

Original Research Article

Pangenomics and Comparative Biosynthetic Diversity of Malaysian Mangrove-Derived *Streptomyces albidoflavus* MUM256 among *S. albidoflavus*-Assigned Genomes

Loh Teng-Hern Tan^{1,2*}, Jodi Woan-Fei Law^{1,2}, Vengadesh Letchumanan³, Kah-Ooi Chua⁴, Ren-You Gan⁵, Lihong Wu⁶, Kok-Gan Chan^{1,2,7}, Learn-Han Lee^{1,2*}

Article History

Received: 01 September 2026;

Received in Revised Form: 30 September 2026;

Accepted: 05 October 2026;

Available Online: 07 October 2026

¹RII Centre for Molecular and Microbiome Sciences, Research Innovation Institute (RII), School of Health and Life Sciences, University of Nottingham Ningbo China, Ningbo 315000, China; jodi-woan-fei.law@nottingham.edu.cn (JW-FL)

²Microbiome and Bioresource Research Strength (MBRS), Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Bandar Sunway 47500, Selangor Darul Ehsan, Malaysia

³Pathogen Resistome Virulome and Diagnostic Research Group (PathRiD), Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Bandar Sunway 47500, Selangor Darul Ehsan, Malaysia; vengadesh.letchumanan1@monash.edu (VL)

⁴Centre for Research in Biotechnology for Agriculture (CEBAR), University of Malaya, Kuala Lumpur 50603, Malaysia; kahooi@um.edu.my (K-OC)

⁵Department of Food Science and Nutrition, The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR, China; renyou.gan@polyu.edu.hk (R-YG)

⁶Department of Basic Oral Medicine, School and Hospital of Stomatology, Guangdong Engineering Research Center of Oral Restoration and Reconstruction, Guangzhou Medical University, Guangzhou, 510812, China; wcanhong@gzhmu.edu.cn (LW)

⁷Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur 50603, Malaysia; kokgan@um.edu.my (K-GC)

*Corresponding author: Loh Teng-Hern Tan, Learn-Han Lee; RII Centre for Molecular and Microbiome Sciences, Research Innovation Institute (RII), School of Health and Life Sciences, University of Nottingham Ningbo China, Ningbo 315000, China; loh-teng-hern.tan@nottingham.edu.cn (LT-HT), learn-han.lee@nottingham.edu.cn (L-HL)

Abstract: *Streptomyces* sp. MUM256, isolated from Malaysian mangrove soil, has previously shown antioxidant and selective anticancer activities, including induction of cell-cycle arrest and apoptosis in colon cancer cells. However, its precise taxonomic position and biosynthetic potential remained unresolved at the whole-genome and population levels. Here, we sequenced a high-quality long-read genome of MUM256 and integrated genome-

based taxonomy, pangenomics and comparative biosynthetic gene cluster (BGC) analysis. The final assembly comprised a single 6.89-Mb contig with 73.4% G+C content, 98.5% BUSCO completeness and 5,897 predicted protein-coding sequences. Genome-based analyses consistently assigned MUM256 to *Streptomyces albidoflavus*, with NRRL B-1271 identified as its closest GTDB-Tk placement reference. A quality-filtered cohort of 151 genomes revealed an open pangenome comprising 21,082 gene clusters, including 2,684 strict-core clusters, with continued gene-family accumulation supported by Heaps' law modelling ($\gamma = 0.307$). Population-scale genome mining of 119 quality-filtered genomes identified 2,651 BGC records grouped into 178 gene cluster families (GCFs), of which 84.3% were rare or singleton. MUM256 contained 18 predicted BGCs representing 18 GCFs, all shared with other members of the analysed *S. albidoflavus*-assigned genomes. Comparative analysis of fredericamycin A-like and SF2768-like loci further revealed population-level variation in BGC prevalence and architecture. Together, these findings establish MUM256 as *S. albidoflavus* and provide a population-genomic framework for prioritising biosynthetic loci for future metabolomic and functional validation.

Keywords: *Streptomyces albidoflavus*; whole-genome sequencing; genome annotation; comparative; biosynthetic gene clusters; mangrove; secondary metabolites

1. Introduction

Members of the genus *Streptomyces* are filamentous, high-G+C Gram-positive bacteria renowned for their capacity to produce structurally and functionally diverse specialised metabolites ^[1]. Compounds derived from *Streptomyces* have contributed extensively to antimicrobial, anticancer, immunosuppressive, antiparasitic and agricultural applications, establishing the genus as one of the most important microbial resources for natural-product discovery ^[2–4]. The continuing relevance of this chemical diversity is underscored by systematic evidence that *Streptomyces*-derived secondary metabolites retain activity against a broad range of bacterial pathogens, including multidrug-resistant organisms ^[5, 6]. Advances in whole-genome sequencing have further revealed that the biosynthetic capacity encoded by *Streptomyces* genomes substantially exceeds that detected through conventional cultivation-based screening. Individual genomes commonly harbour numerous biosynthetic gene clusters (BGCs), many of which remain silent or weakly expressed under standard laboratory conditions ^[3, 7]. Genome mining has consequently become an important complement to activity-guided natural-product discovery by enabling systematic identification and prioritisation of biosynthetic loci whose products may otherwise remain undetected. Importantly, comparative genomic analyses have also demonstrated substantial variation in BGC repertoires among closely related *Streptomyces* strains, indicating that biosynthetic potential is not adequately represented by a single reference genome ^[2, 8].

Ecologically distinctive and comparatively underexplored environments represent valuable reservoirs of such microbial diversity. Mangrove ecosystems are characterised by tidal fluctuation, variable salinity, high organic-matter turnover and steep physicochemical gradients and harbour diverse actinobacterial communities with considerable biosynthetic potential ^[9–12]. Malaysian mangrove exploration has yielded novel *Streptomyces* taxa and biologically active isolates, illustrating that these environments provide complementary opportunities for microbial diversity discovery and natural-product research. Examples include antioxidant-producing strain MUSC 5 and the more recently described mangrove species *Streptomyces griseiviridis* and *Streptomyces learnhanii* ^[13–15]. This broader strategy has been conceptualised within the “Modern Actinobacteria” framework, which emphasises environmentally derived actinobacteria with potential biomedical, agricultural and biotechnological applications ^[16, 17]. Within this context, strain MUM256 was isolated from Malaysian mangrove soil and was initially designated *Streptomyces* sp. MUM256. Its organic extract exhibited antioxidant activity and selective cytotoxic effects against several human cancer cell lines, with HCT116 colorectal carcinoma cells showing particularly high sensitivity ^[18]. Subsequent investigation demonstrated that metabolites extracted from MUM256 inhibited HCT116 proliferation through cell-cycle arrest and induction of intrinsic apoptosis, while chemical profiling identified several putative bioactive constituents ^[19]. These observations established MUM256 as a biologically active mangrove-derived *Streptomyces* strain; however, its taxonomic identity and biosynthetic potential had not been investigated within a comprehensive whole-genome and population-level framework.

Accurate species placement is particularly important in *Streptomyces*, in which closely related taxa may exhibit highly similar 16S rRNA gene sequences and considerable phenotypic overlap. Although 16S rRNA gene analysis remains valuable for preliminary taxonomic identification, its discriminatory power can be insufficient for resolving closely related *Streptomyces* species ^[20]. Genome-wide approaches now provide substantially greater resolution, with average nucleotide identity (ANI), digital DNA-DNA hybridisation (dDDH), phylogenomics and other overall genome-relatedness indices increasingly forming the basis of prokaryotic species delineation ^[21–23]. This is particularly relevant to strains such as MUM256 that were initially characterised before the current availability of extensive genome collections and contemporary genome-taxonomy frameworks. Establishing its genome-based species identity therefore provides more than a revised taxonomic designation: it defines the appropriate biological population against which its genomic composition, evolutionary relationships and biosynthetic repertoire can be interpreted.

Streptomyces albidoflavus, to which MUM256 is assigned in the present study, comprises a genetically diverse and geographically distributed population recovered from a range of environmental sources. Early population-genetic analysis of 41 *S. albidoflavus* strains from predominantly soil, marine, and insect-associated habitats revealed frequent homologous recombination together with strong habitat-associated population structure and reduced gene flow between ecological groups, suggesting that ecological differentiation can impose barriers to genetic exchange within the species ^[24]. Subsequent population-genomic analysis of 30 *S. albidoflavus* genomes extended these observations at the whole-genome level, identifying three well-defined phylogenomic clades and demonstrating habitat-associated variation in both core and accessory genomic components ^[25]. Collectively, these studies indicated that the genomic structure of *S. albidoflavus* reflects an interplay between recombination, ecological adaptation and lineage differentiation. Subsequent studies have further demonstrated functional and biosynthetic heterogeneity among *S. albidoflavus* isolates. Genome analysis of strain MGMM6 revealed an extensive repertoire of genes associated with specialised metabolism and environmentally relevant functions ^[26], while comparative analysis of rhizosphere-derived isolates identified an open pangenome and strain-level differences in predicted secondary-metabolite loci despite close genomic relationships ^[27]. More recently, analysis of 34 *S. albidoflavus* genomes demonstrated considerable accessory-gene and functional variation within the species, even among strains sharing high overall nucleotide identity ^[28]. Similarly, investigation of the bioactive marine strain VIP-1 illustrated how genome-based taxonomy, environmental adaptation and biosynthetic potential can be integrated to characterise an ecologically distinct *S. albidoflavus* isolate ^[29]. Together, these findings show that species assignment alone does not capture the extent of functional diversity within *S. albidoflavus* and highlight the need for broader population-scale comparisons.

Population context is especially important when interpreting specialised metabolism. In a genus-wide analysis of 2,371 quality-filtered *Streptomyces* genomes, ^[8] demonstrated that biosynthetic diversity is distributed unevenly across genomic populations and that assessment of individual genomes can incompletely represent the biosynthetic repertoire present at species or lineage level. Notably, *S. albidoflavus* was one of the most highly represented species in that dataset, with 109 genomes assigned using genome-based taxonomy. Their analysis also showed conservation of BGC order and genomic position within related lineages while identifying gain, loss and structural variation among individual loci ^[8]. More recently, Wang *et al.* ^[30] examined the distribution and evolutionary dynamics of BGCs within *S. albidoflavus* and demonstrated that biosynthetic loci span a continuum from broadly conserved to sparsely distributed gene cluster families. Detailed examination

of a rare kosinostatin biosynthetic locus further implicated horizontal acquisition, subsequent vertical inheritance and recurrent gene loss in shaping its contemporary distribution within the species ^[30]. These studies emphasise that BGC prevalence and genomic context are informative evolutionary characteristics and reinforce the value of interpreting the biosynthetic repertoire of an individual strain against a species-wide background.

In the present study, we sequenced a high-quality PacBio long-read genome assembly of MUM256 and integrated genome-based taxonomy, pangenomics and comparative biosynthetic analysis to place this previously characterised bioactive mangrove isolate within the genomic diversity of *S. albidoflavus*. Genome-based classification incorporating GTDB-Tk, ANI, dDDH and phylogenomic approaches was first used to establish the taxonomic position of MUM256. A quality-controlled species cohort was then constructed to characterise gene-content diversity, core and accessory genome structure, population relationships and pangenome dynamics. In parallel, predicted BGCs from a quality-filtered genome cohort were grouped into gene cluster families (GCFs) to determine their prevalence across the species and to contextualise the 18 biosynthetic regions identified in MUM256. Selected MUM256 loci were subsequently examined at the gene-architecture level to assess their conservation and structural variation among related genomes. By integrating a high-quality genome of a phenotypically characterised environmental isolate with expanded cohort-wide comparative datasets, this study provides a genomic framework for understanding the position of MUM256 within *S. albidoflavus* and for distinguishing conserved from variable components of its biosynthetic repertoire.

2. Methods

2.1. Bacterial Strain, Cultivation and Whole-Genome Sequencing

Strain MUM256 was originally isolated from mangrove soil collected in Malaysia and was maintained on International *Streptomyces* Project (ISP) 2 medium at 28°C. The extraction of high-quality genomic DNA, SMRTbell library construction and whole-genome sequencing were performed on the PacBio Sequel II platform by NovogeneAIT Genomics (Novogene, Singapore).

2.2. PacBio Read Processing, De Novo Assembly, Polishing and Terminal Curation

The PacBio sequencing data were obtained in BAM format and subsequently processed for de novo genome assembly. The raw PacBio BAM data were converted to FASTQ format using bam2fastx (v3.1.1). De novo genome assembly was performed using Flye (v2.9.6) with 5 polishing iterations and PacBio raw-read mode (--pacbio-raw) through the web-based Galaxy platform (Galaxy version 2.9.6+galaxy0) ^[31, 32].

The original PacBio reads were aligned to the Flye assembly using pbmm2 from the v26.2.0 release, a PacBio wrapper around minimap2, with the subread preset (<https://github.com/PacificBiosciences/pbmm2>)^[33]. The resulting alignment was used for an additional consensus-polishing step with GCpp (v2.0.2) using the Arrow algorithm (<https://github.com/PacificBiosciences/gcpp>). Following Arrow polishing, the PacBio raw reads were remapped to the polished assembly and base-level sequencing depth was assessed using SAMtools (v1.23)^[34]. Read support at both chromosomal termini was examined relative to the chromosome-wide coverage profile. Poorly supported terminal sequence was removed by coverage-guided curation, yielding the final 6,892,379-bp assembly.

2.3. Genome Assembly and Quality Assessment

Assembly statistics were evaluated using QUAST (v5.3.0+galaxy1)^[35]. Genome completeness was evaluated by Benchmarking Universal Single-Copy Orthologs (BUSCO) (v6.1.0)^[36, 37] against the *streptomyces_odb12.2* lineage datasets^[36]. Genome quality was additionally assessed using CheckM2 (v1.1.0) to estimate genome completeness and contamination^[38].

2.4. Genome Annotation and Functional Characterisation

The assembled genome of *S. albidoflavus* MUM256 was annotated using Rapid Annotations using Subsystems Technology (RAST) (<https://rast.nmpdr.org/>)^[39]; Prokka (v1.14.6+galaxy2)^[40] and NCBI Prokaryotic Genome Annotation Pipeline (PGAP)^[41]. Prokka was used to provide a standardised local structural annotation of protein-coding sequences, structural RNAs and other genomic features, whereas RAST was used to obtain complementary subsystem-based functional annotation. NCBI PGAP annotation was generated as part of the genome submission and deposition workflow and was additionally used for downstream functional classification.

Functional classification was performed using COGclassifier (v2.0.0) against the Cluster of Orthologous Genes (COG) resource (<https://github.com/moshi4/COGclassifier>)^[42]. The PGAP-annotated protein FASTA was used as input. Protein sequences were queried using RPS-BLAST (v2.17.0) against COG conserved-domain profiles, and proteins with supported matches were assigned to their corresponding COG functional categories^[43]. COG assignments were subsequently used to characterise the functional composition of the MUM256 genome and for genome-wide visualisation.

A circular representation of the MUM256 chromosome was generated using pyCirclize (v1.10.1) (<https://github.com/moshi4/pyCirclize>). The NCBI PGAP-annotated genome was used as the genome feature framework, with COG assignments mapped to their corresponding coding sequences using locus-tag identifiers. The circular map incorporated coding sequences, GC content, GC skew, COG functional assignments and biosynthetic gene clusters predicted by antiSMASH, arranged as concentric tracks to provide an integrated overview of chromosome organisation, functional composition and biosynthetic potential.

