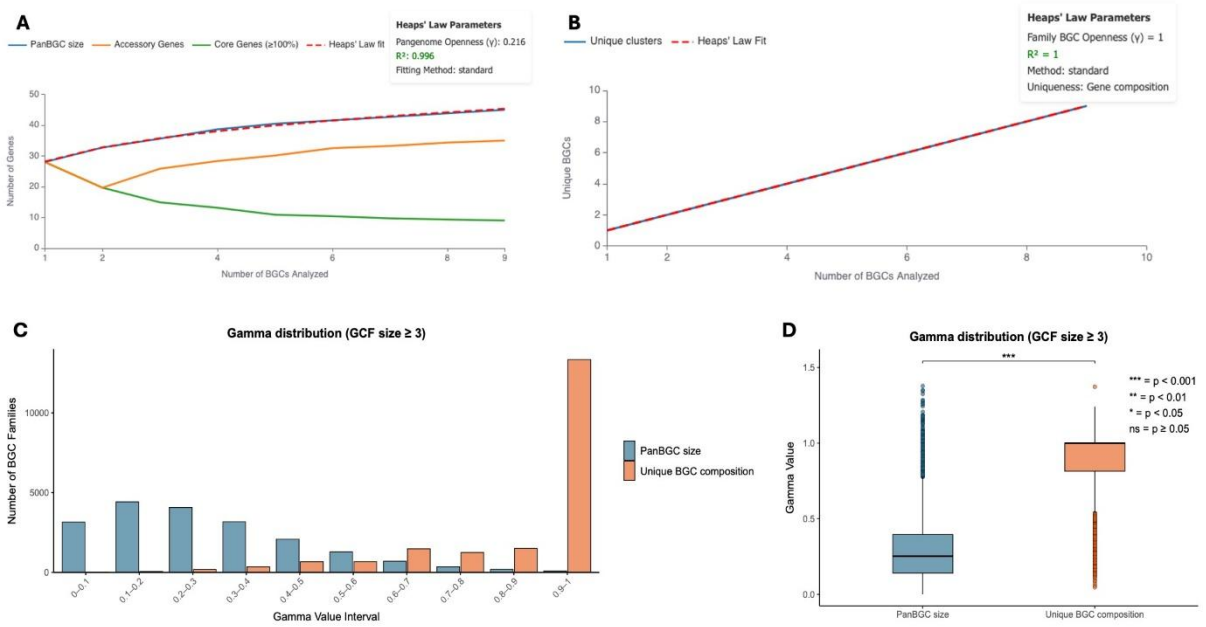


Figure 5: Openness metrics adapted for biosynthetic gene clusters.
a Example Heaps' Law curve showing the accumulation of total (blue curve), core (green curve), and accessory genes (orange curve) as additional BGCs are sampled from a representative GCF. The red dotted line represents the fitted curve. **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 (γ = 0.195) reflecting limited novel gene acquisition, yet a relative high composition-based openness (γ = 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

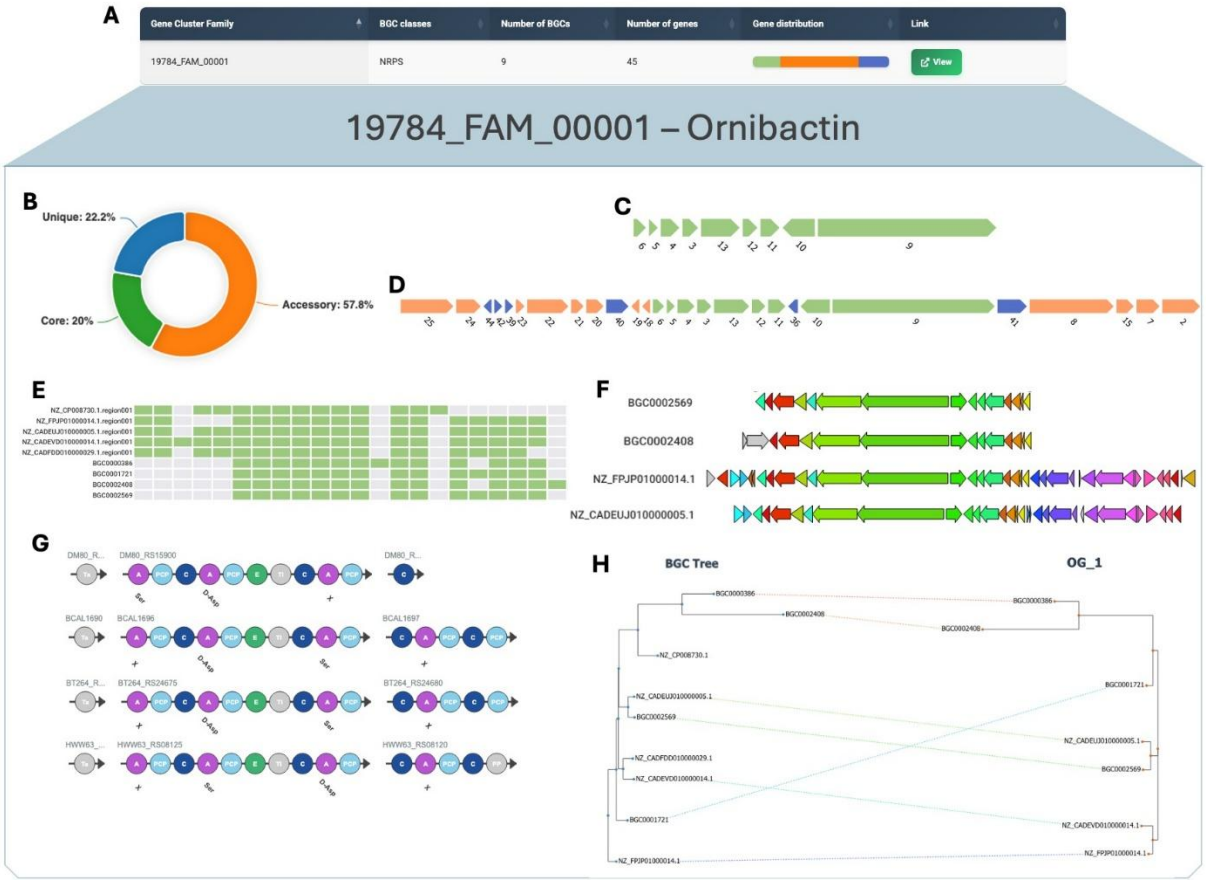


Figure 6: Multiple screenshots from the PanBGC-DB web interface for the Ornibactin and Ornibactin GCF
a Overview table showing metadata of gene cluster families, including class, number of BGCs, total genes and summary of gene distribution (core / accessory / unique). **b** Donut plot illustrating the 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).

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.

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.

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.

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.

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1073/PNAS.2007873118/SUPPL_FILE/PNAS.2007873118.SD05.XLSX

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1073/PNAS.0506758102

12. Medini D et al. The microbial pan-genome. *Curr Opin Genet Dev* 2005;15:589–594. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1038/srep24373

16. Tettelin H et al. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008;11:472–477. https://doi.org/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. https://doi.org/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. https://doi.org/10.3389/FPLS.2025.1539151/BIBTEX

19. Brockhurst MA et al. The Ecology and Evolution of Pangenomes. *Current Biology* 2019;29:R1094–R1103. https://doi.org/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. https://doi.org/10.1128/MSYSTEMS.00281-18/SUPPL_FILE/SYS001192310ST4.XLSX

21. Rouli L et al. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect* 2015;7:72. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/J.FM.2023.104334

37. Zhang C, Mirarab S. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. https://doi.org/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. https://doi.org/10.1093/BIOADV/VBAB016

39. Kautsar SA et al. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res* 2021;49:D490–D497. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1093/NAR/GKAE997

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

44. Nivina A et al. Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem Rev* 2019;119:12524–12547. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1093/BIOINFORMATICS/BTAB007

57. Page AJ et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1002/CHEM.201500757

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

THESIS – DISCUSSION

In this PhD project, the goal was to create a systematic framework that enables insights into diversity occurring in gene cluster families. To achieve this goal, we developed an approach that treats BGC families the same way as pangenomics treats species populations. Just as pangenomic studies group related genomes (typically from the same species) to analyze their collective gene content, we organized biosynthetic gene clusters into gene cluster families based on their structural and functional similarities. In each GCF we used ZOL to form orthologous groups of genes from different BGCs from the same family based on sequence similarity, positional conservation and function. This enabled the classification of orthologous gene groups into core (present in all BGCs), accessory (found in some but not all clusters), and unique (exclusive to single clusters) categories. Such classification provides insights into the conserved biosynthetic machinery that defines each family versus the variable genetic elements that may contribute to functional specialization and subtle chemical diversification. Beyond gene classification, we adapted pangenomic openness metrics to quantify the evolutionary dynamics within BGC families. By applying Heaps' law modelling to BGC families, we could assess whether families operate with relatively stable gene sets (closed) or continue to incorporate new genetic material as more family members are sampled (open). This approach enables systematic analysis of how biosynthetic innovation occurs within families and whether diversity arises primarily through gene acquisition or through modular rearrangement of existing components.

Early studies in which I collaborated provided critical observations that inspired the development of the PanBGC framework. In the caprazamycin biosynthesis investigation (manuscript 1)[123], my analysis of the *liu* cluster distribution revealed that 52% of *Streptomyces* strains harbor the complete cluster, and 48% of the *Streptomyces* strains contain also a second acyl-CoA dehydrogenase, thus raising questions about functional and structural variation of these widespread gene families. Similarly, in the *Rhodococcus* lipopeptide study (manuscript 2)[124], a comparative genomics analysis identified widespread presence of the rhodoheptin and rhodamide pathways distributed among 322 and 478 bacterial species respectively, again highlighting the need for a systematic approach to examine diversity within such widely distributed biosynthetic families. Both studies highlighted comparative genomics as a useful tool in guiding the discovery of natural products while at the same time revealing a gap in analysis: available tools and methods can detect and classify BGC and group them into related cluster families but lack frameworks for analyzing internal diversities within these cluster families.

