ARTICLE

Genome Mining of a Novel Marine Sponge Symbiont *Nocardia* sp. BML-15-R-026U Reveals High Biosynthetic Potential for Secondary Metabolites, Including a Non-Ribosomal Peptide and a Polyketide of High Novelty

Paul Christian T. Gloria[1,2], Elaine M. Romines[3], Marc Jeremie D. Punzalan[3], Christine Marie C. Florece[3], Kreighton E. Cadorna[3], Lilibeth A. Salvador-Reyes[3], and Arturo O. Lluisma*[3]

[1]Natural Sciences Research Institute, University of the Philippines, Diliman, Quezon City
[2]Institute of Biology, University of the Philippines, Diliman, Quezon City
[3]Marine Science Institute, University of the Philippines, Diliman, Quezon City

ABSTRACT

Antibiotic and drug resistance poses serious global public health threats, leading to substantial infections and fatalities annually. Addressing these issues requires the discovery of novel bioactive compounds and a faster and more cost-effective discovery process. However, traditional approaches, which require isolation and multi-step purification of compounds from organisms and running of initial assays, suffer from serious limitations such as the need for substantial amounts of biological material and high rates of compound rediscoveries. Because the biosynthetic capabilities of organisms are encoded in their genomes, genome mining provides a promising solution that would complement traditional approaches. This study conducted long-read whole genome sequencing on a marine sponge symbiont, *Nocardia* sp. BML-15-R-026U, to explore its genomic repertoire of secondary metabolite-encoding Biosynthetic Gene Clusters (BGCs). A four-contig genome

*Corresponding author
Email Address: aolluisma@up.edu.ph
Date received: September 10, 2023
Date revised: October 27, 2023
Date accepted: October 31, 2023

KEYWORDS

Genome mining, *Nocardia*, secondary metabolite prediction, non-ribosomal peptides, polyketides

assembly was generated for this isolate with a high degree of completeness and an estimated genome size of 4.84 Mbp. Its genome displays remarkable biosynthetic potential by containing at least 34 distinct secondary metabolite BGCs, predominantly Non-Ribosomal Peptide Synthetase (NRPS) and Polyketide Synthase (PKS) systems capable of producing novel chemical structures. Further analysis was focused on two genomic regions. In region 3.10, the study predicted a BGC for a novel, serine-rich non-ribosomal peptide with a predicted molecular weight of 2754 g/mol. Region 3.12 contained an iterative type-I PKS BGC, suggesting the potential synthesis of a polyketide compound with oxidoreductase-inhibiting properties. This study highlights genome mining as a productive early-phase approach for identifying promising drug leads and has identified the most promising candidates among this isolate's BGCs for experimental validation.

INTRODUCTION

Antibiotic- and drug-resistant infections are a significant public health burden globally, causing 2.8 million infections and 35,000 deaths annually in the United States (WHO 2020; CDC 2019). This alarming statistic forebodes the rapid increase of more drug-resistant pathogens, hastened community spread, and significant burdens in the future. Thus, there is a pressing need to discover drug leads for more efficacious and novel compounds.

Nature has provided us with abundant natural products that can be exploited as drugs. However, traditional top-down approaches which entail the time-consuming and arduous effort of sampling, screening, and testing various isolates are counterproductive to the drug-resistance problem (Luo et al. 2014). Thus, there is a need to employ rapid screening efforts that may be accomplished by emerging bottom-up techniques like genome mining. The widespread availability of next-generation sequencing platforms along with the abundance of bioinformatics tools has allowed researchers to exploit this strategy for natural product discovery.

Microorganisms are ubiquitous and easy sources of genomic information (Bachmann et al. 2014). Their metabolic and physiological diversity also offers a plethora of drug leads. As such, microbial genome mining has seen significant strides in the recent decade. Identification of biosynthetic gene clusters (BGCs) and predicted compounds has directed scientists into a more reproducible, more efficiently dereplicated, and viable approach for interrogating novel chemistries heterologously.

Actinobacteria produce more than 70% of the natural product scaffolds that are commercially exploited (Gomez-Escribano et al. 2016). They also have a high BGC-to-genome size ratio which makes them good candidates for genome mining. A study by Arulprakasam et al. (2021) highlighted that BGCs are densely found across Actinobacteria and that there is some degree of conservation across the taxa. Researchers also highlighted the utility of various bioinformatics tools for exploring BGCs. Another study on *Streptomyces xinghaiensis* NRRL B-24674 showed that complementary genome mining and *in vitro* studies offer significant strides in the discovery of novel biosynthetic compounds (Chen et al. 2018). Tandem genome-metabolome profiling in *Nocardia* species has been shown to provide efficient discovery of novel non-ribosomal peptides (Chang et al. 2023). Seven new lipopeptides were identified and confirmed with their biosynthesis pathways validated. Baltz (2021) reviews that genome mining for drug discovery has become a major driver for the discovery of new BGCs but it is not an end-all-be-all solution. The review argues that while this technology has expedited discovery there is still a need to do intensive confirmatory and bioactivity assays. Yet, we hope to see the application of genome mining in Philippine marine microbial isolates to further push the frontiers of drug discovery in the country.

In this study, the genome of a marine sponge symbiont *Nocardia sp.* was subjected to long reads whole genome sequencing and mined for secondary metabolite-encoding BGCs to evaluate its biosynthetic potential. In this report, an analysis of two BGC regions is highlighted to reveal their modular organization, backbone composition, and novel biosynthetic logic.

MATERIALS AND METHODS

Sample Provenance and Processing

Nocardia sp. (BML-15-R-026U) was previously isolated from an undescribed marine sponge (BML-15) collected from from Bolinao, Pangasinan by the Marine Pharmacognosy Laboratory under the Discovery and Development of Health Products – Marine Component Program of the Marine Science Institute, University of the Philippines Diliman. Growth of *Nocardia* sp. was first observed in cultures derived from the homogenized inner tissue of the BML-15 marine sponge using the minimal medium R2A (Reasoner's 2A) agar. Hence, propagation and purification of the sponge symbiont were subsequently performed using the same culture medium.

To obtain a sufficient amount of cell pellets, a single colony of BML-15-R-026U was inoculated in 7 mL modified R2A broth (i.e., R2A broth with 1% NaCl, 0.5 g/L L-pyroglutamic acid, and 0.5 g/L peptone) and incubated at 30 °C for 48 h. After incubation, 10 mL of the 48 h culture was transferred to a fresh 100 mL modified R2A broth and incubated again at 30 °C for 48-72 h. Cell pellets were then collected by centrifugation at 5000 RPM, 4 °C for 30 min. High molecular weight (HMW) genomic DNA (gDNA) was extracted and purified from the cell pellets using a modified xanthogenate protocol (Tillet et al. 2000).

Genome Sequencing and Assembly and Annotation

The extracted gDNA together with other marine microorganisms (MMOs) gDNAs were pooled to perform mock metagenome sequencing using the PacBio Sequel I platform at the DNA Core Sequencing Facility of the Philippine Genome Center. Firstly, the genomic library was prepared following the SMRTbell express template preparation kit version 2.0 for large insert gDNA libraries. SMRT cell version 3 for long reads (LR) was used for sequencing with a movie time of 15 h. The sequencing yielded 686,665 reads with an average read length of 10,930 bp for a total of 7.51 Gbps. An in-house binning strategy was conducted to construct the genome assembly of BML-15-R-026U. Briefly, the PacBio raw BAM data was sorted and converted to FASTQ (Andrews 2010) using samtools v1.10 (htslib 1.10.2-3). Metaflye v.2.8.3-b1695 (Kolmogorov et al. 2020) was used to assemble the PacBio raw reads with the parameters –pacbio-raw –meta –plasmids. Barrnap v.09 (Seemann 2013) was used to extract the ribosomal RNA features from the assembly. 16S rRNA data from the previous PCR experiments and the barrnap results were used as the basis for generating the list of reference genomes. A total of 11 high-quality reference genomes were downloaded from NCBI's genome resources (https://www.ncbi.nlm.nih.gov/genome/) (Supplementary Table 1). Binning of the assembled contigs to their respective MMO was conducted by a combination of MetaBat2 v2.15 (Kang et al. 2019) and BLAST search. Briefly, the Continuous Long Reads (CLRs) were mapped to the assembly using minimap2 with the parameter -ax map-pb then MetaBat2 v2.15 with the parameter –percentIdentity 91. BLAST search against a local database comprised of the reference genomes was then used to evaluate the MetaBat2 binned contigs

using a command line installation of Nucleotide-Nucleotide BLAST v2.9.0+ with the parameters -evalue 1E-5 -outfmt 6 -max_target_seqs 5. Quality assessment of the binned bacterial genome assemblies was performed using the QUAST tool v5.0.2 (Gurevich et al. 2013) and CheckM v1.1.9 (Parks et al. 2015).

Afterward, the final genome was annotated using the tool Rapid Annotation using Subsystem Technology (RAST v 2.0) (Brettin et al. 2015) and Prokka v 1.14.5 (Seeman 2014) under default settings. Genome quality was assessed using QUAST v 5.2 (Gurevich et al. 2013) while genome completeness was checked using Benchmarking Universal Single-Copy Orthologs (BUSCO v 5.0) (Simao et al. 2015) and CheckM v 1.1.9 (Parks et al. 2015).

Whole-genome sequence map was generated in Proksee (https://proksee.ca/) (Grant et al. 2023). Additional annotation was done using the Proksee built-in tool Alien_hunter v 1.0.0 (Vernikos and Parkhill 2006) to predict genome regions where Horizontal Gene Transfer (HGT) events may have occurred. Functional annotation from Prokka and BGC prediction results from antibiotics and Secondary Metabolite Analysis SHell (antiSMASH) v 7.0 (Blin et al. 2023) were highlighted in the genome map. Finally, the taxonomic classification of the genome was identified using the Genome Database Taxonomy Toolkit (GTDB-Tk) v 2.0.0 (Chaumeil et al. 2019) and Fast Average Nucleotide Identity (FastANI) v 1.33 (Jain et al. 2018) analysis using all the 490 Nocardia Genbank genomes downloaded in NCBI Genomes database as of August 2023.

Genome Mining for Secondary Metabolites

Antibiotics and secondary metabolite analysis shell (antiSMASH v 7.0) (Blin et al. 2023), using a relaxed detection strictness with all extra features turned on, was employed to detect secondary metabolites BGCs in the genome. Prediction Informatics for Secondary Metabolomes (PRISM v 4.4.5) (Skinnider et al. 2020) using default settings was also used for BGC prediction. Analyses of results mainly relied on antiSMASH but PRISM was occasionally checked for validation.

NCBI BLASTp under default settings was used to identify the closest homolog of each BGC member. For the synthases, NCBI domain enhanced lookup time accelerated (DELTA) BLASTp was used to investigate the adenylation/acylation domain homologs. NRPSPredictor 2 (Rottig et al. 2011) was used to re-validate adenylation/acylation domain-specificity as predicted from antiSMASH or as predicted from the closest homolog if no annotation is available for the domain. HMMsearch was used from HMMER v.3.3.2 (Eddy 2011) to ascertain the condensation domain sub-type of the modules. The profile HMM used was obtained from Raush et al. (2007).

BiG-Fam v 1.0.0 (https://bigfam.bioinformatics.nl/) was used to ascertain the gene cluster family associated with the investigated BGC regions. Afterward, the core and putative members of the GCF were noted. Lastly, Antibiotic Resistance Target Seeker v.2.0 (Mungan et al. 2020) was used to check for antibiotic resistance elements and gene duplications present within the BGCs.

RESULTS

Assembly and Annotation of the *Nocardia* sp. BML-15-R-026U genome

Genome assembly yielded 4,843,841 bp from four contigs and scaffolds namely: scaffold_136 with 4,140,870 bp, scaffold_221 with 300,960 bp, contig_218 with 209,047bp, and contig_111 with 192,964 bp with an average sequencing coverage of 36X. Assessment of genome completeness using BUSCO revealed two missing, three fragmented, and 353 complete orthologs out of 355 single-copy orthologs under Class Actinomycetes. Genome quality estimated in CheckM v1.1.9 (Parks et al. 2015) using the 574 marker genes of Order Actinomycetales reported 98.41% completeness and only 0.75% redundancy. Genome annotation using Prokka revealed 4,237 coding sequences, 39 miscellaneous RNAs, nine ribosomal RNAs, five repeat regions, 54 transfer RNAs, and one transfer-messenger RNA. RAST, on the other hand, detected 4,683 coding sequences and 58 RNA encoding genes. Significant genome features are reported in Figure 1.

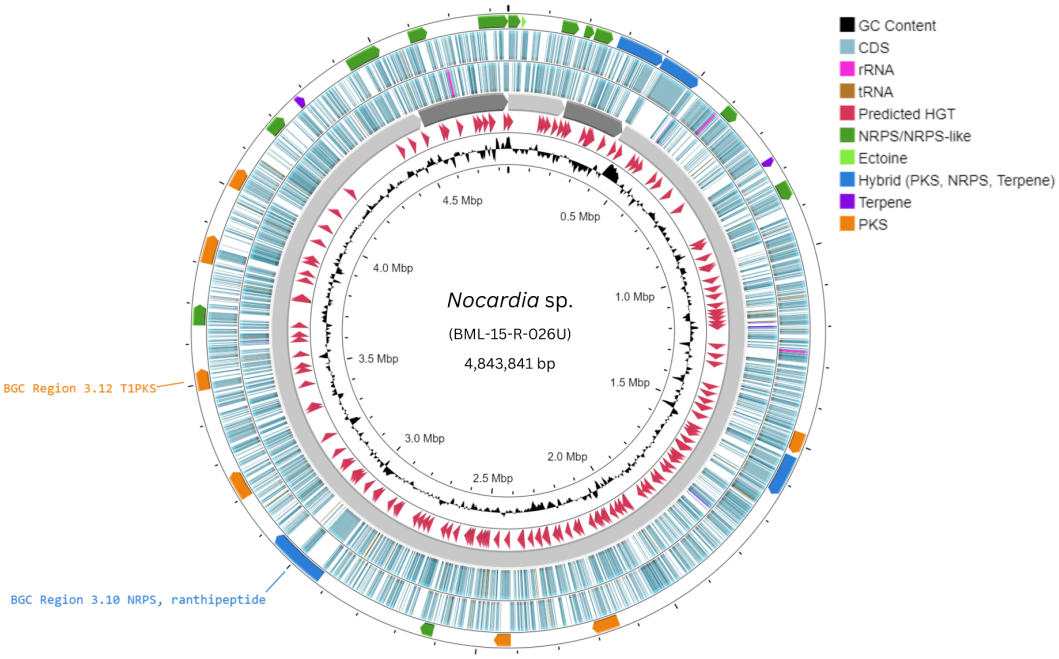


Figure 1: Genome Map of *Nocardia* sp. (BML-15-R-026U) using Proksee (Grant et al. 2023). Tracks indicate GC content, coding regions (CDS), ribosomal RNA (rRNA), transfer RNA (tRNA), predicted horizontal gene transfer (HGT), non-ribosomal peptide synthases (NRPS), and NRPS-like, ectoine, hybrid BGCs, terpenes, and polyketide synthases (PKS).