2.5. 16S rRNA Gene Identification and Preliminary Taxonomic Placement

Ribosomal RNA genes were identified from the MUM256 genome assembly using barrnap v0.8 through Galaxy platform (Galaxy wrapper v1.2.2) ^[44] using the default parameters with the kingdom set to Bacteria. Predicted rRNA sequences were exported in FASTA format, and exact duplicate 16S rRNA sequences were collapsed before taxonomic comparison. Unique 16S rRNA gene sequences were queried independently against EzBioCloud 16S rRNA gene databases (<https://www.ezbiocloud.net/>) ^[45]. Given the preliminary placement within the closely related *S. albidoflavus* group and the historical classification of MUM256, publicly available genomes annotated as *S. albidoflavus* were retrieved from NCBI for comparative genome analysis. Definitive taxonomic placement was subsequently evaluated using GTDB-Tk, ANI and digital DNA-DNA hybridisation (dDDH) analysis.

2.6. Comparative-Genome Collection and Cohort Construction

Publicly available RefSeq assemblies annotated as *Streptomyces albidoflavus* were retrieved from NCBI following preliminary 16S rRNA-based placement of MUM256 within a closely related *Streptomyces* group (retrieval date: 26 August 2026). After deduplication of corresponding GCA/GCF records, genomes were screened for assembly quality using CheckM ^[46], retaining assemblies with estimated completeness $\geq 95\%$ and contamination $\leq 5\%$. This yielded 165 *S. albidoflavus* reference genomes, which were combined with MUM256 to form the initial 166-genome working collection.

For the primary pangenome analysis, three engineered strains were excluded. Genome annotations were subsequently evaluated during PGAP2 preprocessing ^[47], 12 additional genomes in which $>10\%$ of genes were classified as incomplete or discarded were removed. This filtering yielded a primary pangenome cohort of 151 genomes, comprising 150 publicly available genomes and MUM256. The $>10\%$ cutoff was applied as a study-specific quality control criterion to reduce the influence of incomplete gene models on comparative gene-content analysis.

2.7. Genome-Based Taxonomy and Genomic Relatedness

Genome-based taxonomic classification of MUM256 was performed using GTDB-Tk (v2.5.2) classify_wf against GTDB-Tk database release_226 ^[48, 49] through the Galaxy wrapper v1. Average nucleotide identity (ANI) between MUM256 genome and the 165 publicly available reference genomes was calculated using FastANI ^[21]. A 166-genome ANI matrix was additionally visualised using ANIclustermap v2.0.1 (<https://github.com/moshi4/ANIclustermap>). Species-level interpretation followed established genome-based taxonomic criteria ^[50].

Whole-genome taxonomic placement was further assessed using the Type (Strain) Genome Server (TYGS) ^[51] with a 15-genome type/reference comparison panel. The dDDH values were obtained through the Genome-to-Genome Distance Calculator ^[52], with Formula

2(d4) used as the principal metric and 70% dDDH applied as the conventional species-level threshold ^[53].

2.8. Core-Genome Phylogeny, Phylogroup Delineation and Pangenome Analysis

Pangenome analysis was performed using PGAP2 v2.4.1 ^[47] (<https://github.com/bucongfan/PGAP2>) on the 151-genome primary cohort. The prep and main workflows were used to construct homologous gene cluster and generate gene presence/absence data. The resulting pgap2.partition.gene_content.pav matrix comprised 21,082 gene clusters across the 151 genomes. Clusters were classified by PGAP2 into strict-core, core, soft-core, shell and cloud partitions.

Pangenome statistics and accumulation profiles were generated using pgap2 post stat and pgap2 post profile, with the latter using the gene presence/absence matrix as input. Pangenome and core-genome accumulation models were fitted using the PGAP2 implementation. For core-genome phylogenetic reconstruction, pgap2 post singletree was run with a core-presence threshold of 0.95 and para_strategy drop. Individual core-gene alignments were generated using MAFFT ^[54] and processed using ClipKIT ^[55]. Maximum-likelihood phylogenetic inference was performed using IQ-TREE v3.1.3 ^[56] under the GTR-G substitution model. Branch support was estimated using 1,000 ultrafast bootstrap replicates ^[57], with bootstrap nearest-neighbour-interchange optimisation.

To identify major phylogenetic groupings, pairwise patristic distances between genomes were calculated from the branch lengths of the maximum-likelihood tree using cophenetic.phylo in the R package ape (v5.8-1) ^[58]. Agglomerative hierarchical clustering was performed on the distance matrix using the average-linkage method as implemented in hclust(). Candidate partitions from $k = 2$ to 15 groups were evaluated using the mean silhouette width calculated from the original patristic-distance matrix ^[59], and the partition with the highest mean silhouette coefficient was selected as the optimal phylogroup resolution.

2.9. Biosynthetic Gene Cluster Prediction and Comparative Analysis

Biosynthetic gene clusters (BGCs) were predicted using antiSMASH v8.0.4 ^[60], with the bacterial taxon setting, Prodigal gene prediction and strict HMM-detection mode. Enabled analyses included ClusterBlast, KnownClusterBlast, SubClusterBlast, comparison against MIBiG ^[61], active-site finder, Cluster HMMer, Pfam-to-GO mapping and smCOG trees.

Gene cluster family (GCF) analysis was performed using BiG-SCAPE v2.0.3 ^[62, 63] on antiSMASH-predicted BGCs from two comparative cohorts. Following exclusion of three engineered strains, 163 natural-strain genomes (162 publicly available genomes and MUM256) were retained as the broad cohort. To reduce potential effects of genome fragmentation on BGC reconstruction, a contiguity-filtered cohort was generated by retaining genomes with available antiSMASH results and ≤ 100 contigs, yielding 119 genomes (118

publicly available genomes and MUM256). The ≤ 100 -contig threshold was applied as a study-specific criterion because fragmented assemblies can split BGCs across contig boundaries and compromise reconstruction of complete biosynthetic loci [64–66]. Analyses were performed using region-level records, category-based classification, inclusion of singletons and a gene cluster family (GCF) cutoff of 0.3. GCF prevalence was calculated as the proportion of genomes represented within each GCF, with multiple BGC records from the same genome counted once. For the primary cohort, GCFs were classified as cohort-wide (100%), near-core (≥ 95 to $< 100\%$), common (≥ 50 to $< 95\%$), accessory (≥ 15 to $< 50\%$), rare (> 1 genome to $< 15\%$) or singleton (one genome).

The BiG-SCAPE similarity network for the 119-genome cohort was visualised in Cytoscape [67]. Nodes were annotated according to genome, antiSMASH region, biosynthetic category, GCF assignment and MUM256 membership, with MUM256-associated BGCs highlighted within the species-level network. Pairwise BiG-SCAPE edges were filtered for visualisation using a distance threshold of 0.3.

2.10. Comparative Analysis of Focal Biosynthetic Gene Clusters

Selected MUM256 BGCs were further examined using clinker v.0.0.32 [68] (<https://github.com/gamcil/clinker>) to compare gene organisation and biosynthetic architecture. Detailed analyses focused on region008 and region018, with region014 included as a supporting comparison. In the 119-genome BiG-SCAPE analysis, these regions were assigned to FAM_00156, FAM_00116 and FAM_00112, respectively.

Comparisons incorporated corresponding reference BGCs from MIBiG [61]: BGC0000224.4 for the fredericamycin-like locus [69], BGC0001574.5 for the SF2768-like locus [70] and BGC0002358.3 for the cyclofaulknamycin-associated locus [71], together with selected representative loci from *S. albidoflavus* genomes. Protein groups used for consistent locus colouring were generated using MMseqs2 [72] easy-cluster with a minimum sequence identity of 0.5, minimum coverage of 0.8 and coverage mode 0. Gene-function annotations and locus-specific colour maps were subsequently applied to facilitate comparison of homologous genes, gene-content variation and biosynthetic architecture.

3. Results

3.1. Genome Assembly, Quality and Annotation of MUM256

The final MUM256 assembly comprised a single 6,892,379-bp contig with a GC content of 73.44% and an N50 of 6,892,379 bp (Table 1). This sequence was obtained after terminal curation of the 6,892,461-bp Arrow-polished assembly, representing a total reduction of 82 bp. Read mapping across the final assembly demonstrated complete sequence coverage, with 389,720 reads contributing to 100% breadth of coverage and a mean sequencing depth of approximately 500 $\times$. The mean mapping quality was 59. The sequence was used throughout the comparative analyses and has also been deposited at DDBJ/ENA/GenBank under the accession JCCJPS000000000; CM195964. The data are

publicly available at NCBI GenBank under the BioProject accession number PRJNA1517283, and the BioSample accession number SAMN62642883. Raw sequencing data are available through the Sequence Read Archive under accession SRR40366859.

Both conserved-orthologue and genome-quality assessments supported the completeness of the final assembly. BUSCO recovered 98.5% complete markers, comprising 98.0% single-copy and 0.5% duplicated markers; 0.3% were fragmented, and 1.3% were missing ($n = 1,438$). The reference genome (*S. albidoflavus* strain NBC 01110, CP108647.1) analysed using the same lineage dataset produced 98.5% complete BUSCOs [S: 98.1%, D: 0.4%], indicating essentially equivalent recovery of conserved *Streptomyces* orthologues. CheckM2 estimated 100.0% completeness and 0.1% contamination. Prokka annotated 5,996 genes, including 5,897 protein-coding sequences, 77 tRNAs, 21 rRNAs and one tmRNA (Table 1). Among the predicted coding genes (CDS), 2,821 (47.84%) were annotated as hypothetical proteins, while 1,889 (32.03%) and 1,858 (31.51%) were associated with COG and EC annotations, respectively.

Complementary subsystem-based annotation using RAST identified 6,335 genomic features, including 6,162 protein-coding sequences, 86 RNA features and 87 repeat-associated features. A total of 1,272 features were associated with at least one RAST subsystem, including 1,224 CDS, corresponding to approximately 19.9% of the RAST-predicted protein-coding repertoire (Supplementary Material, Table S1). After resolving features assigned to multiple subsystems, 312 distinct subsystem terms were represented. Functional assignments encompassed central metabolism, respiration, lipid metabolism, DNA repair and stress-response processes, together with features associated with siderophore biosynthesis, metal homeostasis and chitin/N-acetylglucosamine utilisation (Supplementary Material, Figure S1). RAST additionally assigned 1,939 CDS to annotations containing a hypothetical-protein designation.

Genome-wide organisation was visualised by integrating protein-coding sequences, RNA genes, functional annotations, predicted biosynthetic loci, GC content, and GC skew (Figure 1A). CDSs were densely distributed on both strands throughout the assembly, while COG-classified genes occurred across the genome without restriction to particular genomic regions. The 18 predicted BGCs were similarly distributed across the genome and represented multiple biosynthetic classes, including NRPS, PKS, hybrid, terpene, RiPP, siderophore, and ectoine clusters. Local variation in GC content and GC skew was apparent across the assembly, together with discrete rRNA loci and more broadly distributed tRNA genes.

Of the 5,842 predicted proteins subjected to COG classification, 4,690 (80.28%) were assigned to a functional category (Figure 1B). Transcription (K) represented the most abundant category, accounting for approximately 500 proteins, followed by amino acid transport and metabolism (E) and carbohydrate transport and metabolism (G), each comprising approximately 400 proteins. Other prominent functional groups included general function prediction only (R), signal transduction mechanisms (T), translation and ribosomal

biogenesis (J), coenzyme transport and metabolism (H), lipid transport and metabolism (I), energy production and conversion (C), and inorganic ion transport and metabolism (P). Notably, nearly 200 proteins were assigned to secondary metabolite biosynthesis, transport, and catabolism (Q), providing a genome-wide functional context for the biosynthetic potential examined in subsequent comparative BGC analyses.

Table 1. Genomic features of *Streptomyces albidoflavus* MUM256

Strain	MUM256
Assembly characteristics	
Number of contigs	1
Genome size (bp)	6,892,379
GC content (%)	73.44
N50	6,892,379
Number of mapped reads	389,720
Coverage breadth (%)	100
Mean sequencing depth	500×
BUSCO: C: 98.5% [S: 98.0%, D: 0.5%], F: 0.3%, M: 1.3%, n: 1,438	
Total BUSCO groups searched	1438
Complete BUSCOs (C)	1416
Single-copy BUSCOs (S)	1409
Duplicated BUSCOs (D)	7
Fragmented BUSCOs (F)	4
Missing BUSCOs (M)	18
CheckM2 assessment	
Completeness (%)	100
Contamination (%)	0.1
Prokka annotation	
Total genes	5,996
Protein-coding sequences	5,897
Hypothetical proteins	2,821
tRNA genes	77
rRNA genes	21
TmRNA genes	1
BioProject	PRJNA1517283
BioSample	SAMN62642883
GenBank Accession Number	CM195964; JCCJPS010000000

3.2. Genome-Based Taxonomic Placement of MUM256 within *Streptomyces albidoflavus*

Genome-based analyses consistently supported the assignment of MUM256 to *Streptomyces albidoflavus*. GTDB-Tk classified the final genome as *S. albidoflavus* and identified GCF_000719955.1 (NRRL B-1271) as the closest placement reference, with a skani ANI of 98.8% and an alignment fraction of 0.94.