To overcome this limitation, the main focus of this thesis was the development and application of the PanBGC framework (manuscript 3)[125]. This approach treats gene cluster families as closely related populations analogous to how pangenomics analyzes microbial species. When used to inspect over 80.000 gene cluster families derived from over 250,000 BGCs revealed

fundamental patterns of genetic diversity within BGC families. The most significant finding was the difference between gene repertoire stability and compositional plasticity. While most GCFs exhibit closed pangenomes with limited acquisition of entirely novel genes (gene pool saturated), they, on the other hand, undergo high compositional gene variation. This pattern indicates that secondary metabolite diversity is generated primarily through modular reorganization of existing genetic components rather than through incorporation of novel genetic material.

The framework enables a quantitative analysis of biosynthetic diversity in all major secondary metabolite classes, revealing patterns in core versus accessory gene distributions and helps to correlate genetic composition with chemical structures. To make this analysis accessible, the PanBGC-DB was developed as an interactive platform containing precomputed analyses of over 80,000 gene cluster families, complete with visualization, comparative analysis modules, and integration capabilities for custom datasets. The platform offers access to population-level BGC analysis, enabling even researchers with minimal bioinformatics knowledge to conduct GCF investigations.

This work outlines a population-scale framework designed for systematically quantifying and interpreting diversity patterns within biosynthetic gene clusters. It offers a new concept in which biosynthetic gene clusters can be analyzed and also helps fields like genome mining, synthetic biology, and evolutionary studies with direct access to implementable use cases. The results of each manuscript are addressed in detail in their respective discussions. The following discussion aims to give a broader implication of the PanBGC framework and to address its impact on contemporary natural products research.

THE FOUNDATION: PANGENOMICS AS A TRANSFORMATIVE FRAMEWORK IN MICROBIAL RESEARCH

As outlined in the introduction, the pangenomic approach has proven invaluable across diverse areas of microbial research by providing systematic frameworks for analyzing genetic diversity within related populations [109, 110, 126–129]. The conceptual shift from single-genome analysis to population-level thinking has fundamentally changed how we understand microbial adaptation, evolution, and functional potential. This transformation provides strong justification for applying similar principles to biosynthetic gene clusters.

In pathogen research, pangenomic studies have revolutionized our understanding of virulence and antimicrobial resistance mechanisms. Comprehensive analyses of major bacterial pathogens such as *Escherichia coli*, *Staphylococcus aureus*, and *Mycobacterium tuberculosis* revealed that pathogenicity often arises from accessory gene pools rather than core genomic variations[116, 128, 130]. For instance, pangenomic analysis of *Escherichia coli* identified an immense reservoir of accessory genes, with individual strains representing only a fraction of the total genetic diversity. This observation explained the high phenotypic differences between commensal and pathogenic *E. coli* isolates, demonstrating that virulence resulted from acquisition of specific accessory genetic elements rather than fundamental alterations to the

core cellular machinery[119]. Similarly, in *Neisseria meningitidis*, pangenomic studies revealed the role of accessory genes in antigenic variation and immune evasion. The accessory genome encompasses genes responsible for capsule biosynthesis switching and phase-variable surface proteins, allowing the bacterium to evade host immune detection[131–133]. These insights have proven instrumental in identifying conserved antigens and developing broadly protective vaccine targets[134].

The quantitative frameworks developed for pangenomic analysis, particularly the use of Heaps' law to characterize pangenome openness, have provided powerful tools for understanding diversity dynamics[135]. As described in the introduction, these metrics enabled researchers to distinguish between populations with high rates of horizontal gene transfer (open pangenomes) versus those with stable genetic arrangements (closed pangenomes) [136, 137]. These metrics have enabled researchers to quantify diversity strategies and predict the genetic potential of microbial populations. The discoveries made with pangenomic approaches in environmental microbiology have demonstrated how microbial populations maintain functional coherence while adapting to diverse ecological conditions[138]. Analysis of marine *Prochlorococcus* populations showed maintenance of core metabolic functions across oceanic habitats, with accessory genes enabling adaptation to local variations in nutrient availability and environmental stress[122]. These studies established that pangenomic organization allows microbial populations to maintain essential functions while retaining the genetic flexibility necessary for environmental adaptation.

METHODOLOGICAL INNOVATION: ADAPTING PANGENOMICS FOR BGC ANALYSIS

The success of pangenomic approaches in microbial genomics provided strong theoretical justification for their application to BGC analysis. Several key principles from microbial pangenomics directly supported the feasibility of this method for BGC population analysis. First, the modular organization of both bacterial genomes and biosynthetic gene clusters suggested that similar analytical frameworks might reveal comparable diversity patterns. Just as bacterial genomes can be broken down into functional modules that respond independently to selection[139, 140], BGC families might contain core biosynthetic modules alongside variable elements that enable functional diversification. Second, the horizontal mobility of BGCs, well-documented in the literature[46, 141, 142], paralleled the horizontal gene transfer processes that drive pangenomic diversity[143]. The compact, self-contained nature of many BGCs makes them particularly suitable for horizontal transfer, potentially leading to diversity dynamics similar to those observed in bacterial accessory genomes. Third, the functional constraints operating on BGCs resembled those affecting bacterial core genomes. Essential biosynthetic functions, like essential cellular functions, would be expected to show strong conservation across family members, while auxiliary functions that enhance biosynthetic efficiency or enable adaptation to specific chemical environments might show patterns of variability similar to bacterial accessory genes.

TECHNICAL ADAPTATIONS FOR BGC-SPECIFIC ANALYSIS

The adaptation of pangenomic principles to BGC analysis required fundamental reconceptualization of what constitutes a population unit. In traditional pangenomics, populations are defined as collections of genomes from the same species or closely related strains, typically identified through genomic similarity thresholds such as average nucleotide identity (ANI) values above 95%[109, 117]. The PanBGC framework instead treats gene cluster families as population units, where each BGC represents an individual member of the population analogous to how individual genomes represent members of a species population.

This conceptual shift required developing new approaches for population definition. While microbial pangenomics relies on established taxonomic frameworks and genomic similarity metrics, BGC families needed to be defined based on biosynthetic function and architectural similarity. Traditional pangenomic studies typically use single-step clustering approaches based on overall genomic similarity[112]. In contrast, the PanBGC framework implemented a two-step clustering strategy that first groups clusters based on broad global similarity using BiG-SLiCE[144], then refines these groupings using BiG-SCAPE[145] to ensure that family members share specific architectural features. This approach addresses the unique challenge that BGCs can share overall function while differing significantly in specific enzymatic domains or gene organization.

Traditional pangenomic analysis employs ortholog detection methods designed for complete genomes, typically using approaches like OrthoMCL[146], OrthoFinder[147], or bidirectional best hits that focus primarily on sequence similarity. These methods are optimized for large gene sets spanning thousands of genes per genome and can afford to ignore gene order information when the primary goal is identifying homologous relationships across diverse functional categories. The PanBGC framework required a more stringent approach due to the smaller scale of gene clusters and the functional importance of every gene in biosynthetic contexts. With BGCs containing only 10-30 genes, applying a 95% threshold could misclassify functionally essential genes as accessory simply due to small differences in cluster boundaries or annotation quality. The framework therefore adopted a 100% presence threshold for core gene definition, ensuring that only truly universal genes within a family are classified as core. This stricter definition preserves meaningful distinctions between conserved and variable functions while providing clearer insights into the genetic diversity within each family.

Traditional pangenomic openness analysis focuses on a single dimension: whether the total gene repertoire continues to expand as more genomes are added to the analysis. This approach uses Heaps' law parameters to classify pangenomes as open ($\gamma > 0.6$), intermediate ($0.3 \leq \gamma \leq 0.6$), or closed ($\gamma < 0.3$) based on gene accumulation curves[110]. While this framework works well for large genomic datasets with hundreds to thousands of genomes, it required modification for BGC analysis due to typically smaller family sizes and the unique characteristics of biosynthetic systems. The PanBGC framework introduced two complementary openness metrics that capture different aspects of biosynthetic diversity. Gene-based openness measures the traditional expansion of gene repertoire, adapted for smaller datasets through modified curve fitting strategies including weighted regression and

non-linear optimization to provide robust estimates even for families with few members. More importantly, the framework introduced compositional openness analysis, which measures how consistently orthologous groups are reused across BGCs within a family, indicating variability in how subsets of the total gene pool are deployed in individual clusters.

This dual-metric approach addresses a question specific to biosynthetic systems: whether diversity arises primarily through acquisition of entirely novel genes or through modular rearrangement of existing genetic components within the family. Traditional pangenomic openness analysis cannot distinguish between these mechanisms, but the distinction is crucial for understanding biosynthetic innovation. The compositional openness metric revealed that while most BGC families maintain closed gene repertoires (limited novel gene acquisition), they exhibit high compositional plasticity, suggesting that modular reorganization rather than gene acquisition drives most biosynthetic diversity.

KEY FINDINGS

GENE REPERTOIRE STABILITY VERSUS COMPOSITIONAL PLASTICITY

The most significant finding from applying the PanBGC framework to over 80,000 gene cluster families was the discovery of a fundamental difference between gene repertoire stability and compositional plasticity. Analysis of gene-based openness revealed that the majority of BGC families exhibit closed pangenomes with average γ-values of 0.286, indicating limited acquisition of entirely novel genes as more clusters are added to families. This finding suggests that successful BGC families reach evolutionary equilibrium in terms of their core genetic content, with established biosynthetic pathways representing combinations of genetic elements that have been optimized over evolutionary time. However, compositional openness analysis revealed a different pattern, with average γ-values of 0.841 indicating high structural variability in how genes, that are part of the same family, are combined within individual clusters. This compositional plasticity demonstrates that BGC families retain significant potential for generating diversity through recombination and rearrangement of existing genes, even when they are not acquiring novel genes. The statistical significance of this difference was confirmed, establishing that these represent genuinely distinct aspects of biosynthetic diversity [125]. This pattern suggests that biosynthetic innovation in natural systems operates primarily through modular reorganization of evolutionarily validated components rather than through incorporation of entirely novel genetic material. This finding has important implications for understanding how secondary metabolite diversity is generated and maintained in microbial populations, suggesting that the chemical diversity observed in natural products is established by the combination of different tailoring genes already present in similar clusters rather than continuous innovation through novel gene acquisition. The predominance of compositional plasticity over gene repertoire expansion suggests that biosynthetic innovation operates under different constraints than general microbial evolution. While bacterial genomes often can incorporate novel genes from environmental gene pools, BGCs appear to operate under

stronger functional constraints that favor recombination of proven genetic components over experimentation with entirely novel elements. This strategy may be particularly advantageous in competitive microbial environments where maintaining chemical defenses or signaling capabilities is essential for survival, but chemical novelty provides competitive advantages [148].