Furthermore, RAST annotation showed that 20% of the genome can be classified into a subsystem while 80% cannot (Figure 2). A total of 904 genes were binned into a subsystem of which 868 are non-hypothetical proteins while 36 are hypothetical. A total of 3,709 genes cannot be classified into a subsystem of which 1903 genes are non-hypothetical while 1876 are hypothetical proteins. The most abundant subsystem is that of the amino acid and derivatives (240 genes), followed by carbohydrates (160 genes), cofactors, vitamins, prosthetic groups, and pigments (156 genes), and fatty acids, lipids, and isoprenoids (130 genes).

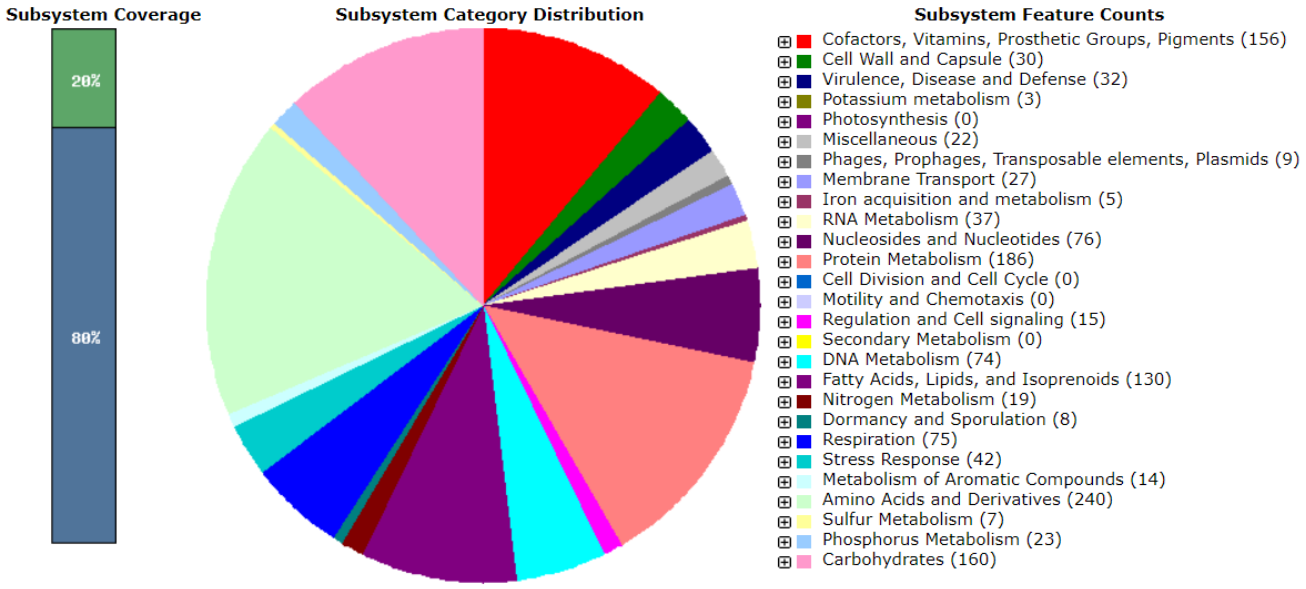


Figure 2: RAST-predicted (Brettin et al. 2015) subsystems and subsystems feature distribution in the annotated *Nocardia* sp. BML-15-R-026U genome.

Taxonomic classification of the genome using GTDB-Tk reported *Nocardia* sp. (GCF_000482385) as its closest match based on the 80.37% ANI value (average nucleotide identity) and *Nocardia pneumoniae* NBRC 100136 based on the 95.65% MSA AA identity value (multiple-sequence alignment amino acid). Moreover, FastANI analysis between the BML-15-R-026U genome and all 490 Nocardia Genbank genomes reported only around 78-82% ANI value with *Nocardia pneumoniae* NBRC 100136 (GCA_000308755.1) at 82.28% being the highest. Consequently, the BML-15-R-026U genome was not assigned to the closest species as it falls outside the suggested ANI threshold for pairwise Nocardia and prokaryotic genomes of at least 95%-96% (Xu et al. 2022; Konstantinidis et al. 2005). Together, these taxonomic results indicate that the isolate represents a novel *Nocardia* species based on molecular data; the isolate shall be designated as *Nocardia* sp. BML-15-R-026U from hereon. Being a novel species, this isolate is thus a promising candidate for genome exploration as novel species, may contain unique drug leads, as further suggested by Supplemental Figure 1.

Biosynthetic Gene Clusters encoding Secondary Metabolites

AntiSMASH predicted 26 unique BGC regions while PRISM predicted 29 BGCs. Non-ribosomal peptides (13-16 NRPS) dominate the BGCs predicted by both platforms followed by polyketides (8-10 PKSs) (Table 1). After collating the predicted BGCs and their nucleotide positions, 34 unique BGCs (Table 2) predicted from both platforms were identified.

Table 1: Summary of BGC type predicted from the *Nocardia* sp. (BML-15-R-026U) genome using antiSMASH 7.0 (Blin et al. 2023) and PRISM 4.4.5 (Skinnider et al. 2020).

	AntiSMASH	PRISM
NRPS	13	16
PKS	10	8
NRPS-PKS hybrid	0	2
Terpene	3	0
NRPS-like (RiPPs)	3	0
PKS-like	1	0
Ectoine	1	1
Unknown thiotemplated cluster	0	1

Table 2: Summary of pertinent information on the different biosynthetic gene clusters encoding secondary metabolites as predicted by antiSMASH 7.0 (Blin et al. 2023) and PRISM 4.4.5 (Skinnider et al. 2020).

		antiSMASH			PRISM		
Contig/Scaffold	**No. of "Unique" BGC**	**Cluster Name (Region)**	**Predicted Class of BGC**	**Nucleotide Position**	**Cluster Name**	**Predicted Class of BGC**	**Nucleotide Position**
Contig_111 (192,964bp)	1	1.1	NRPS	1-30996	1	NRPS	2 – 10,996
	2	1.2	Ectoine	33,907 - 44,305	2	Ectoine	38,906 – 39,305
	3	1.3	NRPS-like	135,407 - 178,613			
Contig_218 (209,047bp)	4	2.1	NRPS	1 - 24,520	3	NRPS	0 – 8,494
	5	2.2	NRPS	26,144 - 73,930	4	NRPS	46,091 – 57,123
	6				5	Angucycline-type PKS	104,237 – 135,250

	7	2.3	T2PKS, PKS-like, T1PKS, NRPS	88,399 - 209,047	6	NRPS	174,019 – 209,047
Scaffold_136 (4,140,870bp)	8	3.1	T1PKS, NRPS	1 - 105,145	7	PKS-NRPS	0 – 85,889
	9	3.2	NRPS-like	193,722 - 233,899			
	10	3.3	Terpene	358,373 - 378,127			
	11	3.4	NRPS	425, 381 – 475, 006	8	NRPS	445,148 - 455,021
	12				9	PKS	468,122 - 470,826
	13	3.5	T1PKS	1,066,936 - 1,121,863	10	PKS	1,086,197 - 1,102,194
	14				11	PKS-NRPS	1,086,197 - 1,102,194
	15	3.6	NRPS, terpene	1,129,695 - 1,241,496	12	NRPS	1,192,859 - 1,223,123
	16	3.7	T1PKS	1,739,895 - 1,808,109	13	PKS	1,759,416 - 1,765,767
	17				14	Unknown thiotemplated cluster	1,783,522 - 1,795,106
	18	3.8	T1PKS	2,013,789 - 2,056,568	15	PKS-NRPS	2,028,996 - 2,038,727
	19				16	NRPS	2,092,590 - 2,095,876
	20	3.9	NAPAA	2,208,993 - 2,242,916			
	21	3.10	NRPS, ranthipeptide	2,521,200 - 2,672,080	17	NRPS	2,541,054 - 2,588,365
	22				18	NRPS	2,601,679 – 2,659,767
	23	3.11	T2PKS	2,794,498 - 2,865,546	19	Angucycline-type PKS	2824140 - 2834477
	24	3.12	T1PKS	3,082,502 - 3,135,854	20	PKS	3101121 - 3122535
	25	3.13	NRPS	3,244,367 - 3,297,666	21	NRPS	3264059 to 3277349
	26	3.14	T2PKS	3,399,882 - 3,472,361	22	Angucycline-type PKS	3419192 to 3440618
	27	3.15	T1PKS	3,599,728 - 3,653,391	23	PKS	3619727 to 3633391
	28	3.16	NRPS	3,766,149 - 3,814,545	24	NRPS	3785681 to 3798024
	29	3.17	Terpene	3,864,858 - 3,885,955			
	30	3.18	NRPS	4,022,416 - 4,113,040	25	NRPS	4,042,092 – 4,051,306
	31				26	NRPS	4,086908 - 4,095,547
Scaffold_221 (300,960bp)	32	4.1	NRPS	44,066 - 96,771	27	NRPS	63,421 - 78.244
	33	4.2	NRPS	223,635 - 300,960	28	NRPS	245,097 - 268,829

Antibiotic Resistance and Core Gene Duplications in *Nocardia sp.* Genome

A total of 32 known resistance hits and 19 gene duplication events were identified after running ARTS 2.0. Carboxyl transferase, biotin lipoyl, ABC efflux proteins, and DNA replication proteins were the most common resistance targets. For the duplicated gene, dapA (4-hydroxy-tetrahydrodipicolinate synthase) is the most common with three duplications.

Region 3.10 harbors two resistance hits associated with biotin lipoyl and carboxyl transferase (Figure 3) and duplication for ribosomal protein Bs18. Meanwhile, region 3.12 had no resistance hits but harbors a gene duplication for oxidoreductase (ycfH).

Sequence id	Location (start-end)	Type	Gene	Description	Function
2633	2607700 - 2608459	Core	TIGR01465	cobM_cbiF: precorrin-4 C11-methyltransferase	Biosynthesis of cofactors, prosthetic groups, and carriers
2624	2596626 - 2598576	Core	TIGR01217	ac_ac_CoA_syn: acetoacetate-CoA ligase	Central intermediary metabolism
2643	2658939 - 2659767	Core	TIGR03284	thym_sym: thymidylate synthase	Purines, pyrimidines, nucleosides, and nucleotides
2628	2601300 - 2603337	ResModel	PF00364.17	Biotin_lipoyl: Biotin-requiring enzyme	N/A
2629	2603342 - 2604941	ResModel	PF01039.17	Carboxyl_trans: Carboxyl transferase domain	N/A
2620	2594283 - 2594784	DUF	PF07336.6		
2621	2594909 - 2595773	DUF	PF00892.15		
2631	2605635 - 2606976	DUF	PF13635.1		

Figure 3: ARTS 2.0 (Mungan et al. 2020) results for core genes and resistance targets associated with Region 3.10.

REGION 3.10: Manual BGC annotations

Boundaries

Region 3.10 is 150,881 nucleotides long with 64 open reading frames and is putatively classified by antiSMASH as containing clusters for ranthipeptide and NRP. It also coincides with clusters 17 and 18 predicted by PRISM.

The AntiSMASH results indicated that Region 3.10 contains multiple, possibly overlapping, neighboring clusters. Figure 4 illustrates that the ranthipeptide cluster spans from ctg3_2245 to ctg3_2260. The downstream part of this cluster overlaps with the NRPS cluster. By noting the annotations of the ORFs and annotations derived from the nearest homologous protein using BlastP, ctg3_2254 to ctg3_2258 were determined to be part of a RiPP BGC. Table 3 summarizes the annotation for these ORFs. For these reasons, ctg3_2259 is designated as the first ORF belonging to the NRPS cluster in Region 3.10.

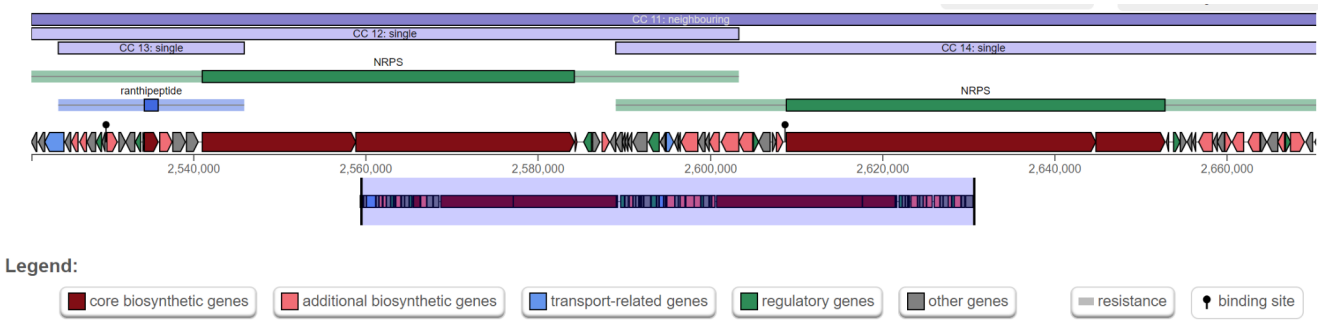


Figure 4: Schematic representation of Region 3.10 displaying ORFs and their annotations and putative BGC boundaries from antiSMASH 7.0 (Blin et al. 2023).

