Biosynthetic Gene Cluster Identification: From Foundational Concepts to Advanced Applications in Drug Discovery

Generated at: 2026-04-03
Source URL: https://www.biosynthchem.com/posts/biosynthetic-gene-cluster-identification-from-foundational-concepts-to-advanced-applications-in-drug-discovery

Biosynthetic Gene Cluster Identification: From Foundational Concepts to Advanced Applications in Drug Discovery

Abstract

This comprehensive review explores the rapidly evolving field of biosynthetic gene cluster (BGC) identification, a cornerstone technology in modern natural product discovery. We examine foundational principles of BGCs as physical clusters of co-regulated genes governing secondary metabolite biosynthesis, which provide the genetic blueprint for countless therapeutics. The article details cutting-edge computational methodologies including genome mining tools, machine learning approaches, and database resources that have revolutionized BGC detection and characterization. We address significant challenges in the field, particularly the activation of silent gene clusters and optimization of heterologous expression systems. Furthermore, we cover advanced validation techniques that connect genetic predictions to chemical structures, highlighting successful applications in discovering novel bioactive compounds with clinical potential. This resource provides researchers, scientists, and drug development professionals with both theoretical knowledge and practical methodologies to advance their natural product discovery pipelines.

Understanding Biosynthetic Gene Clusters: The Genetic Blueprint of Natural Products

Biosynthetic Gene Clusters (BGCs) are sets of genes that are physically clustered on the genome and work in concert to encode the production of specialized microbial metabolites [1]. These tightly linked sets of mostly non-homologous genes participate in a common, discrete metabolic pathway, with their expression often being co-regulated [1]. BGCs represent nature's blueprint for producing a vast array of chemically diverse compounds that serve essential ecological functions, from mediating microbial interactions to providing survival advantages in competitive environments.

The study of BGCs has revolutionized natural product discovery, shifting the paradigm from traditional activity-based screening to targeted genome mining approaches. This transition has been driven by the recognition that microbial genomes contain far greater biosynthetic potential than previously observed through cultivation-based methods [2]. The systematic identification and analysis of BGCs across diverse taxa and environments have revealed an immense, largely untapped reservoir of chemical diversity with significant implications for pharmaceutical development, agricultural applications, and understanding microbial ecology.

Core Structural and Functional Components of BGCs

Genetic Architecture and Key Components

BGCs typically contain core biosynthetic genes that create the basic molecular scaffold, along with additional genes that enhance functionality and regulation. The core biosynthetic genes vary by metabolite class but generally include enzymes responsible for building the fundamental chemical structure. For polyketides, these are polyketide synthases (PKSs); for non-ribosomal peptides, non-ribosomal peptide synthetases (NRPSs); and for ribosomally synthesized and post-translationally modified peptides (RiPPs), the precursor peptides and modification enzymes [3] [4].

Beyond the core biosynthetic machinery, BGCs typically contain several auxiliary genetic elements that support the production and functionality of the metabolite:

- **Regulatory genes**: Control the timing and level of BGC expression in response to environmental or cellular cues
- **Transporters**: Facilitate the export of the final product out of the cell or uptake of precursors; ABC transporters and major facilitator superfamily (MFS) transporters are commonly found in BGCs [5]
- **Resistance genes**: Protect the host organism from the toxic effects of its own metabolite
- **Tailoring enzymes**: Modify the core scaffold through reactions such as hydroxylation, glycosylation, or methylation to enhance bioactivity or alter physicochemical properties

Major Classes of BGCs and Their Products

BGCs are categorized based on the biosynthetic logic of the pathways they encode and the chemical classes of their metabolic products. The table below summarizes the major BGC classes, their key biosynthetic enzymes, and representative compounds.

Table 1: Major Classes of Biosynthetic Gene Clusters and Their Products

BGC Class	Key Biosynthetic Enzymes	Representative Compounds	Biological Functions
Non-Ribosomal Peptide (NRPS)	Non-ribosomal peptide synthetases	Penicillin, Vancomycin, Daptomycin	Antibiotics, siderophores, virulence factors
Polyketide (PKS)	Polyketide synthases (Type I, II, III)	Erythromycin, Tetracycline, Rapamycin	Antimicrobial, immunosuppressant, antitumor
Ribosomally synthesized and Post-translationally Modified Peptide (RiPP)	Precursor peptides, modification enzymes	Nisin, Thiopeptides, Cyanobactins	Antimicrobial, signaling molecules
Terpene	Terpene synthases, prenyltransferases	Taxol, Artemisinin, Carotenoids	Anticancer, antimalarial, antioxidants
Saccharide	Glycosyltransferases, sugar modifiers	Vancomycin, Acarbose, Avilamycin	Antibiotics, antidiabetic drugs
NRPS-Independent Siderophore	NIS synthetases	Vibrioferrin, Aerobactin	Iron chelation, virulence
Hybrid	Combinations of above systems	Bleomycin, Staurosporine	Anticancer, antibacterial

The distribution of these BGC classes varies significantly across bacterial taxa and ecological niches. Systematic analyses have revealed that saccharide BGCs are remarkably abundant, constituting approximately 40% of all predicted BGCs in prokaryotic genomes—more than twice the size of the next largest class [3]. Notably, NRPS and PKS clusters, while less abundant overall, produce some of the most clinically valuable natural products and are therefore disproportionately important in drug discovery.

Biological Significance and Ecological Roles

BGCs and their metabolic products play essential roles in mediating ecological interactions and providing competitive advantages to their producers. These functions span multiple biological domains, from microbial warfare to symbiotic relationships.

Intermicrobial Interactions and Competition

Many BGCs produce antimicrobial compounds that inhibit the growth of competing microorganisms in shared environments. This chemical warfare allows producers to secure ecological niches and resources. For instance, in the human microbiome, thiopeptide BGCs are widely distributed and produce antibiotics with potent activity against Gram-positive pathogens [6]. Similarly, marine bacteria harbor diverse BGCs encoding compounds with antibacterial properties that help them compete in nutrient-limited environments [7] [8].

The ecological significance of these compounds is particularly evident in particle-associated marine communities, where microbes exist in close physical proximity and competition is intense. Metagenomic studies of the Cariaco Basin revealed that particle-associated bacteria harbor greater diversity of BGCs—particularly NRPS, PKS, and RiPP clusters—compared to free-living communities, reflecting the heightened need for chemical mediation of interactions in these dense microbial aggregates [8].

Host-Microbe Interactions

BGCs play crucial roles in establishing and maintaining host-microbe relationships, ranging from pathogenicity to mutualism. In pathogenic interactions, BGC products may function as virulence factors that facilitate host colonization, tissue damage, or immune evasion. For example, siderophore BGCs enable pathogens to acquire iron from host tissues, which is essential for bacterial growth and virulence [5].

In mutualistic symbioses, BGC products often protect both host and microbial partner from external threats. The mutualistic bacteria *Xenorhabdus* and *Photorhabdus* (XP) living in symbiosis with entomopathogenic nematodes provide a striking example. These bacteria produce a cocktail of natural products that help kill insect prey, suppress insect immunity, and protect the insect carcass from colonization by competing microorganisms [9]. Systematic analysis of XP BGCs revealed remarkable conservation of certain clusters, suggesting their essential role in maintaining the tripartite nematode-bacterium-insect relationship [9].

Nutrient Acquisition and Stress Response

Many BGCs produce metabolites that help microorganisms adapt to nutrient limitation and other environmental stresses. Siderophores represent a well-characterized class of such compounds, chelating environmental iron and making it available for microbial uptake [7] [5]. The ecological importance of siderophores is evident in iron-limited marine environments, where bacteria employ diverse siderophore-mediated iron acquisition systems [7].

Beyond iron acquisition, BGCs produce compounds that help microbes cope with oxidative stress, pH fluctuations, and other environmental challenges. For instance, carotenoid BGCs produce pigments that protect against UV damage, while ectoine BGCs generate compatible solutes that help maintain cellular integrity under osmotic stress [8].

Genomic Distribution and Evolutionary Dynamics

Taxonomic Distribution of BGCs

BGCs are distributed across diverse taxonomic groups but are particularly prominent in certain bacterial phyla. They are common features of bacterial and most fungal genomes but are less frequently found in other organisms [1]. Among bacteria, *Actinomycetes*, *Bacillus*, *Pseudomonas*, and *Burkholderia* species are renowned for their rich BGC content and chemical diversity [3] [6].

In the human microbiome, BGCs are particularly abundant in members of the genera *Bacteroides*, *Parabacteroides*, *Corynebacterium*, *Rothia*, and *Ruminococcus*, with some genomes harboring up to seven BGCs despite their relatively small genome size (2-3 Mb) [6]. This concentration of biosynthetic potential in abundant host-associated taxa suggests that BGC products play important roles in microbe-host interactions.

Striking differences in BGC abundance exist between environments and taxonomic groups. For example, *Xenorhabdus* and *Photorhabdus* species harbor an average of 22 BGCs per genome, which is two- to tenfold higher than the average BGC content of other Enterobacteriaceae [9]. This exceptional biosynthetic capacity reflects their ecological strategy centered on chemical mediation of complex multi-organism interactions.

Evolutionary Mechanisms

BGCs evolve through several mechanisms that generate chemical diversity and enable adaptation to ecological niches:

- **Horizontal Gene Transfer**: Entire BGCs can be transferred between organisms, potentially conferring new ecological functions. This process is particularly important for the spread of adaptive traits among bacterial populations [1] [7]. Horizontal gene cluster transfer has been linked to ecological niches where the encoded pathways provide a selective advantage [10].
- **Gene Duplication and Divergence**: Duplication of biosynthetic genes followed by functional divergence can lead to the emergence of new catalytic capabilities and novel compounds.
- **Module Shuffling and Domain Rearrangements**: In modular PKS and NRPS systems, recombination events can generate new module arrangements and thus new chemical structures.
- **Cluster Rearrangement and Genome Dynamics**: Local genome rearrangements can alter cluster composition, regulation, or gene content, modifying the metabolic output.

These evolutionary processes have resulted in the formation of BGC families that can be classified into Gene Cluster Families (GCFs) based on sequence similarity [1] [7]. The global analysis of BGC relationships has revealed large families with hundreds or thousands of members, the vast majority of which remain uncharacterized [3].

Research Methodologies and Experimental Approaches

Computational Identification and Analysis

Modern BGC discovery relies heavily on computational tools that can identify these clusters in genomic sequences. Two complementary approaches dominate the field: rule-based detection and machine learning methods.

Table 2: Major Bioinformatics Tools for BGC Identification and Analysis

Tool Name	Approach	Key Features	Applications
antiSMASH	Rule-based	Detects known BGC classes based on curated rules and databases	Comprehensive BGC annotation in genomes and metagenomes [7] [4]
ClusterFinder	Probabilistic/HMM	Uses Pfam domain frequencies to identify known and novel BGCs	Discovery of novel BGC classes beyond known templates [3] [2]
PRISM	Rule-based	Predicts chemical structures from NRPS/PKS gene sequences	Structure-based prioritization of BGCs [10] [4]
BiG-SCAPE	Similarity-based	Groups BGCs into Gene Cluster Families (GCFs)	Comparative analysis of BGC diversity and evolution [7] [8]

Tool Name	Approach	Key Features	Applications
BiG-SLiCE	Super-linear clustering	Clusters massive numbers of BGCs using Euclidean space representation	Analysis of millions of BGCs across diverse taxa [1]
DeepBGC	Machine learning	Uses deep learning to identify BGCs and predict bioactivity	BGC discovery and functional prediction [10]

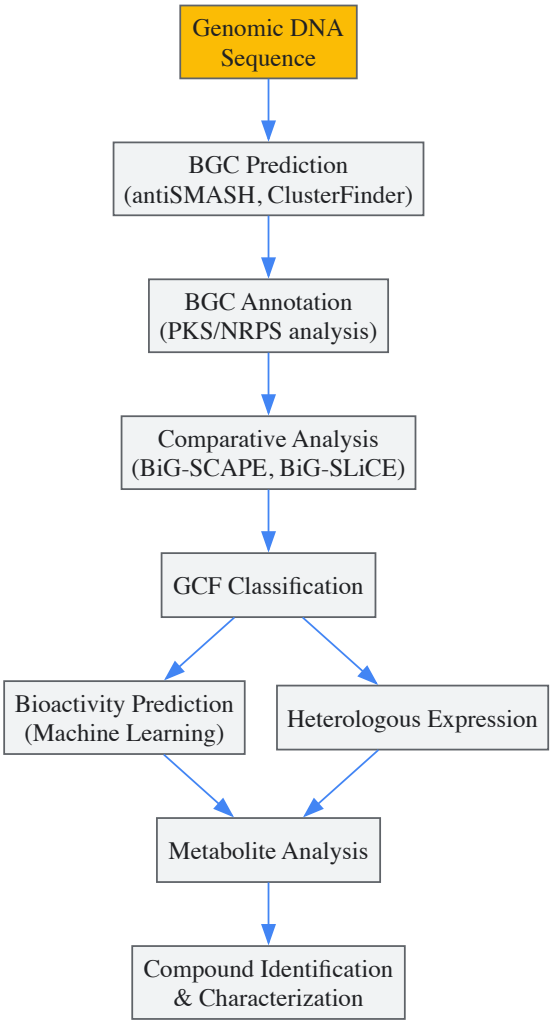


Diagram 1: BGC Identification and Characterization Workflow

Experimental Characterization and Validation

Computational predictions of BGCs require experimental validation to confirm the chemical structures and biological activities of their encoded metabolites. Several key approaches are employed:

Heterologous Expression: BGCs are cloned and expressed in tractable host organisms (e.g., *E. coli*, *S. coelicolor*, *S. cerevisiae*) to produce the encoded compounds [9]. This approach is particularly valuable for studying BGCs from unculturable organisms or those that are silent under laboratory conditions.