FUNCTIONAL PATTERNS IN CORE VERSUS ACCESSORY GENES

Domain-level analysis of core, accessory, and unique orthologous groups revealed systematic patterns in the functional roles of genes across different categories. Core genes were predominantly associated with essential enzymatic activities required for metabolite biosynthesis[125], such as condensation domains in NRPS systems and ketosynthase domains in PKS systems [149]. These findings align with our expectations that core biosynthetic machinery would be universally conserved within families, representing the minimal genetic requirements for producing the characteristic chemical scaffold of each family. Accessory genes displayed greater functional variability and were often associated with tailoring reactions, regulatory functions, or transport activities[125]. The presence or absence of one or multiple accessory genes across multiple BGCs within families highlighted how they contribute to structural diversification or pathway regulation without being universal requirements. This pattern suggests that accessory genes represent functions that can be gained or lost to fine-tune biosynthetic output for specific ecological or physiological contexts.

PANBGC-DB IN THE CONTEXT OF EXISTING BGC RESOURCES

While several resources exist for organizing biosynthetic gene clusters at scale, including BiG-FAM database[150] and antiSMASH-DB[151], these platforms have primarily focused on high-level classification and dereplication of BGCs across global datasets. BiG-FAM organizes clusters into families based on similarity but has only limited tools for analysing internal diversity within those families. The database was designed to provide a broad overview of secondary metabolite diversity and therefore employed relatively high similarity cutoffs to generate gene cluster families[150]. antiSMASH-DB, on the other hand, serves as a comprehensive repository of predicted BGCs but lacks frameworks for comparative analysis of gene cluster families[151].

PanBGC-DB addresses a complementary but distinct analytical need by providing tools for in-depth analysis of variation within gene cluster families. The platform's two-step clustering strategy generates more granular GCFs than existing approaches, allowing the creation of biologically meaningful family definitions that capture both broad functional relationships and specific structural features essential for understanding biosynthetic mechanisms. This means that with stricter similarity cutoffs compared to BiG-FAM, the created families are less biased to include clusters that may encode unrelated secondary metabolites within the same family, ensuring greater biochemical coherence within each GCF. This finer resolution enables

researchers to investigate specific genes that create chemical diversity within biosynthetic families.

PanBGC-DB is designed to complement rather than replace existing BGC analysis tools, fitting seamlessly into established analytical workflows. Researchers can use antiSMASH for initial BGC detection, BiG-SCAPE[103] for family assignment and then employ PanBGC-DB for detailed intra-family analysis. This integration capability ensures that the population-level insights provided by PanBGC-DB enhance rather than duplicate existing analytical results. The platform's query interface enables researchers to contextualize their own BGCs within the broader landscape of biosynthetic diversity by identifying the most similar families in the reference dataset. This functionality allows researchers to rapidly assess whether their clusters represent novel biosynthetic approaches or variations within known families, guiding experimental priorities and informing hypotheses about potential chemical products.

METHODICAL APPLICATIONS

GENOME MINING

The PanBGC framework helps genome mining approaches by enabling systematic prioritization of BGC families based on quantitative diversity metrics. Traditional genome mining strategies often struggle with the overwhelming number of predicted BGCs in genomic databases[152], lacking systematic approaches for assessing which clusters are most likely to yield novel bioactive compounds. The framework addresses this challenge by providing quantitative metrics for assessing biosynthetic novelty potential based on diversity patterns within families.

Families exhibiting unusual combinations of core and accessory genes, or those showing high compositional diversity, may often represent promising targets for natural product discovery. The framework can also help with chemical novelty prediction based on accessory gene combinations that have not been previously observed. By cataloging which accessory genes co-occur in characterized families, researchers can identify clusters with unprecedented gene combinations and prioritize them for chemical characterization. This approach, in addition to tools like ARTS (Antibiotic Resistance Target Seeker)[153], provides a rational basis for predicting chemical novelty without requiring extensive experimental screening.

Beyond basic prioritization, the PanBGC framework enables several sophisticated genome mining strategies. Incomplete cluster detection becomes possible by identifying BGCs that lack expected core genes compared to other family members, potentially revealing clusters that have undergone gene loss or require complementation from elsewhere in the genome. This capability can be used to assess the completeness of assembled genomes with their respective clusters.

SYNTHETIC BIOLOGY

The PanBGC framework provides a systematic approach to identify tailoring enzymes for synthetic biology applications. By cataloging accessory genes that co-occur with specific core biosynthetic machinery, the platform provides strong evidence for functional compatibility. Tailoring enzymes found within the same gene cluster family have been evolutionarily validated to work with similar core pathways, significantly increasing the likelihood of successful integration compared to traditional trial-and-error approaches that often combine enzymes from distantly related systems[154]. This data-driven strategy addresses a major bottleneck in biosynthetic engineering by predicting enzyme compatibility. The framework enables researchers to identify underexploited combinations of tailoring activities by comparing accessory gene profiles across families, transforming enzyme selection from an empirical process into one guided by evolutionary precedents. This systematic cataloging reveals novel tailoring strategies proven effective in natural systems that can be applied to target clusters. Recent successful demonstrations in biosynthetic gene cluster engineering demonstrated how the tailoring enzyme swapping between related biosynthetic gene clusters is a viable option to generate new-to-nature compounds. A compelling proof-of-concept was provided by Ye and colleagues in their work with argimycin P alkaloids [155], where they successfully expressed an isomerase and a dehydrogenase gene from the later biosynthetic steps of coelimycin P1 into *Streptomyces argillaceus* strains which produces the structural similar argimycins P. This combinatorial approach generated five novel hybrid argimycins with unprecedented scaffolds, one of which demonstrated improved antibiotic activity compared to the parental compounds. The success of this enzyme swapping was enabled by the fact that coelimycin P1 and argimycins P are two polyketide alkaloids whose biosynthesis pathways share some early steps, providing sufficient biochemical compatibility between the donor and recipient clusters [155]. Beyond cataloging possible tailoring genes established in nature, the compositional openness metrics calculated for each family provide valuable insights for synthetic biology applications. Families exhibiting high γ-values for compositional diversity indicate greater tolerance to accept genes from similar clusters and engineering interventions, suggesting these clusters are naturally more suitable for modification and represent promising targets for biosynthetic engineering [154]. Conversely, families with low compositional γ-values reflect evolutionary constraint and structural rigidity, indicating a stable cluster which most likely does not accept new genes. This quantitative assessment of family plasticity enables researchers to tailor their engineering strategies appropriately, maximizing the likelihood of successful pathway modification while minimizing the risk of introducing nonfunctional changes to well-conserved biosynthetic systems.

Even when specific tailoring enzymes are not found within the same gene cluster family, PanBGC-DB enables broader analysis across related biosynthetic categories to identify functionally compatible enzymes. The organization in biosynthetic categories allows researchers to examine tailoring enzyme distributions across related biosynthetic systems, such as all major biosynthetic classes and their hybrids. For instance, if a methyltransferase might be commonly found across diverse NRPS families, there is a higher likelihood to integrate this successful into novel NRPS contexts [154]. The PanBGC-DB shows these

patterns, providing researchers a comprehensive list of possible tailoring enzymes suited for different clusters. This category analysis is particularly valuable for identifying tailoring enzymes that operate on chemical moieties that are structurally conserved across different biosynthetic pathways. For example, tailoring enzymes that modify amino acid residues could potentially function across both NRPS and hybrid systems, while enzymes that operate on polyketide chains might be transferable between different PKS families. The PanBGC framework enables systematic identification of these transferable activities by analyzing enzyme distribution patterns across biosynthetic categories, revealing the scope of potential applications for specific tailoring functions.

Furthermore, the addition or removal of one or more tailoring enzymes is not the only way to create new compounds by designing a new BGC. The modularity of biosynthetic systems such as non-ribosomal peptide synthetases and polyketide synthases makes them particularly interesting for genetic modification of the core genes [56, 156–158]. In these systems, the substrate specificity of adenylation (A) domains in NRPS, responsible to recruit different amino acids, and acyltransferase (AT) domains in PKS, responsible for the recruitment of acyl-CoA extender units, plays a central role in determining the chemical structure of the final compound. By analyzing the conservation and variability of these domains across gene cluster families, it becomes possible to identify positions within the assembly line that tolerate substrate diversity. Such flexible positions are prime targets for domain swapping strategies aimed at altering or expanding the chemical output, while positions under strong evolutionary constraint may be less suitable for modification. In PanBGC-DB, we provide a visual representation of the domain composition of NRPS and PKS modules of each family, enabling researchers to easily identify variable positions.

Several successful examples demonstrate the value of understanding positional flexibility in modular biosynthetic systems [159–163]. In PKS engineering, AT domain swapping in module 7 of the rapamycin polyketide synthase successfully altered the extender unit specificity from methylmalonyl-CoA to malonyl-CoA, producing 23-desmethyl rapamycin[164]. This substitution shows how AT domain engineering can systematically remove or add specific chemical features to complex natural products. The success of this swap was enabled by the compatibility between the donor and recipient AT domains and their ability to maintain proper interactions with the surrounding PKS architecture while changing substrate specificity. This suggests that when using AT domains present in similar clusters from the same gene cluster family, the probability that the swap is accepted by the recipient cluster is much higher due to evolutionary validation of compatibility within the family.

Similarly, in NRPS engineering, successful domain swapping has been achieved when guided by understanding of positional conservation patterns. In daptomycin biosynthesis, exchange of C-A domains led to production of novel active antibiotics [165]. Some were as active as daptomycin while one compound was more potent against *E. coli.* This demonstrates that swapping domains is a viable option to create new-to-nature compounds with increased bioactivity.