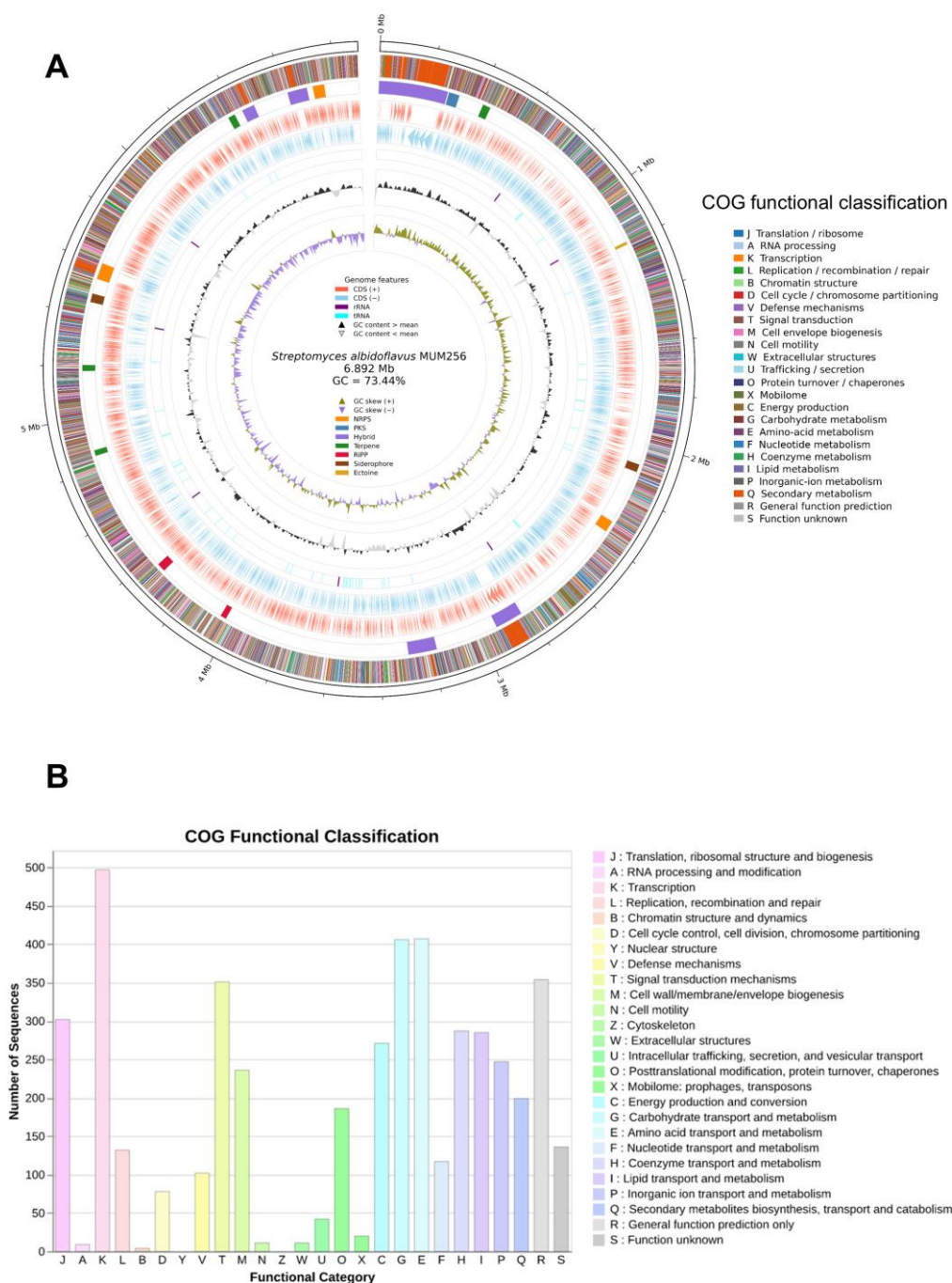


Figure 1. Genome architecture and functional profile of *Streptomyces albidoflavus* MUM256. (A) Circular representation of the 6.892-Mb MUM256 genome (GC content, 73.44%). From the outermost to innermost tracks: genomic coordinates, protein-coding genes coloured according to Clusters of Orthologous Groups (COG) functional category, 18 antiSMASH-predicted biosynthetic gene clusters (BGCs) classified by biosynthetic type, coding sequences (CDSs) on the forward and reverse strands, rRNA and tRNA genes, GC content relative to the genome-wide mean, and GC skew. BGC classes comprise non-ribosomal peptide synthetase (NRPS), polyketide synthase (PKS), hybrid, terpene, ribosomally synthesised and post-translationally modified peptide (RiPP), siderophore, and ectoine clusters. (B) Distribution of COG functional classifications among MUM256 proteins. Bars indicate the number of sequences assigned to each functional category; COG category abbreviations and definitions are provided in the panel key.

To further evaluate genomic relatedness, 165 publicly available genomes annotated as *S. albidoflavus* were retrieved from the NCBI assembly database on 26 August 2026 (Supplementary Material, Table S2) after applying quality criteria of $\geq 95\%$ CheckM completeness and $\leq 5\%$ contamination. Independent pairwise FastANI comparisons between MUM256 and these genomes yielded ANI values ranging from 95.4018% to 99.2086%, indicating substantial genomic diversity among strains currently assigned to the species. MUM256 shared 98.6955% ANI with NRRL B-1271, with 2,145 of 2,297 genomic fragments successfully mapped. Complementary all-versus-all ANI analysis of the complete 166-genome working collection, comprising MUM256 and the 165 NCBI genomes, was visualised using ANIclustermap v2.0.1 and is presented as a genome-wide ANI heatmap in Supplementary Material (Figure S2).

Digital DNA–DNA hybridisation (dDDH) analyses provided concordant evidence of species-level relatedness. Against NRRL B-1271, GGDC Formula 2 yielded a dDDH estimate of 88.4% (95% confidence interval, 85.9–90.5%) and a 95.22% probability of exceeding the conventional 70% species threshold (Table 2). Similarly high genomic relatedness was observed between MUM256 and several historically named *Streptomyces* reference strains, including *S. limosus* NBRC 12790, *S. sampsonii* NBRC 13083, *S. coelicolor* DSM 40233, and *S. albidoflavus* DSM 40455, all of which showed ANI values of approximately 98.7% and dDDH values of 88.3–89.0%. In contrast, the more distantly related *S. koyangensis* VK-A60 and *S. intermedius* JCM 4483 showed lower ANI ($\sim 95.9\%$) and dDDH (~ 63 – 64%) values.

The genome-based TYGS phylogeny further resolved the taxonomic context of MUM256 (Figure 2A). MUM256 was positioned within a strongly supported lineage comprising genomes currently or historically designated as *S. albidoflavus*, *S. limosus*, *S. sampsonii*, and *S. coelicolor* DSM 40233. Despite their differing deposited species names, these genomes were assigned by TYGS to the same species and subspecies clusters. Together with their high ANI and dDDH values, this clustering highlights the persistence of legacy nomenclature among closely related *Streptomyces* reference strains (Table 2). The MUM256-containing lineage was clearly separated from more distant type or reference strains, including *S. intermedius*, *S. koyangensis*, *S. gougerotii*, and related taxa.

To further resolve the position of MUM256 within the genomic diversity of *S. albidoflavus*, the initial comparative genome collection was subsequently refined to a quality-filtered cohort of 151 genomes. This cohort was used both for core-genome phylogenetic reconstruction and for the pangenome analyses described in Section 3.3, with the rationale and consequences of cohort filtering examined further in that section.

Core-genome phylogenetic reconstruction of the 151-genome cohort revealed substantial but structured intraspecific diversity (Figure 2B). Quantitative partitioning of the phylogeny identified five phylogroups comprising 125, 14, 7, 3, and 2 genomes, respectively. These groups formed contiguous and phylogenetically coherent lineages, with the optimal

five-group solution yielding a mean silhouette coefficient of 0.746. MUM256 was assigned to PG1, the predominant 125-genome phylogroup.

Table 2. Representative genome relatedness comparisons between *Streptomyces albidoflavus* MUM256 and selected *Streptomyces* reference strains.

Reference strain	FastANI (%)	dDDH d4 (%)	Probability DDH > 70%
<i>Streptomyces limosus</i> NBRC 12790	98.7034	89.0	95.43
<i>Streptomyces sampsonii</i> NBRC 13083	98.7023	89.0	95.43
<i>Streptomyces coelicolor</i> DSM 40233	98.7021	88.7	95.34
<i>Streptomyces albidoflavus</i> NRRL B-1271	98.6955	88.4	95.22
<i>Streptomyces albidoflavus</i> DSM 40455	98.7058	88.3	95.19
<i>Streptomyces koyangensis</i> VK-A60	95.9184	63.9	64.28
<i>Streptomyces intermedius</i> JCM 4483	95.9414	63.4	62.97

Pairwise ANI comparisons further supported this genomic subdivision. The 151-genome ANI matrix resolved five corresponding genomic groups comprising 125, 14, 7, 3, and 2 genomes, closely mirroring the independently inferred core-genome phylogroups. The largest 125-genome group contained both MUM256 and the *S. albidoflavus* reference strain NRRL B-1271 and showed 97.64–100% ANI to NRRL B-1271, with within-group ANI values of 97.51–99.999%. In contrast, the four smaller groups showed lower ANI to NRRL B-1271, ranging from approximately 95.63% to 96.30%, while retaining comparatively high within-group ANI values. Notably, Group 3 exhibited particularly strong internal genomic coherence, with within-group ANI values of 99.15–99.74%, despite showing only 95.91–96.07% ANI to NRRL B-1271 (Supplementary Material, Table S3).

MUM256 exhibited 98.70% ANI to NRRL B-1271, consistent with its placement within the predominant reference-associated lineage. Together, these results indicate that the lower ANI values observed across the broader cohort reflect structured genomic divergence among distinct lineages rather than uniformly distributed variation.

Collectively, the GTDB-Tk, ANI, dDDH, TYGS and core-genome phylogenetic analyses firmly support the assignment of MUM256 to *S. albidoflavus* and place it within the major NRRL B-1271-associated genomic lineage. At the same time, the concordance between ANI-defined grouping and core-genome phylogroup structure indicates substantial genomic heterogeneity among genomes currently assigned to *S. albidoflavus*. Formal taxonomic delineation of the four smaller divergent lineages was beyond the scope of the present study; accordingly, these genomes are treated here as distinct genomic lineages within the broader *S. albidoflavus*-assigned comparative cohort.

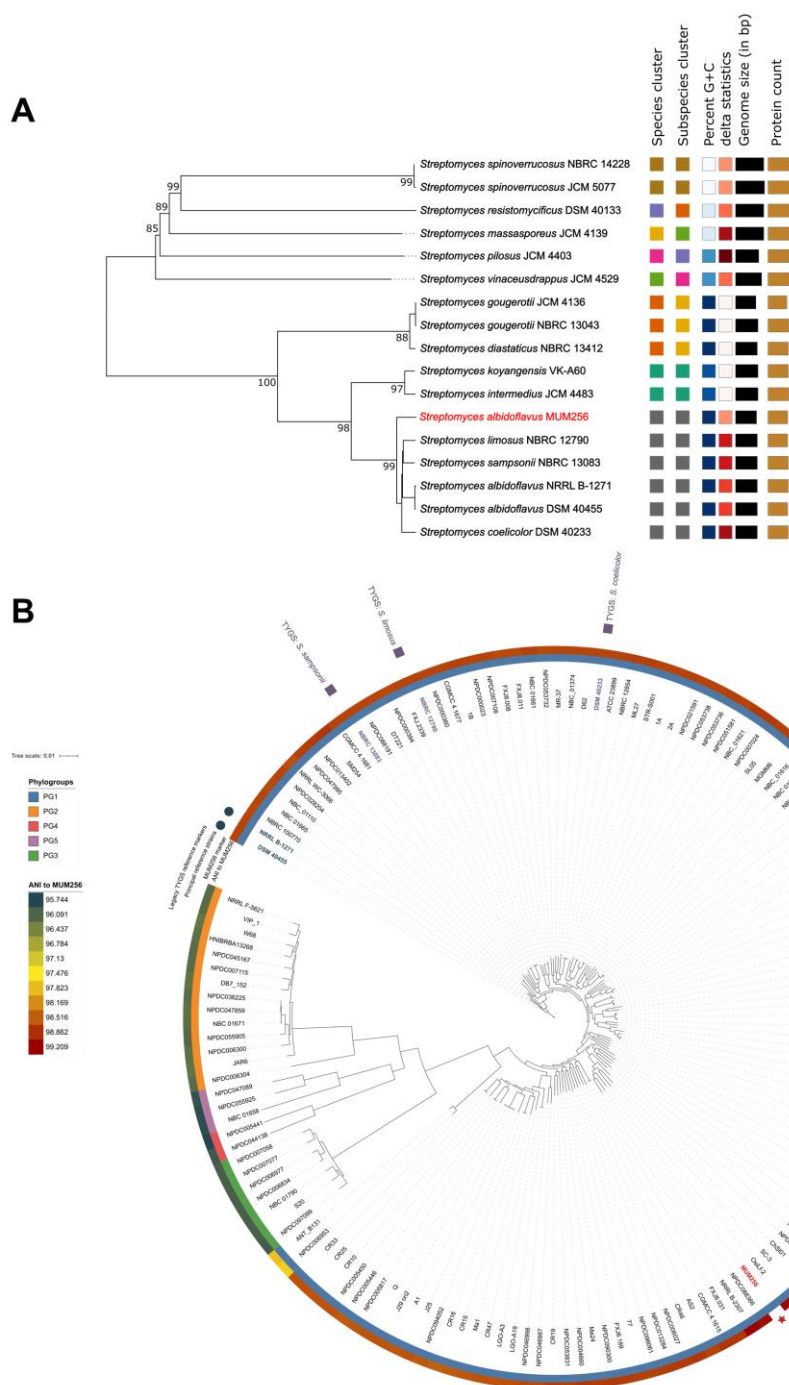


Figure 2. Genome-based taxonomic placement and phylogenomic context of *Streptomyces albidoflavus* MUM256. (A) Genome-based phylogeny generated using the Type Strain Genome Server (TYGS), showing the placement of MUM256 relative to selected type/reference strains. Branch support values are indicated at internal nodes. TYGS species- and subspecies-level cluster assignments are shown together with genome G+C content, δ statistics, genome size, and predicted protein count. MUM256 is highlighted in red. (B) Maximum-likelihood core-genome phylogeny of 151 quality-filtered genomes reconstructed from the concatenated alignment of 4,287 selected core gene clusters. The inner annotation ring denotes five quantitatively defined phylogroups (PG1–PG5), while the outer ring represents pairwise average nucleotide identity (ANI) to MUM256. MUM256 is indicated by a red star. Reference genomes carrying legacy species designations that were included in the TYGS analysis are marked for comparison between the two phylogenetic frameworks. Branch lengths are proportional to sequence divergence; the scale bar indicates substitutions per site.

3.3. Population Pangenome Structure of *S. albidoflavus*

A total of 166 genomes, comprising MUM256 and 165 publicly available genomes annotated as *S. albidoflavus*, were initially considered for comparative genomic analysis. Before pangenome reconstruction, the dataset was filtered to reduce potential biases arising from artificial genome manipulation and incomplete gene models. Three engineered genomes were excluded^[73], followed by removal of 12 genomes in which more than 10% of predicted genes were classified by PGAP2 as incomplete or discarded during preprocessing. The resulting primary cohort therefore comprised 151 genomes, including MUM256. Compared with the unfiltered 166-genome analysis, the quality-filtered cohort showed a larger strict-core component and a smaller total pangenome, consistent with reduced inflation of apparent gene absence and accessory diversity by lower-quality gene predictions. Accordingly, all primary pangenome partitioning, accumulation analyses and core-genome phylogenetic reconstruction were based on the 151-genome cohort, while the 166-genome analysis was retained as a sensitivity comparison (Supplementary Material, Figure S3).

Pangenome analysis of the 151 quality-filtered *S. albidoflavus* genomes identified a total of 21,082 non-redundant gene clusters (Figure 3A). Of these, 2,684 (12.73%) were assigned to the strict-core partition, whereas a further 904 (4.29%) and 922 (4.37%) were classified as core and soft-core clusters, respectively. Collectively, these three highly conserved partitions accounted for 4,510 gene clusters (21.39% of the pangenome). In contrast, the shell fraction comprised 2,125 clusters (10.08%), while 14,447 clusters (68.53%) were assigned to the cloud genome. Thus, more than two-thirds of the total gene-cluster repertoire consisted of low-frequency genes, revealing extensive accessory genomic diversity within the analysed cohort.

Genome-accumulation analysis further demonstrated that the size of the pangenome continued to increase as additional genomes were incorporated (Figure 3B). At the complete 151-genome sampling depth, the observed pangenome reached 21,082 gene clusters. The cumulative pangenome followed the fitted Heaps' law relationship $P(n) = 4.49 \times 10^3 n^{0.307}$. The positive exponent ($\gamma = 0.307$) is consistent with an open pangenome, in which additional genomic sampling continues to expand the collective gene repertoire. Conversely, the number of conserved gene clusters decreased progressively with increasing genome sampling. The core-genome trajectory was described by an exponential-decay model with an asymptote, $C(n) = 2.42 \times 10^3 e^{-0.023n} + 2.74 \times 10^3$, reflecting progressive stabilisation of the conserved genomic component as cohort size increased.