Metabolite Analysis and Structure Elucidation: Advanced analytical techniques including LC-MS, NMR, and MS/MS are used to purify and characterize the chemical structures of metabolites produced by BGCs [6] [9]. These approaches can be coupled with genomic data to link BGCs to their products.

Metatranscriptomics: Sequencing of community RNA reveals which BGCs are actively expressed in natural environments [6] [8]. This approach has demonstrated that BGCs are expressed in situ in diverse habitats, from the human microbiome to marine environments.

Gene Inactivation and Mutational Analysis: Targeted gene knockouts or mutations are used to establish genotype-phenotype relationships and confirm the roles of specific genes in BGCs [9].

Table 3: Key Research Reagents and Experimental Solutions for BGC Research

Reagent/Solution	Composition/Type	Function in BGC Research
antiSMASH Database	Curated BGC repository	Reference for BGC annotation and comparison [7] [4]
MIBiG Database	Minimum Information about BGCs	Standardized BGC data with experimental validation [1] [10]
Bacterial Artificial Chromosomes (BACs)	High-capacity cloning vectors	Heterologous expression of large BGCs [2]
Gateway or Gibson Assembly	DNA assembly methods	Cloning of complete BGCs for expression [9]
Induction Media	Various inductors (e.g., antibiotics, signaling molecules)	Activation of silent or poorly expressed BGCs [2]
Solid Phase Extraction Columns	Various resins (C18, HLB, etc.)	Metabolite purification from culture broths [6]
Mass Spectrometry Databases	Spectral libraries (GNPS, etc.)	Dereplication and identification of known compounds [8]

Applications and Future Directions

Pharmaceutical and Biotechnological Applications

BGCs encode the biosynthetic machinery for most clinically used antibiotics, anticancer agents, immunosuppressants, and other therapeutics. Systematic analysis of BGC diversity has become a cornerstone of modern drug discovery, enabling targeted identification of novel bioactive compounds [4]. The pharmaceutical potential of BGCs is exemplified by the discovery of lactocillin, a thiopeptide antibiotic from the human vaginal microbiota that exhibits potent activity against Gram-positive pathogens [6].

Beyond pharmaceuticals, BGCs have applications in agriculture (e.g., biopesticides, growth promoters), industry (e.g., enzymes, biocatalysts), and biotechnology (e.g., biofuel production, biomaterials). The identification of BGCs in diverse microbial communities provides access to enzymatic tools with novel catalytic properties and specificities.

Emerging Technologies and Research Frontiers

Several emerging technologies are shaping the future of BGC research:

Machine Learning and Artificial Intelligence: Advanced algorithms are being developed to predict BGCs, their chemical structures, and even their biological activities directly from sequence data [10] [4]. These approaches show promise in overcoming the limitations of rule-based detection methods and prioritizing BGCs for experimental characterization.

Single-Cell Metagenomics: This approach enables the recovery of complete BGCs from uncultured microorganisms, providing access to the biosynthetic potential of microbial "dark matter" [8] [2].

CRISPR-Cas Genome Editing: Precise genome editing facilitates the manipulation of BGCs in native hosts, activation of silent clusters, and engineering of novel pathways for optimized compound production [4].

Synthetic Biology and Pathway Refactoring: Complete synthesis and redesign of BGCs enable optimized expression, production of novel analogs, and transfer of pathways to industrial hosts [4].

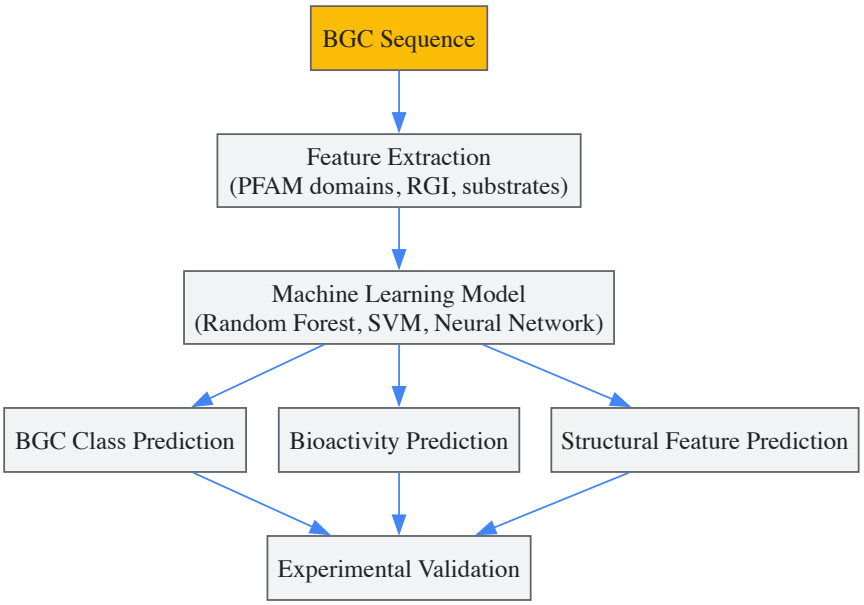


Diagram 2: Machine Learning Approach for BGC Analysis

As these technologies mature, they will accelerate the discovery and characterization of novel natural products, expanding our understanding of chemical biology and enhancing our ability to harness nature's biosynthetic potential for human benefit.

Biosynthetic Gene Clusters represent one of nature's most sophisticated strategies for chemical innovation. These genetically encoded assembly lines produce an extraordinary diversity of specialized metabolites that mediate ecological interactions, provide competitive advantages, and enable organisms to adapt to diverse environments. The systematic study of BGCs through genomic, computational, and experimental approaches has revealed a vast untapped reservoir of chemical diversity with significant implications for medicine, biotechnology, and fundamental biology.

Ongoing advances in sequencing technologies, bioinformatics, and synthetic biology are transforming BGC research, enabling comprehensive mapping of biosynthetic diversity across the microbial world and providing powerful tools for accessing and engineering novel natural products. As these capabilities continue to evolve, BGCs will undoubtedly yield new solutions to pressing challenges in human health, agriculture, and industrial biotechnology while deepening our understanding of the chemical language of life.

Natural products, also known as secondary metabolites, are molecules naturally synthesized by organisms to confer evolutionary advantages and represent an invaluable source for therapeutic drug discovery [11]. Over 60% of drugs approved by the US Food and Drug Administration (FDA) between 1981 and 2019 were derived from natural products [11]. The biosynthetic pathways for these compounds are typically encoded by **biosynthetic gene clusters** (BGCs)—genomic regions containing clustered genes that work coordinately to produce a specific natural product [12] [13].

Advances in genome sequencing and computational tools have revolutionized natural product discovery, enabling systematic analysis of BGC diversity across the tree of life. This technical guide explores the **comparative analysis** of BGCs across three major biological kingdoms: bacteria, fungi, and eukaryotic algae. Understanding the distribution, diversity, and characteristic features of BGCs in these kingdoms provides valuable insights for researchers and drug development professionals engaged in genome mining and natural product discovery [11] [12].

Global Distribution of BGCs Across Kingdoms

Table 1: Comparative BGC Diversity Across Biological Kingdoms

Kingdom	Genomes Analyzed	Total BGCs Identified	Gene Cluster Families (GCFs)	Prediction Tool	Key BGC Types
Bacteria	163,269 [13]	~1,008,546 [13]	Not specified	antiSMASH	NRPS, T1PKS, Hybrid NRPS-PKS, Terpene, RiPPs
Fungi	11,598 [13]	293,926 [13]	26,825 [13]	fungiSMASH	NRPS, T1PKS, NRPS-like, Terpene, DMAT
Eukaryotic Algae	212 [11]	2,762 [11]	Not analyzed	Multiple antiSMASH versions	Trans-AT PKS, T3PKS, Terpene, NRPS

Kingdom-Specific BGC Features

Bacterial BGCs demonstrate remarkable diversity, with entomopathogenic bacteria such as *Xenorhabdus* and *Photorhabdus* showing particular abundance. In these strains, **non-ribosomal peptide synthetases** (NRPS) constitute the predominant class of BGCs (51% of total), followed by hybrid BGCs and type I polyketide synthases (T1PKS) [14]. Bacterial genomes average approximately 24 BGCs per genome, though this varies significantly between taxa [14].

Fungal BGCs display dramatic variation in distribution across taxonomic classes. Eurotiomycetes average 48 BGCs per genome, with approximately five each of NRPS, hybrid NRPS-PKS, and terpene BGCs per genome [12]. In contrast, Basidiomycota possess far fewer BGCs, with terpene BGCs being most abundant in Agaricomycotina [12]. The number of BGCs per genome decreases significantly outside Pezizomycotina, with non-Dikarya phyla averaging fewer than 15 BGCs per genome [12].

Eukaryotic Algal BGCs remain largely unexplored compared to bacterial and fungal systems. Analysis of 212 eukaryotic algal genomes revealed 2,762 putative BGCs [11]. Algal genomes predominantly encode **polyketide synthases** (PKSs), mostly of the trans-acyltransferase type, with very few non-ribosomal peptide synthetases (NRPSs) reported [15]. The complex genomic structure of eukaryotes can present challenges in BGC detection [11].

Table 2: Characteristic BGC Profiles Across Major Fungal Taxa

Fungal Taxon	Average BGCs per Genome	Dominant BGC Types	Noteworthy Genera
Eurotiomycetes	48 [12]	NRPS, T1PKS, Terpene, DMAT	*Aspergillus*, *Penicillium*
Sordariomycetes	~35 [12]	NRPS, NR-PKS, Terpene	*Fusarium*, *Beauveria*
Dothideomycetes	~30 [12]	NRPS, HR-PKS, Terpene	Not specified
Agaricomycotina	<15 [12]	Terpene	Not specified
Non-Dikarya	<15 [12]	Various, limited repertoire	Not specified

Computational Methodologies for BGC Analysis

Standardized BGC Detection Pipeline

The computational detection of BGCs relies on specialized tools that identify genomic regions enriched with biosynthetic genes. The most widely employed tool is **antiSMASH** (antibiotics & Secondary Metabolite Analysis Shell), which employs profile Hidden Markov Models (pHMMs) to detect protein domains characteristic of secondary metabolite biosynthesis [11] [13].

Protocol 1: Genome-Wide BGC Detection Using antiSMASH

- **Input Preparation**: Collect genome assemblies in FASTA format with annotation files (GFF format). For eukaryotic genomes with introns, consider using transcriptome data or specialized gene predictors [15].
- **Tool Selection**: Choose the appropriate antiSMASH version:
  - **Standard antiSMASH**: For bacterial genomes [14]
  - **fungiSMASH**: For fungal genomes [12]
  - **Multiple algorithms**: For eukaryotic algae, run both bacterial and fungal versions as cross-validation [15]
- **Parameter Configuration**: Set analysis parameters based on target BGC types. Key parameters include:
  - `--taxon` (fungi/bacteria) to optimize domain detection
  - `--cb-known-clusters` for comparison with MIBiG database
  - Gene finding tool selection (e.g., `glimmerhmm` for fungi) [13]
- **Execution and Output**: Run antiSMASH to generate BGC predictions with genomic coordinates, domain architectures, and similarity to known BGCs.

Comparative Analysis of BGCs

To manage the substantial number of BGCs identified through genome mining, researchers employ clustering approaches that group similar BGCs into **gene cluster families** (GCFs) [12] [13]. This approach dramatically reduces dataset complexity and enables automated annotation based on experimentally characterized reference BGCs [12].

Protocol 2: BGC Clustering and Comparative Analysis

- **Domain Architecture Vectorization**: Convert BGCs to arrays of protein domains (biosynthetic domain architectures, BDAs) to enable quantitative comparison [11].
- **Similarity Calculation**: Compute pairwise distances between BGCs using metrics such as:
  - Fraction of shared protein domains

    - Backbone protein domain sequence identity [12]
- **Clustering Implementation**: Apply clustering algorithms to group BGCs:
    - **BiG-SLiCE**: For large-scale datasets (>10,000 genomes) [13]
    - **DBSCAN**: Density-based clustering to handle heterogeneous data [12]
- **Threshold Selection**: Choose appropriate clustering thresholds (e.g., T=550 for fungal BGCs) based on reference datasets from MIBiG [13].
- **Network Visualization**: Construct similarity networks to visualize relationships between BGCs and GCFs [14] [12].

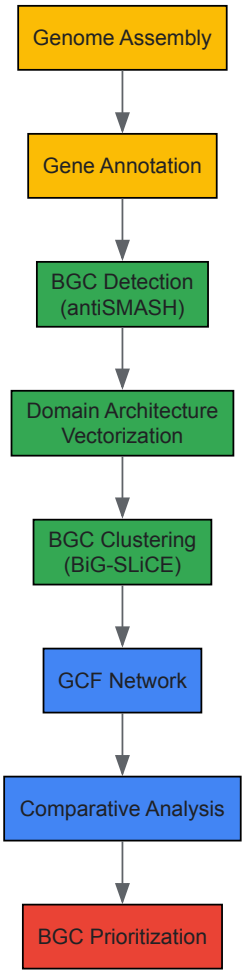


Figure 1: Computational workflow for cross-kingdom BGC analysis, from genome assembly to BGC prioritization.