The mentioned principle of evolutionary compatibility is further illustrated by subdomain swapping guided by understanding of evolutionary relationships within biosynthetic families. In

gramicidin S synthetase engineering, researchers transplanted nine subdomains encoding diverse amino acid specificities into the *GrsA* initiation module[166]. The most successful construct combined a subdomain with a module from the same gramicidin S biosynthetic system, demonstrating how subdomain engineering can systematically alter substrate specificity while maintaining enzymatic function. The success of this combination was enabled by the compatibility between components from the same gramicidin S synthetase family, where both donor and recipient elements share evolutionary origins and structural constraints. This suggests that when using NRPS subdomains and modules present in similar clusters from the same gene cluster family, the probability that the engineering approach is accepted by the recipient system is much higher due to evolutionary validation of compatibility within the family.

POSSIBLE BROADER IMPACT

PHARMACEUTICAL APPLICATIONS

The organization of BGCs in families and the analysis of internal diversity patterns could provide pharmaceutical researchers with evolutionary insights that can inform drug discovery and development strategies. As mentioned in the introduction, a lot of compounds used in modern medicine are derived or inspired by chemical structures found in nature [1, 167]. Using the strategies outlined in the previous chapter, researchers can redesign BGCs responsible for producing specific pharmaceuticals to enhance bioactivity by incorporating accessory genes identified in related clusters, thereby diversifying or optimizing the core structure. In addition to increasing bioactivity, these modifications can be directed to reduce toxicity, improve compound stability, or adapt the molecules to better suit the conditions of their intended application . This approach could also lead to the development of novel antibiotics derived from known compounds, with structural variations that enable them to overcome existing resistance mechanisms and restore their clinical effectiveness.

ENVIRONMENTAL APPLICATIONS

The core versus accessory gene classification within BGC families offers insights into biosynthetic flexibility that may correlate with adaptability to different environmental conditions. Families with high compositional diversity, as indicated by compositional openness metrics, may represent lineages with greater potential for adaptation to diverse environmental contexts, while families with low compositional diversity may represent lineages with strong adaptation to a specific niche. This information can guide researchers in selecting target organisms for environmental applications by prioritizing taxa that belong to BGC families with demonstrated compositional plasticity. For applications involving biosurfactants and other environmentally relevant compounds, the systematic organization of BGC families enables researchers to identify related pathways across diverse bacterial taxa and assess the distribution of specific biosynthetic capabilities. The framework's ability to reveal accessory gene patterns within

families can help predict structural variations in products that may be relevant for specific environmental applications. The systematic organization of biosynthetic capabilities across microbial diversity also supports the development of environmental monitoring approaches. BGC family distributions within microbial communities can potentially serve as indicators of biosynthetic potential and community functional capacity, providing researchers with systematic frameworks for interpreting complex metagenomic datasets in environmental contexts.

LIMITATIONS AND CONSIDERATIONS

CREATION OF MEANINGFUL GENE CLUSTER FAMILIES

The validity of all population-level analyses presented in this framework fundamentally depends on the accurate clustering of BGCs into biologically meaningful families. While BiG-SCAPE represents the current gold standard for creating functionally coherent gene cluster families based on domain architecture and sequence similarity [145], it was not designed to handle datasets of the scale used in this study. This computational limitation necessitated the development of a two-step clustering approach that introduces potential sources of bias in family formation.

Our initial implementation employed strict similarity cutoffs for both BiG-SLiCE and BiG-SCAPE clustering steps, which resulted in an excessive number of singleton families and the inappropriate separation of clusters that should logically belong to the same family. For example, the malleobactin and ornibactin clusters, which produce structurally related siderophores through similar biosynthetic logic [168–170], were assigned to different families during the initial BiG-SLiCE clustering step. Similarly, desferoxamine-producing clusters[171], despite their clear functional relationship, were distributed across multiple distinct families. These separations occurred even before the more refined BiG-SCAPE analysis could assess their architectural similarities, indicating that the initial broad clustering was too restrictive. Several factors likely contributed to this overly restrictive initial clustering. First, functionally related siderophore clusters often exhibit significant sequence divergence while maintaining similar biosynthetic logic, particularly when they have evolved in different bacterial lineages or undergone extensive horizontal gene transfer [172]. Second, BiG-SLiCE's k-mer-based methodology may be sensitive to differences in gene arrangement, enzyme variants, and species-specific sequence features that do not reflect fundamental biosynthetic differences. To address this limitation, we adapted our approach by implementing much higher similarity cutoffs for the BiG-SLiCE step, allowing the more accurate BiG-SCAPE algorithm to work with larger but still computationally manageable family sizes. This modification successfully resolved the separation issues observed with the malleobactin/ornibactin and desferoxamine families, demonstrating that the adjusted parameters could capture biologically meaningful relationships that were previously obscured.

However, even with these methodological improvements, multiple limiting factors continue to affect the creation of meaningful families. Since the majority of BGCs in public databases derive from the automated prediction of antiSMASH, cluster boundaries are inferred based on the position of core biosynthetic genes and domain architecture but lack experimental validation[173]. Inaccurate boundary predictions can significantly impact clustering outcomes by incorporating genes that are not functionally related to the biosynthetic pathway or by excluding relevant genes that fall outside the predicted boundaries. The inclusion of non-functional flanking genes can artificially increase similarity between unrelated clusters that happen to share similar genomic neighborhoods, while the exclusion of genuine cluster components can prevent the recognition of true family relationships. This boundary uncertainty represents a source of potential error that affects the interpretation of core versus accessory gene classifications and compositional diversity metrics throughout the analysis.

ORTHOLOGOUS COMPLETE GROUPING OF THE GENE POOL OF A GCF

The accurate identification of orthologous relationships between genes across BGCs within each family is critical for meaningful core, accessory, and unique gene classification. For this purpose, we employed ZOL[174], which offers several advantages over traditional ortholog detection methods such as OrthoFinder[147]. ZOL was specifically designed for the analysis of gene clusters and incorporates both sequence similarity and positional conservation (synteny) in its clustering algorithm. This dual approach is particularly valuable for BGC analysis, where gene order often reflects functional relationships and evolutionary constraints within biosynthetic pathways. By considering positional information alongside sequence similarity, ZOL can better distinguish between genes that may share similar sequences but serve different functional roles depending on their genomic context within the cluster architecture.[174]

However, despite these methodological advantages, ZOL's reliance on sequence similarity as a primary clustering criterion introduces potential limitations that can affect downstream analyses. The tool may group structurally similar yet functionally divergent genes into the same orthologous group, particularly when these genes share high sequence identity but have evolved distinct substrate specificities or catalytic properties. The grouping of functionally divergent genes has several downstream impacts on the PanBGC analysis. In core and accessory gene classification, the misassignment of functionally distinct genes to the same orthologous group can artificially inflate the apparent conservation of certain functions across a family, potentially masking important functional diversity. Conversely, genes that represent genuine functional equivalents but differ slightly in sequence may be incorrectly separated into different orthologous groups, leading to underestimation of core functions and overestimation of accessory diversity. These classification errors also directly impact the calculation of compositional openness metrics, as the γ-values depend fundamentally on accurate ortholog identification to assess how consistently gene combinations appear across BGCs within families. Systematic errors in ortholog assignment could therefore lead to incorrect interpretations of compositional plasticity and evolutionary dynamics within BGC families.

DATASET BIAS AND TAXONOMIC REPRESENTATION

The comprehensive nature of our analysis depends heavily on the representativeness of the underlying BGC dataset, which exhibits significant taxonomic and ecological biases that may affect the generalizability of our findings. The antiSMASH database, while representing the most comprehensive collection of predicted BGCs available, reflects the historical sampling biases inherent in microbial genome sequencing efforts [151]. Certain taxonomic groups, particularly *Streptomyces* and other well-studied Actinomycetes, are heavily over-represented in public genome databases due to their recognized importance in natural product discovery and their relative ease of cultivation [175]. Conversely, many bacterial lineages remain under-sampled, particularly those from extreme environments, marine ecosystems, and unculturable organisms that may harbor significant biosynthetic diversity. This taxonomic bias has several implications for our population-level analyses. Apparent diversity patterns and openness metrics may be skewed toward the evolutionary dynamics of well-studied lineages, potentially missing important modes of biosynthetic innovation present in under-sampled groups. The core versus accessory gene classifications within families may reflect the genetic organization typical of intensively studied organisms rather than representing universal patterns across bacterial diversity. Additionally, the compositional openness metrics calculated for BGC families may be influenced by the depth of sampling within particular taxonomic groups, as families dominated by closely related organisms may show different diversity patterns compared to those spanning broader phylogenetic distances.

STATISTICAL POWER LIMITATIONS FOR OPENNESS ANALYSIS

The statistical robustness of our openness calculations is constrained by the size distribution of gene cluster families in our dataset. While our analysis encompasses over 80,000 GCFs, the majority of these families consist of very few members, with ~6% of the total BGC dataset being singletons and most multi-member families containing fewer than ten BGCs. Openness analysis requires families with at least three BGCs for meaningful γ-value estimation, which restricts this analysis to only 14,716 families and excludes a substantial portion of the total diversity from quantitative assessment. Even among families that meet the minimum size threshold, the small number of BGCs per family reduces the statistical confidence of γ-value estimates and limits the reliability of curve fitting procedures. Traditional pangenomic studies typically employ hundreds to thousands of genomes to establish robust openness metrics[112, 135, 176], while our BGC families rarely exceed several dozen members. To address this limitation, we implemented modified curve fitting strategies including weighted regression and non-linear optimization approaches specifically adapted for smaller datasets. However, these adaptations do not completely reduce the bias based on the size of the families.