Table 3: Summary of ORFs, their homologs, and annotations associated with Region 3.10 extracted from antiSMASH 7.0 (Blin et al. 2023).

ORF	Classification	TIGR annotation	PFAM annotation	GO Term	Nucleotide length	Protein homology	Organism	Accession Number
ctg3_2254	Regulatory genes		Bacterial regulatory proteins, tetR family	DNA binding	624	MULTISPECIES: mycofactocin system transcriptional regulator MftR2	unclassified *Nocardia*	WP_159924964.1
ctg3_2255	Other genes	mycofactocin: mycofactocin precursor			150	mycofactocin precursor MftA	*Nocardia anaemiae*	WP_063813266.1
ctg3_2256	Core biosynthetic genes	mycofactocin radical SAM maturase	Radical SAM superfamily	iron-sulfur cluster binding	1611	mycofactocin radical SAM maturase	*Nocardia pneumoniae*	WP_040774178.1
ctg3_2257	Additional biosynthetic genes	mycofactocin system glycosyltransferase	Glycosyl transferase family 2		1437	mycofactocin biosynthesis glycosyltransferase MftF	*Nocardia amamiensis*	WP_195129891.1
ctg3_2258	Other genes	Rv0697: dehydrogenase, Rv0697 family	GMC oxidoreductase	oxidation-reduction	1452	mycofactocin system GMC family oxidoreductase MftG	*Nocardia beijingensis*	WP_195044320.1

The predicted genes in the downstream region of the NRPS BGC were inspected more closely to delimit the downstream border of the BGC. The predicted gene, ctg3_2305, was identified to be at the downstream end of the BGC, which putatively encodes for a thioredoxin-containing protein. Another predicted gene, Ctg3_2306, which is downstream of ctg3_2305, was deemed unlikely to be part of the BGC as this encodes for a glucose-6-phosphate (G6P) isomerase. Literature search did not provide any evidence of a BGC utilizing G6P isomerase in their cluster. ARTS also did not designate the G6P isomerase as a resistance element or a gene.

Regulation, resistance, and transport

Region 3.10 contains 5 ORFs encoding for transcriptional regulators like the IcIR family, YbaB/EbfC DNA-binding family, TetR family, GntR family, and a putative stress-induced transcription regulator. TetR and GntR are negative regulators (Deng et al. 2013; Hillerich and Westpheling 2006) while IcIR and YbaB/EbfC can be both negative and positive regulators (Moline-Henarez et al. 2006; Riley et al. 2019). NRP biosynthesis in this region could be under negative control considering the presence of a stress-induced regulatory factor encoded in ctg3_2273.

For transport-related genes, Region 3.10 contains ORFs encoding for an EamA-like family transporter, a hydrolase signal peptide, and a twin-arginine translocation (TAT) signal domain-containing protein. EamA belongs to the protein family drug/metabolite transporter and has been recently characterized as a transporter of nucleotide sugars (Vastermark et al. 2011). However, it also has been shown to function as an L-lysine exporter in another study (Malla et al. 2020). A TAT signal domain is also present in the region, suggesting a potential TAT system-based export of the NRP in region 3.10.

Core biosynthetic enzymes: Adenylation domain specificity

Table 4 summarizes the module architecture of synthases and their adenylation domain specificity. A 26 amino acid-long peptide is expected if all modules are used and arranged in modules. The logical order of the product backbone was determined as described in the next section.

Table 4: Summary of the pertinent information on non-ribosomal peptide synthases encoded in Region 3.10 from antiSMASH (Blin et al. 2023).

Synthase	ORF	Notes	Backbone
1	Ctg3_2260	6 modules with 6th module only containing a condensation domain	Ser-Ser-Ser-Thr-Ser
2	Ctg3_2261	8 modules; First module contains no condensation domain while last module contains a thioesterase domain	(Ser)- Ser-Ser-Orn-Ser-Ser-Tyr-Ser
3	Ctg3_2277	Only an adenylation domain	(Phe)
4	Ctg3_2287	11 modules; last module has no PCP	Ser-Ser-Ser-Ser-Thr-D-Phe-Ser-Ser-Ser-Ser-(Ser)
5	Ctg3_2288	Encodes a PCP, complete module for D-Glu, and a hanging condensation domain.	(D-Glu)

Core biosynthetic enzymes: Module order

Rausch et al. (2007) studied the phylogenetic relationships of condensation domains found in NRPSs. Their analysis showed the presence of four functional sub-types: (1) LCL domains which catalyze the peptide formation between two L-amino acids, (2) DCL domains which link an L-amino acid to a growing peptide ending with a D-amino acid, (3) a starter C domain which adenylates the first amino acid, and (4) heterocyclization domains which catalyze peptide bond formation and subsequent cyclization. The profile Hidden Markov Model (pHMM) they provided was used to ascertain the condensation domain sub-type of the predicted modules in this genome.

The predicted genes Ctg3_2260 and ctg3_2287 were the logical suspects for initiator modules as these are the ones that start with a condensation domain. HMM search revealed that ctg3_2260's first C domain is a starter domain (E-value: 5.2x10-119) while ctg3_2287's first C domain is of the heterocyclization subtype (E-value: 1.8x10). However, antiSMASH annotated these C domains as both starters. Such prediction is supportive of the result in PRISM but the very close proximity of two groups of synthases encoding for entirely different secondary metabolites seems counterintuitive. antiSMASH detection rules also show an entire NRPS protocluster with core elements overlapping all the synthases. ctg3_2260 can thus be identified as the starter module and all the synthases within region 3.10 contribute to a single NRP backbone.

An alternative explanation for the existence of two C starter domains is a two-step colinear production of two peptides that are joined together for the final product. Given this probability, researchers hypothesized a 13 amino acid long peptide as the product of ctg3_2260 and ctg3_2261 and a 12 amino acid long peptide as the product of ctg3_2287 and ctg3_2288. The presence of a thioesterase domain in ctg3_2261 suggests that the 13 amino acid long peptide is cleaved from the PCP and joined head-to-tail on the 12 amino acid product of ctg3_2287 and ctg3_2288. The proposed mechanisms are summarized in Table 5.

Table 5: Peptide Backbone predictions and proposed mechanism for Region 3.10 from antiSMASH (Blin et al. 2023).

Predictions	Core Peptide Backbone	Proposed Mechanism
Case 1: Consecutive linear synthesis	Ser-Ser-Ser-Thr-Ser-Ser-Ser-Ser-Orn-Ser-Ser-Tyr-Ser-Ser-Ser-Ser-Ser-Thr-D-Phe-Ser-Ser-Ser-Ser-Ser-D-Glu-Phe	Backbone production entails the contiguous synthesis of the peptide in a modular arrangement as follows: ctg3_2260, ctg3_2261, ctg3_2287, ctg3_2288, ctg3_2277. In the second synthase, the thioesterase domain is skipped. The last module ends with a Phe and is cleaved from the PCP by borrowing the TE domain from ctg3_2261. This case yields a 26 amino acid long peptide and makes use of all the encoded synthases.

Case 2: Co-linear synthesis followed by condensation	(Ser-Ser-Ser-Ser-Thr-D-Phe-Ser-Ser-Ser-Ser-Ser-D-Glu)+ (Ser-Ser-Ser-Thr-Ser-Ser-Ser-Ser-Orn-Ser-Ser-Tyr-Ser)	Backbone production entails two simultaneously produced peptides. The first peptide is produced by ctg3_2287 and ctg3_2288 while the second peptide is produced by ctg3_2260 and ctg3_2261. The TE domain from ctg3_2261 frees up the second peptide from its PCP and is added head to tail to the first peptide through the condensation domain of ctg3_2261. This case yields a 25 amino acid long peptide with a different backbone arrangement from case 1.

Molecular weight range

Given the proposed backbones, the molecular weight range of the putative product can be predicted (Table 6). It can have a minimum of 2754.54 g/mol if biosynthesis follows the mechanism illustrated for case 2 or a minimum of 2919.78 g/mol for case 1.

Table 6: Summary of amino acid residues and molecular weight ranges of possible peptide backbone for Region 3.10.

Amino Acid/Monomer	Molecular Weight (g/mol)	No. of times used					
		Case 1	Cumulative MW	Case 2	Cumulative MW	PRISM	Cumulative MW
Serine	105.09	19	1891.62	19	1891.62	20	2101.8
Threonine	119.12	2	238.24	2	238.24	2	238.24
Ornithine	132.16	1	132.16	1	132.16	0	0
Tyrosine	181.19	1	181.19	1	181.19	0	0
Phenylalanine	165.19	2	330.38	1	165.19	0	0
Glutamine	146.14	1	146.14	1	146.14	0	0
Tryptophan	204.23	0	0	0	0	3	612.69
		26	**2919.78**	25	**2754.54**		**2952.73**

Additional biosynthetic enzymes

Thirteen ORFs are putatively identified as tailoring enzymes. AntiSMASH annotates tailoring enzymes as "additional biosynthetic genes" but other possible enzymes were annotated as "other genes" and were included in this category by considering their Pfam and TIGR annotations.

A *Mycobacterium tuberculosis* H (mtbH)-like protein was predicted for the predicted gene ctg3_2262. This protein is surmised to play a role in modulating biosynthesis and/or act as a chaperone to adenylation domains in dictating substrate specificity through protein-protein interactions (Baltz 2011). Hence, it may contribute to structural diversification in this NRPS through the creation of congeners.

A carboxylesterase protein is associated with ctg3_2298. Carboxylesterases might harbor thioesterase activity Cantu et al. (2010) and the ctg3_2298-encoded-enzyme might therefore act as a trans-thioesterase domain in the co-linear synthesis of the backbone since the 12 amino acid long part is synthesized from a module not harboring a thioesterase domain.

Other tailoring reactions that may occur include methylation, decarboxylation, addition of a glutamyl residue, and dehydration. One interesting reaction is the addition of an amino acid residue through a PLP-dependent aminotransferase protein family member (ctg3_2299). This presents a mechanism to add an amino acid not contributed by an adenylation domain. There is also an enzyme associated with the conversion of glutamine to a carbamoyl phosphate through a carbamoyl phosphate synthase enzyme (ctg3_2281). This reaction is highly likely as the predicted backbone harbors a D-Glutamine residue.

Additional potential tailoring enzymes were also found and annotated. These include enoyl reductase, tetrapyrrole methyltransferase, and acyl-CoA dehydrogenase which could potentially modify the peptide backbone. However, their specific roles cannot be established yet given just the peptide backbone. Alternatively, they might be non-functional as only certain domains are present and not the holoenzyme. Lastly, a trans-PKS may contribute to the final product which is modified by these enzymes. The current analysis, however, shows no evidence for this.

REGION 3.10: ASSESSMENT OF NOVELTY

Peptide Backbone

As highlighted in Table 5, a serine-rich, 25 amino acid-long peptide backbone can be expected as the product of Region 3.10 BGC. There are two possible backbone structures which are essentially just a rearrangement of the (12/13) +13 amino acid long backbone arising from two distinct groups of NRPSs. To ascertain novelty, we use the NORINE (Caboche et al. 2008) database to search for similar compounds.

Pyoverdine and isopyoverdines are the most common results for the NORINE search (Table 7). Upon checking the structure of each hit, researchers think that the hit is associated with the serine-rich nature of the query. However, the pyoverdines and isopyoverdines are much smaller in size and have a lower number of residues compared to our predicted compound.

Table 7: Monomer backbone search for compound similarity predicted from Region 3.10 using NORINE (Caboche et al 2008). Backbones were based either on the antiSMASH (Blin et al. 2023) default or from the putatively predicted backbone.

Ser*,Ser*,Thr*@1@0,2@1,3@2,4@3 Ser*,Ser*,Ser*@1@0,2@1,3@2,4@3,5@4,6@5,7@6 x,Ser*,Ser*,Thr**,Ser*,Ser*,Ser*@1@0,2@1,3@2,4@3,5@4,6@5,7@6,8@7,9@8 d-x*@1@0			Ser*,Ser*,Ser*,Thr*,Ser*,Ser*,Ser*,Ser*,Orn*,Ser*,Ser*,Ser*,Tyr*,Ser*,Ser*,Ser*,Ser*,Thr*,[Phe*|D-Phe],Ser*,Ser*,Ser*,Ser*,Ser*,D-Glu		
Anti-SMASH default (substrate specificity X)			Defined Substrate Specificity		
Similarity	**Peptide**	**NORINE ID**	**Similarity**	**Peptide**	**NORINE ID**
1.033	Pyoverdin 2912	NOR00182	0.286	Pyoverdin 2192	NOR00182
0.868	Isopyoverdin 90-44	NOR00183	0.224	Pyoverdin 95-275	NOR00191

0.826	Pyoverdine 2908	NOR00167	0.200	Isopyoverdine 90-44	NOR00183
0.819	Pseudobactin A225-I	NOR00908	0.200	Pyoverdine 1.2	NOR00180
0.819	Pyoverdine BTP16	NOR00206	0.189	Pyoverdin 96-318	NOR00194
0.819	Pyoverdin C	NOR00205	0.189	Pyoverdin C-E	NOR00190
0.808	Pyoverdin Pau	NOR00173	0.183	Pyoverdine PaA	NOR01353
0.795	Pyoverdin 95-275	NOR00191	0.179	Pyoverdin 2908	NOR00167
0.765	Pyoverdin 1547	NOR00186	0.172	Pseudobactin A225-I	NOR00908
0.763	Pyoverdin 1.2	NOR00180	0.172	Pseudobactin A225-II	NOR00909

MiBiG cluster comparisons

Based on the MiBiG protocluster-to-region comparisons, syringopeptin 25A is most similar to our predicted product (Figure 5). They both have 25 amino acid residues but syringopeptin has a fatty acid residue in its C-terminal region. Its BGC has regions of homology to three of the NRPSs in Region 3.10, but it is only second to virginiafactin BGC in terms of percent similarity (Table 8). However, syringopeptin 25A is more closely related to our predicted compound due to it being homologous to three synthases while virginiafactin is only homologous to two.