BGC Prioritization Strategies

The substantial number of BGCs identified through computational methods necessitates prioritization for experimental characterization [11]. Effective strategies include:

- **Phylogenetic Profiling**: Identify GCFs with limited phylogenetic distribution that may produce novel compounds [12] [13].
- **Reference-Based Prioritization**: Select BGCs with similarity to experimentally characterized BGCs from the MIBiG database [11].
- **Taxon-Specific Enrichment**: Target BGCs from taxonomic groups with high biosynthetic potential, such as genera *Xylaria*, *Hypoxylon*, and *Colletotrichum* in fungi [13].

Experimental Characterization of BGCs

Heterologous Expression Systems

A significant challenge in natural product discovery is that many BGCs are "silent" or expressed at low levels under laboratory conditions [16]. **Heterologous expression** provides a powerful solution by transferring BGCs into amenable host organisms [16].

Protocol 3: Heterologous Expression of BGCs

- **Host Selection**: Choose an appropriate heterologous host based on:
    - Phylogenetic proximity to native producer
    - Genetic tractability

    - Codon usage compatibility
    - Substrate availability [16]
- **Vector Construction**: Assemble BGCs in suitable expression vectors:
    - Bacterial Artificial Chromosomes (BACs) for large BGCs
    - Yeast Integration Vectors for fungal BGCs
- **Transformation**: Introduce constructed vectors into selected host:
    - *Escherichia coli* for prokaryotic BGCs
    - *Aspergillus nidulans* or *Saccharomyces cerevisiae* for fungal BGCs
    - *Streptomyces* species for actinobacterial BGCs [16]
- **Metabolite Analysis**: Screen transformants for compound production using LC-MS and NMR.

Validation of BGC Function

Protocol 4: Functional Validation of BGCs

- **Gene Inactivation**: Knock out core biosynthetic genes using CRISPR-Cas9 or homologous recombination.
- **Metabolite Profiling**: Compare metabolic profiles of wild-type and mutant strains.
- **Heterologous Expression**: Express candidate BGCs in surrogate hosts and monitor metabolite production.
- **Enzymatic Assays**: Characterize individual enzymes from the BGC in vitro to confirm predicted functions [11].

The Scientist's Toolkit: Essential Research Reagents

Table 3: Essential Research Reagents for BGC Analysis and Characterization

Reagent/Tool	Function	Application Examples
antiSMASH	BGC detection and annotation	Identification of NRPS, PKS, and hybrid BGCs across kingdoms [11] [13]
BiG-SLiCE	Large-scale BGC clustering	GCF analysis of >10,000 fungal genomes [13]
MIBiG Database	Repository of experimentally characterized BGCs	BGC annotation and prioritization [11] [13]
Heterologous Hosts (E. coli, S. cerevisiae, A. nidulans)	Expression of silent/cryptic BGCs	Production of marine natural products from unculturable microorganisms [16]
CRISPR-Cas9 Systems	Genetic manipulation of BGCs	Gene knockout and pathway engineering [16]

Comparative analysis of BGC diversity across bacterial, fungal, and eukaryotic algal kingdoms reveals dramatic differences in biosynthetic logic and chemical space [12]. While bacterial BGCs demonstrate remarkable abundance and diversity, fungal genomes encode extensive uncharacterized biosynthetic potential, with less than 1% of GCFs mapped to known natural products [13]. Eukaryotic algal BGCs represent a largely untapped resource with unique evolutionary trajectories [11] [15].

Future directions in cross-kingdom BGC research will likely focus on integrating **multi-omics data** to prioritize BGCs for experimental characterization, developing improved **heterologous expression systems** for eukaryotic BGCs, and applying **artificial intelligence** approaches to predict chemical structures from genomic sequences [17]. These advances will accelerate the discovery of novel bioactive compounds from diverse biological kingdoms, expanding the available chemical space for therapeutic development.

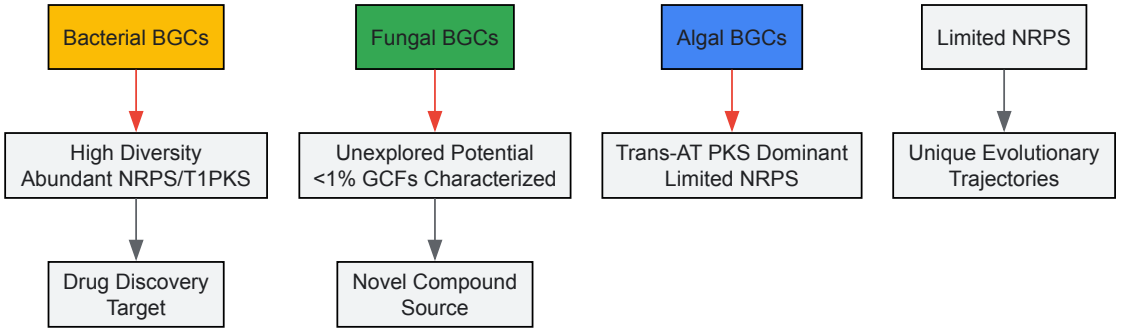


Figure 2: Kingdom-specific BGC features and their implications for natural product discovery.

The **Minimum Information about a Biosynthetic Gene cluster (MIBiG)** repository represents a cornerstone resource in the field of natural product discovery and genomics. Established as a community-driven standard, MIBiG provides a centralized, curated collection of experimentally validated biosynthetic gene clusters (BGCs) from fungal, bacterial, and plant secondary metabolites [18]. In the context of biosynthetic gene cluster identification research, MIBiG serves as an essential reference dataset that enables researchers to interpret the function and novelty of newly identified BGCs through comparative analysis [19]. This repository has become increasingly vital as genome mining technologies have advanced, generating thousands of candidate BGCs from microbial genomes that require functional characterization against a well-documented set of BGCs with known products [19].

The pharmaceutical and biotechnology industries heavily rely on natural products as sources for novel drugs, crop protection agents, and biomaterials [20]. MIBiG accelerates this discovery process by providing standardized annotations that connect genetic information to chemical structures and biological activities, thereby facilitating computational analyses that can predict sequence-structure-function relationships for diverse natural products [20]. By offering a robust framework for data comparison, MIBiG has positioned itself as an indispensable tool for researchers exploring the vast landscape of microbial specialized metabolism.

MIBiG Specifications and Data Architecture

Data Standardization Framework

The MIBiG specification provides a comprehensive data standard that captures the architectural and enzymatic diversity present in characterized BGCs while maintaining flexibility to accommodate future discoveries [19]. The standard encompasses detailed information about biosynthetic pathways, including genomic loci parameters, chemical compound structures, biological activities, biosynthesis steps, and experimental evidence linking BGCs to their molecular products [18]. This standardized approach ensures that data from diverse sources can be computationally analyzed and compared, addressing the previous inconsistency in how BGC information was reported and deposited [21].

To maintain data quality and integrity, MIBiG has adopted JSON schema description and validation technology, which enables embedding validation and dependency rules directly into the schema [19]. This validation framework can be processed programmatically through libraries available in most popular programming languages, ensuring that submissions adhere to the required format and completeness standards. The data schema has undergone significant redesign between versions to improve annotation quality and ensure compliance of future submissions [19].

Database Architecture and Access

MIBiG has evolved from a collection of static HTML pages to a sophisticated relational database architecture that supports complex queries and data retrieval [19]. The current implementation utilizes a PostgreSQL database with a REST-like web API that handles access to the underlying data [19]. This architecture supports multiple user interfaces, including a single-page web application written in AngularJS that allows users to browse the repository, view database statistics, and execute metadata queries through either simple search forms or an interactive query builder for constructing complex Boolean queries [19].

The individual BGC pages within MIBiG are generated using a customized antiSMASH module that sideloads MIBiG annotation files in JSON format [19]. This integration ensures compatibility with one of the most widely used computational tools for BGC identification. Additionally, MIBiG has established cross-links with external chemical structure databases including PubChem, the Natural Products Atlas, and the GNPS spectral library, enabling users to acquire information about specialized metabolites with similar structures and to identify mass spectra linked to specific molecules of interest [19].

Evolution of MIBiG: Version History and Expansion

The MIBiG repository has undergone significant expansion and improvement since its initial release in 2015. The database has grown from an initial collection of 1,170 BGC entries to encompass 2,021 manually curated BGCs in version 2.0 (released in 2019), representing a 73% increase [19]. The most recent update, MIBiG 3.0, added 661 new entries and included large-scale validation and re-annotation of existing entries [20]. This progressive expansion has been driven by both community submissions and organized curation efforts, with particular attention paid to compound structures, biological activities, and protein domain selectivities in the latest version [20].

Table: MIBiG Database Version History and Growth

Version	Release Year	Number of BGC Entries	Key Improvements
Initial Release	2015	1,170	Establishment of data standard and initial repository
MIBiG 2.0	2019	2,021	Schema redesign, manual curation, enhanced query functionality

Version	Release Year	Number of BGC Entries	Key Improvements
MIBiG 3.0	2023	2,682+	Large-scale validation, re-annotation, focus on chemical structures and biological activities

The growth of MIBiG has been facilitated by diverse curation strategies, including crowd-sourced submissions through an online submission form, periodic "Annotathons" where scientists gathered for intensive curation sessions, and educational integration that brought BGC annotation into classroom environments [19]. These efforts have collectively addressed the challenge of keeping the repository current with the rapidly expanding body of literature on characterized biosynthetic pathways. The international scope of these curation efforts is notable, with MIBiG 2.0 involving contributions from 288 scientists across nearly 180 research institutions and companies in 33 countries [18].

Data Content and Composition Analysis

Biosynthetic Class Distribution

MIBiG categorizes BGCs into seven structure-based classes: Alkaloid, Nonribosomal Peptide (NRP), Polyketide, Ribosomally synthesized and Post-translationally modified Peptide (RiPP), Saccharide, Terpene, and Other [19]. These classes are not mutually exclusive, as hybrid clusters encoding multiple biosynthetic pathways are common, such as Polyketide-NRP hybrids like Rapamycin (BGC0001040) and Bleomycin (BGC0000963) [19]. The "Other" category encompasses diverse natural products including cyclitols, indolocarbazoles, and phosphonates [19].

Table: Distribution of BGCs in MIBiG by Biosynthetic Class

Biosynthetic Class	Number of BGCs	Representative Examples
Polyketide (PK)	825	Rapamycin (BGC0001040)
Nonribosomal Peptide (NRP)	627	Bleomycin (BGC0000963)
Ribosomally synthesized and Post-translationally modified Peptide (RiPP)	Information Missing	Information Missing
Terpene	Information Missing	Information Missing
Alkaloid	Information Missing	Information Missing
Saccharide	Information Missing	Information Missing
Other	Information Missing	Fosfomycin (BGC0000938), Rebeccamycin (BGC0000821)

The current distribution reflects historical research emphasis, with polyketides and nonribosomal peptides together comprising more than half (59%) of all entries in MIBiG 2.0 [19]. These classes have been particularly prolific sources of pharmaceutical agents, explaining their disproportionate representation. As detection methods for other classes improve and research interest expands, the distribution across biosynthetic classes is expected to become more balanced in future versions.

Taxonomic Distribution of BGCs

The taxonomic origins of BGCs in MIBiG reflect both scientific interest and the biosynthetic capacity of different organisms. The repository is dominated by bacterial and fungal BGCs, with only 19 entries originating from plants [19]. Within bacteria, the genus *Streptomyces* is particularly prominent with 568 BGCs, followed by *Pseudomonas* with 61 entries [19]. Among fungi, *Aspergillus* leads with 79 BGCs [19]. This distribution mirrors the historical focus on cultivable actinomycetes and fungi as rich sources of bioactive natural products, though recent efforts have expanded to include less studied taxonomic groups.

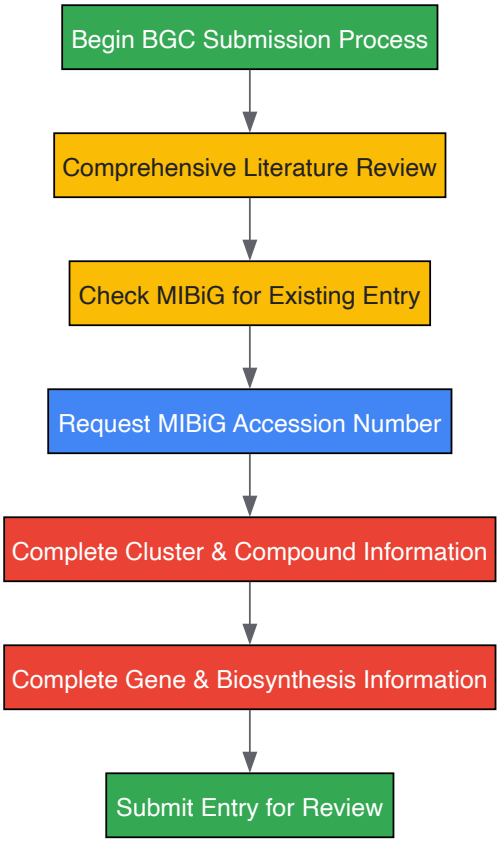
The bias toward certain taxonomic groups also reflects technical considerations in BGC characterization. Organisms with larger genomes and more complex genetics present greater challenges for complete BGC identification and characterization. As sequencing technologies advance and more diverse organisms become genetically tractable, the taxonomic diversity within MIBiG is expected to increase significantly, particularly from uncultivated microorganisms accessed through metagenomic approaches.