CHEMICAL PRODUCT VALIDATION GAP

A fundamental limitation of our population-level framework lies in the disconnect between computational BGC family organization and experimental validation of chemical products. The vast majority of BGC families in our dataset lack experimentally characterized chemical products, with only a small fraction containing clusters from the MIBiG database with validated structures and bioactivities[101]. This validation gap creates uncertainty about whether computationally defined families represent biologically meaningful units that produce structurally or functionally related compounds. While these assumptions are supported by well-characterized examples like the malleobactin/ornibactin and desferoxamine family, the extent to which this relationship holds across the broader diversity of biosynthetic families remains unclear. Families that appear coherent based on genetic architecture may nevertheless produce chemically complete distinct compounds due to subtle differences in enzymatic activities or substrate specificities that are not captured by sequence-based clustering approaches. Furthermore, the functional roles attributed to accessory genes remain largely theoretical for most families. Since the majority of BGC families lack characterized chemical products, the specific contributions of accessory genes to structural diversification is not yet experimentally validated and must instead be inferred from functional annotations. This reliance on computational predictions rather than experimental evidence introduces additional uncertainty into what accessory genes can be used in synthetic biology.

Despite these limitations, the PanBGC framework represents a significant methodological advance in the systematic analysis of biosynthetic gene cluster diversity. Many of these challenges are common to large-scale computational analyses of predicted BGCs and affect all current approaches in the field. The limitations identified here should be viewed as areas for future improvement rather than fundamental flaws that invalidate the framework's utility. As genome sequencing efforts expand to include more diverse microbial lineages[173, 177], as BGC prediction and annotation tools improve[175, 178], and as more natural products are experimentally characterized and linked to their biosynthetic origins[101, 179], the accuracy and biological relevance of population-level BGC analyses will continue to increase. Most importantly, the systematic organization of BGC families and the quantitative metrics for assessing diversity patterns provide a robust foundation for hypothesis generation and experimental design, even when individual family assignments or gene classifications may be imperfect. The framework's value lies not in providing definitive answers about every BGC family, but in establishing a systematic approach for analyzing biosynthetic diversity that can guide more targeted experimental investigations and inform rational engineering strategies based on evolutionary principles.

FUTURE DIRECTIONS

The systematic framework established by PanBGC opens several promising avenues for future development that can address current limitations while expanding the scope and utility of population-level BGC analysis. The most immediate and impactful advancement lies in the

integration of multi-omics data to bridge the gap between computational BGC organization and experimental validation of chemical products. Metabolomics data represents the most direct approach for validating the chemical coherence of computationally defined BGC families. By systematically linking mass spectrometry-based metabolic profiles to BGC family assignments[180], researchers can test whether families that appear genetically coherent also produce chemically related compounds. This integration would be particularly valuable for validating the functional significance of accessory gene variations within families, as correlations between specific accessory gene combinations and distinct metabolic features could provide experimental evidence for the chemical diversification mechanisms inferred from genetic analysis. Large-scale metabolomics datasets from projects like the Global Natural Products Social Molecular Networking (GNPS) platform[181] could be mapped to BGC families, creating a comprehensive resource linking genetic and chemical diversity. Additionally, transcriptomics data could indicate co-dependence of accessory genes revealing that multiple of them are needed in order to work and perform adaptations to the final compound.

Advances in artificial intelligence and machine learning present transformative opportunities for improving the technical foundations of BGC family analysis[182, 183]. Deep learning approaches could significantly enhance BGC boundary prediction accuracy, addressing one of the limitations affecting family formation and gene classification. Neural networks trained on experimentally validated cluster boundaries from the MIBiG database could learn to recognize subtle patterns in gene organization and expression that current rule-based approaches miss, potentially reducing the systematic errors introduced by inaccurate boundary predictions[182]. Machine learning could also revolutionize the clustering process itself by developing algorithms that incorporate chemical similarity alongside genetic similarity when forming families. Such approaches could integrate structural information from characterized natural products to guide the grouping of uncharacterized BGCs, creating families that are coherent both genetically and chemically. Large language models, already showing promise in protein function prediction[184], could be adapted for automated functional annotation of accessory genes by training on the growing amount of biochemical literature and experimental data. Furthermore, predictive models could be developed to estimate chemical product structures based on BGC family composition, enabling researchers to prioritize families for experimental characterization based on predicted chemical novelty or bioactivity.

Community contribution and collaborative validation represent essential components for realizing the full potential of population-level BGC analysis. The development of crowdsourced curation platforms could harness the expertise of the global natural products research community to validate and refine BGC family assignments. Such platforms could enable researchers to submit experimental data linking characterized compounds to specific BGCs, gradually building a comprehensive database of validated gene-to-chemical relationships. Community-driven annotation of accessory gene functions could leverage the collective knowledge to provide more accurate functional assignments than automated prediction methods alone[101].

The integration of these developments will contribute to the continued evolution of population-level BGC analysis as genome sequencing efforts expand, multi-omics technologies become more accessible, and computational methods advance. While significant challenges remain in accurately predicting chemical products from genetic information and validating computational family assignments experimentally, the systematic framework established in this thesis provides a foundation for incremental improvements that can enhance our understanding of biosynthetic diversity.

CONCLUSION

This thesis establishes the PanBGC framework as a systematic approach for understanding diversity within biosynthetic gene cluster families by adapting proven pangenomic principles to secondary metabolite biosynthesis. By treating gene cluster families as evolutionary populations rather than isolated genomic islands, this work provides the first comprehensive framework for analyzing the internal diversity patterns that drive chemical innovation in microbial natural products. The most significant finding from this population-level analysis is the contrasting patterns of gene repertoire stability and compositional plasticity within BGC families. While most families exhibit closed pangenomes with limited acquisition of novel genes, they demonstrate high compositional diversity in how existing genetic components are combined across individual clusters. This reveals that biosynthetic innovation operates primarily through modular reorganization of evolutionarily validated components rather than continuous incorporation of novel genetic material. This insight suggests that chemical complexity emerges from recombination of proven biosynthetic elements rather than endless genetic innovation. Through the PanBGC-DB website, these population-level analyses become accessible to researchers from different fields. The framework enables researchers to make informed decisions about experimental priorities and engineering strategies based on comprehensive evolutionary context rather than limited precedent. Most importantly, this work establishes a new strategy for natural product research that emphasizes population-level thinking over individual case studies thereby enabling more insight into GCF diversity.

THESIS - BIBLIOGRAPHY

1. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J Nat Prod* 2020;**83**:770–803. https://doi.org/10.1021/ACS.JNATPROD.9B01285
2. Demain AL. Importance of microbial natural products and the need to revitalize their discovery. *J Ind Microbiol Biotechnol* 2014;**41**:185–201. https://doi.org/10.1007/S10295-013-1325-Z
3. Demain AL. Antibiotics: Natural products essential to human health. *Med Res Rev* 2009;**29**:821–842. https://doi.org/10.1002/MED.20154
4. Atanasov AG et al. Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery 2021 20:3* 2021;**20**:200–216. https://doi.org/10.1038/s41573-020-00114-z
5. Vicente MF et al. Microbial natural products as a source of antifungals. *Clinical Microbiology and Infection* 2003;**9**:15–32. https://doi.org/10.1046/J.1469-0691.2003.00489.X
6. Fleming A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *Br J Exp Pathol* 1929;**10**:226.
7. Mushtaq S et al. Natural products as reservoirs of novel therapeutic agents. *EXCLI J* 2018;**17**:420. https://doi.org/10.17179/EXCLI2018-1174
8. Cragg GM, Pezzuto JM. Natural Products as a Vital Source for the Discovery of Cancer Chemotherapeutic and Chemopreventive Agents. *Medical Principles and Practice* 2016;**25**:41–59. https://doi.org/10.1159/000443404
9. Fitzgerald JB et al. Systems biology and combination therapy in the quest for clinical efficacy. *Nature Chemical Biology 2006 2:9* 2006;**2**:458–466. https://doi.org/10.1038/nchembio817
10. Metsä-Ketelä M et al. Anthracycline Biosynthesis: Genes, Enzymes and Mechanisms. *Top Curr Chem* 2007;**282**:101–140. https://doi.org/10.1007/128_2007_14
11. Weiss RB. The anthracyclines: will we ever find a better doxorubicin? *Semin Oncol* 1992;**19**:670–686.
12. Ramírez-Rendon D et al. Impact of novel microbial secondary metabolites on the pharma industry. *Appl Microbiol Biotechnol* 2022;**106**:1855. https://doi.org/10.1007/S00253-022-11821-5
13. Pham J V. et al. A review of the microbial production of bioactive natural products and biologics. *Front Microbiol* 2019;**10**:449147. https://doi.org/10.3389/FMICB.2019.01404/XML/NLM
14. Murray CJ et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* 2022;**399**:629–655. https://doi.org/10.1016/S0140-6736(21)02724-0
15. Park D et al. Complex natural product production methods and options. *Synth Syst Biotechnol* 2021;**6**:1. https://doi.org/10.1016/J.SYNBIO.2020.12.001

16. Kunakom S, Eustáquio AS. Natural Products and Synthetic Biology: Where We Are and Where We Need To Go. *mSystems* 2019;**4**:e00113-19. https://doi.org/10.1128/MSYSTEMS.00113-19

17. Simoben C V. et al. Challenges in natural product-based drug discovery assisted with in silico-based methods. *RSC Adv* 2023;**13**:31578. https://doi.org/10.1039/D3RA06831E

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

19. Pacwa-Płociniczak M et al. Environmental Applications of Biosurfactants: Recent Advances. *International Journal of Molecular Sciences 2011, Vol 12, Pages 633-654* 2011;**12**:633–654. https://doi.org/10.3390/IJMS12010633

20. Marchant R, Banat IM. Biosurfactants: A sustainable replacement for chemical surfactants? *Biotechnol Lett* 2012;**34**:1597–1605. https://doi.org/10.1007/S10529-012-0956-X/FIGURES/3

21. Sharma N et al. A comprehensive review on microbial production and significant applications of multifunctional biomolecules: biosurfactants. *Biodegradation* 2025;**36**:1–23. https://doi.org/10.1007/S10532-025-10121-9/METRICS

22. Nikolova C, Gutierrez T. Biosurfactants and Their Applications in the Oil and Gas Industry: Current State of Knowledge and Future Perspectives. *Front Bioeng Biotechnol* 2021;**9**:626639. https://doi.org/10.3389/FBIOE.2021.626639/XML

23. Eras-Muñoz E et al. Microbial biosurfactants: a review of recent environmental applications. *Bioengineered* 2022;**13**:12365–12391. https://doi.org/10.1080/21655979.2022.2074621