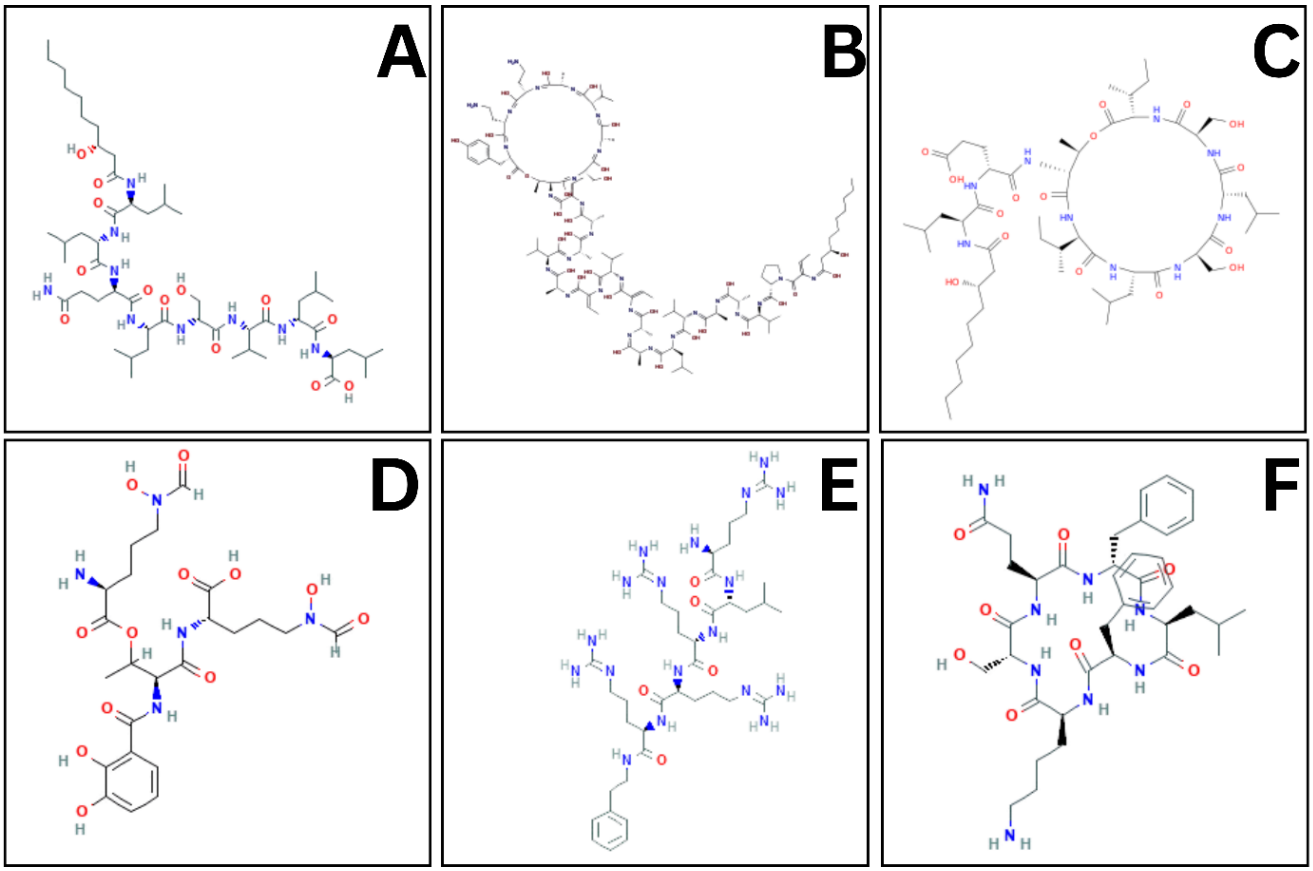


Figure 5: Structures of similar peptides based on MiBiG comparisons. (A) virginiafactin (B) syringopeptin 25A (C) massetolide A (D) rhodochelin (E) bicornutin (F) ambactin. All structures are derived from Pub Chem (Kim et al. 2023).

Table 8: Protocluster to region comparisons of Region 3.10 to MiBiG BGC entries as extracted from antiSMASH 7.0 (Blin et al. 2023).

Reference	Similarity Score	Peptide	Microorganism	Region of similarity
BGC0002071.1	0.47	Virginiafactin	*Pseudomonas sp.* QS1027	Ctg3_2260 Ctg3_2261
BGC0000438.1	0.38	syringopeptin 25A	*Pseudomonas syringae pv. syringae*	Ctg3_2261 Ctg3_2287 Ctg3_2288
BGC0000389.1	0.37	massetolide A	*Pseudomonas fluorescens* SS101	Ctg3_2261 Ctg3_2288
BGC0000417.1	0.36	rhodochelin	*Rhodococcus jostii* RHA1	Ctg3_2288
BGC0001135.1	0.35	bicornutin A1, bicornutin A2	*Xenorhabdus budapestensis*	Ctg3_2288
BGC0001131.1	0.35	ambactin	*Xenorhabdus miraniensis*	Ctg3_2288

Gene Cluster Family

Region 3.10 is classified under GCF_13372. Currently, this GCF has only 1 known member, *Nocardia shimofuensis* (GCF_001613125.1/NZ_BDBT01000002.region003). This GCF is 165.96kb long. Such results highlight the significant novelty of our predicted BGC.

Region 3.10: Insights from PRISM Predictions

Region 3.10 corresponds to clusters 17 and 18 in PRISM. For cluster 17, three ORFs associated with core biosynthetic genes were annotated: orf2265 which encodes for a five-module NRPS, orf2266 eight-module NRPS with a TE domain, and orf2270 TE domain. These two NRPSs correspond to the genes ctg3_2260 and ctg3_2261 predicted by antiSMASH. For cluster 18, four ORFs with core functions were identified: orf30944 for a dehydratase, orf2293 for an eleven-module NRPS, orf2294 for a single-module NRPS, and orf2304 for thymidylate synthase. The two NRPS genes in this cluster correspond to ctg3_2287 and ctg3_2288 in antiSMASH. Table 9 also shows the adenylation domain specificity comparisons between the two programs.

Table 9: Summary of peptide backbone comparisons between antiSMASH 7.0 (Blin et al. 2023). and PRISM v 4.4.5 (Skinnider et al. 2020) predictions.

	ORF (PRISM)	Contig (antiSMASH)	Backbone	Cyclization domains (module location)
17		2260	Ser-Ser-Ser-Thr-Ser	
	2265		Ser-Ser-Ser-Thr-Ser	3rd
		2261	(Ser)- Ser-Ser-Orn-Ser-Ser-Tyr-Ser	
	2266		Ser-Ser-Ser-Ser-Ser-Ser-Trp-Ser	1st, 8th
18		2287	Ser-Ser-Ser-Ser-Thr-D-Phe-Ser-Ser-Ser-Ser-(Ser)	
	2293		Ser-Ser-Ser-Ser-Thr-Trp-Ser-Ser-Ser-Ser-Ser	11th
		2288	(D-Glu)	
	2294		D-Trp	

PRISM predictions also show the presence of cyclization domains which were annotated as normal condensation domains in antiSMASH. Figure 6 highlights possible ring systems arising from these cyclization events. Finally, the predicted product in PRISM has a molecular weight of 2952.73g/mol, just slightly larger than what was predicted by AntiSMASH (2919.78 g/mol for case 1; see Table 6).

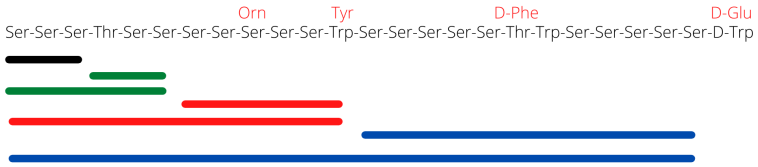


Figure 6: Schematic representation of potential backbone cyclization events. Texts in black represent the backbone predicted in PRISM while red texts are residues predicted in antiSMASH 7.0 (Blin et al. 2023). Lines of the same color represent plausible cyclization events stemming from the same cyclization domain.

Region 3.10: Insights into potential biosynthetic mechanisms from MiBiG hits

So far, results indicate that the BGC in Region 3.10 provides the enzymes for the biosynthesis of a serine-rich 25 to 26 amino acid-long non-ribosomal peptide. The BGC is under negative regulation by multiple transcription factors which might be activated in stressful conditions. Upon induction, two potential biosynthetic mechanisms are proposed: (1) consecutive linear synthesis of a 26 amino acid long peptide that uses a trans-module in the termination module (illustrated in Figure 7) or (2) co-linear synthesis which yields a 25 amino acid long peptide. These two mechanisms result in two similar backbones which are only different by 1 amino acid residue and in the order of backbone constituents.

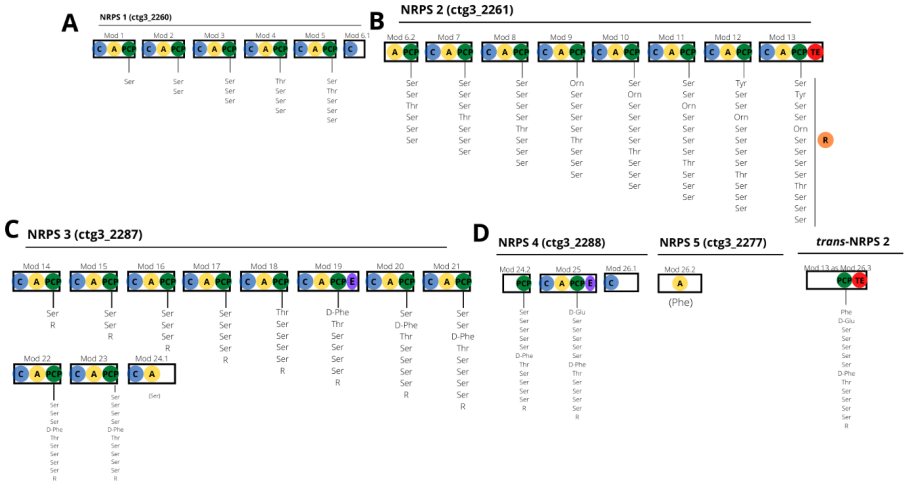


Figure 7: Proposed biosynthetic pathway for NRP compound in Region 3.10. Figures A-D illustrate the linear biosynthesis of the peptide backbone. The R group represents the shorthand notation of amino acids encoded from modules 1 to 13. In Figure D, module 13 from NRPS2 acts as a *trans*-module to supply the PCP and TE domains for the terminal module.

Post-synthesis modification can include methylation, decarboxylation, the addition of glutamyl residues, and amination as highlighted by ORFs encoding for such enzymes. However, the backbone residue that is the target for modification and the extent of modification are yet to be determined.

Biotin lipoyl, carboxyl transferase, and/or ribosomal protein Bs18 are suspected as potential targets of the compound. These were identified using ARTS 2.0 where the former two are resistance elements while the latter is a gene duplication within Region 3.10. Lastly, it is surmised that the peptide is exported out of the cell in its active form through a TAT-based system.

In this section, researchers used MiBiG hits associated with individual synthases and explored the biosynthetic logic of the BGCs which can point to the congruence or departure from researchers' predictions. Table 10 highlights the MiBiG hits for each NRPSs and the associated BGCs. Ctg3_2261 and ctg3_2287 are homologous to speC (46%) and speB (47%) of sarpeptin A BGC, respectively. Sarpeptin A is a hexapeptide encoded by a thirteen-module NRP megasynthase (Koomsiri et al. 2019). However, the underlying genetic elements for this apparent peptide truncation seemed to be absent from region 3.10 and no other similar logic can be gleaned from the comparison.

Table 10: Summary of pertinent BGC information on MiBiG hits based on individual Region 3.12 NRPS from antiSMASH (Blin et al. 2023).