MIBiG Submission and Curation Workflow

Standardized Submission Process

The MIBiG repository employs a structured workflow for submitting new BGC entries, designed to ensure data completeness and standardization [21]. The process begins with researchers requesting a MIBiG accession number by providing preliminary information including the product name(s) and sequence information for the BGC, preferably as coordinates within an NCBI GenBank accession [19]. Following approval by MIBiG staff, submitters complete an extended entry form with detailed information about the BGC and its products [19].

To support thorough and accurate submissions, MIBiG provides a Standard Operating Procedure, Excel templates, and tutorial videos that guide researchers through the annotation process [21]. These resources scaffold the annotation procedure, ensuring that all required information is collected and appropriately formatted. The submission workflow emphasizes comprehensive literature review, recommending searches across multiple platforms (Google Scholar, PubMed, etc.) using the natural product name along with "biosynthetic gene cluster" or "biosynthesis" as key terms [21]. Citation tracking and bibliography mining of key authors help identify relevant publications that might otherwise be overlooked.



Data Curation and Quality Control

MIBiG employs multiple strategies to ensure data quality and annotation completeness. The repository utilizes both crowd-sourced contributions and organized curation efforts to expand and maintain its contents [19]. Since 2015, the open submission form has garnered 140 new entries from experts worldwide, while organized "Annotathons" have produced 702 new entries and annotation quality improvements for over 600 BGCs [19].

More recently, MIBiG has integrated curation activities into educational settings, using detailed guidelines and expert validation to generate high-quality BGC entries [19]. This approach has successfully produced annotations for important compounds including actinomycin, daptomycin, nocardicin A, and salinosporamide, among others [19]. The classroom integration serves dual purposes: expanding MIBiG's coverage while providing valuable research experiences for students. For entries that are partially annotated, researchers can update existing information using the 'Update form' on the MIBiG website, ensuring the repository evolves with new scientific discoveries [21].

MIBiG Integration in Research Workflows

Computational Tool Integration

MIBiG serves as a foundational resource for numerous computational tools in natural product discovery. Most notably, it provides the reference dataset for the KnownClusterBlast module in antiSMASH (Automated identification of Biosynthetic Gene Clusters), which enables comparative analysis of newly identified BGCs against characterized clusters [19]. This integration allows researchers to quickly assess the novelty of identified BGCs and generate hypotheses about their products based on similarity to known pathways.

Beyond antiSMASH, MIBiG data supports specialized tools like ClusterCAD, which sources BGC information as starting points for computer-aided design of new biochemical pathways [19]. The standardized annotations in MIBiG enable these computational approaches by providing reliably curated information about gene functions, substrate specificities, and pathway architecture. As machine learning approaches become increasingly important in natural product discovery, MIBiG's comprehensively annotated dataset provides essential training data for models predicting sequence-structure-function relationships [20].

Experimental Design and Validation

For laboratory researchers, MIBiG provides critical guidance for experimental design and validation of BGC functions. The repository includes information about the nature of evidence linking BGCs to their products, helping researchers design appropriate validation experiments [18]. By examining well-characterized examples, scientists can identify common experimental approaches for demonstrating gene functions and pathway assignments, avoiding redundant efforts while ensuring rigorous validation of new discoveries.

The biological activity data incorporated in MIBiG entries helps researchers prioritize BGCs for further investigation based on potential applications [20]. For example, Bahram et al. used homology searches against MIBiG to identify fungal BGCs associated with antibacterial activity across soil metagenomes, demonstrating how the repository can guide targeted discovery efforts for specific bioactivities [19]. This application highlights how MIBiG transcends its role as a simple database to become an active tool for hypothesis generation and experimental planning.

Table: Key Research Reagent Solutions for BGC Research

Resource/Tool	Function/Application	Relevance to MIBiG
antiSMASH	Automated identification and analysis of biosynthetic gene clusters in genomic data	Provides comparative analysis against MIBiG entries via KnownClusterBlast module
NCBI GenBank	Repository for nucleotide sequence data	Source of genomic sequences for BGC deposition and MIBiG submissions
PubMed/Google Scholar	Literature databases for scientific publications	Essential for comprehensive research during MIBiG annotation
ClusterCAD	Computer-aided design tool for engineered biochemical pathways	Uses MIBiG data as starting points for pathway design
GNPS Spectral Library	Repository for mass spectrometry data	Enables cross-linking of chemical structures with analytical data
Natural Products Atlas	Database of known natural product structures	Provides complementary chemical structure information

The resources listed in the table above represent essential components of the biosynthetic researcher's toolkit when working with MIBiG data. These tools collectively support the complete workflow from BGC identification and characterization to comparative analysis and engineering. The integration between these resources creates a powerful ecosystem for natural product discovery, with MIBiG serving as the central reference point for connecting genetic information with chemical structures and biological activities.

Future Directions and Development

The MIBiG repository continues to evolve in response to technological advances and community needs. Future developments will likely focus on improving the representation of less common biosynthetic classes, expanding taxonomic diversity, and enhancing connections to analytical data. The success of classroom integration for curation activities suggests potential for scaling this approach to address the thousands of partially characterized BGCs that remain to be fully annotated [19]. As single-cell sequencing and metagenomic approaches reveal BGCs from previously inaccessible microbial dark matter, MIBiG will play an increasingly important role in contextualizing these discoveries.

The MIBiG consortium has demonstrated a strong commitment to maintaining and expanding this critical resource through community-driven efforts [20]. By continuing to refine data standards, improve annotation quality, and foster integrations with computational tools, MIBiG will remain an indispensable resource for researchers exploring the chemical diversity encoded in microbial genomes. The repository's evolution reflects the dynamic nature of natural product research, adapting to new technologies while maintaining its core mission of providing standardized, accessible information about biosynthetic pathways.

Biosynthetic Gene Clusters (BGCs) represent a fundamental genomic architecture underlying the production of specialized metabolites across the tree of life. These physically clustered sets of non-homologous genes encode discrete metabolic pathways that provide organisms with adaptive advantages in specific ecological contexts. This review examines the evolutionary forces shaping BGC formation, maintenance, and diversification, with particular emphasis on their roles in microbial defense, nutrient acquisition, and inter-organismal communication. We synthesize current understanding of the evolutionary mechanisms driving BGC diversity, including horizontal gene transfer, gene duplication, and genome

rearrangement. Additionally, we provide a comprehensive toolkit for BGC identification and analysis, featuring standardized experimental protocols, computational workflows, and reagent solutions to support ongoing research in natural product discovery and biosynthetic engineering.

Biosynthetic gene clusters (BGCs) are physically linked groups of mostly non-homologous genes that participate in a common, discrete metabolic pathway, often with coregulated expression [1]. These genomic arrangements are widespread features in bacterial and fungal genomes, though they also occur less frequently in other organisms including plants [1] [22]. BGCs are most renowned for encoding the biosynthetic machinery for **secondary metabolites** - the source of most pharmaceutical compounds, natural toxins, and chemical mediators of ecological interactions [1].

The specialized metabolites produced by BGCs are not typically essential for basic growth and development but provide significant **adaptive advantages** in specific environmental contexts [22]. These compounds serve as weapons in microbial warfare (antibiotics), facilitators of nutrient acquisition (siderophores), and mediators of complex communication networks within ecosystems [1] [7]. The evolutionary persistence of the clustered genomic architecture across diverse lineages suggests significant selective advantages despite the potential for horizontal gene transfer to disseminate individual genes [1].

Ecological Roles of BGCs

BGCs encode sophisticated chemical solutions to ecological challenges, effectively serving as genomic toolkits for environmental adaptation. The table below summarizes the primary ecological functions of BGCs and representative examples.

Table 1: Ecological Functions of Biosynthetic Gene Clusters

Ecological Function	Key BGC Types	Representative Example	Ecological Context
Chemical Defense	Non-ribosomal peptide synthetases (NRPS), Polyketide synthases (PKS)	Palmerolide A from Antarctic ascidian microbiome [23]	Defense against competitors; potent V-ATPase inhibitor targeting melanoma cells
Nutrient Acquisition	NRPS-independent siderophore (NIS) clusters	Vibrioferrin BGCs in marine bacteria [7]	Iron chelation in iron-limited marine environments (0.1-2 nM iron in ocean surface waters)
Microbial Interactions	NRPS, Betalactone, Terpene	*Bacillus velezensis* BGCs with surfactin, fengycin, macrolactin [24]	Phytopathogen inhibition and plant root colonization; plant growth promotion
Ecosystem Adaptation	Hybrid PKS/NRPS, Terpene	Siderophore BGCs with varying iron-binding affinities [7]	Strategy for survival in competitive environments with limited resources

Chemical Mediation of Microbial Interactions

BGCs play pivotal roles in structuring microbial communities through the production of bioactive compounds. In rhizosphere ecosystems, certain bacterial BGCs enable beneficial plant-microbe interactions through **antiphytopathogen activity**. For instance, *Bacillus velezensis* strain 2A-2B contains 13 unique BGCs—including those for surfactin, fengycin, and macrolactin—that correlate with its superior antifungal capacity and benign interaction with chili pepper roots [24]. Pre-inoculation of chili pepper plants with this strain reduced disease indices from 3.8 to 0.6 (on a 4-point scale) when challenged with the pathogens *Phytophthora capsici* and *Rhizoctonia solani* [24].

Nutrient Acquisition in Resource-Limited Environments

In oligotrophic environments like marine ecosystems, BGCs encoding siderophores provide crucial adaptations for survival. Marine bacteria face extreme **iron limitation**, with ocean surface waters containing only 0.1-2 nM iron while bacterial growth requires micromolar concentrations [7]. Vibrioferrin-producing NI-siderophore BGCs enable marine bacteria to overcome this limitation through the production of carboxylate-class siderophores with specific iron-binding properties [7]. Genomic analyses reveal that many bacteria maintain multiple siderophore BGCs with varying iron-binding affinities as a strategy to reserve gene clusters for both high and low affinity iron acquisition systems [7].

Evolutionary Mechanisms of BGC Formation and Diversification

The origin and evolution of metabolic gene clusters has been debated since the 1990s, with several mechanisms identified as drivers of BGC diversity [1]. The dynamic nature of BGC genomes reflects multiple evolutionary processes operating across different timescales.

Table 2: Evolutionary Mechanisms Driving BGC Diversity

Evolutionary Mechanism	Process Description	Evolutionary Outcome	Example
Horizontal Gene Transfer	Cross-species transfer of gene clusters	Rapid acquisition of adaptive traits; explains discontinuous distribution	Horizontal transfer of hallucinogenic mushroom BGCs [1]
Gene Duplication & Divergence	Duplication of biosynthetic genes followed by functional specialization	Expansion of metabolic diversity; emergence of new chemical structures	Tandem arrays of terpene synthases and cytochrome P450s in plants [22]
Genome Rearrangement	Reorganization of genomic architecture bringing functionally related genes into proximity	*De novo* formation of BGCs from previously unlinked genes	Convergence of galactose utilization clusters in fungi [1]
Module Skipping & Branching	Unconventional assembly line programming in PKS/NRPS systems	Increased structural diversity without additional genetic material	Branched assembly lines in polyketide and nonribosomal peptide biosynthesis [25]

Horizontal Gene Transfer and Niche Adaptation

Horizontal gene cluster transfer has been strongly linked to ecological niches where the encoded pathways provide selective advantages [1]. This process enables the rapid acquisition of complex metabolic traits without *de novo* pathway evolution. For example, the horizontal transfer of BGCs for specialized tyrosine metabolism modules in fungi correlates with specific ecological niches [1]. The **reproductive trends** among microbial populations further promote this mechanism, as clustering of genes for ecological functions contributes to accelerated adaptation by refining complex functions in the population pangenome [1].

Structural Plasticity and Functional Diversification

BGCs exhibit remarkable structural variability that directly influences their functional outputs. Marine bacterial vibrioferrin BGCs demonstrate this principle, with high genetic variability in **accessory genes** while core biosynthetic genes remain conserved [7]. This structural plasticity affects the resulting metabolites' iron-chelation properties and consequently influences microbial interactions [7]. Clustering analysis reveals that at 10% sequence similarity, vibrioferrin BGCs form 12 distinct families, while at 30% similarity they merge into a single gene cluster family, indicating a continuum of diversity [7].

Research Methodologies for BGC Identification and Analysis

Experimental Workflow for BGC Identification

The following diagram illustrates the integrated computational and experimental pipeline for BGC identification and characterization from environmental samples:

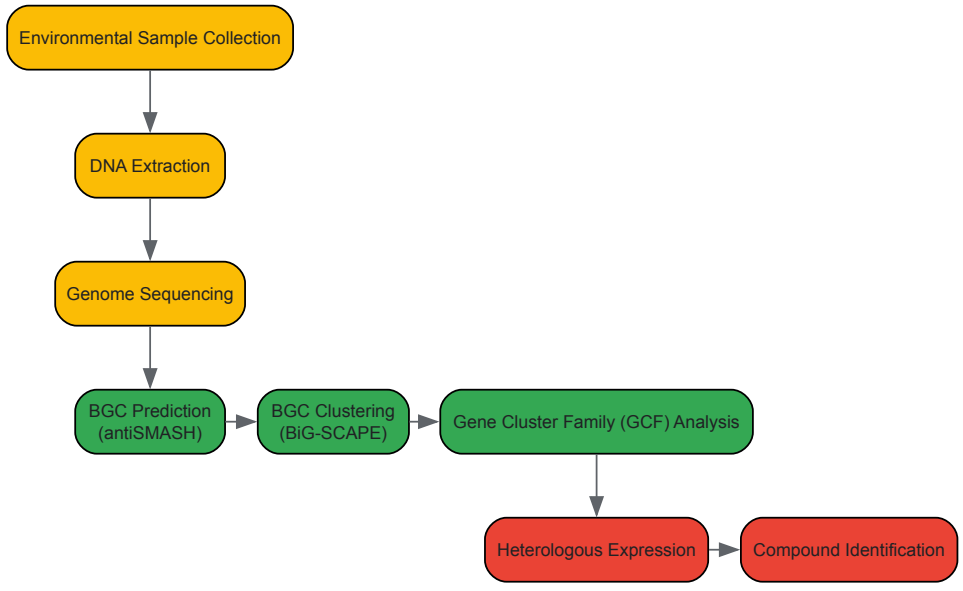


Diagram 1: BGC Identification and Characterization Workflow

Computational Protocols for BGC Identification

BGC Prediction Using AntiSMASH

Purpose: To identify and annotate biosynthetic gene clusters in genomic or metagenomic data [7] [26].