24. Thomas GE et al. Effects of dispersants and biosurfactants on crude-oil biodegradation and bacterial community succession. *Microorganisms* 2021;**9**:1200. https://doi.org/10.3390/MICROORGANISMS9061200/S1

25. Zhang P et al. Recent advances in the natural products-based lead discovery for new agrochemicals. *Advanced Agrochem* 2023;**2**:324–339. https://doi.org/10.1016/J.AAC.2023.09.004

26. Marks BB, Nogueira MA, Hungria M. Microbial Secondary Metabolites and Their Use in Achieving Sustainable Agriculture: Present Achievements and Future Challenges. *Agronomy* 2025;**15**:1350. https://doi.org/10.3390/AGRONOMY15061350

27. Dayan FE, Cantrell CL, Duke SO. Natural products in crop protection. *Bioorg Med Chem* 2009;**17**:4022–4034. https://doi.org/10.1016/J.BMC.2009.01.046

28. Fouillaud M, Dufossé L. Microbial Secondary Metabolism and Biotechnology. *Microorganisms* 2022;**10**:123. https://doi.org/10.3390/MICROORGANISMS10010123

29. Singh R et al. Microbial metabolites in nutrition, healthcare and agriculture. *3 Biotech* 2017;**7**:1–14. https://doi.org/10.1007/S13205-016-0586-4/TABLES/2

30. Sajjad W et al. Pigment production by cold-adapted bacteria and fungi: colorful tale of cryosphere with wide range applications. *Extremophiles 2020 24:4* 2020;**24**:447–473. https://doi.org/10.1007/S00792-020-01180-2

31. Rather LJ et al. Research progress, challenges, and perspectives in microbial pigment production for industrial applications - A review. *Dyes and Pigments* 2023;**210**:110989. https://doi.org/10.1016/J.DYEPIG.2022.110989

32. Dinglasan JLN et al. Microbial secondary metabolites: advancements to accelerate discovery towards application. *Nature Reviews Microbiology 2025 23:6* 2025;**23**:338–354. https://doi.org/10.1038/s41579-024-01141-y

33. Marks BB, Nogueira MA, Hungria M. Microbial Secondary Metabolites and Their Use in Achieving Sustainable Agriculture: Present Achievements and Future Challenges. *Agronomy 2025, Vol 15, Page 1350* 2025;**15**:1350. https://doi.org/10.3390/AGRONOMY15061350

34. Voigt CA. Synthetic biology 2020–2030: six commercially-available products that are changing our world. *Nature Communications 2020 11:1* 2020;**11**:1–6. https://doi.org/10.1038/s41467-020-20122-2

35. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics* 2010;**26**:449–457. https://doi.org/10.1016/J.TIG.2010.07.001

36. Cimermancic P et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 2014;**158**:412. https://doi.org/10.1016/J.CELL.2014.06.034

37. Chavali AK, Rhee SY. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Brief Bioinform* 2017;**19**:1022. https://doi.org/10.1093/BIB/BBX020

38. Egan S et al. Transfer of Streptomycin Biosynthesis Gene Clusters within Streptomycetes Isolated from Soil. *Appl Environ Microbiol* 1998;**64**:5061. https://doi.org/10.1128/AEM.64.12.5061-5063.1998

39. Murakami T et al. A system for the targeted amplification of bacterial gene clusters multiplies antibiotic yield in Streptomyces coelicolor. *Proc Natl Acad Sci U S A* 2011;**108**:16020–16025. https://doi.org/10.1073/PNAS.1108124108/SUPPL_FILE/PNAS.201108124SI.PDF

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

41. Barnard AML et al. Quorum sensing, virulence and secondary metabolite production in plant soft-rotting bacteria. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2007;**362**:1165–1183. https://doi.org/10.1098/RSTB.2007.2042

42. Jousset A, Scheu S, Bonkowski M. Secondary metabolite production facilitates establishment of rhizobacteria by reducing both protozoan predation and the competitive effects of indigenous bacteria. *Funct Ecol* 2008;**22**:714–719. https://doi.org/10.1111/J.1365-2435.2008.01411.X

43. Santamaria G et al. Evolution and regulation of microbial secondary metabolism. *Elife* 2022;**11**:e76119. https://doi.org/10.7554/ELIFE.76119

44. Romano S et al. Phosphate limitation induces drastic physiological changes, virulence-related gene expression, and secondary metabolite production in Pseudovibrio sp.

strain FO-BEG1. *Appl Environ Microbiol* 2015;**81**:3518–3528. https://doi.org/10.1128/AEM.04167-14/SUPPL_FILE/ZAM999116258SO1.PDF

45. Zhang X, Hindra, Elliot MA. Unlocking the trove of metabolic treasures: activating silent biosynthetic gene clusters in bacteria and fungi. *Curr Opin Microbiol* 2019;**51**:9–15. https://doi.org/10.1016/J.MIB.2019.03.003

46. Chase AB et al. Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. *mBio* 2021;**12**. https://doi.org/10.1128/MBIO.02700-21/ASSET/8008E855-E38C-481A-9F4C-4AB9A369DD5E/ASSETS/IMAGES/MEDIUM/MBIO.02700-21-F004.GIF

47. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. *Proc Natl Acad Sci U S A* 2008;**105**:4601. https://doi.org/10.1073/PNAS.0709132105

48. Jensen PR. Natural Products and the Gene Cluster Revolution. *Trends Microbiol* 2016;**24**:968–977. https://doi.org/10.1016/J.TIM.2016.07.006

49. Geller-McGrath D et al. Diverse secondary metabolites are expressed in particle-associated and free-living microorganisms of the permanently anoxic Cariaco Basin. *Nature Communications 2023 14:1* 2023;**14**:1–12. https://doi.org/10.1038/s41467-023-36026-w

50. Dufour N, Rao RP. Secondary metabolites and other small molecules as intercellular pathogenic signals. *FEMS Microbiol Lett* 2011;**314**:10–17. https://doi.org/10.1111/J.1574-6968.2010.02154.X

51. Krauss S et al. Horizontal Transfer of Bacteriocin Biosynthesis Genes Requires Metabolic Adaptation To Improve Compound Production and Cellular Fitness. *Microbiol Spectr* 2022;**11**:e03176-22. https://doi.org/10.1128/SPECTRUM.03176-22

52. Tyc O et al. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends Microbiol* 2017;**25**:280–292. https://doi.org/10.1016/J.TIM.2016.12.002

53. Wlodek A et al. Diversity oriented biosynthesis via accelerated evolution of modular gene clusters. *Nature Communications 2017 8:1* 2017;**8**:1–10. https://doi.org/10.1038/s41467-017-01344-3

54. Salamzade R, Kalan LR. Context matters: assessing the impacts of genomic background and ecology on microbial biosynthetic gene cluster evolution. *mSystems* 2025;**10**. https://doi.org/10.1128/MSYSTEMS.01538-24/ASSET/D97FB42C-03B5-4962-B1CC-8B7A391220C5/ASSETS/IMAGES/LARGE/MSYSTEMS.01538-24.F002.JPG

55. Beck C, Garzón JFG, Weber T. Recent Advances in Re-engineering Modular PKS and NRPS Assembly Lines. *Biotechnology and Bioprocess Engineering* 2020;**25**:886–894. https://doi.org/10.1007/S12257-020-0265-5/METRICS

56. Awakawa T et al. Reprogramming of the antimycin NRPS-PKS assembly lines inspired by gene evolution. *Nature Communications 2018 9:1* 2018;**9**:1–10. https://doi.org/10.1038/s41467-018-05877-z

57. Strieker M, Tanović A, Marahiel MA. Nonribosomal peptide synthetases: structures and dynamics. *Curr Opin Struct Biol* 2010;**20**:234–240. https://doi.org/10.1016/J.SBI.2010.01.009

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

59. Risdian C, Mozef T, Wink J. Biosynthesis of Polyketides in Streptomyces. *Microorganisms* 2019;**7**:124. https://doi.org/10.3390/MICROORGANISMS7050124

60. Lal R et al. Regulation and manipulation of the gene clusters encoding type-I PKSs. *Trends Biotechnol* 2000;**18**:264–274. https://doi.org/10.1016/S0167-7799(00)01443-8

61. Okamoto S et al. Biosynthesis of actinorhodin and related antibiotics: discovery of alternative routes for quinone formation encoded in the act gene cluster. *Chem Biol* 2009;**16**:226–236. https://doi.org/10.1016/J.CHEMBIOL.2009.01.015

62. Funabashi M, Funa N, Horinouchi S. Phenolic lipids synthesized by type III polyketide synthase confer penicillin resistance on Streptomyces griseus. *J Biol Chem* 2008;**283**:13983–13991. https://doi.org/10.1074/JBC.M710461200

63. Chopra I, Roberts M. Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol Mol Biol Rev* 2001;**65**:232–260. https://doi.org/10.1128/MMBR.65.2.232-260.2001

64. Li J, Kim SG, Blenis J. Rapamycin: one drug, many effects. *Cell Metab* 2014;**19**:373–379. https://doi.org/10.1016/J.CMET.2014.01.001

65. Tacar O, Sriamornsak P, Dass CR. Doxorubicin: an update on anticancer molecular action, toxicity and novel drug delivery systems. *J Pharm Pharmacol* 2013;**65**:157–170. https://doi.org/10.1111/J.2042-7158.2012.01567.X

66. Hudson GA, Mitchell DA. RiPP antibiotics: biosynthesis and engineering potential. *Curr Opin Microbiol* 2018;**45**:61–69. https://doi.org/10.1016/J.MIB.2018.02.010

67. Arnison PG et al. Ribosomally synthesized and post-translationally modified peptide natural products: Overview and recommendations for a universal nomenclature. *Nat Prod Rep* 2013;**30**:108–160. https://doi.org/10.1039/C2NP20085F

68. Li C, Zhang F, Kelly WL. Heterologous production of thiostrepton A and biosynthetic engineering of thiostrepton analogs. *Mol Biosyst* 2011;**7**:82–90. https://doi.org/10.1039/C0MB00129E