The rate at which previously unobserved gene clusters were recovered declined markedly as progressively more genomes were sampled (Figure 3C). The new-cluster trajectory followed $\Delta P(n) = 1.01 \times 10^3 n^{-0.621}$, decreasing from several hundred newly detected clusters during early genome additions to an estimated 40.17 clusters at the 151-genome sampling depth. Importantly, the discovery rate remained above zero despite the substantial cohort size, supporting the inference that the *S. albidoflavus* gene repertoire had not reached complete saturation.

The gene presence/absence landscape was consistent with these accumulation patterns (Supplementary Material, Figure S4). A broadly conserved block corresponding to the combined strict-core, core, and soft-core fractions was retained across most genomes, whereas shell and particularly cloud clusters displayed substantially greater strain-to-strain variability. Numerous cloud clusters occurred in restricted subsets of genomes or individual lineages, producing a sparse accessory-gene landscape beyond the conserved genomic backbone. Hierarchical clustering based on gene-content profiles further revealed structured variation among strains, indicating that the large accessory repertoire is not distributed uniformly across the species cohort.

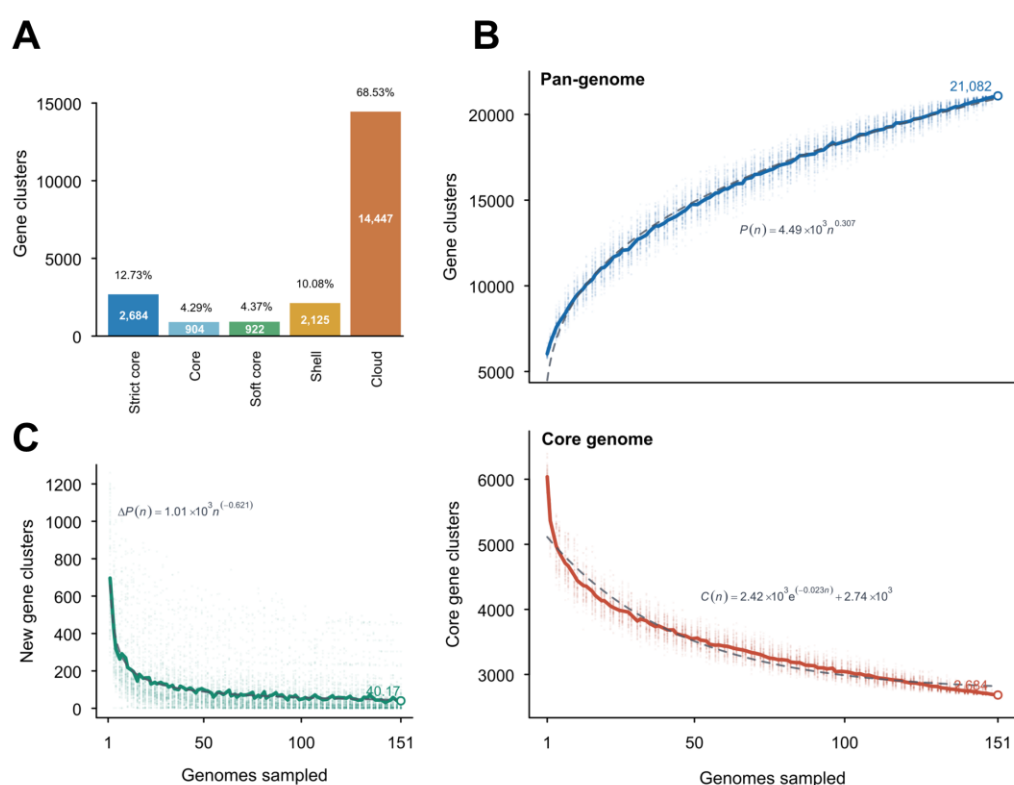


Figure 3. Pangenome composition and accumulation dynamics of the 151-genome *Streptomyces albidoflavus* cohort. (A) Distribution of the 21,082 gene clusters identified by PGAP2 among five frequency-based pangenome partitions: strict core (2,684; 12.73%), core (904; 4.29%), soft core (922; 4.37%), shell (2,125; 10.08%), and cloud (14,447; 68.53%). (B) Cumulative pan-genome and core-genome trajectories with increasing numbers of sampled genomes. The pan-genome was fitted using a Heaps' law model, $P(n) = \kappa n^\gamma$, whereas the core genome was fitted using an exponential-decay model with an asymptote, $C(n) = C_0 e^{-an} + C_\infty$. Pale points/trajectories represent genome-sampling permutations, and the solid lines indicate the fitted trends. (C) Number of newly detected gene clusters contributed with increasing genome sampling, fitted using the power-law model $\Delta P(n) = \kappa_{\text{new}} n^{-\alpha}$. At 151 genomes, the pangenome contained 21,082 gene clusters, while the fitted new-cluster trajectory remained above zero, indicating continued gene-cluster discovery with additional genome sampling.

3.4. Population-Level Biosynthetic Diversity and GCF Structure

Genome mining with antiSMASH v8.0.4 identified 18 putative biosynthetic gene cluster (BGC) regions in the MUM256 genome under the strict detection setting (Table 3). These encompassed a diverse range of biosynthetic classes, including non-ribosomal peptide synthetase (NRPS), type I–III polyketide synthase (PKS), terpene, ribosomally synthesised and post-translationally modified peptide (RiPP), siderophore and ectoine pathways, together with several multi-type regions. This diversity provided the basis for examining how the MUM256 biosynthetic repertoire is positioned within the broader *Streptomyces albidoflavus* population.

Table 3. Predicted biosynthetic gene clusters in *Streptomyces albidoflavus* MUM256 and their closest MIBiG references.

No.	AntiSMASH region	Predicted BGC type	Closest MIBiG reference cluster (biosynthetic class)	Similarity (%)*
1	1.1	T1PKS, NRPS, Lanthipeptide (Class II)	Candicidin (NRPS:Type I+PKS)	100
2	1.2	T3PKS	Naringenin (PKS:Type III)	100
3	1.3	Terpene	Isorenieratene (Terpene)	75
4	1.4	Ectoine	Ectoine (other:other)	100
5	1.5	NI-siderophore	Desferrioxamin B (other:other)	100
6	1.6	NRPS	Dudomycin A (NRPS:Type I)	17
7	1.7	Azole-containing-RiPP, NRPS	Surugamide A/Surugamide D (NRPS:Type I)	100
8	1.8	NRPS, T2PKS	Fredericamycin A(PKS:Type II)	96
9	1.9	Lanthipeptide (Class III)	SAL-2242 (Ribosomal:RiPP:Lanthipeptide)	100
10	1.10	Azole-containing-RiPP	Fluostatin M/fluostatin N/fluostatin O/fluostatin P/fluostatin Q (PKS:Type II aromatic)	4
11	1.11	Terpene	Albaflavenone (Terpene)	100
12	1.12	Terpene	Geosmin (Terpene:Sesquiterpene)	100
13	1.13	NI-siderophore	Kinamycin (PKS)	13
14	1.14	NRPS	Cyclofaulknamycin (PKS)	100
15	1.15	Terpene	Hopene (Terpene)	76
16	1.16	NRPS, T1PKS	SGR PTMs, SGR PTM Compound b, SGR PTM Compound c, SGR PTM Compound d (NRPS:Type I+PKS)	100
17	1.17	NRPS, Terpene	Valinomycin/Montanastatin (Saccharide:hybrid/tailoring+NRPS:Type I)	13
18	1.18	NRPS	SF2768 (NRPS:Type I)	66

*KnownClusterBlast similarity. NRPS - Non-ribosomal peptide synthetase; T1PKS – Type 1 Polyketide synthases; T2PKS – Type 2 Polyketide synthases; T3PKS – Type 3 Polyketide synthases; RiPP – Ribosomally synthesised and post-translationally modified peptides.

Population-level analyses were based primarily on a contiguity-filtered cohort of 119 genomes, comprising 118 publicly available genomes together with MUM256. These genomes contained 2,651 antiSMASH BGC records, which were grouped by BiG-SCAPE into 178 gene cluster families (GCFs) at a distance cutoff of 0.3. To evaluate whether exclusion of highly fragmented assemblies materially altered the inferred biosynthetic landscape, an independent sensitivity analysis was also performed using the broader 163-genome cohort. This larger dataset contained 4,236 BGC records assigned to 315 GCFs, but showed a closely comparable predominance of low-frequency families: rare and singleton GCFs accounted for 272 of 315 families (86.3%), compared with 150 of 178 (84.3%) in the 119-genome cohort (Supplementary Material, Figure S5). The consistency of this population-level pattern supported the use of the 119-genome cohort for the primary prevalence, phylogenetic and network analyses while reducing potential effects of assembly fragmentation on BGC reconstruction.

The 178 GCFs in the primary cohort displayed a strongly skewed prevalence distribution (Figure 4A). Six GCFs were cohort-wide, three were near-core, seven were common and 12 were accessory, whereas 74 were rare and 76 were singletons. Thus, 84.3% of all GCFs occurred in fewer than 15% of genomes or in only a single genome. Despite their numerical predominance, however, low-frequency families accounted for a relatively small fraction of total BGC occurrences. The 16 cohort-wide, near-core and common GCFs represented only 9.0% of all families but contained 1,755 of the 2,651 BGC records (66.2%). When accessory GCFs were included, the 28 families occurring in at least 15% of genomes collectively accounted for 80.8% of all BGC records. The biosynthetic landscape therefore comprised a relatively small set of recurrent GCFs contributing the majority of BGC occurrences, superimposed on a much larger pool of sparsely distributed families; the contribution of these low-frequency families to individual genome repertoires was further examined in relation to the core-genome phylogeny.

Mapping GCF presence and absence onto the core-genome phylogeny revealed substantial variation in biosynthetic repertoire among individual strains and phylogenetic groups (Figure 4A). Widely distributed GCFs extended across much of the species phylogeny, whereas accessory, rare, and singleton families showed progressively more restricted distributions. At the individual-genome level, the 119 genomes contained a median of 20 distinct GCFs per genome (Interquartile range (IQR) 19–24; range 14–40) (Figure 4B). MUM256 contained 18 BGCs assigned to 18 distinct GCFs, placing its GCF repertoire at approximately the 23.5th percentile of the cohort distribution. These comprised six cohort-wide, three near-core, five common and four accessory families, with no rare or singleton GCFs detected.

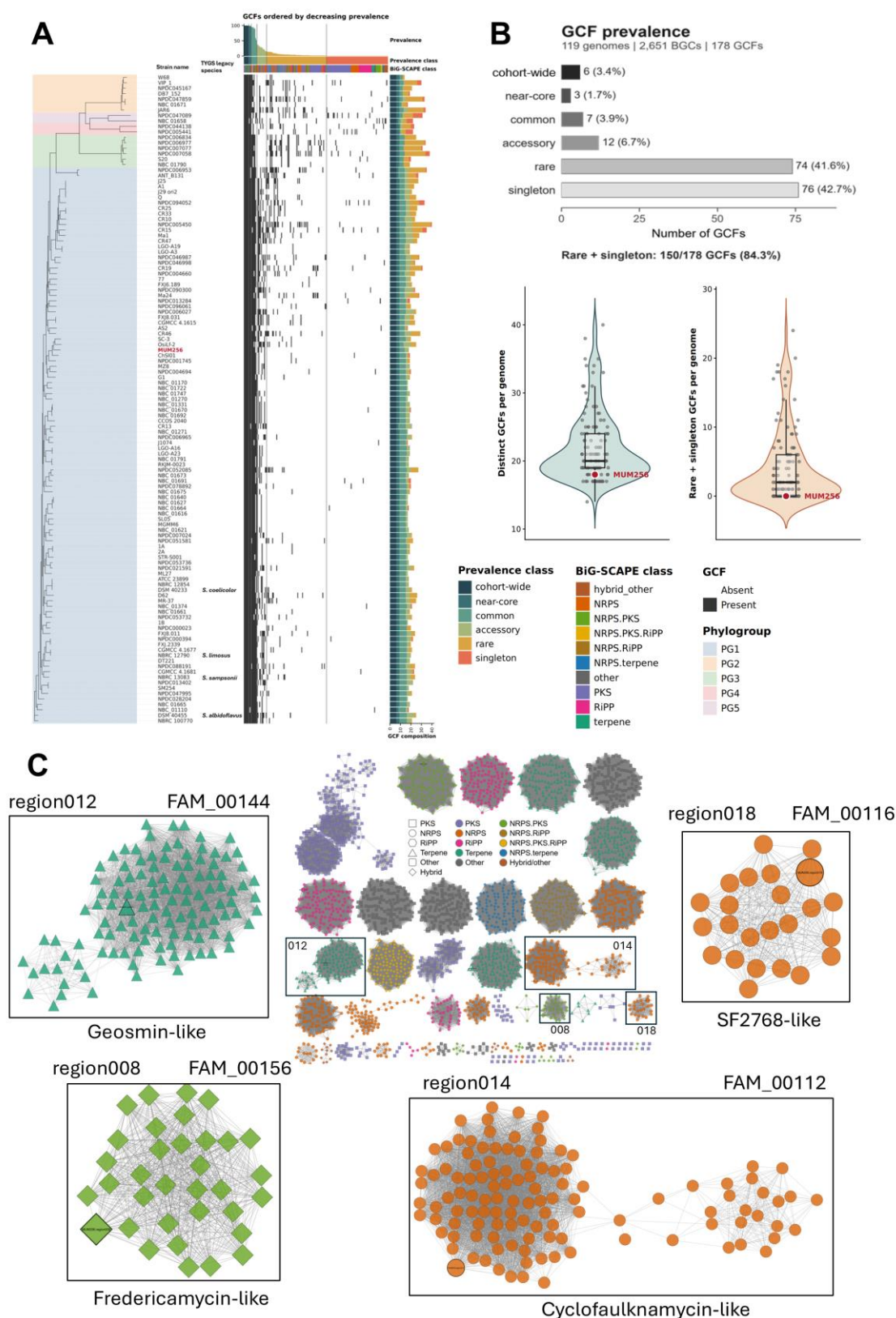


Figure 4. Biosynthetic gene cluster diversity and gene cluster family landscape across *Streptomyces albidoflavus*. **(A)** Distribution of biosynthetic gene cluster families (GCFs) across the 119-genome contiguity-filtered cohort. The core-genome phylogeny is shown at left, with background shading indicating the five

genome-based phylogroups (PG1–PG5); MUM256 is highlighted in red. Rows of the binary matrix represent individual genomes and columns represent the 178 GCFs ordered by decreasing prevalence, with dark cells indicating GCF presence. The upper profile shows the percentage of genomes containing each GCF, with annotation tracks indicating prevalence class and BiG-SCAPE biosynthetic class. Stacked horizontal bars at right show the number and prevalence-class composition of distinct GCFs detected in each genome, ordered from cohort-wide and near-core to rare and singleton families. **(B)** Cohort-level prevalence and genome-level distribution of GCF repertoires. The upper bar chart shows the numbers of GCFs assigned to the six prevalence classes: cohort-wide (6), near-core (3), common (7), accessory (12), rare (74) and singleton (76); rare and singleton GCFs together accounted for 150 of 178 families (84.3%). Violin plots below show the distributions of total distinct GCFs per genome and rare-plus-singleton GCFs per genome across the 119-genome cohort. Individual points represent genomes, embedded boxplots summarise the median and interquartile range, and MUM256 is highlighted in red. **(C)** BiG-SCAPE similarity network of 2,651 BGC records from the 119-genome cohort, with nodes representing individual BGCs and edges representing retained similarity relationships. Node colour and shape encode biosynthetic classification. Enlarged subnetworks show four MUM256-containing GCFs associated with geosmin-, fredericamycin A-, SF2768- and cyclofaulnamycin-like reference clusters. Boxes in the global network indicate the locations of the corresponding enlarged GCFs.