Procedure:

- **Input Preparation**: Prepare genome sequences in FASTA format or use genomic data directly from databases.
- **Analysis Execution**:
  - Run antiSMASH 7.0 (bacterial version) with default detection settings [7].
  - Enable complementary analyses: KnownClusterBlast, ClusterBlast, SubClusterBlast, and Pfam domain annotation [7].
- **Result Compilation**:
  - Export results to a structured format (e.g., Excel spreadsheet).
  - Record total BGC counts and classifications for each genome [7].
  - Compare BGC abundance and diversity across samples.

Technical Notes: antiSMASH employs Hidden Markov Models (HMMs) to identify BGCs based on probabilistic models of core biosynthetic genes [26]. The tool detects predominant BGC classes containing polyketide synthases (PKSs), non-ribosomal peptide synthetases (NRPSs), terpenes, ribosomally synthesized and post-translationally modified peptides (RiPPs), and hybrid BGCs [26].

BGC Clustering and Network Analysis Using BiG-SCAPE

Purpose: To group identified BGCs into Gene Cluster Families (GCFs) based on domain sequence similarity [7] [26].

Procedure:

- **Input Preparation**: Compile antiSMASH-annotated BGCs in GenBank format.
- **Analysis Execution**:
  - Run BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine) version 2.0.
  - Conduct analysis across multiple similarity cutoffs (e.g., 10%, 20%, 30%) [7] [26].
  - Retain singletons and use PFAM database (v37.0) for domain annotation [26].
- **Network Visualization**:
  - Generate similarity networks using distance cutoffs (e.g., 30% and 10%).
  - Import networks into Cytoscape version 3.10.3 for visualization and annotation [7].
- **GCF Interpretation**:
  - At 10% similarity, BGCs form fine-scale families.
  - At 30% similarity, BGCs merge into broader gene cluster families [7].

Technical Notes: The clustering threshold affects family resolution—lower cutoffs imply stricter clustering, leading to fewer connections and vice versa [26]. A cutoff of 0.6 (60% similarity) prevents overestimation of potentially novel BGCs [26].

Experimental Protocols for BGC Functional Characterization

Heterologous Expression and Functional Screening

Purpose: To identify clones containing functional BGCs from environmental DNA (eDNA) libraries [27].

Procedure:

- **Reporter Strain Preparation**:
  - Engineer *Streptomyces albus::bpsA ΔPPTase* reporter strain by deleting the native phosphopantetheinyl transferase (PPTase) gene while maintaining the blue pigment synthase A (bpsA) gene [27].
  - Validate PPTase deletion by PCR screening of genomic DNA.
- **Library Construction**:
  - Clone eDNA into appropriate vectors (cosmid or BAC) to capture complete BGCs.
  - Transfer library from *E. coli* into *S. albus::bpsA ΔPPTase* reporter strain via conjugation [27].
- **Functional Screening**:
  - Screen for clones that restore production of the blue pigment indigoidine, indicating functional PPTase expression [27].
  - Select pigment-producing clones for further analysis.
- **Metabolite Analysis**:
  - Culture positive clones in appropriate media.
  - Extract metabolites and analyze via LC-MS/MS for compound identification.

- Isope and structurally elucidate novel compounds.

Technical Notes: PPTase genes frequently occur in BGCs and are required for activating carrier protein domains in NRPS and PKS systems [27]. This complementation-based screening method identifies BGC-containing clones without prior sequence knowledge.

Table 3: Essential Research Reagents and Resources for BGC Research

Resource/Reagent	Function/Application	Key Features	Reference
antiSMASH	BGC identification and annotation	Hidden Markov Model-based prediction; detects PKS, NRPS, RiPPs, terpenes, hybrids	[7] [26]
BiG-SCAPE	BGC clustering and network analysis	Groups BGCs into Gene Cluster Families (GCFs) based on domain similarity	[7] [26]
MIBiG Repository	Reference database for characterized BGCs	Standardized entries for experimentally validated BGCs; enables comparative analysis	[26] [25]
Streptomyces albus::bpsA ΔPPTase	Reporter strain for functional metagenomics	PPTase-dependent indigoidine production identifies clones with functional BGCs	[27]
Cytoscape	Network visualization and analysis	Visualizes BiG-SCAPE similarity networks and GCF relationships	[7]

Biosynthetic gene clusters represent dynamic genomic architectures that have evolved repeatedly across the tree of life in response to specific ecological challenges. The evolutionary persistence of this clustered organization—despite the potential for horizontal dissemination of individual genes—suggests significant selective advantages in coordinating expression of functionally related genes [1]. The **ecological roles** of BGCs span chemical defense, nutrient acquisition, and mediation of microbial interactions, while their **evolutionary dynamics** are driven by horizontal gene transfer, gene duplication, and genome rearrangement [1] [7].

Future research directions will likely focus on exploiting BGC diversity for natural product discovery through continued development of sophisticated bioinformatic tools and heterologous expression systems. The integration of **metagenomic approaches** with functional screening provides particularly promising avenues for accessing the biosynthetic potential of uncultured microorganisms [27] [23]. Furthermore, the systematic curation of BGC data in standardized repositories like MIBiG will enhance our ability to connect genes to chemistry and elucidate the evolutionary principles governing BGC distribution and diversity [25]. As these efforts mature, they will undoubtedly yield new insights into the evolutionary ecology of specialized metabolism while expanding the repertoire of bioactive compounds available for pharmaceutical and biotechnological applications.

The discovery of bioactive natural products, which have been crucial for developing therapeutics, antibiotics, and agrochemicals, has undergone a fundamental paradigm shift over recent decades. Historically, the identification of biosynthetic gene clusters (BGCs) relied predominantly on traditional biochemical screening methods—a "top-down" approach beginning with phenotypic screening of bioactive natural products in fermentation brods [28] [29]. While this approach yielded many successful drugs, it encountered significant limitations including high rates of rediscovery, low throughput, and insufficiently sensitive technology [28]. The advancement of high-throughput sequencing and bioinformatics has facilitated a transformative new discovery approach: genome mining [28].

Genome mining has emerged as a culture-independent, transformative strategy that enables high-resolution research into secondary metabolite (SM) biosynthesis by directly examining the genetic blueprint of organisms [30] [28]. This "bottom-up" approach starts with sequencing microbial genomes that encode multiple secondary metabolites, then identifies novel secondary metabolite BGCs to pursue [31]. The shift from traditional to genome-based discovery has not only accelerated the identification of novel compounds but has also revealed a previously underestimated biosynthetic potential, particularly through the discovery of cryptic gene clusters that are not expressed under standard laboratory conditions [30] [32].

Historical Perspective: Traditional Biochemical Screening

The Traditional Discovery Workflow

The traditional bioactivity-guided pathway for natural product discovery followed a sequential process beginning with the extraction of compounds from natural sources, followed by isolation and purification of molecules of interest [29]. Biological activity assessments were performed in parallel using various phenotypic assays to confirm bioactivity presence [29]. This process was largely driven by observable phenotypes and was limited by several factors: purification efficiency for targeted molecules, availability of appropriate chemical characterization tools, limited natural resources, and challenges in mass production [29].

Limitations of Conventional Approaches

The constraints of traditional methods became increasingly apparent as the rate of novel compound discovery declined. The high rediscovery rate of known compounds, particularly from commonly studied microbial sources, rendered the approach increasingly inefficient [28]. Furthermore, the traditional pathway provided limited insight into biosynthetic pathways, making structural modifications for optimization challenging. These limitations prompted the exploration of more targeted, genetics-driven approaches that could potentially unlock the vast untapped reservoir of microbial natural products [29].

The Rise of Genome Mining

Conceptual Foundation and Technological Enablers

The conceptual foundation for genome mining emerged from observations in the early 2000s that newly sequenced actinomycete genomes encode many more secondary metabolite BGCs than predicted from known secondary metabolomes [31]. This revelation of "cryptic" or "silent" genetic potential highlighted the limitations of traditional culture-based approaches and stimulated the development of methods to access this hidden chemical diversity [30].

Several technological advancements enabled this paradigm shift. The dramatic decrease in DNA sequencing costs coupled with improvements in bioinformatics tools made large-scale genomic analyses feasible [28]. Critical bioinformatics resources such as antiSMASH (antibiotics and secondary metabolite analysis shell) and the MIBiG (Minimum Information about a Biosynthetic Gene cluster) repository provided standardized platforms for BGC prediction and functional annotation [7] [32]. The development of sophisticated clustering algorithms like BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine) further enabled the organization of BGCs into gene cluster families (GCFs) based on domain sequence similarity [7].

The Genome Mining Value Chain

The genome mining process can be described as a multi-step value chain [31]:

- **Identify promising microbes and sequence their genomes** to finished quality
- **Identify novel BGCs** and their predicted products through bioinformatic analysis
- **Develop fermentation processes** to produce target molecules
- **Isolate and characterize molecules** of interest
- **Determine biological activities** of candidate molecules
- **Optimize pharmacological properties** through medicinal chemistry and combinatorial biosynthesis
- **Pursue clinical development** of promising molecules
- **Obtain regulatory approval** for successful candidates
- **Market approved compounds**

The first two steps represent the most innovative and transformative aspects of the process, where genome mining diverges fundamentally from traditional approaches [31].

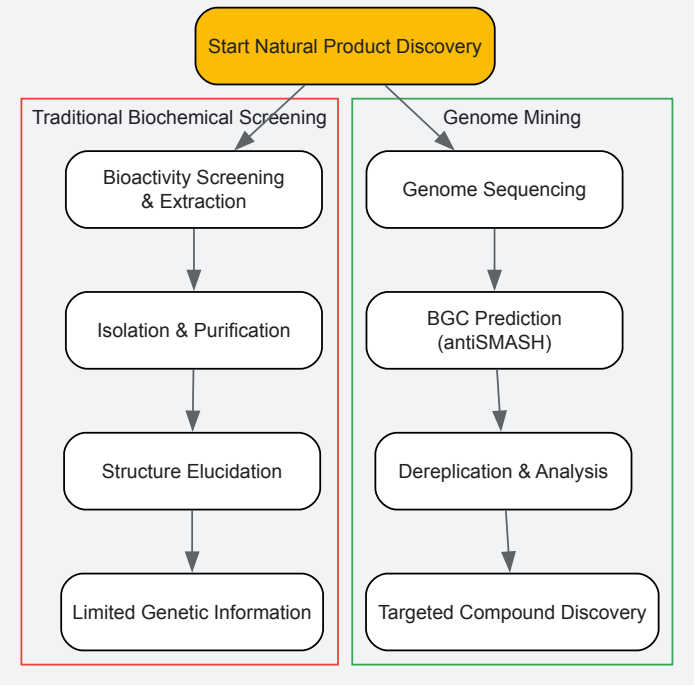


Figure 1: Paradigm Shift from Traditional Biochemical Screening to Genome Mining

Core Methodologies in Genome Mining

Essential Bioinformatics Workflow

A standardized genome mining workflow incorporates multiple bioinformatics tools and databases to progress from raw sequence data to characterized BGCs:

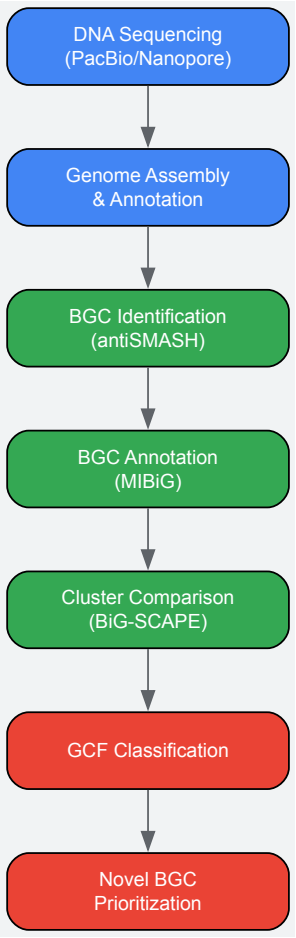


Figure 2: Genome Mining Workflow

Experimental Protocols for BGC Identification and Analysis

BGC Prediction Using antiSMASH

Purpose: To identify and annotate biosynthetic gene clusters in genomic data [7]. **Procedure**:

- Input genomic data in FASTA or GenBank format
- Run antiSMASH analysis using default detection settings
- Enable complementary analyses: KnownClusterBlast, ClusterBlast, SubClusterBlast, and Pfam domain annotation
- Compile results systematically, recording total BGCs and classifications for each genome
- Compare BGC abundance and diversity across multiple strains

Applications: This protocol was used to analyze 199 marine bacterial genomes, identifying 29 BGC types with non-ribosomal peptide synthetases (NRPS), betalactone, and NI-siderophores being most predominant [7].