69. Hegemann JD et al. Lasso Peptides: An Intriguing Class of Bacterial Natural Products. *Acc Chem Res* 2015;**48**:1909–1919. https://doi.org/10.1021/ACS.ACCOUNTS.5B00156/SUPPL_FILE/AR5B00156_SI_001.PDF

70. Avalos M et al. Biosynthesis, evolution and ecology of microbial terpenoids. *Nat Prod Rep* 2022;**39**:249–272. https://doi.org/10.1039/D1NP00047K

71. Ro DK et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature 2006 440:7086* 2006;**440**:940–943. https://doi.org/10.1038/nature04640

72. Miura Y et al. Production of the Carotenoids Lycopene, β-Carotene, and Astaxanthin in the Food Yeast Candida utilis. *Appl Environ Microbiol* 1998;**64**:1226. https://doi.org/10.1128/AEM.64.4.1226-1229.1998

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

74. De Rop AS et al. Novel alkaloids from marine actinobacteria: Discovery and characterization. *Mar Drugs* 2022;**20**:6. https://doi.org/10.3390/MD20010006/S1

75. Fisch KM. Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS. *RSC Adv* 2013;**3**:18228–18247. https://doi.org/10.1039/C3RA42661K

76. Wink M. Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry* 2003;**64**:3–19. https://doi.org/10.1016/S0031-9422(03)00300-5

77. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics* 2010;**26**:449–457. https://doi.org/10.1016/J.TIG.2010.07.001

78. O'Brien J, Wright GD. An ecological perspective of microbial secondary metabolism. *Curr Opin Biotechnol* 2011;**22**:552–558. https://doi.org/10.1016/J.COPBIO.2011.03.010

79. Hibbing ME et al. Bacterial competition: surviving and thriving in the microbial jungle. *Nat Rev Microbiol* 2010;**8**:15. https://doi.org/10.1038/NRMICRO2259

80. Westhoff S et al. Competition sensing changes antibiotic production in streptomyces. *mBio* 2021;**12**:1–13. https://doi.org/10.1128/MBIO.02729-20/SUPPL_FILE/MBIO.02729-20-SF004.PDF

81. Santamaria G et al. Evolution and regulation of microbial secondary metabolism. *Elife* 2022;**11**:e76119. https://doi.org/10.7554/ELIFE.76119

82. Baquero F et al. Evolutionary Pathways and Trajectories in Antibiotic Resistance. *Clin Microbiol Rev* 2021;**34**:e00050-19. https://doi.org/10.1128/CMR.00050-19

83. Klumbys E et al. Discovery, characterization, and engineering of an advantageous Streptomyces host for heterologous expression of natural product biosynthetic gene clusters. *Microb Cell Fact* 2024;**23**:1–12. https://doi.org/10.1186/S12934-024-02416-Y/FIGURES/5

84. Teasdale ME et al. Secondary metabolites produced by the marine bacterium halobacillus salinus that Inhibit quorum sensing-controlled phenotypes in gram-negative bacteria. *Appl Environ Microbiol* 2009;**75**:567–572. https://doi.org/10.1128/AEM.00632-08/ASSET/04C6F130-A453-4F5B-AAB5-BD8ECBA48F2B/ASSETS/GRAPHIC/ZAM0030995780004.JPEG

85. Taga ME, Bassler BL. Chemical communication among bacteria. *Proc Natl Acad Sci U S A* 2003;**100**:14549–14554. https://doi.org/10.1073/PNAS.1934514100/ASSET/B8B5B424-7AE3-40E4-AE08-632FC36612CF/ASSETS/GRAPHIC/PQ1934514005.JPEG

86. Waters CM, Bassler BL. Quorum sensing: cell-to-cell communication in bacteria. *Annu Rev Cell Dev Biol* 2005;**21**:319–346. https://doi.org/10.1146/ANNUREV.CELLBIO.21.012704.131001

87. Liao L et al. An aryl-homoserine lactone quorum-sensing signal produced by a dimorphic prosthecate bacterium. *Proc Natl Acad Sci U S A* 2018;**115**:7587–7592.

https://doi.org/10.1073/PNAS.1808351115/SUPPL_FILE/PNAS.1808351115.SAPP.PDF

88. Sionov RV, Steinberg D. Targeting the Holy Triangle of Quorum Sensing, Biofilm Formation, and Antibiotic Resistance in Pathogenic Bacteria. *Microorganisms* 2022;**10**:1239. https://doi.org/10.3390/MICROORGANISMS10061239

89. Preda VG, Săndulescu O. Communication is the key: biofilms, quorum sensing, formation and prevention. *Discoveries* 2019;**7**:e100. https://doi.org/10.15190/D.2019.13

90. Patel JK, Archana G. Engineered production of 2,4-diacetylphloroglucinol in the diazotrophic endophytic bacterium Pseudomonas sp. WS5 and its beneficial effect in multiple plant-pathogen systems. *Applied Soil Ecology* 2018;**124**:34–44. https://doi.org/10.1016/J.APSOIL.2017.10.008

91. Balthazar C et al. Pyoluteorin and 2,4-diacetylphloroglucinol are major contributors to Pseudomonas protegens Pf-5 biocontrol against Botrytis cinerea in cannabis. *Front Microbiol* 2022;**13**:945498. https://doi.org/10.3389/FMICB.2022.945498/BIBTEX

92. Dong X et al. A vast repertoire of secondary metabolites potentially influences community dynamics and biogeochemical processes in cold seeps. *Sci Adv* 2024;**10**. https://doi.org/10.1126/SCIADV.ADL2281/SUPPL_FILE/SCIADV.ADL2281_TABLES_S1_TO_S12.ZIP

93. Carroll CS, Moore MM. Ironing out siderophore biosynthesis: a review of non-ribosomal peptide synthetase (NRPS)-independent siderophore synthetases. *Crit Rev Biochem Mol Biol* 2018;**53**:356–381. https://doi.org/10.1080/10409238.2018.1476449

94. Sorrels CM, Proteau PJ, Gerwick WH. Organization, Evolution, and Expression Analysis of the Biosynthetic Gene Cluster for Scytonemin, a Cyanobacterial UV-Absorbing Pigment. *Appl Environ Microbiol* 2009;**75**:4861. https://doi.org/10.1128/AEM.02508-08

95. Richter AA et al. Biosynthesis of the Stress-Protectant and Chemical Chaperon Ectoine: Biochemistry of the Transaminase EctB. *Front Microbiol* 2019;**10**:488821. https://doi.org/10.3389/FMICB.2019.02811/XML

96. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery 2015 14:2* 2015;**14**:111–129. https://doi.org/10.1038/nrd4510

97. Jensen PR et al. Challenges and Triumphs to Genomics-Based Natural Product Discovery. *J Ind Microbiol Biotechnol* 2013;**41**:203. https://doi.org/10.1007/S10295-013-1353-8

98. 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. https://doi.org/10.1093/NAR/GKAF334

99. Hannigan GD et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* 2019;**47**:e110–e110. https://doi.org/10.1093/NAR/GKZ654

100. Carroll LM et al. Accurate de novo identification of biosynthetic gene clusters with GECCO. *bioRxiv* 2021;2021.05.03.442509. https://doi.org/10.1101/2021.05.03.442509

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

102. 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. https://doi.org/10.1093/NAR/GKAD984

103. Navarro-Muñoz JC et al. A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology 2019 16:1* 2019;**16**:60–68. https://doi.org/10.1038/s41589-019-0400-9

104. Kautsar SA et al. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience* 2021;**10**:1–17. https://doi.org/10.1093/GIGASCIENCE/GIAA154

105. Gilchrist CLM et al. cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. *Bioinformatics Advances* 2021;**1**:vbab016. https://doi.org/10.1093/BIOADV/VBAB016

106. Salamzade R et al. Evolutionary investigations of the biosynthetic diversity in the skin microbiome using IsaBGC. *Microb Genom* 2023;**9**:000988. https://doi.org/10.1099/MGEN.0.000988/CITE/REFWORKS

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

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

109. 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. https://doi.org/10.1073/PNAS.0506758102

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

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

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

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

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

115. 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. https://doi.org/10.3389/FPLS.2025.1539151/BIBTEX

116. 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. https://doi.org/10.1186/S12864-021-08223-8/FIGURES/7

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

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

119. Rasko DA et al. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol* 2008;**190**:6881–6893. https://doi.org/10.1128/JB.00619-08

120. den Bakker HC et al. Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of Salmonella enterica. *BMC Genomics 2011 12:1* 2011;**12**:1–11. https://doi.org/10.1186/1471-2164-12-425

121. Zhou Z et al. Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Current Biology* 2018;**28**:2420-2428.e10. https://doi.org/10.1016/J.CUB.2018.05.058/ASSET/D381E2FF-C663-4D9D-A3EC-6AD22624C26D/MAIN.ASSETS/GR3.JPG

122. Delmont TO, Eren EM. Linking pangenomes and metagenomes: The Prochlorococcus metapangenome. *PeerJ* 2018;**2018**:e4320. https://doi.org/10.7717/PEERJ.4320/SUPP-8

123. Bär D et al. Origin of the 3-methylglutaryl moiety in caprazamycin biosynthesis. *Microb Cell Fact* 2022;**21**:1–15. https://doi.org/10.1186/S12934-022-01955-6/FIGURES/5

124. Ragozzino C et al. Integrated genome and metabolome mining unveiled structure and biosynthesis of novel lipopeptides from a deep-sea Rhodococcus. *Microb Biotechnol* 2024;**17**:e70011. https://doi.org/10.1111/1751-7915.70011

125. Paccagnella D et al. PanBGC: A Pangenome-inspired framework for comparative analysis of biosynthetic gene clusters. *bioRxiv* 2025;2025.08.11.669102. https://doi.org/10.1101/2025.08.11.669102

126. Anani H et al. Interest of bacterial pangenome analyses in clinical microbiology. *Microb Pathog* 2020;**149**:104275. https://doi.org/10.1016/J.MICPATH.2020.104275

127. Aggarwal SK et al. Pangenomics in Microbial and Crop Research: Progress, Applications, and Perspectives. *Genes 2022, Vol 13, Page 598* 2022;**13**:598. https://doi.org/10.3390/GENES13040598