Contig						
ctg3_2260	MiBiG protein	QED55421.1	AKC91856.1	QED88055.1	ALG65317.1	AAF63833.1
	Description	NRPS (speC)	NRPS (griB)	NRPS (adeH)	Cal19	PstD
	MiBiG cluster	BGC0001984	BGC0001414	BGC0001967	BGC0001297	BGC0000362
	% Identity	46	45	45	45	44
	BGC	sarpeptin A biosynthetic gene cluster from *Streptomyces sp.*	griselimycin biosynthetic gene cluster from *Streptomyces muensis*	acyldepsipeptide 1 biosynthetic gene cluster from *Streptomyces hawaiiensis*	WS9326 biosynthetic gene cluster from *Streptomyces calvus*	glycopeptidolipid biosynthetic gene cluster from *Mycobacterium avium*
ctg3_2261	MiBiG protein	AEG64697.1	CAM56770.1	QBC75022.1	QBG38784.1	AGC09528.1
	Description	LpmC	PstC	NRPS (cdeJ)	Atr23	NRPS
	MiBiG cluster	BGC0000379	BGC0000354	BGC0001968	BGC0001975	BGC0001183
	% Identity	44	43	43	43	43
	BGC	glycinocin A biosynthetic gene cluster from *Streptomyces viridochromogenes*	friulimicin A biosynthetic gene cluster from *Actinoplanes friuliensis*	cadaside A biosynthetic gene cluster from uncultured bacterium	atratumycin biosynthetic gene cluster from *Streptomyces atratus*	lobophorin A biosynthetic gene cluster from *Streptomyces sp.* FXJ7.023
ctg3_2277	MiBiG protein	ADI24927.1	AAF81733.1	ADG27358.1	CAH10130.1	BW17377.1

	Description	VrtB	Putative aryl-CoA ligase EncN	Peptide synthetase (acmB)	Feruloyl-CoA synthetase (schA30)	PsoC
	MiBiG cluster	BGC0000168	BGC0000220	BGC0000296	BGC0000268	BGC0000411
	% Identity	34	26	26	26	26
	BGC	viridicatumtoxin biosynthetic gene cluster from *Penicillium aethiopicum*	enterocin biosynthetic gene cluster from *Streptomyces maritimus*	actinomycin D biosynthetic gene cluster from *Streptomyces anulatus*	Sch-47554 biosynthetic gene cluster from *Streptomyces sp.* SCC 2136	putisolvin biosynthetic gene cluster from *Pseudomonas putida*
ctg3_2287	**MiBiG protein**	QED55421.1	AGZ15458.1	ALV86867.1	WP_016638469.1	QED88054.1
	Description	NRPS (speB)	NRPS	Tlo21	NRPS	NRPS (adeG)
	MiBiG cluster	BGC0001984	BGC0001036	BGC0001406	BGC0001519	BGC0001967
	% Identity	47	47	46	46	45
	BGC	sarpeptin A biosynthetic gene cluster from *Streptomyces sp.*	polyoxypeptin biosynthetic gene cluster from *Streptomyces sp.* MK498-98F14	telomycin biosynthetic gene cluster from *Streptomyces canus*	aurantimycin A biosynthetic gene cluster from *Streptomyces aurantiacus* JA 4570	acyldepsipeptide 1 biosynthetic gene cluster from *Streptomyces hawaiiensis*
ctg3_2288	**MiBiG protein**	EFE73312.1	AAF63833.1	QED55421.1	BBD17759.1	WP_079164394.1
	Description	NRPS	PstD	NRPS (SpeC)	NRPS (natB)	NRPS
	MiBiG cluster	BGC0000431	BGC0000362	BGC0001984	BGC0001919	BGC0001567
	% Identity	49	49	47	47	47
	BGC	stenothricin biosynthetic gene cluster from *Streptomyces filamentosus* NRRL 15998	glycopeptidolipid biosynthetic gene cluster from *Mycobacterium avium*	sarpeptin A biosynthetic gene cluster from *Streptomyces sp.*	neoantimycin biosynthetic gene cluster from *Streptomyces orinoci*	cysteoamide biosynthetic gene cluster from *Streptomyces lincolnensis*

Ctg3_2261 is homologous to *LpmC* (44%) of glycinocin A (also known as laspartomycin C) BGC. Glycinocin is a calcium-dependent cyclic lipopeptide (Wang et al. 2011). It contains a non-proteinogenic amino acid DAP (diamino propionic acid) not encoded within the synthase but added by a PLP aminotransferase. The same strategy can be at play for Region 3.10 where a PLP aminotransferase is annotated in ctg3_2299. Hence, the compound can harbor an additional non-proteinogenic amino acid residue or have one of the existing residues modified.

In relation to the previous discussion, researchers also noticed that most of the adenylation domains show significant similarity to A domains of calcium-dependent antibiotics. However, the compound lacks the conserved Asp-X-Asp-Gly motif for calcium binding (Wood. et al 2019).

REGION 3.12: Manual BGC Annotations

Boundaries

Region 3.12 is 53,353 nucleotides long and is putatively classified by antiSMASH to contain a Type 1 Polyketide Synthase cluster (Figure 8). The region contains 61 open reading frames with the synthases encoded by 3 different ORFs located at positions 3,101,122 to 3,117, 465 of the contig.

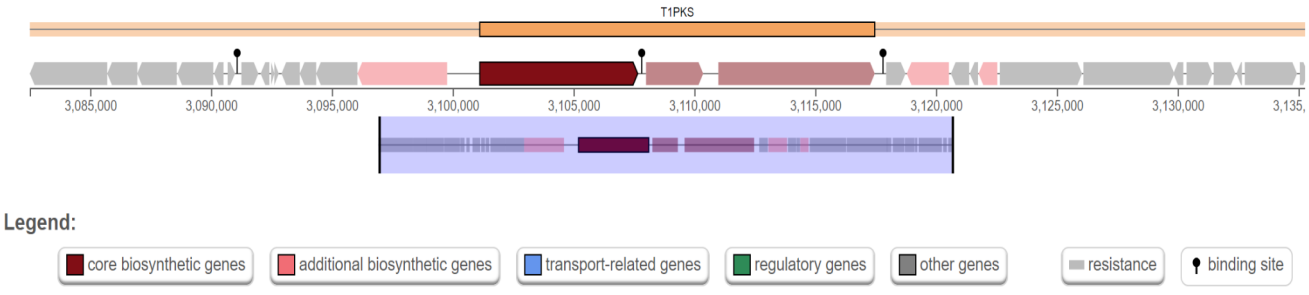


Figure 8: antiSMASH-predicted Region 3.12 BGC contig members and their putative categorical functions (Blin et al. 2023).

Regulation, resistance, and transport

The first three ORFs in the region encode a putative Type I restriction-modification system due to the presence of an endonuclease subunit R (ctg3_2696), endonuclease subunit S (ctg3_2697), and an N-6 DNA methylase (ctg3_2698). This organization mirrors the Type I restriction-modification system in *E. coli* where subunits R and S are responsible for restriction and DNA sequence specification respectively and the methylase for methylation (Mernagh et al. 1998).

The region also encodes three ORFs for nitroreductase subunits and its cofactor chaperone domain (ctg3_2706 to ctg3_2708). A related report suggests that nitrogen reductase is important for *Nocardia*'s tolerance to oxidative stress within host cells (Wu et al. 2006). It was surmised that the presence of these ORFs within the region suggests co-regulation between compound production and response to oxidative stress. The nitrogen reductase ORFs may then serve as a link between primary metabolism and secondary metabolism.

A UvrD domain-containing protein (ctg3_2721) and UvrD2 DNA helicase homolog (ctg3_2724) are also encoded within the region. These DNA helicases unwind DNA for replicative repair (Lee and Yang 2006). One possible reason for UvrD existence within the BGC is that compound production is triggered by DNA damage or reduced genomic integrity. Secondary metabolite production for UV-induced stress has been reported in *Promicromonospora kermanensis* DSM 45485 (Mohammadipanah et al. 2020) and *Aspergilus fumigatus* (Blachowicz et al. 2020).

The gene content of the BGC suggests that the product it produces may be under the influence of phage invasion due to the presence of restriction-modification system ORFs, oxidative stress due to nitrate reductase ORFs, and DNA damage due to UvrD ORFs. A well-established transcriptional family regulator protein is found in ctg3_2726. This ORF encodes for a WhiB-like protein. This family of protein is widespread in actinomycetes and is known to activate transcription in response to different stressors (Geiman et al. 2016). This WhiB-like transcriptional regulator may therefore have a role in the transcriptional induction of the BGC in Region 3.12.

Core biosynthetic enzymes

The region harbors iterative type I polyketide synthases, predicted to be encoded by ctg3_2710, ctg3_2711, and ctg3_2712 (Figure 8). All their ketosynthases have substrate specificity for malonyl-CoA. Ctg3_2710 and ctg3_2712 contain a dehydratase domain, enoyl reductase, and ketoreductase domains along with an acyl carrier protein. The ketoreductase of ctg3_2712 is of the B1-type which creates beta-hydroxylation in 3R configuration and leaves the alpha-substituent unchanged in R orientation while the ketoreductase (KR) domain of ctg3_2710 is of unknown type. Meanwhile, ctg3_2711 harbors no tailoring enzymes and has no acyl carrier protein domain. Another notable feature is the lack of a thioesterase domain in any of the modules.

To predict the product backbone, multiple possibilities in terms of module order and iteration were considered. Chen and Du (2016) reviewed two types of iterative PKSs, (1) unimodular iteration and (2) bimodular (multimodular) iteration. Hence, these two configurations were considered. Additionally, one of the PKSs lacks an ACP domain. One possibility is the use of a trans-ACP from neighboring modules, or it is skipped altogether. Using a combination of these possibilities, the following mechanisms are proposed.

Table 11 shows all the potential backbone combinations for Region 3.12. The n+1 denotes the number of iterations in each module. A variable number was set as there was no computational framework available to predict the number of iterations in this type of PKS. Experiments are needed to establish actual numbers of iterations, but Chen and Du (2016) allude to a small number of repetitions common to bacterial systems. Next, the backbones are represented by malonyl residues which are respectively modified by the tailoring enzymes associated with each module. The molecular weight ranges of the backbone associated with tailoring reactions in the next section were considered. Lastly, this study took into account the level of iteration within or between modules. For case 1 with a multimodular mechanism, all permutations were considered.

Table 11: List of potential polyketide backbone organization for region 3.12.

CASE 1: *trans*-ACP employed; ctg3_2711 included		**CASE 2: ctg3_2711 non-functional**	
Unimodular	**Multimodular**	**Unimodular**	**Bimodular**
(ctg3_2710-mal)$_{n+1}$-(ctg3_2711-mal)$_{n+1}$-(ctg3_2712-mal)$_{n+1}$	(ctg3_2710-mal- ctg3_2711-mal-ctg3_2712-mal)$_{n+1}$ OR (ctg3_2710-mal-ctg3_2711-mal)$_{n+1}$–(ctg3_2712-mal) $_{n+1}$ OR (ctg3_2710-mal)$_{n+1}$– (ctg3_2711-mal-ctg3_2712-mal) $_{n+1}$	(ctg3_2710-mal)$_{n+1}$-(ctg3_2712-mal)$_{n+1}$	(ctg3_2710-mal-ctg3_2712-mal)$_{n+1}$

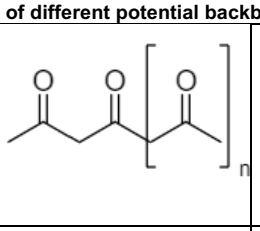
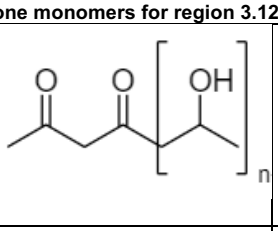
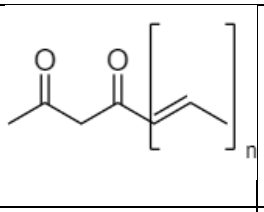
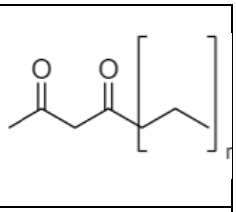
There are six potential backbones for region 3.12. It is possible that while iteration occurs, the added residue may differ slightly as some tailoring reactions may be skipped in an iteration. This further increases the complexity of backbone prediction for this region.

Cluster 20 in PRISM coincides with Region 3.12 in antiSMASH. The annotations are similar. ARTS 2.0 also revealed no resistance elements within this region but a gene duplication for oxidoreductase is present which may allude to the encoded compound's activity against this.

Molecular weight ranges

Table 12 summarizes the minimum molecular weight ranges of the polyketide backbone. Given a single iteration with associated ketoreductase and dehydratase activity, we expect a 252.31 g/mol compound. The table is, however, an oversimplification as each malonyl residue can be modified by different tailoring reactions.

Table 12: Molecular weights of different potential backbone monomers for region 3.12.

Monomer					
Tailoring domains used		None	KR	KR, DH	KR, DH, ER
Molecular weight (g/mol)		142.154	144.17	126.155	128.171
Minimum no. of iterations (n=1)	Ctg3_2711 included	426.462	432.51	378.465	384.513
	Ctg3_2711 not included	284.308	288.34	252.31	256.342

CONCLUSIONS

Additional biosynthetic genes

antiSMASH has also predicted two additional ORFs contributing to post-synthesis modification. Ctg3_2709 encodes for a nitrate reductase subunit alpha homolog. This may be part of the nitrate reductase ORFs discussed earlier. Likely, it functions more on transcriptional regulation than backbone modification. Ctg3_2717 encodes for an alpha/beta fold hydrolase-like protein. This protein likely participates in an enzymatic reaction to modify the initial compound. However, its specific role is yet to be determined as alpha/beta fold-containing proteins are structurally and functionally diverse (Dimitriou et al. 2019).

There are seven ORFs attributed to hypothetical proteins that cannot be traced to any known domains. Determining the function of these proteins would therefore require experimental studies.

REGION 3.12: ASSESSMENT OF NOVELTY

MiBiG Comparisons

Region 3.12 has 8 MiBiG BGC hits which encode for mostly cyclic compounds except for 1-heptadecene (Figure 9). Also, almost all hits are associated with ctg3_2711 which may be accounted for by its homology to ubiquitous PKSs (Table 13). However, the BGC was deemed to have moderately high novelty as the percentage similarity is mostly below 30%.

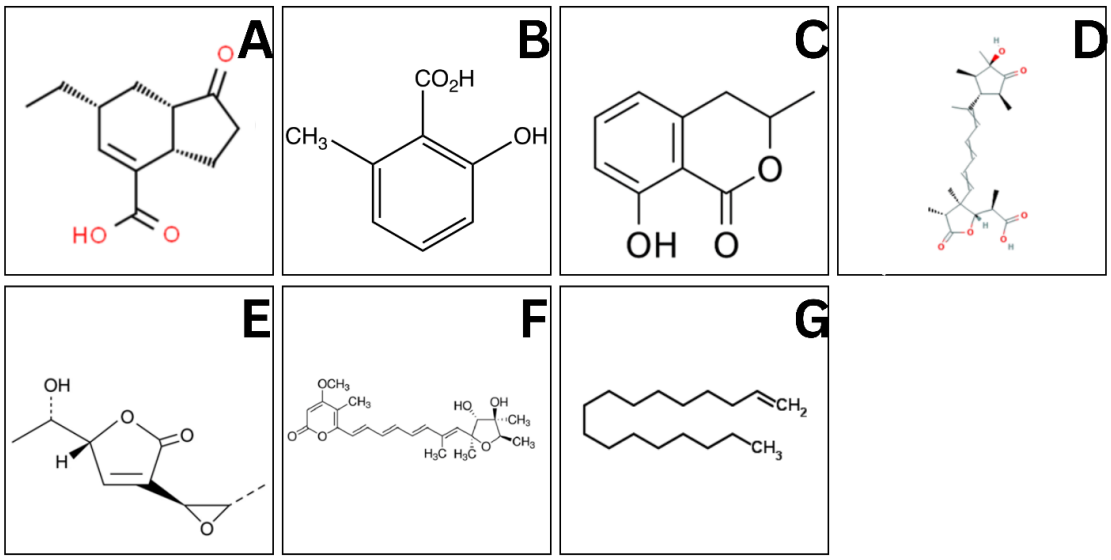


Figure 9: Chemical structure of MiBiG BGC compounds similar to Region 3.12. (A) orofacic acid (B) 6-methylsalicylic acid (C) mullein (D) aspermidgulene (E) asperlactone (F) citreoviridin (G) 1-heptadecene. All structures are derived from Pub Chem (Kim et al. 2023).