The contribution of low-frequency GCFs varied substantially among individual genomes (Figure 4B). Rare and singleton GCFs together ranged from 0 to 24 per genome, with a median of 2 (IQR 0–6). Thirty-four of the 119 genomes (28.6%) contained no rare or singleton GCFs, indicating that this feature was not unique to MUM256. Thus, although rare and singleton families dominated GCF diversity at the cohort level, their contribution to individual biosynthetic repertoires was highly uneven. MUM256 therefore carried a repertoire composed entirely of GCFs represented elsewhere in the sampled genome collection, rather than an unusually enriched set of low-frequency or MUM256-specific families.

BiG-SCAPE network analysis further resolved the relationships among the predicted BGCs, revealing large multi-member components alongside numerous smaller families and isolated clusters (Figure 4C). Selected MUM256-containing GCFs were highlighted to illustrate this range of population distributions, including region012 (FAM_00144; geosmin-like), region008 (FAM_00156; fredericamycin A-like), region018 (FAM_00116; SF2768-like) and region014 (FAM_00112; cyclofaulnamycin-like). Among these, the accessory region008 and region018 families were subsequently selected for detailed gene-level comparison with characterised reference clusters and representative *S. albidoflavus* loci in section 3.5.

3.5. Comparative Architecture of Selected MUM256 Biosynthetic Loci

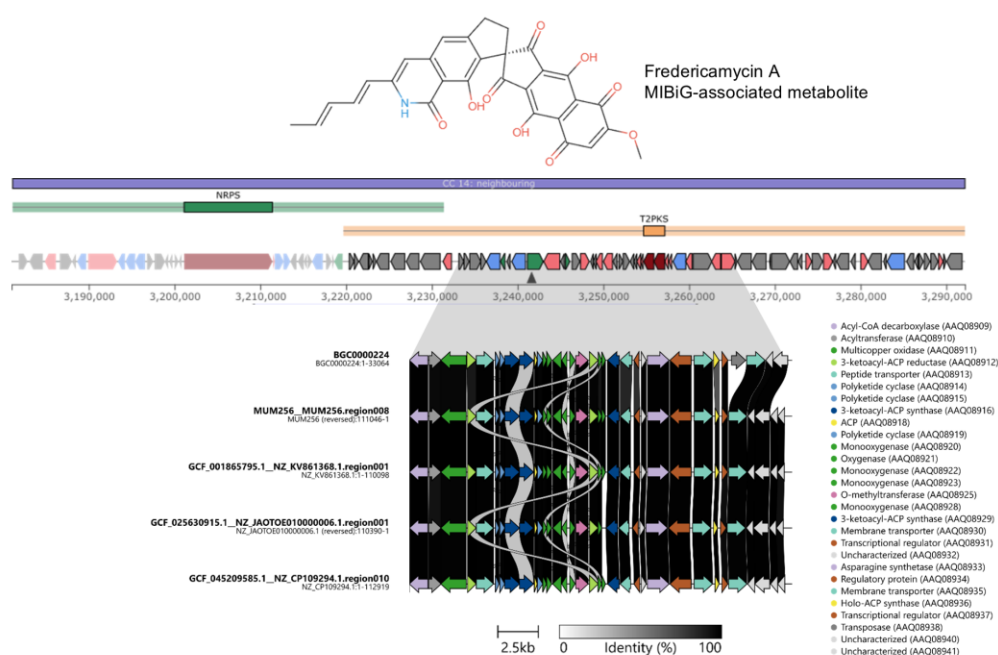
Region008 was predicted as an NRPS-type II polyketide synthase locus and showed a fredericamycin-like association with MIBiG reference BGC0000224.4 [69]. The antiSMASH KnownClusterBlast similarity score was 96%. Quantitative clinker comparison identified homologous links for 32 of 33 assessed reference proteins (97%), with a median amino-acid identity of 98.81% among linked protein pairs (Figure 5A). The conserved components included the principal type II polyketide synthase machinery, comprising ketosynthase-, chain-length factor- and acyl-carrier-protein-associated proteins, together

with cyclisation, tailoring, regulatory and transport-associated components. The single reference without a retained comparison link was annotated as a transposase (AAQ08938).

A

region008 — fredericamycin A-like BGC

FAM_00156 | 34/119 | accessory | KnownClusterBlast similarity 96%



B

region018 — SF2768-like BGC

FAM_00116 | 24/119 | accessory | KnowClusterBlast similarity 66%

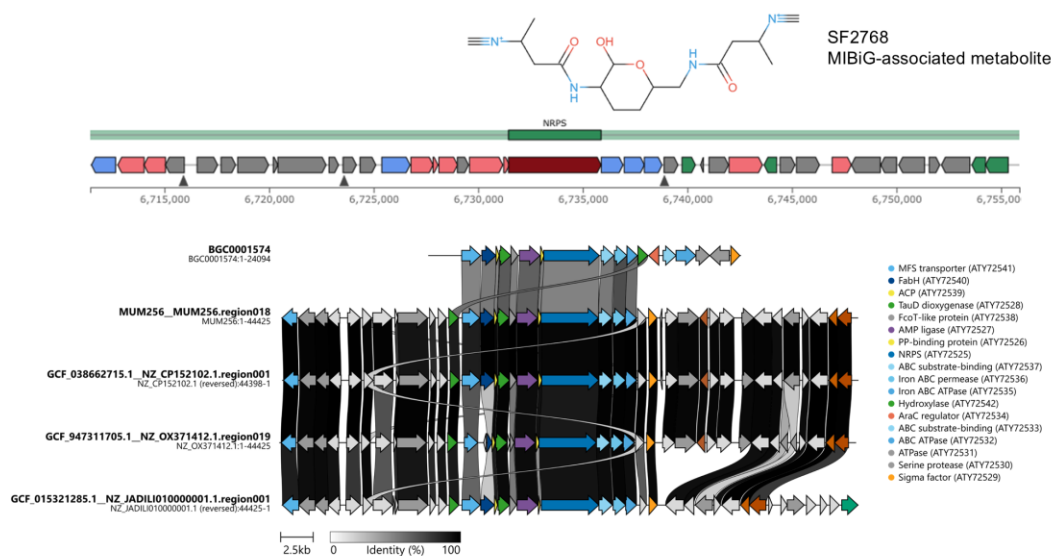


Figure 5. Comparative architecture of accessory fredericamycin A- and SF2768-like biosynthetic loci in *Streptomyces albidoflavus* MUM256. **(A)** MUM256 region008, assigned to FAM_00156 and present in 34 of 119 genomes (28.6%; accessory), showing its genomic context and comparative organisation relative to the characterised fredericamycin A BGC BGC0000224.4 and representative FAM_00156 loci. antiSMASH

KnownClusterBlast reported 96% similarity to the reference cluster. **(B)** MUM256 region018, assigned to FAM_00116 and present in 24 of 119 genomes (20.2%; accessory), compared with the characterised SF2768 BGC BGC0001574.5 and representative homologous loci; KnownClusterBlast similarity to the reference was 66%. Upper tracks show the corresponding antiSMASH-predicted MUM256 regions and the positions of biosynthetic components, whereas lower Clinker comparisons show gene-level organisation of the reference-homologous regions. Arrows represent predicted coding sequences, with homologous functional groups indicated by consistent colours; connecting regions are shaded according to amino-acid sequence identity. Chemical structures represent the reported metabolites associated with the corresponding MIBiG reference BGCs and do not establish production of these compounds by MUM256.

Region018 was predicted as an NRPS locus and showed an SF2768-like association with MIBiG reference BGC0001574.5, with a KnownClusterBlast similarity score of 66% [70]. Homologous links were identified for 12 of 18 assessed reference proteins (66.7%), with a median amino-acid identity of 60.29% among linked protein pairs (Figure 5B). Corresponding components included NRPS-, AMP-ligase-, phosphopantetheinyl-binding-, acyl-carrier-protein-, FabH-, TauD-dioxygenase-, hydroxylase- and transport-associated proteins. Six reference proteins had no retained comparison links and were annotated as a sigma factor, serine protease, ATPase, two ABC-transporter components and an AraC-family regulator.

Collectively, region008 showed substantially greater correspondence to its characterised MIBiG reference than region018 at both the proportion of linked reference components and amino-acid identity levels. The absence of individual clinker comparison links was recorded as a lack of detectable correspondence under the applied comparison criteria and was not treated as evidence of confirmed gene deletion. The region014 comparison with the cyclofaulknamycin-associated BGC was retained in Supplementary Material (Figure S6).

4. Discussion

The present study extends the characterisation of *Streptomyces albidoflavus* MUM256 from that of an individual bioactive mangrove isolate to a comparative genomic framework across *S. albidoflavus*-assigned genomes. Previous investigations demonstrated antioxidant activity and selective cytotoxicity of MUM256 extracts and subsequently showed cell-cycle arrest and induction of apoptosis in HCT116 colon cancer cells [18, 19]. Those studies established the biological interest of the strain but provided limited information regarding its genome-wide taxonomic position, the population context of its gene content or the distribution of its predicted biosynthetic capacity. By integrating a high-quality single-contig genome with a 151-genome pangenome analysis and a quality-controlled 119-genome BGC/GCF analysis, the present study places MUM256 within the genomic and biosynthetic diversity represented by a genetically open and heterogeneous cohort of *S. albidoflavus*-assigned genomes.

The availability of the MUM256 genome substantially strengthens its taxonomic placement. Earlier characterisation of environmental *Streptomyces* frequently relied heavily on 16S rRNA gene similarity, despite the recognised difficulty of resolving closely related

species using this marker alone ^[20]. In the present study, core-genome phylogeny, GTDB-Tk classification, ANI and dDDH analyses converged on assignment of MUM256 to *S. albidoflavus*. MUM256 shared approximately 98.7–98.8% ANI with the principal reference strain NRRL B-1271 and a dDDH value of 88.4%, well above commonly applied genomic species boundaries. Large-scale bacterial genome comparisons have established approximately 95–96% ANI as a useful general species boundary, although *Streptomyces*-specific analyses indicate that species delineation can benefit from somewhat more stringent interpretation and integration with phylogenomics and other taxonomic evidence ^[21, 22]. Current recommendations similarly emphasise the use of multiple overall genome relatedness indices rather than reliance on any single metric ^[23]. The historical species names associated with some closely related reference genomes in TYGS and NCBI should therefore be interpreted primarily as nomenclatural and database-history complexity rather than as evidence contradicting the genome-based placement of MUM256.

The core-genome analysis further revealed structured diversity within the analysed cohort, with five monophyletic phylogroups recovered from the 151-genome cohort. This population structure is consistent with earlier evidence that genetic differentiation within *S. albidoflavus* is not randomly distributed. Cheng *et al.* ^[24] identified habitat-associated population structure and reduced homologous gene flow among soil-, marine- and insect-associated groups, while Li *et al.* ^[25] subsequently resolved three genome-wide phylogenomic clades and linked their differentiation to ecological adaptation, habitat-associated genomic variants and differences in recombination. The five phylogroups identified here were defined solely from core-genome patristic distances rather than environmental metadata and therefore should not be assumed to correspond directly to the ecological groups reported previously. Nevertheless, their recovery in a substantially expanded genome cohort indicates that appreciable lineage structure persists among contemporary genomes assigned to *S. albidoflavus*. Integration of standardised habitat, geographic and host-association metadata will be required to determine whether these phylogroups reflect ecological differentiation similar to that described in earlier populations.

A major finding of the present study is that the pangenome of the analysed *S. albidoflavus*-assigned cohort remains open despite the substantial increase in genome sampling. The 151-genome primary cohort contained 21,082 gene clusters. The cumulative pangenome followed a positive Heaps' law exponent ($\gamma = 0.307$), while new-cluster discovery declined with an exponent of $\alpha = 0.621$; approximately 40 additional clusters were still being detected near the 151-genome sampling point. These patterns are consistent with continued gene-family accumulation rather than convergence towards a closed gene repertoire ^[47]. This finding extends earlier observations rather than establishing pangenome openness for the first time. Li *et al.* ^[25] reported an open pangenome from 30 *S. albidoflavus* strains, Gonzalez-Silva *et al.* ^[27] reached a similar conclusion using a smaller eight-genome dataset, Popov *et al.* ^[28] subsequently identified 11,638 gene families across 34 complete genomes, including 3,867 strict-core families. The present analysis therefore provides an independent confirmation using a substantially larger and explicitly quality-filtered comparative cohort.

Absolute pangenome sizes across these studies should not be compared directly because genome selection, annotation procedures, clustering algorithms and sequence-similarity thresholds differ among analyses. More informative is the recurrent qualitative pattern: increasing genome sampling continues to uncover additional gene families while preserving a substantial accessory component, indicating that genomic diversity represented among *S. albidoflavus*-assigned genomes remains incompletely sampled. This pattern is consistent with broader observations across *Streptomyces*, where comparative analysis of 213 species revealed extensive variability in accessory genes and specialised-metabolite biosynthetic loci, highlighting the evolutionary plasticity of the genus ^[74]. The value of integrating pangenome structure with phylogenomic context has also been demonstrated in other bacterial populations; for example, a recent analysis of 118 clinical *Clostridioides difficile* genomes combined phylogenomics with pangenome analysis to resolve population structure and accessory-genome diversity ^[75]. Although biologically distinct from *S. albidoflavus*, such studies illustrate the broader utility of combining core-genome phylogeny with gene-content variation when interpreting structured bacterial populations.

The internal composition of the pangenome further illustrates the scale of this diversity. Cloud clusters accounted for 14,447 of the 21,082 clusters (68.53%), whereas 4,510 clusters occurred in at least 95% of the retained genomes. The high proportion of cloud clusters should not be interpreted to mean that most genes within an individual *S. albidoflavus* genome are rare. Instead, it reflects the accumulation of a large cohort-level reservoir of individually low-frequency gene families superimposed on a conserved genomic backbone. The comparison between the original 166-genome dataset and the quality-controlled 151-genome cohort also demonstrates the importance of genome curation. Filtering reduced the total pangenome by 1,163 clusters, of which 1,147 belonged to the cloud compartment, while simultaneously increasing the inferred strict-core component. Fragmented assemblies, incomplete gene models and annotation inconsistencies are known to inflate inferred accessory genomes and erode apparent core-genome size in bacterial pangenome analyses ^[76]. The excluded clusters cannot automatically be regarded as artefacts, because genuine rare genes may also be preferentially associated with lower-quality assemblies. Nevertheless, the concentration of the change within the low-frequency compartment indicates that rare-gene interpretations are particularly sensitive to data quality and supports the use of the curated 151-genome cohort as the primary dataset.