128. Kim Y et al. Current status of pan-genome analysis for pathogenic bacteria. *Curr Opin Biotechnol* 2020;**63**:54–62. https://doi.org/10.1016/J.COPBIO.2019.12.001

129. Livingstone PG, Morphew RM, Whitworth DE. Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. *Front Microbiol* 2018;**9**:428423. https://doi.org/10.3389/FMICB.2018.03187/BIBTEX

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

131. Caugant DA, Brynildsrud OB. Neisseria meningitidis: using genomics to understand diversity, evolution and pathogenesis. *Nature Reviews Microbiology 2019 18:2* 2019;**18**:84–96. https://doi.org/10.1038/s41579-019-0282-6

132. Yang Z et al. Pangenome graphs in infectious disease: a comprehensive genetic variation analysis of Neisseria meningitidis leveraging Oxford Nanopore long reads. *Front Genet* 2023;**14**:1225248. https://doi.org/10.3389/FGENE.2023.1225248/BIBTEX

133. Lu QF et al. Genus-Wide Comparative Genomics Analysis of Neisseria to Identify New Genes Associated with Pathogenicity and Niche Adaptation of Neisseria Pathogens. *Int J Genomics* 2019;**2019**:6015730. https://doi.org/10.1155/2019/6015730

134. Pizza M, Rappuoli R. Neisseria meningitidis: pathogenesis and immunity. *Curr Opin Microbiol* 2015;**23**:68–72. https://doi.org/10.1016/J.MIB.2014.11.006

135. Vernikos GS. A Review of Pangenome Tools and Recent Studies. *The Pangenome: Diversity, Dynamics and Evolution of Genomes* 2020;89–112. https://doi.org/10.1007/978-3-030-38281-0_4

136. Reis AC, Cunha M V. The open pan-genome architecture and virulence landscape of mycobacterium bovis. *Microb Genom* 2021;**7**:000664. https://doi.org/10.1099/MGEN.0.000664/CITE/REFWORKS

137. Argemi X et al. Comparative genomic analysis of Staphylococcus lugdunensis shows a closed pan-genome and multiple barriers to horizontal gene transfer. *BMC Genomics* 2018;**19**:621. https://doi.org/10.1186/S12864-018-4978-1

138. Dewar AE et al. Bacterial lifestyle shapes pangenomes. *Proc Natl Acad Sci U S A* 2024;**121**:e2320170121. https://doi.org/10.1073/PNAS.2320170121/SUPPL_FILE/PNAS.2320170121.SAPP.PDF

139. Castro JC, Brown SP. Modular gene interactions drive modular pan-genome evolution in bacteria. *bioRxiv* 2023;2022.11.15.516554. https://doi.org/10.1101/2022.11.15.516554

140. Ballouz S et al. Conditions for the Evolution of Gene Clusters in Bacterial Genomes. *PLoS Comput Biol* 2010;**6**:e1000672. https://doi.org/10.1371/JOURNAL.PCBI.1000672

141. Khaldi N et al. Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome Biol* 2008;**9**:1–10. https://doi.org/10.1186/GB-2008-9-1-R18/FIGURES/3

142. Deng MR et al. Granaticins and their biosynthetic gene cluster from Streptomyces vietnamensis: Evidence of horizontal gene transfer. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* 2011;**100**:607–617. https://doi.org/10.1007/S10482-011-9615-9/FIGURES/4

143. Freschi L et al. The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biol Evol* 2019;**11**:109–120. https://doi.org/10.1093/GBE/EVY259

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

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

146. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res* 2003;**13**:2178–2189. https://doi.org/10.1101/GR.1224503

147. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;**16**:1–14. https://doi.org/10.1186/S13059-015-0721-2/FIGURES/7

148. Kennedy NW, Comstock LE. Mechanisms of bacterial immunity, protection, and survival during interbacterial warfare. *Cell Host Microbe* 2024;**32**:794–803. https://doi.org/10.1016/j.chom.2024.05.006

149. Ansari MZ et al. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res* 2004;**32**:W405. https://doi.org/10.1093/NAR/GKH359

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

151. 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. https://doi.org/10.1093/NAR/GKAD984

152. Medema MH, Fischbach MA. Computational approaches to natural product discovery. *Nature Chemical Biology 2015 11:9* 2015;**11**:639–648. https://doi.org/10.1038/nchembio.1884

153. Alanjary M et al. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res* 2017;**45**:W42–W48. https://doi.org/10.1093/NAR/GKX360

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

155. Ye S et al. Combinatorial biosynthesis yields novel hybrid argimycin P alkaloids with diverse scaffolds in Streptomyces argillaceus. *Microb Biotechnol* 2022;**15**:2905–2916. https://doi.org/10.1111/1751-7915.14167

156. Hwang S et al. Repurposing Modular Polyketide Synthases and Non-ribosomal Peptide Synthetases for Novel Chemical Biosynthesis. *Front Mol Biosci* 2020;**7**:540848. https://doi.org/10.3389/FMOLB.2020.00087/XML

157. Duckworth BP, Wilson DJ, Aldrich CC. Measurement of Nonribosomal Peptide Synthetase Adenylation Domain Activity Using a Continuous Hydroxylamine Release Assay. *Methods Mol Biol* 2016;**1401**:53. https://doi.org/10.1007/978-1-4939-3375-4_3

158. Calcott MJ, Owen JG, Ackerley DF. Efficient rational modification of non-ribosomal peptides by adenylation domain substitution. *Nature Communications 2020 11:1* 2020;**11**:1–10. https://doi.org/10.1038/s41467-020-18365-0

159. Englund E et al. Expanding Extender Substrate Selection for Unnatural Polyketide Biosynthesis by Acyltransferase Domain Exchange within a Modular Polyketide Synthase. *J Am Chem Soc* 2023;**145**:8822–8832. https://doi.org/10.1021/JACS.2C11027/SUPPL_FILE/JA2C11027_SI_001.XLSX

160. Duckworth BP, Wilson DJ, Aldrich CC. Measurement of Nonribosomal Peptide Synthetase Adenylation Domain Activity Using a Continuous Hydroxylamine Release Assay. *Methods Mol Biol* 2016;**1401**:53. https://doi.org/10.1007/978-1-4939-3375-4_3

161. Kalkreuter E et al. Computationally-guided exchange of substrate selectivity motifs in a modular polyketide synthase acyltransferase. *Nature Communications 2021 12:1* 2021;**12**:1–12. https://doi.org/10.1038/s41467-021-22497-2

162. Dunn BJ, Khosla C. Engineering the acyltransferase substrate specificity of assembly line polyketide synthases. *J R Soc Interface* 2013;**10**. https://doi.org/10.1098/RSIF.2013.0297

163. Beck C, Garzón JFG, Weber T. Recent Advances in Re-engineering Modular PKS and NRPS Assembly Lines. *Biotechnology and Bioprocess Engineering* 2020;**25**:886–894. https://doi.org/10.1007/S12257-020-0265-5/METRICS

164. Pang B, Graziani EI, Keasling JD. Acyltransferase domain swap in modular type I polyketide synthase to adjust the molecular gluing strength of rapamycin. *Tetrahedron Lett* 2022;**112**:154229. https://doi.org/10.1016/J.TETLET.2022.154229

165. Nguyen KT et al. Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proc Natl Acad Sci U S A* 2006;**103**:17462. https://doi.org/10.1073/PNAS.0608589103

166. Kries H, Niquille DL, Hilvert D. A Subdomain Swap Strategy for Reengineering Nonribosomal Peptides. *Chem Biol* 2015;**22**:640–648. https://doi.org/10.1016/J.CHEMBIOL.2015.04.015

167. Li L, MacIntyre LW, Brady SF. Refactoring biosynthetic gene clusters for heterologous production of microbial natural products. *Curr Opin Biotechnol* 2021;**69**:145. https://doi.org/10.1016/J.COPBIO.2020.12.011

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

169. 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. https://doi.org/10.1002/ANIE.201303196

170. 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. https://doi.org/10.1002/CHEM.201500757

171. Barona-Gómez F et al. Identification of a cluster of genes that directs desferrioxamine biosynthesis in Streptomyces coelicolor M145. *J Am Chem Soc* 2004;**126**:16282–16283. https://doi.org/10.1021/JA045774K

172. Crosa JH, Walsh CT. Genetics and Assembly Line Enzymology of Siderophore Biosynthesis in Bacteria. *Microbiology and Molecular Biology Reviews* 2002;**66**:223. https://doi.org/10.1128/MMBR.66.2.223-249.2002

173. 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. https://doi.org/10.1093/NAR/GKAF334

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

175. Zhu S et al. Computational advances in biosynthetic gene cluster discovery and prediction. *Biotechnol Adv* 2025;**79**:108532. https://doi.org/10.1016/J.BIOTECHADV.2025.108532

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

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

178. Lai Q et al. Deciphering the biosynthetic potential of microbial genomes using a BGC language processing neural network model. *Nucleic Acids Res* 2025;**53**:305. https://doi.org/10.1093/NAR/GKAF305

179. Louwen JJR, Medema MH, van der Hooft JJJ. Enhanced correlation-based linking of biosynthetic gene clusters to their metabolic products through chemical class matching. *Microbiome* 2023;**11**:1–12. https://doi.org/10.1186/S40168-022-01444-3/TABLES/1

180. Leão TF et al. NPOmix: A machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters. *PNAS Nexus* 2022;**1**:1–15. https://doi.org/10.1093/PNASNEXUS/PGAC257

181. Wang M et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology 2016 34:8* 2016;**34**:828–837. https://doi.org/10.1038/nbt.3597

182. Liu M, Li Y, Li H. Deep Learning to Predict the Biosynthetic Gene Clusters in Bacterial Genomes. *J Mol Biol* 2022;**434**:167597. https://doi.org/10.1016/J.JMB.2022.167597

183. Yan B et al. Recent advances in deep learning and language models for studying the microbiome. *Front Genet* 2024;**15**:1494474. https://doi.org/10.3389/FGENE.2024.1494474/BIBTEX

184. Vitale R et al. Evaluating large language models for annotating proteins. *Brief Bioinform* 2024;**25**:177. https://doi.org/10.1093/BIB/BBAE177