Table 13: MiBiG BGC similarity hits for region 3.12 (Blin et al. 2023).

Reference	Similarity Score	Polyketide	Microorganism	Region of similarity
BGC0000041.1	0.30	oronofacic acid	*Pseudomonas syringae*	Ctg3_2710 Ctg3_2711
BGC0001276.1	0.22	6-methylsalicylic acid	*Aspergillus terreus*	Ctg3_2711
BGC0001244.1	0.22	(-)-Mellein	*Parastagonospora nodorum*	Ctg3_2711
BGC0001275.1	0.22	6-methylsalicylic acid	*Glarea lozoyensis*	Ctg3_2711

BGC0001998.1	0.22	aspernidgulene A1, spernidgulene A2, spernidgulene B1	*Aspergillus nidulans FGSC A4*	Ctg3_2711
BGC0001273.1	0.22	asperlactone	*Aspergillus ochraceus*	Ctg3_2711
BGC0001400.1	0.21	citreoviridin	*Aspergillus terreus NIH2624*	Ctg3_2711
BGC0001161.1	0.21	1-heptadecene	*Moorea bouillonii PNG*	Ctg3_2711
BGC0001161.1	0.21	1-heptadecene	*Moorea producens JHB*	Ctg3_2711

Gene Cluster Family

Region 3.12 is classified under GCF_01292. Under this GCF, there are 37 members where 11 of which are considered core members (distance ≤900) and 16 are putative members (distance >900). Listed in Table 14 are the core members. The presence of a few members highlights BGC similarities within them, but the region may still present peculiarities as highlighted by a model distance of at least 494.

Table 14: List of gene cluster family members for GCF associated with region 3.12.

Identifier (BGC/Genome)	Distance to model	Length (kb)	Organism
GCF_002949635.1/NZ_CP026746.1.region015	494	43.34	*Nocardia cyriacigeorgica*
GCF_002933455.1/NZ_PSZF01000020.region001	494	46.46	*Nocardia cyriacigeorgica*
GCF_005863295.1/NZ_VBUU01000003.region001	512	46.46	*Nocardia*
GCF_005863225.1/NZ_VBUR01000003.region001	531	46.47	*Nocardia cyriacigeorgica*
GCF_010858045.1/c00419_NZ_JAAG...region001	534	46.47	*Nocardia*
GCF_900683635.1/NZ_LR215973.1.region005	560	44.67	*Nocardia cyriacigeorgica*
GCF_005863245.1/NZ_VBUT01000005.region001	575	46.46	*Nocardia*
GCF_000284035.1/NC_016887.1.region010	615	45.97	*Nocardia cyriacigeorgica_B*
GCF_010858005.1/c00199_NZ_JAAG...region001	615	46.46	*Nocardia cyriacigeorgica_B*
GCF_000308835.1/NZ_BAGN01000125.region001	714	15.01	*Nocardia vinacea*
GCF_000308435.1/NZ_BAFR01000109.region001	740	27.42	*Nocardia araoensis*

DISCUSSION

Nocardia species are aerobic, Gram-positive, weakly acid-fast, and non-motile actinomycetes. They are known prolific producers of bioactive secondary metabolites. Phylogenomic analysis of type strains by Tamura et al. (2018) revealed that genome sizes range from 6Mb to 11Mb in length. This puts our isolate below the range with only a size of 4.83Mb. Moreover, whole-genome classification using GTDB-Tk and FastANI suggests that BML-15-R-026U is a new *Nocardia* species reporting only an average of 79.31% ANI value against the 490 *Nocardia* Genbank genomes in NCBI to date. PRISM and antiSMASH annotation predicted at least 34 “unique” BGCs within the isolate which shows a relatively impressive number of clusters packed within a small genome size. As a comparison, *Nocardia sp.* CS682 has a genome size of 8.9Mb with 44 BGCs (Dhakal et al. 2019). Our isolate is almost half the size of this strain yet encodes only ten fewer BGCs.

The high biosynthetic potential of *Nocardia* sp. BML-15-R-026U should not be surprising. A survey reported an average of 16-49 BGCs in *Nocardia* genomes. This shows that the number of BGCs in the BML-15-R-026U genome is comparable to those of other *Nocardia* species although the isolate has a smaller genome size. Among the genomes analyzed by Mannle et al. (2020), all have sizes above 5Mb which shows that BML-15-R-026U has a compactly organized genome with many BGCs.

Out of the 26 BGCs predicted in antiSMASH, two have 100% similarity to nocobactin NA and ectoine. The rest have percent similarities of 59% below to known BGCs. Ten BGCs have a percent similarity of 10% and below. These numbers highlight the plethora of potentially highly novel compounds from the isolate. RAST annotation also supports this claim as the dominant subsystems are associated with amino acids and derivatives, co-factors, prosthetic groups, fatty acids, lipids, and isoprenoids. Such elements are common players in secondary metabolite production.

The potential molecular targets of compounds encoded in the BGCs are revealed by ARTS 2.0. In terms of resistance hits within clusters, biotin lipoyl, carboxyl transferases, proteasome, and DNA topoisomerases are the most common. For gene duplication within clusters, proteins related to DNA replication, protein synthesis, and metabolism are found. These hits provide clues to the mechanism of action of the predicted compounds.

Region 3.10

Region 3.10 was predicted to encode for a negatively regulated BGC which produces a 25-26 amino acid long NRP likely exported via a TAT-dependent pathway. The product mostly contains serine residues and has a predicted minimum weight of 2754.5g/mol. NRPs are usually 3-15 amino acids long (Martinez-Nunez et al. 2016) so the predicted product falls outside of this range. Among the MiBiG hits, syringopeptin 25A is the closest in length. This compound is found in *Pseudomonas syringae* and acts by forming ion channel pores in the lipid bilayer (Carpaneto et al. 2002).

The two proposed mechanisms for product formation, the normal modular synthesis, and the production of two peptides joined together by a condensation domain, are not uncommon. This is seen in the production of fungisporin in *Penicillium chrysogenum* which follows a non-modular arrangement with a *trans*-adenylation domain protein (Ali et al. 2014). However, to the researchers’ knowledge, no NRP biosynthesis has been

reported to follow the proposed mechanism of peptide joining. However, in PKSs, co-linear synthesis and eventual condensation are exhibited in Alpiniamide biosynthesis and are similar to the proposed mechanism (Sigrist et al. 2020).

In terms of transport, the presence of a TAT signal peptide may point to the product's export in a TAT-dependent pathway. Palmer and Berks (2012) explained that one of the rationales for using this export system is when protein folding needs to be accomplished before secretion. This can be applied to our predicted compound which is relatively big for non-ribosomal peptides.

Resistance hits to biotin lipoyl and carboxyl transferase are traced to this region. There is also a gene duplication for ribosomal protein Bs18. Biotin and lipoic acid are essential components of the lipid bilayer (Cronan 2014). Given that the BGC harbors resistance elements to these, it could be because the product attacks the housekeeping versions of these enzymes Hence, one mechanism of action is lipid bilayer biosynthesis disruption. Carboxyl transferase is a ubiquitous enzyme and is essential in fatty acid biosynthesis. Previous antibiotics have already been shown to act against this target, particularly acetyl-CoA carboxylase, which causes cell death (Freiberg et al. 2006). Lastly, the compound may also act as a ribosomal protein inhibitor as a gene duplication for ribosomal protein Bs18 is found within the region.

A BigFam (Kautsar et al. 2021) search revealed that region 3.10 belongs to gene cluster family 13372 with only one current member. This shows the high novelty of the BGC. NORINE database search and MiBiG hits also show low compound similarity to the predicted backbone. In addition, homologous genetic elements as predicted in KnownClusterBlast search harbored less than 40% similarity in some BGC members (Supplementary Table 2).

Region 3.12

Region 3.12 encodes for bacterial iterative type I polyketide synthases. There were three modules found, two of which encode a complete set of domains with ketosynthase (KS), enoyl reductase (ER), and dehydratase (DH) tailoring domains while the other only encodes for a ketosynthase and acylation domain. All acylation domains are specific for malonyl-CoA. Due to the iterative nature of the cluster and a non-complete module, 6 potential ketide backbones were suggested (Table 11). Using the simplest mechanism, the resulting compound is predicted to be at least 252.31g/mol.

Like the previous cluster, this BGC is suspected to be under negative transcriptional regulation. Stresses like DNA damage and oxidative stress might activate the BGC as suggested by ORF annotations. Alternatively, the nitrate reductase machinery found within the region can be a resistance mechanism. Such an element was not detected by ARTS, yet this study still considered it a mechanism of action that needs to be validated experimentally.

The number of iterations in each module cannot be determined computationally. MiBiG hits also did not provide clues as most were associated with modular PKSs. However, ARTS can provide clues on the potential function of the compound. Due to the duplication of an oxidoreductase in the BGC, this may be the target of the compound. Sodium-translocating NADH: ubiquinone oxidoreductase has been targeted in *Chlamydia trachomatis* as a novel antibiotic mechanism (Dibroy et al. 2017). Polymyxin B has also been shown to secondarily act by inhibiting this enzyme in *Escherichia coli* (Deris et al. 2014). Lastly, researchers suspect high BGC novelty has KnownClusterBlast result only highlighted three hits ranging from 7-30% similarity (Supplementary Table 3).

The paper demonstrates the utility of genome mining in evaluating the biosynthetic potential of bacterial isolates in terms of novel BGCs up to the prediction of their biosynthetic logic and potential novel chemistries. While genome mining is an effective and high-throughput endeavor, it is still limited to predictions, and thus other OMICs technologies, such as transcriptomics and metabolomics, must be employed to ascertain expressed compounds and proposed biosynthetic pathways. Another issue is that some BGCs are cryptic, in that they are only expressed in certain conditions. This is a common phenomenon in *Nocardia* species, such as *N. brasiliensis* and *N. vulneris* (Herisse et al. 2019). Thus, future studies must deal with culture optimization for the induction of desired BGCs.

In addition, further *in silico* approaches such as ligand-receptor docking of the compound from Region 3.10 are recommended to ascertain whether the resistance hits and duplicated genes are indeed potential targets. Once a definitive target is identified, it will be fruitful to compare binding energies between existing drugs that operate in a similar mechanism. For region 3.12, it is recommended to heterologously express the BGC and subject the culture to mass spectroscopy for compound structure determination.

Table 15: Summary of pertinent BGC information on MiBiG hits based on individual Region 3.12 PKSs generated from antiSMASH (Blin et al. 2023).

Contig						
ctg3_2710	**MiBiG protein**	CTQ34881.1	AWH12667.1	ABY21539.1	AKQ22681.1	AFN27480.1
	Description	AtcD; PKS, modules 1-4	RmpD1	AngAII	Malonyl CoA ACP transacylase	PKS_BonA
	MiBiG cluster	BGC0001301	BGC0001759	BGC0000018	BGC0001656	BGC0000173
	% Identity	44	43	43	43	43
	BCG	anthracimycin biosynthetic gene cluster from *Streptomyces sp.* T676	rifamorpholine A biosynthetic gene cluster from *Amycolatopsis sp.*	angolamycin biosynthetic gene cluster from *Streptomyces eurythermus*	luminaolide biosynthetic gene cluster from *Planktothrix paucivesiculata* PCC 9631	isobongkrekic acid biosynthetic gene cluster from *Burkholderia gladioli*

ctg3_2711	**MiBiG protein**	APZ78780.1	APZ78793.1	APZ78807.1	APZ78854.1	APZ78678.1
	Description	PKS (mchA)	PKS (mchA)	PKS (mchA)	PKS (mchA)	PKS (mchA)
	MiBiG cluster	BGC0001426	BGC0001427	BGC0001428	BGC0001432	BGC0001417
	% Identity	48	48	48	48	48
	BCG	myxochromide A biosynthetic gene cluster from *Myxococcus xanthus*	myxochromide A biosynthetic gene cluster from *Myxococcus xanthus*	myxochromide A biosynthetic gene cluster from *Myxococcus xanthus*	myxochromide D biosynthetic gene cluster from *Stigmatella erecta*	myxochromide D biosynthetic gene cluster from *Cystobacterineae bacterium*
ctg3_2712	**MiBiG protein**	IF55_RS36525	AWH12667.1	ALD83688.1	AQW44888.1	ACY06289.1
	Description	PKS	RmpD1	tAT PKS (atcF)	PKS (phaC)	T1PKS (sfaH)
	MiBiG cluster	BGC0001348	BGC0001759	BGC0001300	BGC0001737	BGC0001042
	% Identity	44	44	43	42	42
	BCG	JBIR-100 biosynthetic gene cluster from *Streptomyces varsoviensis*	rifamorpholine A biosynthetic gene cluster from *Amycolatopsis sp.*	anthracimycin biosynthetic gene cluster from *Streptomyces sp.* CNH365	phenalamide biosynthetic gene cluster from *Corallococcus coralloides*	sanglifehrin A biosynthetic gene cluster from *Streptomyces flaveolus*

CONCLUSION

This paper reports a novel *Nocardia* sp. with extraordinary biosynthetic capacity. Its 4.84Mb genome encodes for at least 34 unique secondary metabolite BGCs, mostly NRPS and PKS that reflect potentially highly novel biosynthetic chemistries. The analysis was focused on two genomic regions, regions 3.10 and 3.12, which code for a non-ribosomal peptide and polyketide, respectively, to demonstrate the prediction of the structure and potential biochemical function of the BGCs' products.