BGC Clustering and Network Analysis with BiG-SCAPE

Purpose: To group BGCs into Gene Cluster Families (GCFs) based on domain sequence similarity [7]. **Procedure**:

- Prepare annotated BGC files in appropriate format
- Run BiG-SCAPE analysis across multiple similarity cutoffs
- Interpret final clustering results at 10% and 30% similarity cutoffs
- Generate similarity networks using distance cutoffs
- Visualize resulting networks using Cytoscape

- Analyze each GCF to assess structural relationships

Applications: In marine bacteria studies, vibrioferrin BGCs formed 12 families at 10% similarity but merged into a single GCF at 30% similarity, revealing structural plasticity in these siderophore-producing clusters [7].

Phylogenetic Analysis of BGC Distribution

Purpose: To correlate BGC distribution patterns with evolutionary relationships [28]. **Procedure**:

- Select appropriate genetic marker (e.g., rpoB gene for bacterial phylogeny)
- Retrieve and align gene sequences using ClustalW multiple alignment
- Construct maximum likelihood phylogeny with bootstrap replicates
- Export tree as Newick format for visualization in Interactive Tree of Life (iToL)
- Annotate phylogenetic tree with BGC data to explore evolutionary patterns

Applications: This approach revealed that BGC presence/absence patterns generally correlated with phylogenomic patterns at higher taxonomic levels in Alternaria fungi, informing disease management and food safety practices [28].

Table 1: Key Research Reagent Solutions for Genome Mining

Tool/Resource	Function	Application Context
antiSMASH [7] [28]	BGC identification and annotation	Predicts BGCs in genomic data using signature domain detection
MIBiG Repository [7] [32]	BGC functional interpretation	Reference database for comparing target BGCs with known clusters
BiG-SCAPE [7]	BGC clustering and network analysis	Groups homologous BGCs into Gene Cluster Families (GCFs)
PacBio/Nanopore [31]	Long-read DNA sequencing	Generates finished-quality genomes essential for large BGC assembly
funannotate Pipeline [28]	Unified gene prediction	Standardized genome annotation across diverse samples
BiG-FAM Database [32]	Novel BGC identification	Compares query BGCs against ~1.2 million known BGCs

Comparative Analysis: Traditional vs. Genome Mining Approaches

Quantitative Assessment of Discovery Efficiency

Table 2: Performance Comparison of Discovery Approaches

Parameter	Traditional Biochemical Screening	Genome Mining
Starting Point	Bioactivity in fermentation broths [28]	Genetic potential in genomes [28]
BGC Discovery Rate	Limited to expressed compounds	Full biosynthetic landscape including silent clusters [32]
Rediscovery Rate	High [28]	Substantially reduced
Throughput	Low, resource-intensive [28]	High, scalable
Genetic Information	Limited, obtained after characterization	Comprehensive, guides discovery
Novel Compound Potential	Declining	Significant (25-30% novel BGCs in lichenized fungi) [32]
Dereplication Efficiency	Late-stage, after isolation	Early-stage, in silico

Impact on BGC Discovery and Characterization

Genome mining has dramatically expanded the observable biosynthetic universe. Studies of microbial genomes consistently reveal far more BGCs than previously suspected based on traditional methods. For instance, analysis of 187 fungal genomes from Alternaria and related taxa identified 6,323 BGCs, with an average of 34 BGCs per genome—greatly exceeding the number of compounds detected through traditional methods [28]. Similarly, research on lichen-forming fungi of the genus Umbilicaria found that 25-30% of biosynthetic genes showed significant divergence from globally characterized BGCs, indicating substantial untapped structural diversity [32].

The resolution of genome mining also enables fine-scale differentiation of BGC structural variants. Analysis of vibrioferrin-producing BGCs in marine bacteria demonstrated high genetic variability in accessory genes while core biosynthetic genes remained conserved, potentially influencing functional properties like iron-chelation capacity [7].

Advanced Applications and Case Studies

Stereodivergent Enzyme Discovery

Genome mining has proven particularly valuable for identifying enzymes with unusual stereoselectivities, expanding the enzymatic repertoire for constructing complex chiral architectures in pharmaceutical synthesis [30]. Comparative analyses have revealed that subtle variations in sequence and active-site environments produce diverse stereochemical outcomes across enzyme families [30]. This application demonstrates how genome mining directly facilitates the discovery of biocatalysts with tailored properties for synthetic applications.

Taxonomic and Biogeographic Patterns in BGC Distribution

Large-scale genome mining studies have revealed how BGCs are distributed across taxonomic groups and ecosystems. Analysis of Alternaria sections showed that the Infectoriae and Pseudoalternaria sections possessed highly unique GCF profiles compared to other sections, providing ideal candidates for diagnostic marker development [28]. Similarly, the GCF for the Alternaria mycotoxin alternariol (AOH) was found specifically in Alternaria sections Alternaria and Porri, enabling targeted food safety monitoring [28].

Unusual Gene Clusters and Non-Canonical Biosynthesis

Recent genome mining efforts have identified an emerging class of "unusual gene clusters" (uGCs) that lack prominent canonical core enzymes but produce structurally diverse natural products with intriguing biosynthetic logic [33]. These clusters represent a previously overlooked source of chemical diversity that challenges conventional BGC annotation methods and expands the genetic and chemical repertoire of microbial natural products.

Future Perspectives and Concluding Remarks

The paradigm shift from traditional biochemical screening to genome mining has fundamentally transformed natural product discovery. This transition has addressed critical limitations of the former approach while unlocking previously inaccessible chemical diversity. The field continues to evolve with several emerging trends:

The integration of machine learning approaches for BGC prediction and prioritization shows promise for enhancing discovery efficiency [34]. Additionally, the exploration of unconventional gene clusters without prominent core enzymes represents a frontier in expanding the known biosynthetic universe [33]. As the volume of genomic data grows, methods for rapid comparative analysis and GCF classification will become increasingly important for identifying truly novel biosynthetic potential.

Genome mining has firmly established itself as an essential biotechnological tool for discovering novel biosynthetic genes [32]. By linking genetic potential with chemical diversity, this approach has revitalized natural product discovery and will continue to drive innovation in pharmaceutical development, agricultural science, and industrial biotechnology. The ongoing refinement of bioinformatics tools, sequencing technologies, and heterologous expression systems will further enhance our ability to access and exploit the vast reservoir of microbial natural products for years to come.

Computational Tools and Experimental Strategies for BGC Discovery and Characterization

The discovery of novel bioactive natural products has been revolutionized by genome mining, a data-driven approach that leverages microbial genome sequences to identify biosynthetic gene clusters (BGCs). These clusters encode the enzymatic machinery for producing secondary metabolites with diverse pharmaceutical applications, including antibiotics, anticancer agents, and immunosuppressants [35] [36]. The declining cost of DNA sequencing has revealed that typical microbial genomes harbor numerous uncharacterized or "silent" BGCs that are not expressed under standard laboratory conditions [36] [14]. This hidden biosynthetic potential represents a vast resource for drug discovery, particularly crucial in an era of rising antibiotic resistance [37]. Genome mining approaches have subsequently shifted the discovery paradigm from traditional activity-based screening to targeted, in silico prediction of chemical diversity encoded within microbial genomes [35].

Bioinformatics platforms for genome mining share a common goal: to systematically identify BGCs in genomic data and predict their chemical products. Advanced tools now employ sophisticated algorithms including hidden Markov models (HMMs), machine learning, and comparative genomics to annotate these clusters [38] [39]. The integration of genomic predictions with metabolomic data through multi-omics strategies constitutes the current state of the art in "deep mining," enabling researchers to bridge the critical "genome-metabolome gap" where only approximately 25% of predicted BGCs have known products [35]. This technical guide examines core genome mining platforms, their workflows, and integrative approaches for comprehensive BGC identification and characterization.

Core Genome Mining Platforms and Technologies

Platform Architectures and Detection Methodologies

antiSMASH (antibiotics and Secondary Metabolite Analysis SHell) is a versatile, widely-adopted platform for BGC identification. Its modular, plug-and-play architecture employs profile hidden Markov models (pHMMs) to detect signature enzymes across numerous secondary metabolite classes [39]. Initially supporting major classes like polyketides (PKS), non-ribosomal peptides (NRPS), and terpenes, antiSMASH has progressively expanded its detection capabilities. Version 2.0 added support for oligosaccharide antibiotics, phenazines, thiopeptides, and phosphonates [39], while the current version 7.0 integrates over 40 BGC types and incorporates additional functionalities like trans-AT PKS-specific HMMs and CompaRiPPson analysis for novelty assessment in ribosomally synthesized and post-translationally modified peptides (RiPPs) [35] [38]. A key feature is ClusterBlast, which enables comparative analysis against known BGC databases [39].

PRISM (PRediction Informatics for Secondary Metabolomes) adopts a distinct chemical structure-focused approach. Rather than merely identifying cluster boundaries, PRISM connects biosynthetic genes to the enzymatic reactions they catalyze, permitting in silico reconstruction of complete biosynthetic pathways and prediction of final chemical structures [37]. PRISM 4 represents a comprehensive expansion, implementing 618 in silico tailoring reactions and 1,772 HMMs to predict structures for 16 classes of bacterial natural products, including aminoglycosides, nucleosides, β-lactams, and lincosamides [37]. Its combinatorial chemical graph-based approach generates structurally complex, natural product-like predictions that significantly diverge from synthetic chemical space [37].

Specialized and Emerging Tools address specific gaps in BGC analysis. **BGCFlow** provides a systematic Snakemake-based workflow for pangenome-scale mining, integrating multiple tools for functional annotation, phylogenetics, genome mining, and comparative analysis [40]. **DeepBGC** employs machine learning through bidirectional long short-term memory (BiLSTM) networks and random forest classifiers to identify orphan clusters in under-explored phylogenetic groups [35]. **ARTS** 2.0 (Antibiotic Resistant Target Seeker) specializes in identifying resistance genes within BGCs, providing insights into self-resistance mechanisms and potential antibiotic targets [38]. **BAGEL4** and **RODEO** focus specifically on RiPPs, utilizing motif searches and heuristic scoring to identify precursor peptides and associated modification enzymes [38] [41].

Comparative Performance and Capabilities

Table 1: Comparative Analysis of Major Genome Mining Platforms

Platform	Primary Methodology	BGC Classes Detected	Structural Prediction	Key Differentiators
antiSMASH	Profile HMMs, comparative genomics	>40 classes including PKS, NRPS, RiPPs, terpenes, β-lactams	Limited core structure prediction for specific classes	Most comprehensive BGC detection; integrated comparative analysis with ClusterBlast; modular plug-and-play architecture
PRISM 4	HMMs, combinatorial chemical graphs	16 classes including aminoglycosides, nucleosides, β-lactams, alkaloids	Comprehensive chemical structure prediction	Most accurate and complete structural prediction; generates natural product-like chemical space
DeepBGC	BiLSTM neural networks, random forest	Multiple classes with emphasis on novel/orphan clusters	Limited structural prediction	Machine learning approach identifies divergent BGCs in under-explored taxa; reduced false positives
BGCFlow	Integrated workflow (Snakemake)	Dependent on incorporated tools (antiSMASH, GECCO, ARTS2)	Dependent on incorporated tools	Pangenome-scale analysis; reproducible workflow; combines multiple analytics tools

Quantitative evaluations demonstrate PRISM 4's enhanced prediction accuracy, detecting 96% of reference BGCs (1,230 of 1,281) and generating structural predictions for 94% of detected clusters [37]. In comparative analysis, PRISM 4 generated structurally complex predictions with greater molecular weights, more hydrogen bond donors/acceptors, and higher Bertz complexity indices than alternative tools, indicating more natural product-like characteristics [37]. antiSMASH maintains advantages in detection breadth and comparative genomics, while machine learning tools like DeepBGC offer improved capability for identifying novel BGC architectures in phylogenetically distinct organisms [37] [38].

Integrated Workflows and Experimental Protocols

Comprehensive Genome Mining Methodology

Effective genome mining requires a multi-stage analytical process that integrates various bioinformatics tools and experimental validation. The following workflow outlines a systematic approach for BGC identification and characterization:

- **Genome Acquisition and Quality Assessment**: Secure high-quality genome sequences, as assembly completeness significantly impacts BGC prediction accuracy. For Streptomyces genomes, which have large linear chromosomes with high GC content, preferentially use complete

genomes or high-quality drafts assembled from long-read sequencing (PacBio, Nanopore) to avoid fragmentation of large biosynthetic genes [36].

- **BGC Identification and Annotation**: Process genomic data through multiple complementary tools:
  - **Primary Detection**: Run antiSMASH for comprehensive initial BGC identification and boundary prediction [39] [14].
  - **In-Depth Analysis**: Process results through PRISM 4 for detailed structural predictions of encoded metabolites [37].
  - **Specialized Mining**: For specific classes like RiPPs, run dedicated tools (BAGEL4, RODEO); for resistance gene identification, use ARTS 2.0 [38] [41].
- **Comparative Analysis and Dereplication**: Perform sequence similarity networking with BiG-SCAPE to cluster identified BGCs into Gene Cluster Families (GCFs) and compare against known clusters in databases like MIBiG [14] [41]. This facilitates prioritization of novel BGCs and avoids rediscovery of known compounds.
- **Multi-omics Integration**: Correlate genomic predictions with metabolomic data through:
  - **LC-MS/MS Analysis**: Generate feature-based molecular networks (GNPS) from cultured samples [35].
  - **Correlative Linking**: Use tools like NPLinker to statistically associate GCFs with Mass Spectra Molecular Families (MSMFs) [38].
- **Experimental Validation**: Implement heterologous expression (e.g., in *Streptomyces albus* or *E. coli*) or pathway-specific cultivation approaches (OSMAC) to activate silent BGCs [35] [14]. Purify compounds and determine structures using advanced NMR (e.g., 2D NMR with cryogenic probes) and HRMS techniques [35].