The large accessory gene repertoire was paralleled by pronounced heterogeneity in specialised-metabolite potential. Across the 119-genome primary BGC cohort, 2,651 BGC records were grouped into 178 computationally defined GCFs. Only six GCFs were cohort-wide, and three were near-core, whereas 74 were rare and 76 were singletons. Thus, rare and singleton families represented 84.3% of the GCF types. Similar strain-level heterogeneity has been observed across *Streptomyces*, including substantial BGC differences among genomes assigned to the same species ^[2]. Genus-scale pangenome mining has likewise demonstrated that reliance on one representative genome can underestimate population biosynthetic potential and that differences in BGC boundary definition can strongly influence apparent

family diversity [8]. Most directly, the present GCF distribution can also be considered alongside the recent species-level analysis of Wang *et al.* [30], who identified 1,120 BGCs in 48 *S. albidoflavus* genomes and grouped them into 45 GCFs, comprising 20 high-frequency, seven intermediate-frequency and 18 low-frequency families. The substantially larger number of GCFs recovered here should not be interpreted as a direct increase in biological biosynthetic diversity because the two studies differed in genome sampling, antiSMASH versions, BGC boundary definitions, clustering procedures and prevalence thresholds. More informative is the shared qualitative conclusion that the sampled *S. albidoflavus*-assigned genomes contain a relatively conserved set of recurrent BGC families together with a larger pool of variably distributed loci. The experimental demonstration by Wang *et al.* [30] that a rare kosinostatin BGC reflects horizontal acquisition, vertical inheritance, recurrent loss and context-dependent selection further illustrates how low-frequency GCFs may represent genuine evolutionary features rather than merely technical noise.

An important result in this context is that the 18 BGC regions identified in MUM256 were each assigned to GCFs represented elsewhere in the 119-genome cohort. At the individual-genome level, the cohort contained a median of 20 distinct GCFs per genome (IQR 19–24; range 14–40), whereas MUM256 contained 18, placing its repertoire toward the lower part of the observed distribution. Moreover, MUM256 contained no rare or singleton GCFs, but this pattern was not exceptional: 34 of the 119 genomes (28.6%) similarly lacked GCFs from these low-prevalence categories. Thus, although rare and singleton families dominated GCF diversity at the cohort level, their contribution to individual genome repertoires was highly uneven. MUM256 is therefore better characterised by the particular combination and structural variation of recurrent GCFs it carries than by an unusually large or strain-exclusive repertoire.

This distinction is biologically important. Functional differentiation between strains can arise through variation within shared GCFs, including allelic differences, acquisition or loss of accessory biosynthetic genes, and changes in cluster architecture, without requiring the presence of a wholly unique cluster family [77, 78]. Such variation can alter pathway function or chemical output even when clusters remain sufficiently similar to be grouped within the same broader biosynthetic family [77, 78]. It therefore cautions against equating "not unique at the GCF level" with "functionally identical". The locus-level comparisons performed here provide examples of how meaningful structural variation persists within families that are distributed across multiple genomes. Recent chemical investigations of *S. albidoflavus* provide direct support for interpreting shared biosynthetic families at a finer level than simple presence or absence. Marine-derived *S. albidoflavus* RKJM-0023 was shown to produce acyl-surugamide A2, a previously unreported cyclic non-ribosomal peptide related to the established surugamide biosynthetic system [79]. Thus, even a recognisable biosynthetic context can generate previously undescribed chemical variants. Moreover, chemical elicitation of the same strain with rhamnolipids induced a previously silent molecular family and led to the discovery of alubactins A–H and related congeners [80]. These observations reinforce the interpretation that absence of MUM256-specific GCFs does

not imply an absence of strain-specific chemical potential; differences in pathway sequence, tailoring functions, regulation and cultivation-dependent expression may generate chemical diversity within broadly shared biosynthetic families.

The use of a quality-controlled BGC cohort was particularly important for this analysis. BGCs frequently contain large modular genes and repetitive sequence, making their reconstruction especially sensitive to assembly fragmentation. antiSMASH flags regions extending to contig boundaries because such predictions may represent incomplete or divided biosynthetic loci ^[81], empirical sequencing studies of actinomycetes have similarly demonstrated that fragmented assemblies can truncate BGCs and impede reconstruction of complete specialised-metabolite pathways ^[65]. Importantly, the ≤ 100 -contig threshold adopted here has precedent in large-scale BGC resources: both versions 3 and 4 of the antiSMASH database excluded bacterial and archaeal assemblies containing >100 contigs specifically to reduce the adverse effects of genome fragmentation on BGC prediction quality ^[66, 82]. Large-scale *Streptomyces* genome mining has likewise shown that BGCs terminating at contig edges are disproportionately associated with fragmented assemblies and can artificially inflate predicted BGC counts when different fragments of the same locus are reported separately ^[8].

Accordingly, exclusion of the 44 genomes containing >100 contigs from the 163-genome broad cohort was applied as a pragmatic, BGC-focused quality-control criterion rather than as a universal definition of bacterial genome quality. Comparison of the broad and contiguity-filtered cohorts showed that this filtering substantially reduced the magnitude of inferred biosynthetic diversity while preserving the overall cohort-level prevalence pattern. The broad 163-genome cohort contained 4,236 BGC records distributed among 315 GCFs, whereas the contiguity-filtered 119-genome cohort contained 2,651 BGC records and 178 GCFs. Thus, removal of 27.0% of the genomes was accompanied by reductions of 37.4% in BGC records and 43.5% in inferred GCF richness, indicating that highly fragmented assemblies contributed disproportionately to the apparent biosynthetic-family diversity. Nevertheless, the prevalence structure remained similar between cohorts, with rare and singleton families comprising 86.3% of GCFs in the broad cohort and 84.3% in the contiguity-filtered cohort. These results indicate that assembly fragmentation affects the estimated magnitude of BGC/GCF diversity without altering the broader conclusion that the *S. albidoflavus*-assigned biosynthetic landscape is numerically dominated by low-frequency families.

A complementary pangenome sensitivity analysis further supported the robustness of the broader genomic interpretation to assembly contiguity. Re-analysis of the 119-genome contiguity-filtered subset reduced the cloud fraction from 68.53% in the 151-genome PGAP2 cohort to 64.83%, consistent with a contribution of fragmented assemblies to estimates of rare gene content. However, the cumulative pangenome exponent remained comparable ($\gamma=0.289$) for the 119-genome subset versus ($\gamma=0.307$) for the 151-genome cohort, indicating that the inference of an open pangenome was retained after stringent contiguity filtering (Supplementary Material, Figure S7). Collectively, these sensitivity analyses suggest that

assembly quality influences quantitative estimates of rare genomic and biosynthetic diversity, but does not account for the principal conclusions regarding pangenome openness or predominance of low-frequency biosynthetic families.

Because the 163- and 119-genome cohorts were clustered independently by BiG-SCAPE, individual GCF assignments and prevalence estimates are inherently cohort-dependent, and the observed reduction in GCF richness cannot be attributed exclusively to assembly artefacts. Rather, the concordance of the cohort-level prevalence pattern across independently clustered cohorts supports use of the 119-genome set as the more conservative primary dataset, while retaining the 163-genome analysis as a broader sensitivity comparison.

The present comparative genomic strategy also extends an established progression of genome-enabled and phenotype-guided investigation of Malaysian mangrove-derived *Streptomyces*. Earlier studies examined *Streptomyces antioxidans* MUSC 164^T, whose extract had demonstrated antioxidative and neuroprotective activity, and the anti-MRSA strain *Streptomyces pluripotens* MUSC 135^T, using genome information to investigate biosynthetic potential underlying previously observed phenotypes [83, 84]. Related mangrove bioprospecting subsequently encompassed antioxidant and chemical profiling of strains like MUM212, MUSC 5, as well as formal description of novel mangrove taxa including *S. griseiviridis* and *S. learnhanii* [13–15, 85]. More recent work has linked phenotype and genome information in *S. pluripotens* MUM 16J and the anti-*Vibrio* strain MUM 178J, while functional studies of MUM 178J and MUM 195J further demonstrated how mangrove-derived *Streptomyces* can progress from strain discovery to antimicrobial and aquaculture applications [86–90]. Collectively, these studies reflect a broader Malaysian MOD-ACTINO research strategy in which environmentally derived actinobacteria are investigated as sources of antimicrobial, antioxidant, anticancer and other biologically active metabolites [16, 17, 91, 92]. The present study advances this trajectory by adding a population-genomic dimension; rather than asking only which biosynthetic loci occur in MUM256, it determines how those loci are distributed and structurally conserved across the broader collection of *S. albidoflavus*-assigned genomes.

Region008 exemplifies the value of integrating similarity to a characterised reference BGC with population prevalence and within-family locus conservation. antiSMASH identified this region as highly similar to the experimentally characterised fredericamycin A BGC, with a KnownClusterBlast similarity of 96%. Quantitative comparison further showed that 32 of 33 assessed MIBiG reference proteins (97.0%) retained homologous links to the MUM256 locus, with a median amino-acid identity of 98.81% among the linked proteins. The only reference component without a retained link was annotated as a transposase, although the present analysis does not establish whether this element is genuinely absent, replaced, or simply falls below the homology criterion. The conserved correspondence encompassed the principal type II PKS biosynthetic machinery together with multiple tailoring, regulatory and transport-associated genes. This close architectural resemblance is consistent with the experimentally established fredericamycin A pathway of *Streptomyces griseus*, which was demonstrated to function as a type II PKS system through gene disruption,

complementation and heterologous-expression studies ^[69]. Despite the high similarity of the MUM256 locus to this reference, its associated GCF occurred in only 34 of the 119 genomes. Thus, strong correspondence to a characterised biosynthetic pathway does not imply cohort-wide conservation. The presence of this locus is particularly noteworthy in view of the previously reported cytotoxic and apoptosis-inducing activity of MUM256 extracts ^[18, 19]. Fredericamycin A itself has documented antibacterial, antifungal and cytotoxic properties and showed antitumour activity in early in vivo tumour models ^[93]. Subsequent biochemical studies demonstrated inhibition of DNA topoisomerases I and II in vitro, although these observations did not establish a definitive mechanism for its antitumour effects ^[94]. Together, the strong reference correspondence and conserved within-GCF architecture make region008 a rational target for targeted metabolomic and functional investigation. Nevertheless, the biological activities of fredericamycin A cannot be attributed to MUM256 on genomic evidence alone. Demonstration of fredericamycin A, a related congener or another product derived from region008, followed by experimental linkage of that metabolite to the reported MUM256 phenotype, would be required.

Region018 illustrates a contrasting relationship between reference similarity and conservation across the analysed genomic cohort. This locus displayed a lower KnownClusterBlast similarity to the SF2768 reference cluster and retained comparison links for 12 of 18 assessed reference proteins (66.7%), yet selected *S. albidoflavus* homologues preserved a closely related NRPS-domain architecture, including an eight-domain organisation and, in one comparator, the same domain order observed in MUM256. In the experimentally characterised *Streptomyces thioluteus* pathway, SF2768 is encoded by the NRPS-associated *sfa* cluster and functions as a copper-binding chalkophore involved in copper acquisition ^[70]. The MUM256 locus is therefore most appropriately described as SF2768-like, rather than as an SF2768-producing pathway. Its occurrence in 24 of 119 genomes indicates that the architecture is recurrent but not widespread across the analysed cohort. This distinction is relevant to the potential ecological interpretation of region018. SF2768 is unusual in combining antibiotic activity with highly specific copper-chelating chalkophore activity, illustrating how a specialised metabolite can contribute both to antagonistic interactions and micronutrient acquisition ^[95]. The occurrence of an SF2768-like architecture in a subset of the analysed *S. albidoflavus* genomes is therefore of ecological interest, but neither production of SF2768 nor conservation of its copper-acquisition function can be inferred from sequence similarity alone. Experiments examining region018 transcription and metabolite production under different copper conditions, coupled with chemical characterisation and functional assays, would be required to determine whether the MUM256 pathway retains a comparable physiological role.

The analysis of region014 provides a further example of this distinction (Supplementary Material, Figure S6). The locus showed a strong association with the cyclofaulknamycin reference BGC and conserved NRPS-domain organisation among selected *S. albidoflavus* genomes. Cyclofaulknamycin was originally identified through genome-guided investigation of the *flk* locus and contains the uncommon non-proteinogenic

amino acid D-capreomycinidine ^[71]. However, the apparent prevalence of the region014-associated GCF differed substantially between the broad and primary BiG-SCAPE analyses. This discrepancy resulted not only from removal of fragmented genomes but also from reassignment of BGCs when the two datasets were clustered independently. This phenomenon is expected because GCFs are computational groupings determined by the input collection, similarity calculations and clustering parameters rather than immutable biological entities ^[63]. Family identifiers and prevalence estimates from independent BiG-SCAPE runs should therefore not be assumed to be directly interchangeable. Region014 also illustrates why structural similarity to a characterised BGC should not be used as a proxy for a particular biological activity. Although cyclofaulknamicin contains the unusual D-capreomycinidine residue and was recovered through targeted genome mining of the *flk* locus, Horbal *et al.* ^[71] detected no activity against their tested panel of bacteria, yeasts and moulds at concentrations of 0.03–64 µg/mL. The high reference similarity and conserved architecture observed for MUM256 region014 therefore provide information about biosynthetic relatedness, but do not imply a predictable antimicrobial phenotype. Region014 consequently serves as both a biological and methodological example: high similarity to a known cluster, structural conservation and apparent population prevalence are related but distinct properties.

Several limitations should be considered when interpreting the biosynthetic findings. First, antiSMASH and BiG-SCAPE describe genomic potential, not pathway activity or metabolite production. Many *Streptomyces* BGCs are weakly expressed or transcriptionally silent under standard laboratory conditions, and genetic activation can reveal products that are absent from routine culture extracts. CRISPR-Cas9 promoter replacement, for example, has activated multiple silent BGC classes across several *Streptomyces* species and induced previously undetected metabolites ^[96]. Metabolite production is also strongly dependent on cultivation conditions; fermentation variables, including medium composition, temperature and pH, can substantially alter secondary-metabolite output from the same strain ^[97]. In addition, developmental state, morphology and phenotypic heterogeneity can influence specialised-metabolite production, introducing a further regulatory layer between genome content and measured chemistry ^[98]. Consequently, failure to detect a MUM256 metabolite under one culture condition would not establish that the corresponding BGC is inactive. At a finer biological scale, phenotypic heterogeneity within *Streptomyces* cultures may further uncouple genome content from average population-level metabolite production, motivating future application of single-cell approaches to identify high-producing subpopulations ^[98].

Second, the population-level analyses are constrained by the composition of currently available public genomes. These datasets are neither geographically nor ecologically sampled at random, and environmental metadata are often incomplete or inconsistently reported. Accordingly, the observed GCF prevalence represents the analysed genome collection rather than an unbiased estimate of the global *S. albidoflavus* population. The same limitation applies to interpretation of the five phylogroups, because their possible associations with habitat, geography or host origin cannot be evaluated rigorously without standardised metadata. Expanded sampling of complete genomes from under-represented environments

will therefore be important for determining whether the present phylogenomic and biosynthetic patterns persist across the broader species population.