ACKNOWLEDGEMENTS

The study was funded by the Philippine Council for Health Research and Development – Department of Science and Technology under the "Anti-infective and Anticancer Drug Candidates from Marine Microorganisms and Sponges: Discovery and Development" project, Marine Science Institute – UP Diliman. The authors would like to thank the researchers of the Marine Genomics and Molecular Genetics Laboratory, MSI. The authors would also like to thank the researchers of the Discovery and Development of Health Products – Marine Component Phase I and researchers of the Marine Pharmacognosy Laboratory for the collection and initial analysis of the sample used in this study and storage and maintenance of the bacterial cultures. Sample collection was done under Gratuitous Permit No. GP-0084-15.

CONFLICT OF INTERESTS

The authors declare no conflict of interest in the publication of this article.

CONTRIBUTIONS OF INDIVIDUAL AUTHORS

PCTG: Conceptualization, some Methodology, Formal Analysis, Data Curation, Writing, Visualization; EMR and MJDP: some Methodology, some Formal Analysis, Data Curation, and Writing; CMCF, KEC: some Methodology, Validation; LASR and AOL: Manuscript revision, Funding Acquisition, Project Administration, Funding Acquisition, Resources, Conceptualization, and Data Curation.

REFERENCES

Ali, H., Ries, M. I., Lankhorst, P. P., van der Hoeven, R. A., Schouten, O. L., Noga, M., ... & Driessen, A. J. (2014). A non-canonical NRPS is involved in the synthesis of fungisporin and related hydrophobic cyclic tetrapeptides in Penicillium chrysogenum. *PloS one*, *9*(6), e98212.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arulprakasam, K. R., & Dharumadurai, D. (2021). Genome mining of biosynthetic gene clusters intended for secondary metabolites conservation in actinobacteria. Microbial Pathogenesis, 161, 105252.

Bachmann, B. O., Van Lanen, S. G., & Baltz, R. H. (2014). Microbial genome mining for accelerated natural products discovery: is a renaissance in the making?. *Journal of Industrial Microbiology and Biotechnology*, *41*(2), 175-184.

Baltz, R. H. (2011). Function of MbtH homologs in nonribosomal peptide biosynthesis and applications in secondary metabolite discovery. *Journal of Industrial Microbiology and Biotechnology*, 38(11), 1747.

Baltz, R. H. (2021). Genome mining for drug discovery: progress at the front end. Journal of Industrial Microbiology and Biotechnology, 48(9-10), kuab044.

Blachowicz, A., Raffa, N., Bok, J. W., Choera, T., Knox, B., Lim, F. Y., ... & Keller, N. P. (2020). Contributions of spore secondary metabolites to UV-C protection and virulence vary in different Aspergillus fumigatus strains. MBio, 11(1), 10-1128.

Blin, K., Shaw, S., Augustijn, H. E., Reitz, Z. L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B. R., Metcalf, W. W., Helfrich, E. J. N., van Wezel, G. P., Medema, M. H., & Weber, T. (2023). antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures, and visualisation. *Nucleic acids research*, 51(W1), W46–W50. https://doi.org/10.1093/nar/gkad344

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., ... & Xia, F. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports*, *5*(1), 1-6.

Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., & Kucherov, G. (2008). NORINE: a database of nonribosomal peptides. *Nucleic acids research*, *36*(Database issue), D326–D331. https://doi.org/10.1093/nar/gkm792

Cantu, D. C., Chen, Y., & Reilly, P. J. (2010). Thioesterases: a new perspective based on their primary and tertiary structures. *Protein Science*, 19(7), 1281-1295.

Carpaneto, A., Dalla Serra, M., Menestrina, G., Fogliano, V., & Gambale, F. (2002). The phytotoxic lipodepsipeptide syringopeptin 25A from Pseudomonas syringae pv syringae forms ion channels in sugar beet vacuoles. *The Journal of membrane biology*, *188*(3), 237-248.

CDC Centers for Disease Control and Prevention (2019). Antibiotic/Antimicrobial Resistance. Retrieved from https://www.cdc.gov/drugresistance/biggest-threats.html

Chang, S., Luo, Y., Wang, M., He, N., Chen, M., Huang, X., ... & Xie, Y. (2023). Pairing comparative genomics with tandem mass-based molecular networking allows to highly efficient discovery of nonribosomal peptides from *Nocardia* spp. *Journal of Chromatography A*, 1708, 464343.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics (Oxford, England), 36(6), 1925–1927. Advance online publication. https://doi.org/10.1093/bioinformatics/btz848

Chen, H., & Du, L. (2016). Iterative polyketide biosynthesis by modular polyketide synthases in bacteria. *Applied microbiology and biotechnology*, *100*(2), 541-557.

Chen, L. Y., Wang, X. Q., Wang, Y. M., Geng, X., Xu, X. N., Su, C., ... & Zhao, X. Q. (2018). Genome mining of Streptomyces xinghaiensis NRRL B-24674 T for the discovery of the gene cluster involved in anticomplement activities and detection of novel xiamycin analogs. Applied microbiology and biotechnology, 102, 9549-9562.

Cronan J. E. (2014). Biotin and Lipoic Acid: Synthesis, Attachment, and Regulation. *EcoSal Plus*, *6*(1), 10.1128/ecosalplus.ESP-0001-2012. https://doi.org/10.1128/ecosalplus.ESP-0001-2012

Deng, W., Li, C., & Xie, J. (2013). The underling mechanism of bacterial TetR/AcrR family transcriptional repressors. *Cellular signalling*, 25(7), 1608-1613.

Deris, Z. Z., Akter, J., Sivanesan, S., Roberts, K. D., Thompson, P. E., Nation, R. L., ... & Velkov, T. (2014). A secondary mode of action of polymyxins against Gram-negative bacteria involves the inhibition of NADH-quinone oxidoreductase activity. *The Journal of antibiotics*, *67*(2), 147-151.

Dhakal, D., Rayamajhi, V., Nguyen, H. T., Poudel, P. B., & Sohng, J. K. (2019). Complete genome sequence of *Nocardia sp*. strain CS682, a producer of antibacterial compound Nargenicin A1. *Microbiology resource announcements*, *8*(49), e01098-19.

Dibrov, P., Dibrov, E., Maddaford, T. G., Kenneth, M., Nelson, J., Resch, C., & Pierce, G. N. (2017). Development of a novel rationally designed antibiotic to inhibit a nontraditional bacterial target. *Canadian journal of physiology and pharmacology*, *95*(5), 595-603.

Dillon, S. C., & Bateman, A. (2004). The Hotdog fold: wrapping up a superfamily of thioesterases and dehydratases. *BMC bioinformatics*, 5(1), 1-14.

Dimitriou, P. S., Denesyuk, A. I., Nakayama, T., Johnson, M. S., & Denessiouk, K. (2019). Distinctive structural motifs coordinate the catalytic nucleophile and the residues of the oxyanion hole in the alpha/beta-hydrolase fold enzymes. *Protein Science*, *28*(2), 344-364.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput Biol*, *7*(10), e1002195.

Freiberg, C., Pohlmann, J., Nell, P. G., Endermann, R., Schuhmacher, J., Newton, B., Otteneder, M., Lampe, T., Häbich, D., & Ziegelbauer, K. (2006). Novel bacterial acetyl coenzyme A carboxylase inhibitors with antibiotic efficacy in vivo. *Antimicrobial agents and chemotherapy*, *50*(8), 2707–2712. https://doi.org/10.1128/AAC.00012-06

Geiman, D. E., Raghunand, T. R., Agarwal, N., & Bishai, W. R. (2006). Differential gene expression in response to exposure to antimycobacterial agents and other stress conditions among seven Mycobacterium tuberculosis whiB-like genes. *Antimicrobial agents and chemotherapy*, *50*(8), 2836-2841.

Gomez-Escribano, J. P., Alt, S., & Bibb, M. J. (2016). Next generation sequencing of actinobacteria for the discovery of novel natural products. *Marine drugs*, *14*(4), 78.

Grant, J. R., Enns, E., Marinier, E., Mandal, A., Herman, E. K., Chen, C. Y., Graham, M., Van Domselaar, G., & Stothard, P. (2023). Proksee: in-depth characterization and visualization of bacterial genomes. Nucleic acids research, 51(W1), W484–W492. https://doi.org/10.1093/nar/gkad326

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Herisse, M., Ishida, K., Porter, J. L., Howden, B., Hertweck, C., Stinear, T. P., & Pidot, S. J. (2019). Identification and mobilization of a cryptic antibiotic biosynthesis gene locus from a human-pathogenic Nocardia isolate. ACS Chemical Biology, 15(5), 1161-1168.

Hillerich, B., & Westpheling, J. (2006). A new GntR family transcriptional regulator in Streptomyces coelicolor is required for morphogenesis and antibiotic production and controls transcription of an ABC transporter in response to carbon source. *Journal of Bacteriology*, 188(21), 7477-7487.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M. *et al.* (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9, 5114. https://doi.org/10.1038/s41467-018-07641-9

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ, 7, e7359. https://doi.org/10.7717/peerj.7359

Kautsar, S. A., Blin, K., Shaw, S., Weber, T., & Medema, M. H. (2021). BiG-FAM: the biosynthetic gene cluster families database. *Nucleic acids research*, *49*(D1), D490-D497.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., ... & Bolton, E. E. (2023). PubChem 2023 update. *Nucleic acids research*, *51*(D1), D1373-D1380.

Kolmogorov, M., Bickhart, D.M., Behsaz, B. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods 17, 1103–1110 (2020). https://doi.org/10.1038/s41592-020-00971-x

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102:2567–2572. doi:10.1073/pnas.0409727102.

Koomsiri, W., Inahashi, Y., Leetanasaksakul, K., Shiomi, K., Takahashi, Y., Ōmura, S., ... & Nakashima, T. (2019). Sarpeptins A and B, lipopeptides produced by Streptomyces sp. KO-7888 overexpressing a specific SARP regulator. *Journal of natural products*, *82*(8), 2144-2151.

Lee, J. Y., & Yang, W. (2006). UvrD helicase unwinds DNA one base pair at a time by a two-part power stroke. *Cell*, *127*(7), 1349-1360.

Luo, Y., Cobb, R. E., & Zhao, H. (2014). *Recent advances in natural product discovery. Current Opinion in Biotechnology, 30, 230–237.* doi:10.1016/j.copbio.2014.09.002

Männle, D., McKinnie, S. M., Mantri, S. S., Steinke, K., Lu, Z., Moore, B. S., ... & Kaysser, L. (2020). Comparative genomics and metabolomics in the genus Nocardia. Msystems, 5(3), 10-1128.

Malla, S., van der Helm, E., Darbani, B., Wieschalka, S., Forster, J., Borodina, I., & Sommer, M. O. (2020). A novel efficient L-lysine exporter identified by functional metagenomics. *bioRxiv*.

Martínez-Núñez, M. A., & y López, V. E. L. (2016). Nonribosomal peptides synthetases and their applications in industry. *Sustainable Chemical Processes*, *4*(1), 1-8.

Mernagh, D. R., Janscak, P., Firman, K., & Kneale, G. G. (1998). Protein-Protein and Protein-DNA Interactions in the Type I. Restriction Endonuclease R. EcoR124I.

Milano, T., Paiardini, A., Grgurina, I., & Pascarella, S. (2013). Type I pyridoxal 5′-phosphate dependent enzymatic domains embedded within multimodular nonribosomal peptide synthetase and polyketide synthase assembly lines. *BMC structural biology,* 13(1), 1-18.

Mohammadipanah, F., Kermani, F., & Salimi, F. (2020). Awakening the secondary metabolite pathways of Promicromonospora kermanensis using physicochemical and biological elicitors. Applied Biochemistry and Biotechnology, 192, 1224-1237.

Molina-Henares, A. J., Krell, T., Eugenia Guazzaroni, M., Segura, A., & Ramos, J. L. (2006). Members of the IclR family of bacterial transcriptional regulators function as activators and/or repressors. *FEMS Microbiology Reviews*, 30(2), 157–186. doi:10.1111/j.1574-6976.2005.00008.x

Moretto, L., Vance, S., Heames, B., & Broadhurst, R. W. (2017). Dissecting how modular polyketide synthase ketoreductases interact with acyl carrier protein-attached substrates. *Chemical communications (Cambridge, England)*, *53*(83), 11457–11460. https://doi.org/10.1039/c7cc04625a

Mungan, M. D., Alanjary, M., Blin, K., Weber, T., Medema, M. H., & Ziemert, N. (2020). ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic acids research*, *48*(W1), W546-W552.

Palmer, T., & Berks, B. C. (2012). The twin-arginine translocation (Tat) protein export pathway. *Nature Reviews Microbiology*, *10*(7), 483-496.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research, 25(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Protein family database (2021). CPSase_L_D2 (PF02786). Retrieved from https://pfam.xfam.org/family/CPSase_L_D2

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*(1), 1-15.

Riley, S. P., Bykowski, T., Cooley, A. E., Burns, L. H., Babb, K., Brissette, C. A., Bowman, A., Rotondi, M., Miller, M. C., DeMoll, E., Lim, K., Fried, M. G., & Stevenson, B. (2009). *Borrelia burgdorferi* EbfC defines a newly identified, widespread family of bacterial DNA-binding proteins. *Nucleic acids research*, 37(6), 1973–1983. https://doi.org/10.1093/nar/gkp027

Röttig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C., & Kohlbacher, O. (2011). NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic acids research*, *39*(suppl_2), W362-W367.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068â2069. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btu153

Seemann, T. (2013). Barrnap 0.9: rapid ribosomal RNA prediction. https://github.com/tseemann/barrnap

Sigrist, R., Luhavaya, H., McKinnie, S. M. K., Ferreira Da Silva, A., Jurberg, I. D., Moore, B. S., & Gonzaga De Oliveira, L. (2020). Nonlinear Biosynthetic Assembly of Alpiniamide by a Hybrid cis/ trans-AT PKS-NRPS. ACS Chemical Biology,

15(4), 1067–1077. https://doi.org/10.1021/acschembio.0c00081

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.