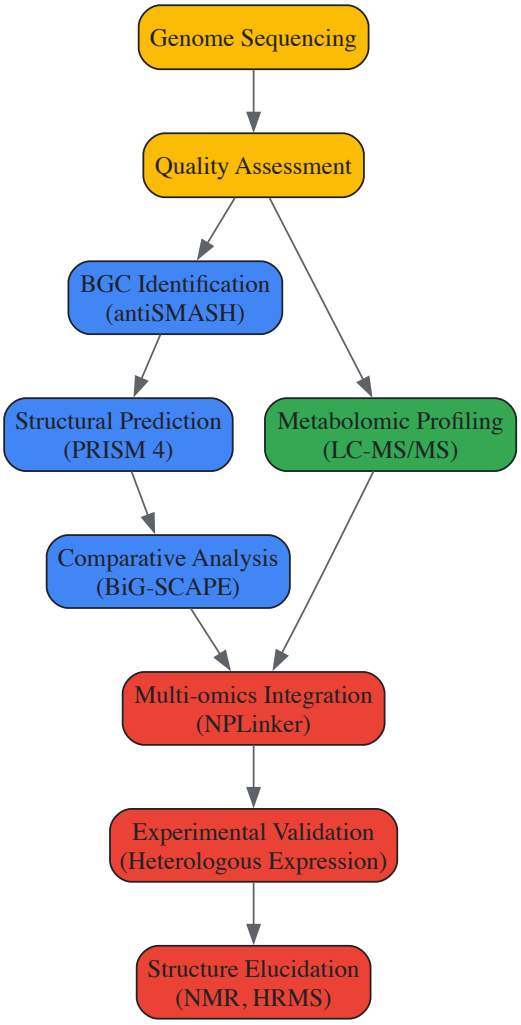


Diagram 1: Integrated Genome Mining and Validation Workflow. The process begins with genome sequencing and progresses through bioinformatic analysis to experimental validation.

Case Study: Identification of Novel P450-Modified RiPPs

A sophisticated workflow for discovering novel P450-modified ribosomally synthesized and post-translationally modified peptides (RiPPs) demonstrates the power of integrated computational approaches [35]:

- **Initial Genome Screening**: Analyze 20,399 actinomycete genomes using the SPECO (short peptide and enzyme co-localization) algorithm to identify co-localized precursor peptide and P450 enzyme genes [35].
- **Structural Binding Prediction**: Process candidate pairs through AlphaFold-Multimer to predict protein complex structures, focusing on conserved binding modes where precursor peptides embed C-termini within P450 pockets while extending core peptides toward the heme center [35].
- **Network Analysis**: Construct multilayer sequence similarity networks (MSSN) of precursor peptide-P450 pairs using EFI-EST (Enzyme Function Initiative-Enzyme Similarity Tool), validated against characterized families like tryptorubin A and cittilin A [35].
- **Heterologous Expression**: Clone and express prioritized BGCs (kst, mci, scn, sgr) in suitable bacterial hosts (*E. coli* or *S. albus* J1074) [35].
- **Compound Characterization**: Purify and structurally elucidate macrocyclic peptides (kitasatides, micitide, strecintide, gristide) using HRMS and NMR, confirming P450-mediated cyclization [35].

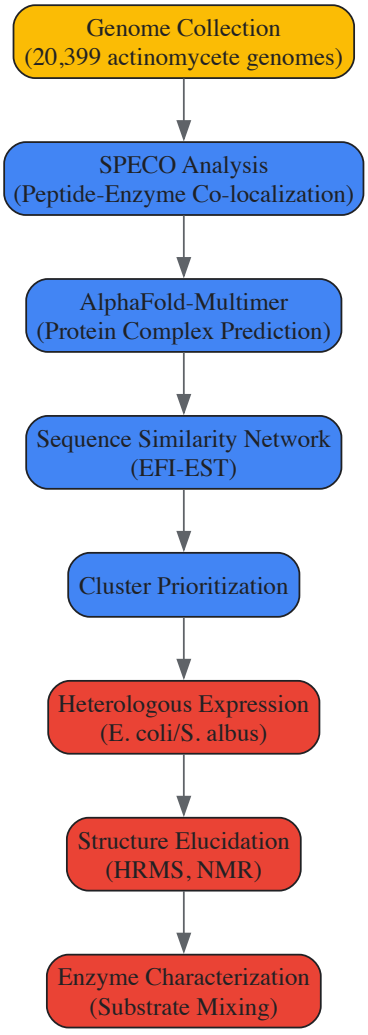


Diagram 2: specialized workflow for P450-modified RiPP discovery, integrating genomic screening with protein structure prediction.

Essential Research Reagents and Computational Tools

Table 2: Key Research Reagents and Computational Resources for Genome Mining

Category	Tool/Resource	Specific Function	Application in Workflow
Core Mining Platforms	antiSMASH 7.0	Comprehensive BGC detection and annotation	Initial BGC identification and boundary prediction [35] [38]
	PRISM 4	Chemical structure prediction from BGC sequences	In-depth structural prediction of encoded metabolites [37]
	DeepBGC	Machine learning-based BGC identification	Detection of novel BGCs in under-explored taxa [35] [38]

Category	Tool/Resource	Specific Function	Application in Workflow
Comparative Analysis	BiG-SCAPE	Gene cluster family network analysis	Dereplication and novelty assessment of BGCs [14] [41]
	BiG-FAM	BGC database and family analysis	Large-scale BGC classification and comparison [38]
Specialized Tools	BAGEL4	RiPPs and bacteriocin identification	Detection of ribosomally synthesized peptides [38] [41]
	ARTS 2.0	Resistance gene identification within BGCs	Prediction of self-resistance mechanisms and mode of action [38]
Database Resources	MIBiG	Repository of known BGCs	Reference for comparative analysis and dereplication [41]
	GNPS	Mass spectrometry data repository and analysis	Metabolomic profiling and correlation with genomic data [35]
Experimental Resources	Heterologous Hosts (*S. albus*, *E. coli*)	BGC expression systems	Activation and production of compounds from silent BGCs [35] [14]
	Cryogenic NMR Probes	High-sensitivity structural analysis	Determination of compound structures and stereochemistry [35]

Genome mining platforms have fundamentally transformed natural product discovery by enabling targeted, data-driven identification of biosynthetic gene clusters. antiSMASH provides the most comprehensive BGC detection framework, while PRISM offers unparalleled accuracy in predicting chemical structures of encoded metabolites [35] [37]. The integration of these tools with machine learning approaches like DeepBGC and workflow management systems like BGCFlow represents the current state of the art in computational BGC identification [40] [38].

The emerging paradigm emphasizes multi-omics integration, combining genomic predictions with metabolomic profiling through platforms like GNPS and NPLinker to bridge the critical gap between genetic potential and chemical expression [35] [38]. Future advancements will likely involve increased incorporation of artificial intelligence for both BGC identification and structural prediction, expanded databases of characterized clusters, and improved algorithms for activating silent BGCs through heterologous expression or precise cultivation strategies [35] [36]. As these technologies mature, genome mining will continue to accelerate the discovery of novel bioactive compounds with therapeutic potential, addressing critical needs in drug development and microbial chemistry.

Machine Learning and Deep Learning Applications in Novel BGC Prediction

Biosynthetic Gene Clusters (BGCs) are groups of co-localized genes in microbial genomes that encode the biosynthesis of natural products, which are chemical compounds that form the basis of many pharmaceuticals, including antimicrobial drugs and anticancer therapies [42] [43]. The identification of BGCs, a process known as genome mining, has been revolutionized by computational tools [42]. Traditional rule-based algorithms, such as antiSMASH, use profile hidden Markov models (pHMMs) and expert-defined heuristics to identify BGCs [44] [42]. However, these methods can struggle with atypical or hybrid clusters and are limited by their reliance on pre-defined rules, which may miss novel BGC classes [44] [43].

Machine learning (ML) and deep learning (DL) have emerged as powerful approaches to overcome these limitations. By learning complex patterns directly from genomic data, these models can identify BGCs with reduced false positive rates and an improved ability to extrapolate and discover novel BGC classes [45] [46]. This technical guide explores the core ML/DL methodologies, experimental protocols, and reagent solutions that underpin modern, computationally-driven natural product discovery.

Core Machine Learning Frameworks for BGC Prediction

The workflow for applying ML to BGC discovery involves several standardized steps: data preparation, feature engineering, model training, and evaluation [42]. Different models make distinct assumptions and are suited to particular tasks.

Feature Engineering and Molecular Representations

A critical step in BGC prediction is converting biological sequences into a numerical format that ML models can process, a step known as featurization [42]. BGCs are typically represented as sequences of protein family (Pfam) domains, which are families of evolutionarily related proteins [43]. Common feature engineering approaches include:

- **Pfam2Vec Embeddings**: Inspired by Natural Language Processing (NLP), this method represents each Pfam domain as a dense vector in a continuous space, allowing the model to capture semantic relationships between domains [45] [47].

- **One-Hot Encoding**: This method converts each sequence of Pfams into a single binary vector, where each bit represents the presence or absence of a specific Pfam domain [47].

Key Machine Learning Models and Algorithms

Table 1: Key Machine Learning Models for BGC Prediction and Their Characteristics

Model Category	Example Algorithms	Core Principle	Application in BGC Prediction
Supervised Classifiers	Random Forest (RF), Support Vector Machine (SVM), Logistic Regression [10] [44]	Learns a mapping function from labeled training data (known BGCs and non-BGCs) to make predictions on new data.	Used for both BGC detection and classification of BGC product classes (e.g., PKS, NRPS). RF often serves as a high-performance benchmark [44].
Recurrent Neural Networks (RNNs)	Bidirectional Long Short-Term Memory (Bi-LSTM) [45] [47]	A type of neural network designed for sequence data that can remember long-term dependencies, processing sequences in both forward and backward directions.	Effective for determining BGC boundaries in a genomic sequence by capturing dependencies between adjacent and distant Pfam domains [45].
Self-Supervised Learning	Masked Language Models (e.g., BiGCARP) [43]	A model is trained to reconstruct parts of its input that have been randomly masked, learning meaningful representations without explicit labels.	Learns rich, contextualized representations of BGCs from large volumes of unlabeled data, which can be fine-tuned for specific tasks like detection and classification [43].
Convolutional Neural Networks (CNNs)	Parallel CNN, Stacked BiLSTM with CNN [47] [42]	Uses convolutional filters to detect local patterns and hierarchical features within data, commonly used in image processing.	Applied to identify RiPP precursor peptides and to improve the stability and accuracy of BGC detection frameworks [47] [42].

The following diagram illustrates a generalized workflow integrating these models for BGC discovery:

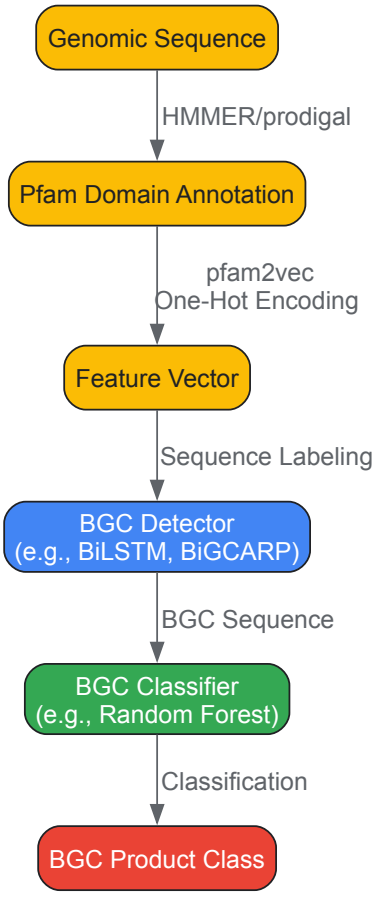


Figure 1: A Generalized ML Workflow for BGC Discovery and Classification. This workflow shows the pipeline from raw genomic sequence to BGC product class prediction, highlighting the roles of different ML models.

Experimental Protocols and Performance Benchmarks

Implementing ML models for BGC prediction requires careful experimental design. Below are detailed methodologies for key tasks.

Protocol: Training a DeepBGC-style BiLSTM Detector

Objective: To train a BiLSTM model to identify BGC regions within a bacterial genome sequence [45] [47].

- **Data Collection:**
  - **Positive Data:** Obtain known BGC sequences from a curated database such as MIBiG (Minimum Information about a Biosynthetic Gene cluster) [10] [45].
  - **Negative Data:** Generate non-BGC sequences. This can be done by extracting genomic regions from reference bacteria that are dissimilar to known BGCs, or by randomly swapping genes in known BGCs with non-BGC genes from a reference genome [45].
- **Feature Extraction:**
  - Annotate all sequences (both positive and negative) using HMMER (hmmscan) against the Pfam database to identify protein domains within each gene [45] [47].
  - Convert the ordered list of Pfam domains for each sequence into a numerical representation using pfam2vec embeddings [45].
- **Model Training and Configuration:**
  - **Architecture:** Implement a Bidirectional LSTM (BiLSTM) network. A typical configuration uses a stateful BiLSTM layer with 128 units and a dropout rate of 0.2, followed by a time-distributed dense layer with sigmoid activation [45] [47].
  - **Training Parameters:** Use the Adam optimizer with a learning rate of 0.0001. Train the model for a sufficient number of epochs (e.g., 1000) with early stopping configured to monitor the validation AUC-ROC with a patience of 20 epochs [47].
  - **Input Format:** Sequences are processed in batches (e.g., size 64) with a fixed number of timesteps (e.g., 256) [47].