These limitations define several direct experimental priorities. Targeted LC-MS/MS, comparative metabolomics and molecular networking should be used to determine whether fredericamycin-, SF2768- or other predicted congeners are produced by MUM256 under standard or elicited cultivation conditions. Transcriptomic profiling could establish whether prioritised loci are transcriptionally active, whereas pathway disruption, promoter activation and heterologous expression would provide progressively stronger evidence linking individual BGCs to specific metabolites ^[99]. Species-specific precedent further supports an elicitation-based strategy: chemical elicitation of *S. albidoflavus* RKJM0023 with rhamnolipids induced a molecular family that was not prominent under untreated conditions and enabled isolation of the previously undescribed alubactins ^[80]. An “one strain many compounds” (OSMAC) experimental design combining alternative culture media, chemical or co-culture elicitation, transcript profiling and LC-MS/MS molecular networking could therefore be particularly informative for MUM256 ^[100]. Such experiments would help distinguish biosynthetic loci that are transcriptionally inactive under routine cultivation from those that produce metabolites only under specific environmental or physiological conditions.

At the population level, additional complete *S. albidoflavus* genomes accompanied by standardised ecological and geographic metadata would enable phylogeny-aware analyses of BGC gain, loss and habitat association. Integration of these datasets with the five phylogroups identified here could determine whether low-frequency GCFs preferentially track evolutionary lineages, ecological niches or both. Such analyses would also help distinguish recent horizontal acquisition from lineage-specific retention or loss, an evolutionary framework that has already proved informative for rare *S. albidoflavus* BGCs ^[30]. More broadly, recent population-genomic analyses of *Streptomyces griseus* and *Streptomyces pratensis* have shown that horizontal gene transfer is strongly structured by phylogenetic relatedness and biogeographic context, suggesting that accessory-gene flux in *Streptomyces* is constrained by both evolutionary history and ecology ^[101]. Although these observations were derived from species other than *S. albidoflavus*, they provide a useful comparative framework for testing whether the phylogroups identified here similarly differ in their propensity to acquire, retain or lose accessory genes and biosynthetic loci.

Computational prioritisation could further integrate BGC sequence features, phylogenetic distribution, predicted bioactivity and experimentally observed production profiles. Machine-learning approaches are increasingly being explored for *Streptomyces* genome mining, biological-activity prediction and optimisation of natural-product discovery workflows ^[102]. Combining such approaches with phylogeny-aware GCF analysis could help distinguish broadly conserved biosynthetic functions from lineage-associated or recently acquired loci and thereby focus experimental validation on BGCs with the strongest evolutionary, chemical and biological rationale.

Collectively, these considerations reinforce the central interpretation of the present study: the biosynthetic potential of MUM256 is best understood not simply from the presence of individual predicted BGCs, but from their population prevalence, structural variation, evolutionary context and experimentally determined expression. This framework provides a rational basis for prioritising MUM256 biosynthetic loci for subsequent chemical and functional validation.

5. Conclusions

This study establishes *Streptomyces* sp. MUM256 as *S. albidoflavus* through an integrated genome-based taxonomic framework and places this previously characterised bioactive mangrove isolate within the broader genomic and biosynthetic diversity represented by publicly available *S. albidoflavus*-assigned genomes. The high-quality single-contig genome provided a robust foundation for comparative analysis, while the expanded 151-genome dataset demonstrated an open pangenome across the analysed cohort, characterised by a conserved genomic backbone together with extensive accessory-gene diversity. Population-scale BGC analysis further revealed pronounced biosynthetic heterogeneity, with most GCF types occurring at low frequency across the 119-genome primary cohort.

Notably, the 18 BGCs identified in MUM256 were all associated with GCFs present elsewhere in the analysed cohort, indicating that its biosynthetic potential is shaped primarily by a distinct combination and architecture of shared biosynthetic families rather than by strain-exclusive GCFs. Comparative examination of the fredericamycin A-like, SF2768-like and cyclofaulknamycin-associated loci further demonstrated that similarity to characterised reference clusters, population prevalence and within-cohort structural conservation represent complementary but non-equivalent dimensions of biosynthetic diversity. These findings provide a population-genomic context for interpreting the previously reported biological activities of MUM256 and identify several loci suitable for targeted metabolomic, transcriptional and functional validation. More broadly, the study demonstrates how integrating high-quality genome sequencing with comparative pangenomics and comparative biosynthetic analysis can refine the prioritisation of environmentally derived *Streptomyces* strains and their specialised-metabolite pathways for future natural-product discovery.

Supplementary Materials: Table S1. RAST subsystem-based functional annotation of the *S. albidoflavus* MUM256 genome. Table S2. The 165 publicly available genomes annotated as *S. albidoflavus* were retrieved from the NCBI assembly database. Table S3. ANI-defined genomic groups within the 151-genome *S. albidoflavus*-assigned cohort. Figure S1. RAST subsystem assignments of *S. albidoflavus* MUM256. Figure S2. A genome-wide ANI heatmap comprising MUM256 and 165 *S. albidoflavus* genomes from NCBI, generated by ANIclustermap. Figure S3. Cohort curation and sensitivity of PGAP2 pangenome partition structure. Figure S4. Gene-cluster presence/absence landscape across the 151-genome *S. albidoflavus* cohort. Figure S5. Cohort selection and sensitivity of population BGC/GCF structure to assembly-contiguity filtering. Figure S6. Comparative architecture of accessory cyclofaulknamycin-like biosynthetic loci in *S. albidoflavus* MUM256. Figure S7. Sensitivity analysis of pangenome structure in the 151- and 119-genome PGAP2 cohorts.

Author Contributions: Conceptualisation, LT-HT and L-HL; methodology, LT-HT; formal analysis, LT-HT, K-OC, L-HL; investigation, LT-HT, JW-FL; resources, LT-HT, L-HL; writing, review and editing, LT-HT, JW-FL, VL, K-OC, R-YG, LW, K-GC, and L-HL.

Funding: This work was supported by the Jeffrey Cheah School of Medicine and Health Sciences Early Career Researcher Grant 2022 (Vote Number: ECR-000039) awarded to LT-HT. The author also thanks the Zhejiang Provincial Youth Thousand Talents Program for financial support.

Acknowledgement: Authors acknowledge the continuous support from the RII Centre for Molecular and Microbiome Sciences, under the Research Innovation Institute (RII) of the University of Nottingham Ningbo China.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pusparajah P, Letchumanan V, Law JW-F, *et al.* Streptomyces sp.—A Treasure Trove of Weapons to Combat Methicillin-Resistant Staphylococcus aureus Biofilm Associated with Biomedical Devices. *International Journal of Molecular Sciences* 2021; 22(17): 9360. doi: 10.3390/ijms22179360.
2. Belknap KC, Park CJ, Barth BM, *et al.* Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. *Scientific Reports* 2020; 10: 2023. doi: 10.1038/s41598-020-58904-9.
3. Lee N, Hwang S, Kim J, *et al.* Mini review: Genome mining approaches for the identification of secondary metabolite biosynthetic gene clusters in Streptomyces. *Computational and Structural Biotechnology Journal* 2020; 18: 1548–1556. doi: 10.1016/j.csbj.2020.06.024.
4. Beroigui O and Errachidi F. Streptomyces at the Heart of Several Sectors to Support Practical and Sustainable Applications: A Review. *Progress In Microbes & Molecular Biology* 2023; 6(1): a0000345. doi: 10.36877/pmmb.a0000345.
5. Elsalam RM, Goh KW, Mahadi M, *et al.* The Antibacterial Activities of Secondary Metabolites Derived from Streptomyces sp. *Progress In Microbes & Molecular Biology* 2022; 5(1): a0000281. doi: 10.36877/pmmb.a0000281.
6. Kemung HM, Tan LT-H, Khan TM, *et al.* Streptomyces as a Prominent Resource of Future Anti-MRSA Drugs. *Frontiers in Microbiology* 2018; Volume 9 - 2018. doi: 10.3389/fmicb.2018.02221.
7. Liu Z, Zhao Y, Huang C, *et al.* Recent Advances in Silent Gene Cluster Activation in Streptomyces. *Frontiers in Bioengineering and Biotechnology* 2021; 9: 632230. doi: 10.3389/fbioe.2021.632230.
8. Mohite OS, Jørgensen TS, Booth TJ, *et al.* Pangenome mining of the Streptomyces genus redefines species' biosynthetic potential. *Genome Biology* 2025; 226: 9. doi: 10.1186/s13059-024-03471-9.
9. Law JW-F, Pusparajah P, Ab Mutalib N-S, *et al.* A Review on Mangrove Actinobacterial Diversity: The Roles of Streptomyces and Novel Species Discovery. *Progress In Microbes & Molecular Biology* 2019; 2(1): a0000024. doi: 10.36877/pmmb.a0000024.
10. Ab Mutalib N-S, Wong SH, Ser H-L, *et al.* Bioprospecting of Microbes for Valuable Compounds to Mankind. *Progress In Microbes & Molecular Biology* 2020; 3(1): a0000088. doi: 10.36877/pmmb.a0000088.
11. Tan LT-H, Mahendra CK, Yow Y-Y, *et al.* Streptomyces sp. MUM273b: A mangrove-derived potential source for antioxidant and UVB radiation protectants. *MicrobiologyOpen* 2019; 8(10): e859. doi: 10.1002/mbo3.859.

12. Mangzira Kemung H, Tan LT-H, Chan K-G, *et al.* Streptomyces sp. Strain MUSC 125 from Mangrove Soil in Malaysia with Anti-MRSA, Anti-Biofilm and Antioxidant Activities. *Molecules* 2020; 25(15): 3545. doi: 10.3390/molecules25153545.
13. Kemung HM, Tan LT-H, Chan K-G, *et al.* Streptomyces sp. strain MUSC 5 from mangrove forest in Malaysia: Identification, antioxidant potential and chemical profiling of its methanolic extract. *Progress In Microbes & Molecular Biology* 2020; 3(1): a0000087. doi: 10.36877/pmmb.a0000087.
14. Law JW-F, Tan LT-H, Letchumanan V, *et al.* Streptomyces griseiviridis sp. nov., a Novel “Modern Actinobacteria” isolated from Malaysia Mangrove Soil. *Progress In Microbes & Molecular Biology* 2023; 6(1): a0000270. doi: 10.36877/pmmb.a0000270.
15. Law JW-F, Letchumanan V, Hong K-W, *et al.* Streptomyces learnhianii sp. nov., unveiling a Mangrove-Derived Novel “Modern Actinobacteria” in Malaysia. *Progress In Microbes & Molecular Biology* 2023; 6(1): a0000342. doi: 10.36877/pmmb.a0000342.
16. Law JW-F, Letchumanan V, Tan LT-H, *et al.* The Rising of “Modern Actinobacteria” Era. *Progress In Microbes & Molecular Biology* 2020; 3(1): a0000064. doi: 10.36877/pmmb.a0000064.
17. Thye AY-K, Letchumanan V, Tan LT-H, *et al.* Malaysia’s Breakthrough in Modern Actinobacteria (MOD-ACTINO) Drug Discovery Research. *Progress In Microbes & Molecular Biology* 2022; 5(1): a0000275. doi: 10.36877/pmmb.a0000275.
18. Tan LT-H, Ser H-L, Yin W-F, *et al.* Investigation of Antioxidative and Anticancer Potentials of Streptomyces sp. MUM256 Isolated from Malaysia Mangrove Soil. *Frontiers in Microbiology* 2015; 6: 1316. doi: 10.3389/fmicb.2015.01316.
19. Tan LT-H, Chan C-K, Chan K-G, *et al.* Streptomyces sp. MUM256: A Source for Apoptosis Inducing and Cell Cycle-Arresting Bioactive Compounds against Colon Cancer Cells. *Cancers* 2019; 11(11): 1742. doi: 10.3390/cancers11111742.
20. Law JW-F, Tan K-X, Wong SH, *et al.* Taxonomic and Characterization Methods of Streptomyces: A Review. *Progress In Microbes & Molecular Biology* 2018; 1(1): a0000009. doi: 10.36877/pmmb.a0000009.
21. Jain C, Rodriguez-R LM, Phillippy AM, *et al.* High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature communications* 2018; 9: 5114. doi: 10.1038/s41467-018-07641-9.
22. Hu S, Li K, Zhang Y, *et al.* New Insights Into the Threshold Values of Multi-Locus Sequence Analysis, Average Nucleotide Identity and Digital DNA-DNA Hybridization in Delineating Streptomyces Species. *Frontiers in Microbiology* 2022; 13: 910277. doi: 10.3389/fmicb.2022.910277.
23. Riesco R and Trujillo ME. Update on the proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 2024; 74: 006300. doi: 10.1099/ijsem.0.006300.
24. Cheng K, Rong X, Pinto-Tomás AA, *et al.* Population genetic analysis of Streptomyces albidoflavus reveals habitat barriers to homologous recombination in the diversification of streptomycetes. *Applied and Environmental Microbiology* 2015; 81(3): 966–75. doi: 10.1128/aem.02925-14.
25. Li Y, Pinto-Tomás AA, Rong X, *et al.* Population Genomics Insights into Adaptive Evolution and Ecological Differentiation in Streptomycetes. *Applied and Environmental Microbiology* 2019; 85(7): e02555–18. doi: 10.1128/AEM.02555-18.

26. Diabankana RGC, Frolov M, Keremli S, *et al.* Genomic Insights into the Microbial Agent *Streptomyces albidoflavus* MGMM6 for Various Biotechnology Applications. *Microorganisms* 2023; 11(12): 2872. doi: 10.3390/microorganisms11122872.
27. Gonzalez-Silva A, San Juan-Mendo M, Delgado-Prudencio G, *et al.* Comparative Genomics and Biosynthetic Cluster Analysis of Antifungal Secondary Metabolites of Three Strains of *Streptomyces albidoflavus* Isolated from Rhizospheric Soils. *Microorganisms* 2024; 12(12): 2637. doi: 10.3390/microorganisms12122637.
28. Popov IV, Chikindas ML, and Popov IV. Genomic insights into *Streptomyces albidoflavus* SM254: tracing the putative signs of anti-Pseudogymnoascus destructans properties. *Brazilian Journal of Microbiology* 2025; 56(3): 2121–2131. doi: 10.1007/s42770-025-01740-8.
29. Sedeek AM, Elfeky H, Hanora AS, *et al.* Genomic insights into biosynthesis and adaptation in the bioactive marine bacterium *Streptomyces albidoflavus* VIP-1 from the Red Sea. *BMC Microbiology* 2025; 25: 372. doi: 10.1186/s12866-025-04109-x.
30. Wang J, Liu N, Liu M, *et al.* Eco-evolutionary dynamics sustain a potent yet rare antibiotic gene cluster in *Streptomyces*. *The ISME journal* 2026; 20(1): wrag060. doi: 10.1093/ismejo/wrag060.
31. Lin Y, Yuan J, Kolmogorov M, *et al.* Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences* 2016; 113(52): E8396–E8405. doi: 10.1073/pnas.1604560113.
32. Kolmogorov M, Yuan J, Lin Y, *et al.* Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 2019; 37: 540–546. doi: 10.1038/s41587-019-0072-8.
33. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018; 34(18): 3094–3100. doi: 10.1093/bioinformatics/bty191.
34. Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; 25: 2078–2079. doi: 10.1093/bioinformatics/btp352.
35. Gurevich A, Saveliev V, Vyahhi N, *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013; 29(8): 1072–1075. doi: 10.1093/bioinformatics/btt086.
36. Tegenfeldt F, Kuznetsov D, Manni M, *et al.* OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Research* 2025; 53(D1): D516–D522. doi: 10.1093/nar/gkae987.
37. Manni M, Berkeley MR, Seppely M, *et al.* BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* 2021; 38(10): 4647–4654. doi: 10.1093/molbev/msab199.
38. Chklovski A, Parks DH, Woodcroft BJ, *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods* 2023; 20(8): 1203–1212. doi: 10.1038/s41592-023-01940-w.
39. Aziz RK, Bartels D, Best AA, *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 2008; 9: 75. doi: 10.1186/1471-2164-9-75.
40. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; 30(14): 2068–9. doi: 10.1093/bioinformatics/btu153.