Skinnider, M. A., Johnston, C. W., Gunabalasingam, M., Merwin, N. J., Kieliszek, A. M., MacLellan, R. J., ... & Magarvey, N. A. (2020). Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature communications*, *11*(1), 1-9.

Tamura, T., Ohji, S., Ichikawa, N., Hosoyama, A., Yamazoe, A., Hamada, M., ... & Fujita, N. (2018). Reclassification of Nocardia species based on whole genome sequence and associated phenotypic data. *The Journal of Antibiotics*, *71*(7), 633-641.

Tillett, D. and Neilan, B.A. (2000), Xanthogenate nucleic acid isolation from cultured and environmental cyanobacteria. Journal of Phycology, 36: 251-258. https://doi.org/10.1046/j.1529-8817.2000.99079.x

UniProt Consortium (2021). Pyridoxine/pyridoxamine 5'-phosphate oxidase. Retrieved from https://www.uniprot.org/uniprot/A0A0B8N6E8

Västermark, Å., Almén, M. S., Simmen, M. W., Fredriksson, R., & Schiöth, H. B. (2011). Functional specialization in nucleotide sugar transporters occurred through differentiation of the gene cluster EamA (DUF6) before the radiation of Viridiplantae. *BMC evolutionary biology*, *11*(1), 1-18.

Vernikos, G. S., & Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics (Oxford, England), 22(18), 2196–2203. https://doi.org/10.1093/bioinformatics/btl369

Wang, Y., Chen, Y., Shen, Q., & Yin, X. (2011). Molecular cloning and identification of the laspartomycin biosynthetic gene cluster from Streptomyces viridochromogenes. *Gene*, *483*(1-2), 11-21.

Wood, T. M., & Martin, N. I. (2019). The calcium-dependent lipopeptide antibiotics: structure, mechanism, & medicinal chemistry. *MedChemComm*, *10*(5), 634–646. https://doi.org/10.1039/c9md00126c

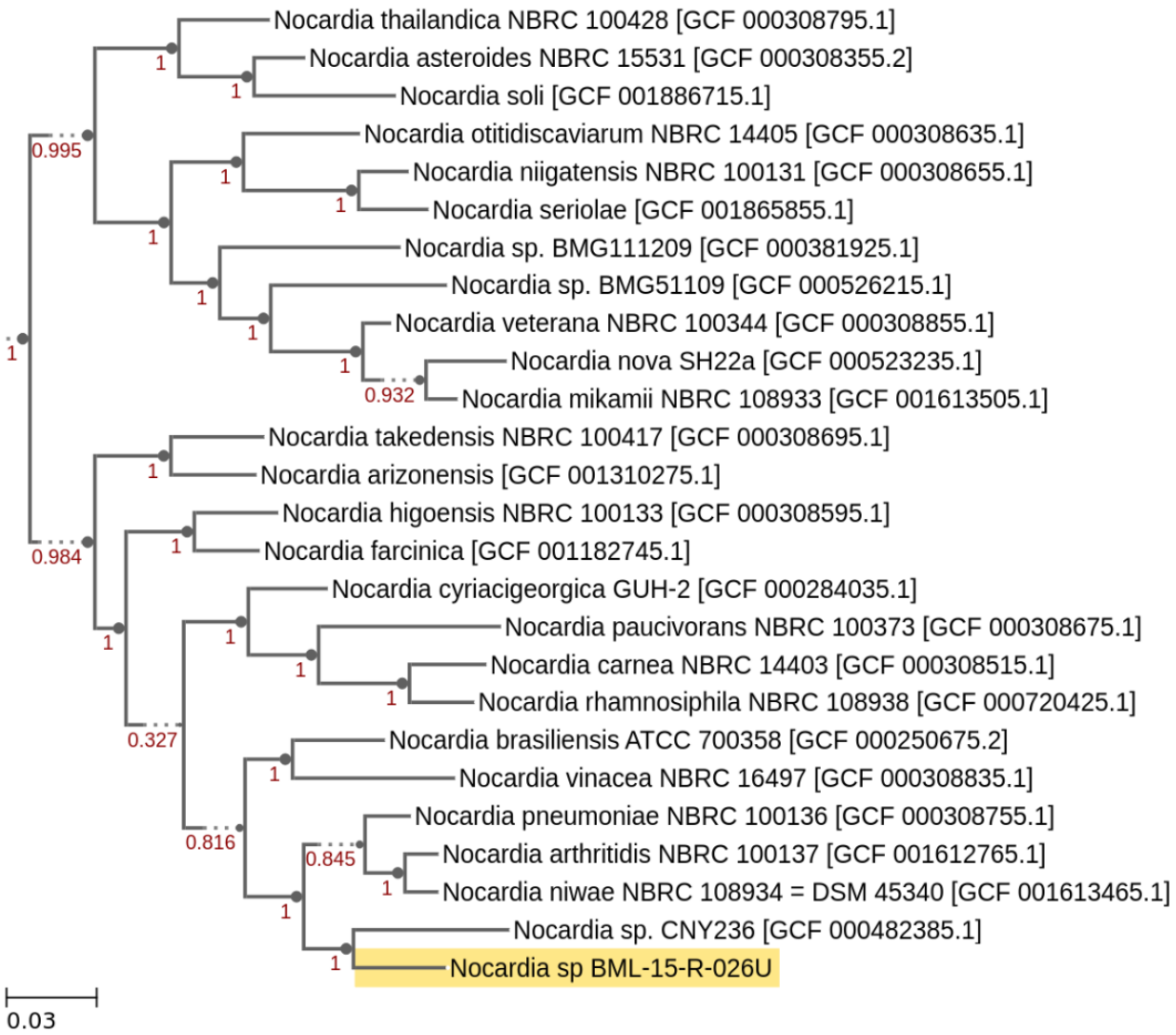
World Health Organization (2020). Antibiotic Resistance. Retrieved from https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance

Wu, G., Nie, L., & Zhang, W. (2006). Predicted highly expressed genes in Nocardia farcinica and the implication for its primary metabolism and nocardial virulence. *Antonie Van Leeuwenhoek*, *89*(1), 135-146.

Xu, S., Wei, M., Li, G., Li, Z., Che, Y., Han, L., Jia, W., Li, F., Li, D., & Li, Z. (2022). Comprehensive Analysis of the Nocardia cyriacigeorgica Complex Reveals Five Species-Level Clades with Different Evolutionary and Pathogenicity Characteristics. *mSystems*, *7*(3), e0140621. https://doi.org/10.1128/msystems.01406-21

Yuwen, L., Zhang, F. L., Chen, Q. H., Lin, S. J., Zhao, Y. L., & Li, Z. Y. (2013). The role of aromatic L-amino acid decarboxylase in bacillamide C biosynthesis by *Bacillus atrophaeus* C89. *Scientific reports*, 3(1), 1-10.

SUPPLEMENTARY INFORMATION

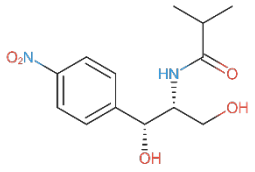
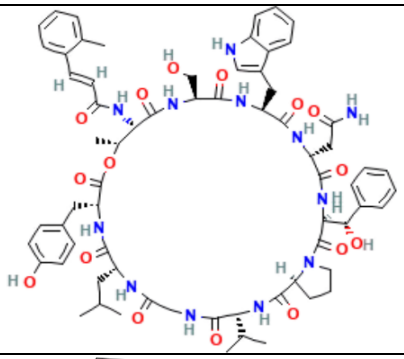
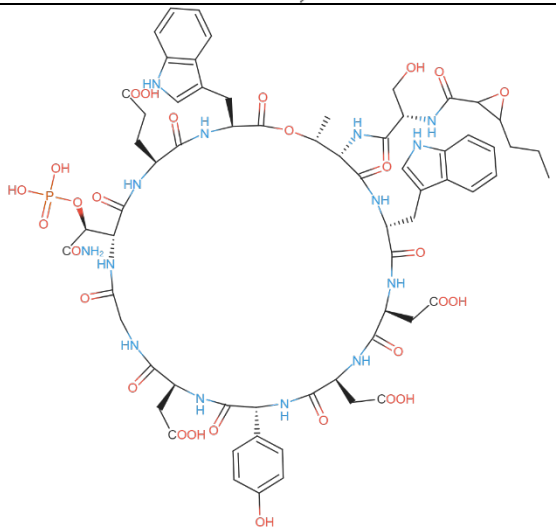


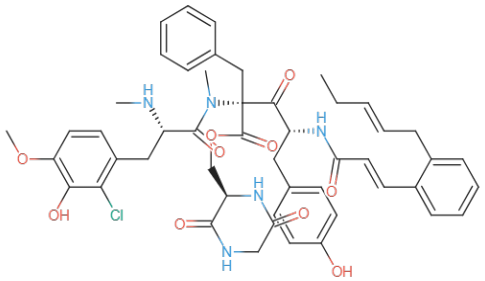
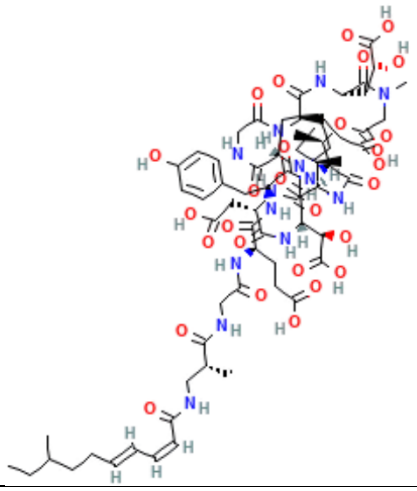
Supplemental Figure 1: Maximum-Likelihood Tree based on the concatenated 49 core, universal genes established by the Clusters of Orthologous Groups gene families of the BML-15-R-026U genome and 25 top closely related RefSeq genomes constructed in the SpeciesTree tool in KBase (https://www.kbase.us/). Values on nodes represent the percentage bootstrap support values out of 1000 bootstrap replicates. Scale bar at the bottom of the tree represents 3% genetic variations for every scale length.

Supplementary Table 1: Reference genomes used in the manual binning of the mock metagenome sequencing assembly.

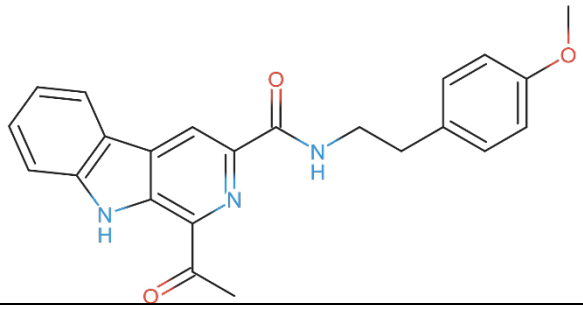
Organism Name	Strain	RefSeq category	GenBank Assembly Accession
Paenibacillus lautus NBRC 15380 (firmicutes)	NBRC 15380	representative genome	GCA_004000945.1
Mycolicibacterium fortuitum (high GC Gram+)	CT6	representative genome	GCA_001307545.1
Microbulbifer agarilyticus (g-proteobacteria)	GP101	representative genome	GCA_001999945.1
Glutamicibacter nicotianae (high GC Gram+)	OTC-16	representative genome	GCA_003687415.1
Staphylococcus sciuri subsp. *sciuri* (firmicutes)	NCTC12103	representative genome	GCA_900474615.1
Streptomyces cavourensis (high GC Gram+)	1AS2a	representative genome	GCA_002804165.1
Tersicoccus solisilvae (high GC Gram+)	CGMCC 1.15480	representative genome	GCA_014643255.1
Arthrobacter agilis (high GC Gram+)	NCTC2676_1	representative genome	GCA_900631605.1
Nocardia brasiliensis (high GC Gram+)	FDAARGOS_352	representative genome	GCA_002209125.2
Bacillus paramycoides (firmicutes)	LB_RP2	representative genome	GCA_013391375.1
Gordonia bronchialis DSM 43247 (high GC Gram+)	DSM 43247	representative genome	GCA_000024785.1

Supplementary Table 2: Summary of top 5 KnownClusterBlast results related to Region 3.10 as predicted in antiSMASH (Blin et al. 2023). All compound structures and specific details were pulled from the corresponding MiBiG database.

BGC Code (MiBiG 3.1)	BGC-encoded compound	% of Genes Showing Similarity	Structure of Compound
BGC0002284	corynecin III biosynthetic gene cluster from *Rhodococcus* sp. H-CA8f	40%	
BGC0001975	atratumycin biosynthetic gene cluster from	7%	
BGC0000315	CDA1b biosynthetic gene cluster from *Streptomyces coelicolor* A3(2)	5%	

BGC0002014	pepticinnamin E biosynthetic gene cluster from *Actinobacteria* bacterium OK006	6%	
BGC0001968	cadaside A biosynthetic gene cluster from uncultured bacterium	9%	

Supplementary Table 3: Summary of all KnownClusterBlast results related to Region 3.12 as predicted in antiSMASH (Blin et al. 2023). All compound structures and specific details were pulled from the corresponding MiBiG database.

BGC Code (MiBiG 3.1)	BGC-encoded compound	% of Genes Showing Similarity	Structure of Compound
BGC0002666	legonindolizidine A6 biosynthetic gene cluster from *Streptomyces* sp. MA37	16%	No known structure
BGC0001137	marinacarboline A biosynthetic gene cluster from *Marinactinospora thermotolerans*	30%	
BGC0001048	tallysomycin A biosynthetic gene cluster from *Streptoalloteichus hindustanus*	7%