41. Li W, O'Neill KR, Haft DH, *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Research* 2021; 49(D1): D1020–D1028. doi: 10.1093/nar/gkaa1105.
42. Galperin MY, Vera Alvarez R, Karamycheva S, *et al.* COG database update 2024. *Nucleic Acids Research* 2025; 53: D356–D363. doi: 10.1093/nar/gkae983.
43. Galperin MY, Wolf YI, Makarova KS, *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research* 2021; 49(D1): D274–D281. doi: 10.1093/nar/gkaa1018.
44. Seemann T. Barrnap 0.8: rapid ribosomal RNA prediction. <https://github.com/tseemann/barrnap> 2013.
45. Chalita M, Kim YO, Park S, *et al.* EzBioCloud: a genome-driven database and platform for microbiome identification and discovery. *International Journal of Systematic and Evolutionary Microbiology* 2024; 74(6): 006421. doi: 10.1099/ijsem.0.006421.
46. Parks DH, Imelfort M, Skennerton CT, *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 2015; 25(7): 1043–1055. doi: 10.1101/gr.186072.114.
47. Bu C, Zhang H, Zhang F, *et al.* PGAP2: A comprehensive toolkit for prokaryotic pan-genome analysis based on fine-grained feature networks. *Nature communications* 2025; 16: 9865. doi: 10.1038/s41467-025-64846-5.
48. Chaumeil P-A, Mussig AJ, Hugenholtz P, *et al.* GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2020; 36(6): 1925–1927. doi: 10.1093/bioinformatics/btz848.
49. Shaw J and Yu YW. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods* 2023; 20: 1661–1665. doi: 10.1038/s41592-023-02018-3.
50. Richter M and Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America* 2009; 106: 19126–19131. doi: 10.1073/pnas.0906412106.
51. Meier-Kolthoff JP and Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature communications* 2019; 10: 2182. doi: 10.1038/s41467-019-10210-3.
52. Meier-Kolthoff JP, Auch AF, Klenk H-P, *et al.* Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013; 14: 60. doi: 10.1186/1471-2105-14-60.
53. Goris J, Konstantinidis KT, Klappenbach JA, *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 2007. doi: 10.1099/ijss.0.64483-0.
54. Katoh K and Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 2013; 30: 772–780. doi: 10.1093/molbev/mst010.
55. Steenwyk JL, Buida TJ, III, Li Y, *et al.* ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biology* 2020; 18: e3001007. doi: 10.1371/journal.pbio.3001007.

56. Wong TKF, Ly-Trong N, Ren H, *et al.* IQ-TREE 3: phylogenomic inference software using complex evolutionary models. *Molecular Biology and Evolution* 2026; 43: msag117. doi: 10.1093/molbev/msag117.
57. Hoang DT, Chernomor O, von Haeseler A, *et al.* UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* 2018; 35: 518–522. doi: 10.1093/molbev/msx281.
58. Paradis E and Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019; 35(3): 526–528. doi: 10.1093/bioinformatics/bty633.
59. Struyf A, Hubert M, and Rousseeuw P. Clustering in an Object-Oriented Environment. *Journal of Statistical Software* 1997; 1(4): 1 – 30. doi: 10.18637/jss.v001.i04.
60. Blin K, Shaw S, Vader L, *et al.* antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Research* 2025; 53(W1): W32–W38. doi: 10.1093/nar/gkaf334.
61. Zdouc MM, Blin K, Louwen NLL, *et al.* MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Research* 2025; 53(D1): D678–D690. doi: 10.1093/nar/gkae1115.
62. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology* 2020; 16: 60–68. doi: 10.1038/s41589-019-0400-9.
63. Draisma A, Loureiro C, Louwen NLL, *et al.* BiG-SCAPE 2.0 and BiG-SLiCE 2.0: scalable, accurate and interactive sequence clustering of metabolic gene clusters. *Nature communications* 2026; 17: 2000. doi: 10.1038/s41467-026-68733-5.
64. Blin K, Pascal Andreu V, de Los Santos ELC, *et al.* The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research* 2019; 47(D1): D625–D630. doi: 10.1093/nar/gky1060.
65. Rajwani R, Ohlemacher SI, Zhao G, *et al.* Genome-Guided Discovery of Natural Products through Multiplexed Low-Coverage Whole-Genome Sequencing of Soil Actinomycetes on Oxford Nanopore Flongle. *mSystems* 2021; 6(6): e01020–21. doi: 10.1128/mSystems.01020-21.
66. Blin K, Shaw S, Medema MH, *et al.* The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more. *Nucleic Acids Research* 2024; 52(D1): D586–D589. doi: 10.1093/nar/gkad984.
67. Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 2003; 13: 2498–2504. doi: 10.1101/gr.1239303.
68. Gilchrist CLM and Chooi Y-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* 2021; 37: 2473–2475. doi: 10.1093/bioinformatics/btab007.
69. Wendt-Pienkowski E, Huang Y, Zhang J, *et al.* Cloning, sequencing, analysis, and heterologous expression of the fredericamycin biosynthetic gene cluster from *Streptomyces griseus*. *Journal of the American Chemical Society* 2005; 127(47): 16442–16452. doi: 10.1021/ja054376u.
70. Wang L, Zhu M, Zhang Q, *et al.* Diisonitrile Natural Product SF2768 Functions As a Chalkophore That Mediates Copper Acquisition in *Streptomyces thioluteus*. *ACS Chemical Biology* 2017; 12(12): 3607–3075. doi: 10.1021/acschembio.7b00897.

71. Horbal L, Stierhof M, Paluszczak A, *et al.* Cyclofaulknamycin with the Rare Amino Acid D-capreomycinidine Isolated from a Well-Characterized *Streptomyces albus* Strain. *Microorganisms* 2021; 9: 1609. doi: 10.3390/microorganisms9081609.
72. Steinegger M and Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 2017; 35: 1026–1028. doi: 10.1038/nbt.3988.
73. Zaburannyi N, Rabyk M, Ostash B, *et al.* Insights into naturally minimised *Streptomyces albus* J1074 genome. *BMC Genomics* 2014. doi: 10.1186/1471-2164-15-97.
74. Nikolaidis M, Hesketh A, Frangou N, *et al.* A panoramic view of the genomic landscape of the genus *Streptomyces*. *Microbial Genomics* 2023; 9(6): mgen001028. doi: 10.1099/mgen.0.001028.
75. Tan LT-H, Law JW-F, Chua K-O, *et al.* Phylogenomics, antimicrobial resistance, and mobile genetic elements in clinical *Clostridioides difficile* isolates from Hong Kong, China. *Progress In Microbes & Molecular Biology* 2026; 9(1): a0000651. doi: 10.36877/pmmb.a0000651.
76. Tonkin-Hill G, MacAlasdair N, Ruis C, *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* 2020; 21: 180. doi: 10.1186/s13059-020-02090-4.
77. Choudoir MJ, Pepe-Ranne C, and Buckley DH. Diversification of Secondary Metabolite Biosynthetic Gene Clusters Coincides with Lineage Divergence in *Streptomyces*. *Antibiotics (Basel)* 2018; 7(1). doi: 10.3390/antibiotics7010012.
78. Nayeib GHG and Barona-Gómez F. Intra- and inter-biosynthetic gene cluster allelic variation as drivers of chemical diversification in *Streptomyces*. *Essays in Biochemistry* 2026; 70(2): 283–293. doi: 10.1042/ebc20250013.
79. Maw ZA, Haltli B, Guo JJ, *et al.* Discovery of Acyl-Surugamide A2 from Marine *Streptomyces albidoflavus* RKJM-0023—A New Cyclic Nonribosomal Peptide Containing an N- ϵ -acetyl-L-lysine Residue. *Molecules* 2024; 29(7): 1482. doi: 10.3390/molecules29071482.
80. Maw ZA, Grunwald AL, Haltli BA, *et al.* Discovery of the Lipopeptides Albubactins A-H from *Streptomyces albidoflavus* RKJM0023 via Chemical Elicitation with Rhamnolipids and Synthesis of Albubactin A. *Journal of Natural Products* 2024; 87(7): 1682–1693. doi: 10.1021/acs.jnatprod.3c01234.
81. Blin K, Shaw S, Steinke K, *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research* 2019; 47(W1): W81–W87. doi: 10.1093/nar/gkz310.
82. Blin K, Shaw S, Kautsar SA, *et al.* The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Research* 2021; 49(D1): D639–d643. doi: 10.1093/nar/gkaa978.
83. Ser H-L, Ab Mutalib N-S, Yin W-F, *et al.* Genome sequence of *Streptomyces antioxidans* MUSC 164T isolated from mangrove forest. *Progress In Microbes & Molecular Biology* 2018; 1(1): a0000001. doi: 10.36877/pmmb.a0000001.
84. Ser H-L, Chan K-G, Tan W-S, *et al.* Complete genome of mangrove-derived anti-MRSA streptomycete, *Streptomyces pluripotens* MUSC 135T. *Progress In Microbes & Molecular Biology* 2018; 1(1): a0000004. doi: 10.36877/pmmb.a0000004.
85. Tan LT-H, Chan K-G, Khan TM, *et al.* *Streptomyces* sp. MUM212 as a Source of Antioxidants with Radical Scavenging and Metal Chelating Properties. *Frontiers in Pharmacology* 2017; Volume 8 - 2017. doi: 10.3389/fphar.2017.00276.

86. Pusparajah P, Law JW-F, Chan K-G, *et al.* Whole-Genome Sequence of *Streptomyces pluripotens* strain MUM 16J, a Potential Resource of Glycopeptide Antibiotic and Biocontrol Agent against Biofilm-forming Bacteria. *Progress In Microbes & Molecular Biology* 2023; 6(1): a0000330. doi: 10.36877/pmmb.a0000330.
87. Loo K-Y, Tan LT-H, Law JW-F, *et al.* Detection of multidrug resistant *Vibrio parahaemolyticus* and anti-*Vibrio* *Streptomyces* sp. MUM 178J. *Progress In Microbes & Molecular Biology* 2023; 6(1): a0000347. doi: 10.36877/pmmb.a0000347.
88. Loo K-Y, Tan LT-H, Chua K-O, *et al.* Complete Whole-Genome Sequence of *Streptomyces* sp. MUM 178J, a Potential Anti-*Vibrio* Agent. *Progress In Microbes & Molecular Biology* 2024; 7(1): a0000399. doi: 10.36877/pmmb.a0000399.
89. Goh JXH, Tan LT-H, Law JW-F, *et al.* *Streptomyces* sp. MUM 195J: A Promising Probiotic for Controlling *Vibrio parahaemolyticus* Infection in Aquaculture. *Progress In Microbes & Molecular Biology* 2024; 7(1): a0000443. doi: 10.36877/pmmb.a0000443.
90. Tan LT-H, Lee L-H, and Goh B-H. The Bioprospecting of Anti-*Vibrio* *Streptomyces* species: Prevalence and Applications. *Progress In Microbes & Molecular Biology* 2019; 2(1): a0000034. doi: 10.36877/pmmb.a0000034.
91. Ser H-L, Tan LT-H, Law JW-F, *et al.* Focused Review: Cytotoxic and Antioxidant Potentials of Mangrove-Derived *Streptomyces*. *Frontiers in Microbiology* 2017. doi: 10.3389/fmicb.2017.02065.
92. Tan LT-H, Chan K-G, Pusparajah P, *et al.* Mangrove derived *Streptomyces* sp. MUM265 as a potential source of antioxidant and anticolon-cancer agents. *BMC Microbiology* 2019; 19(1): 38. doi: 10.1186/s12866-019-1409-7.
93. Warnick-Pickle DJ, Byrne KM, Pandey RC, *et al.* Fredericamycin A, a new antitumor antibiotic. II. Biological properties. *The Journal of Antibiotics* 1981; 34(11): 1402–1407. doi: 10.7164/antibiotics.34.1402.
94. Latham MD, King CK, Gorycki P, *et al.* Inhibition of topoisomerases by fredericamycin A. *Cancer Chemotherapy and Pharmacology* 1989; 24(3): 167–171. doi: 10.1007/BF00300237.
95. Ueda K. From antibiotic to chalkophore: the biology and evolution of SF2768 in *Streptomyces*. *Essays in Biochemistry* 2026; 70(2): 219–227. doi: 10.1042/EBC20250022.
96. Zhang MM, Wong FT, Wang Y, *et al.* CRISPR-Cas9 strategy for activation of silent *Streptomyces* biosynthetic gene clusters. *Nature Chemical Biology* 2017; 13: 607–609. doi: 10.1038/nchembio.2341.
97. Tan LT-H, Lee L-H, and Goh B-H. Critical review of fermentation and extraction of anti-*Vibrio* compounds from *Streptomyces*. *Progress In Microbes & Molecular Biology* 2020; 3(1): a0000051. doi: 10.36877/pmmb.a0000051.
98. Tan LT-H. Decoding Multicellularity in *Streptomyces*: Single-Cell Technologies Reveal the Link between Phenotypic Heterogeneity and Secondary Metabolite Production. *Progress In Microbes & Molecular Biology* 2025; 8(1): a0000477. doi: 10.36877/pmmb.a0000477.
99. Ahmed Y, Rebets Y, Rodríguez Estévez M, *et al.* Engineering of *Streptomyces lividans* for heterologous expression of secondary metabolite gene clusters. *Microbial cell factories* 2020; 19(1): 5. doi: 10.1186/s12934-020-1277-8.
100. Romano S, Jackson SA, Patry S, *et al.* Extending the "One Strain Many Compounds" (OSMAC) Principle to Marine Microorganisms. *Marine Drugs* 2018; 16(7). doi: 10.3390/md16070244.

101. Hariharan J, Andam CP, and Buckley DH. Biogeographical and phylogenetic constraints on horizontal gene transfer and genome evolution in *Streptomyces*. *Microbiology Spectrum* 2026; 14(2): e02958–25. doi: 10.1128/spectrum.02958-25.
102. Ang W-S, Law JW-F, Kumari Y, *et al.* Unleashing the Power of Artificial Intelligence-Driven Drug Discovery in *Streptomyces*. *Progress In Microbes & Molecular Biology* 2024; 7(1): a0000453. doi: 10.36877/pmmb.a0000453.



Author(s) shall retain the copyright of their work and grant the Journal/Publisher right for the first publication with the work simultaneously licensed under:

Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). This license allows for the copying, distribution and transmission of the work, provided the correct attribution of the original creator is stated. Adaptation and remixing are also permitted.