Protocol: Predicting Bioactivity from BGC Sequence

Objective: To train a classifier that predicts the biological activity (e.g., antibacterial, antifungal) of a natural product directly from its BGC sequence [10].

- **Data Set Preparation:**
  - Assemble a dataset of BGCs with experimentally verified bioactivities from databases like MIBiG and literature curation [10].
  - Annotate BGCs with a wide set of features, including Pfam domains, sub-PFAM classifications from Sequence Similarity Networks (SSNs), and predictions from the Resistance Gene Identifier (RGI) [10]. This can result in a feature vector with over 1800 dimensions [10].
- **Model Training and Evaluation:**
  - Train multiple binary classifiers (e.g., Random Forest, SVM, Logistic Regression) using scikit-learn for different activity types (e.g., antibacterial, antifungal) [10].
  - Optimize model parameters using 10-fold cross-validation to maximize the balanced accuracy, which is crucial for imbalanced datasets [10].
  - Evaluate the model on a held-out test set and compare its performance against a classifier trained on scrambled features to ensure it performs better than random guessing [10].

Table 2: Performance Benchmarks of Selected BGC Prediction Tools and Models

Tool / Model	Primary Model Architecture	Reported Performance	Key Application / Strength
DeepBGC [45]	BiLSTM RNN, pfam2vec, Random Forest	Improved BGC detection accuracy and identification of novel BGC classes compared to ClusterFinder.	A deep learning genome-mining strategy for BGC identification and product class classification.
Bioactivity Predictor [10]	Random Forest, SVM, Logistic Regression	Balanced accuracy up to 80% for predicting antibacterial activity. Accuracies for anti-Gram-negative and antifungal classifiers were lower (57-70%).	Predicts natural product antibiotic activity directly from BGC sequence, bypassing the need for structure prediction.
RFBGCpred [44]	Random Forest with Word2Vec and UMAP	Reported high predictive performance (Accuracy: 98.02%, MCC: 0.9752, AUC: 0.9928) on a focused set of BGC classes.	A complementary classifier focused on five major BGC classes (PKS, NRPS, RiPPs, Terpenes, Hybrids).

Tool / Model	Primary Model Architecture	Reported Performance	Key Application / Strength
BiGCARP [43]	Self-supervised Masked Language Model (CNN-based)	Shows improvements in BGC prediction and natural product classification over DeepBGC in a comparative analysis.	Leverages self-supervised learning to create meaningful BGC representations without relying on curated negative examples.

Successful implementation of the protocols above depends on a suite of key software tools and databases.

Table 3: Essential Research Reagents and Resources for ML-based BGC Discovery

Resource Name	Type	Function in the Workflow
antiSMASH [7] [42]	Software Tool / Database	The gold-standard, rule-based platform for BGC identification and analysis; often used for generating initial annotations and training data.
MIBiG [10] [48]	Database	A curated repository of known BGCs with experimentally characterized metabolites; serves as a vital source of positive training data.
Pfam Database [45] [47]	Database	A large collection of protein family models; used with HMMER to identify and annotate domains within protein sequences, forming the basic vocabulary for BGC representation.
HMMER [45] [47]	Software Tool	A toolkit for profiling with profile hidden Markov models; essential for scanning genomic sequences against the Pfam database.
prodigal [45]	Software Tool	A software for prokaryotic gene prediction; used to identify open reading frames (ORFs) in genomic sequences prior to Pfam annotation.
Word2Vec / pfam2vec [45] [44]	Algorithm	An NLP-inspired method for generating distributed vector representations of Pfam domains, capturing functional similarities.
scikit-learn [10] [44]	Software Library	A fundamental Python library for machine learning; provides implementations of Random Forest, SVM, and other classifiers used for BGC and bioactivity prediction.
TensorFlow / Keras [45] [47]	Software Library	Open-source libraries for building and training deep learning models, including the BiLSTM networks used in DeepBGC.

Architectural Diagrams of Key Models

To deepen understanding, the architectures of two pivotal models are detailed below.

DeepBGC BiLSTM Architecture

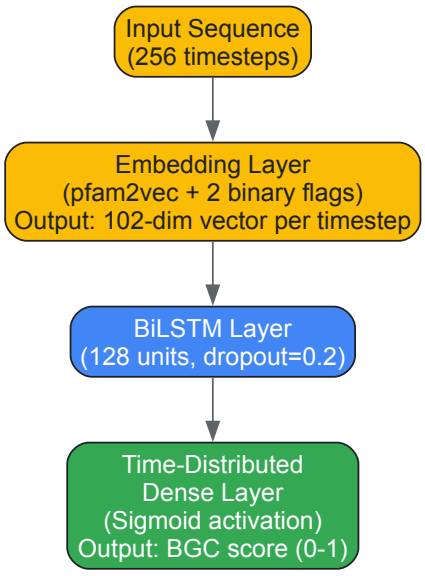


Figure 2: DeepBGC's BiLSTM detector architecture processes Pfam sequences to score BGC regions.

BiGCARP Self-Supervised Model

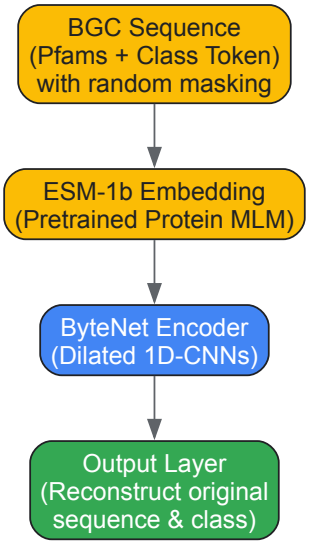


Figure 3: BiGCARP self-supervised model uses masked language modeling to learn BGC representations.

Machine learning and deep learning have fundamentally transformed the field of BGC prediction. The movement from rule-based systems to models that learn directly from data has enabled the discovery of novel BGCs with greater accuracy and less human bias. Current research is pushing the boundaries with self-supervised learning, which reduces the dependency on meticulously curated labeled data. As these computational methods continue to evolve, integrating more diverse biological contexts and multi-omics data, they promise to further accelerate the discovery of novel natural products, fueling future innovations in drug development and biotechnology.

The systematic identification of biosynthetic gene clusters (BGCs) has become a cornerstone of natural product discovery and drug development research. BGCs are physically clustered groups of genes that encode the enzymatic pathways for specialized metabolite biosynthesis, representing valuable blueprints for antibiotics, anticancer agents, insecticides, and immunosuppressants [49]. The exponential growth of genomic data has necessitated the development of specialized database ecosystems that catalog, annotate, and facilitate analysis of these genetic determinants of chemical diversity. These resources provide life scientists with convenient access to extensive data while simultaneously offering computer scientists curated datasets for artificial intelligence development [50]. This technical guide categorizes and characterizes the current landscape of BGC databases, providing researchers with a structured framework for selecting appropriate resources based on specific research objectives.

BGC databases can be systematically classified into three primary categories based on scope and specialization: comprehensive repositories, organism-specific resources, and metabolite-focused collections. The table below summarizes the key characteristics and representative examples of each database category.

Table 1: Classification Framework for BGC Database Ecosystems

Database Category	Primary Focus	Key Characteristics	Representative Examples
Comprehensive Resources	Broad BGC cataloging across diverse taxa	Extensive metadata, cross-references, standardized annotations	PLSDB, MIBiG, BiG-FAM
Organism-Specific Resources	BGCs from particular taxonomic groups	Tailored to phylogenetic constraints, specialized annotation pipelines	XP-focused resources, Marine BGC databases
Metabolite-Focused Resources	Specific metabolite classes or pathways	Chemistry-driven, structural annotation, pathway mapping	Terpenoid databases, NRPS/PKS resources

Comprehensive BGC databases aim to provide extensive collections of gene clusters across diverse taxonomic ranges. PLSDB represents a paradigm of this category, hosting 72,360 curated plasmid entries as of its 2025 update, with enriched annotations including BGC identification, protein-coding genes, antimicrobial resistance genes, and mobility typing [50]. The database architecture incorporates sophisticated filtering capabilities, allowing selection based on host ecosystems, associated diseases, geographical information, and functional structures. The MIBiG (Minimum Information about a Biosynthetic Gene Cluster) database provides a complementary comprehensive resource, with its version 4.0 offering standardized annotations for understanding BGC organization and function through global collaborative curation [7].

The BiG-FAM database represents another comprehensive approach, specializing in gene cluster family (GCF) explorations that enable researchers to identify evolutionarily related BGCs across taxonomic boundaries. This resource has demonstrated that approximately 58% of GCFs in certain bacterial genera are exclusive, highlighting the importance of comprehensive databases for revealing taxonomic-specific biosynthetic capacity [9].

Organism-specific databases focus on BGCs from particular taxonomic groups, enabling specialized investigations into phylogenetic patterns and evolutionary trajectories. Research on entomopathogenic Xenorhabdus and Photorhabdus (XP) bacteria exemplifies this approach, with dedicated resources cataloging 1,000 BGCs across 45 bacterial strains, categorized into 176 families and 11 conserved BGC classes [9]. These organism-specific collections reveal striking patterns, such as NRPS BGCs accounting for 59% of total clusters in XP bacteria, with an average of 22 BGCs per species—two to tenfold higher than average BGC levels in other Enterobacteriaceae [9].

Marine bacteria represent another focal point for organism-specific BGC databases, with resources dedicated to taxa such as Pseudovibrio species and other marine isolates. These databases leverage the unique biosynthetic potential of marine microorganisms, which have been shown to encode diverse BGC types including non-ribosomal peptide synthetases (NRPS), betalactone, and NI-siderophores as predominant cluster types [7] [51].

Metabolite-focused databases organize BGC information around specific chemical classes or biosynthetic pathways, facilitating compound discovery and engineering. These resources are particularly valuable for investigating prominent natural product families such as polyketides, non-ribosomal peptides, ribosomally synthesized and post-translationally modified peptides (RiPPs), terpenoids, and saccharides [49] [52]. Specialized tools like NRPSpredictor and SBSPKS complement these databases by inferring substrate specificity for identified gene clusters, enabling more accurate prediction of metabolic output [49].

The strategic value of metabolite-focused databases lies in their ability to connect genetic architecture with chemical structure, illuminating the relationship between gene organization and compound properties. This is particularly evident in studies of siderophore diversity, where databases catalog the approximately 500 structurally diverse molecules produced through NRPS or NRPS-independent siderophore (NIS) enzymes, with variations evolving due to microbial competition for Fe^{3+} uptake [7].

Experimental Methodologies for BGC Analysis

Standardized BGC Identification Workflow

The computational identification of BGCs follows a standardized workflow incorporating multiple bioinformatics tools and database resources. The following diagram illustrates the core pipeline for BGC discovery and characterization:

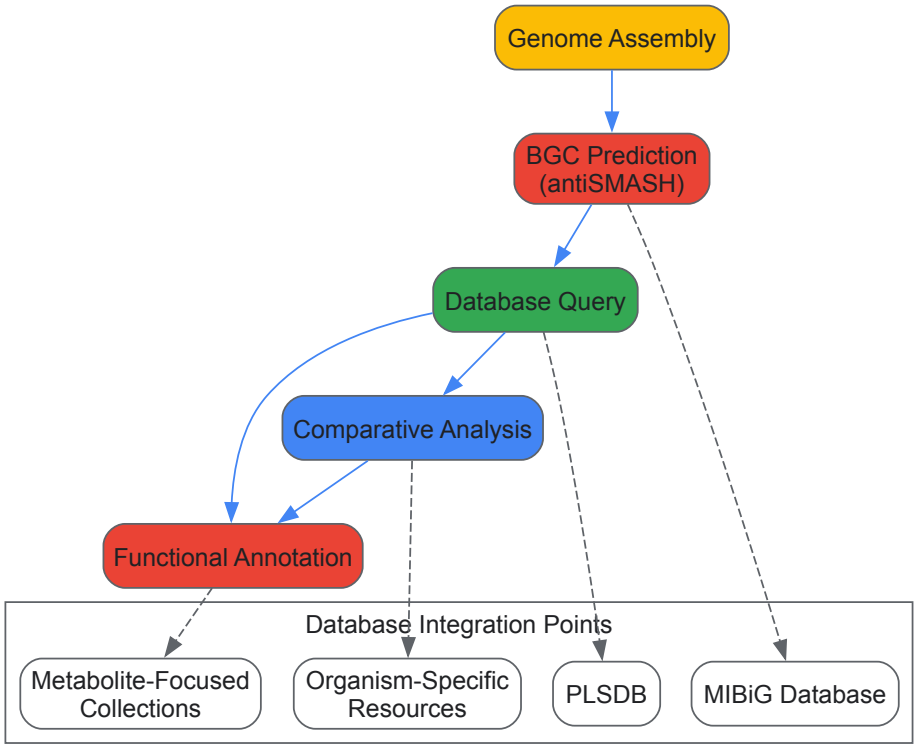


Diagram 1: BGC Analysis Workflow

The experimental protocol for comprehensive BGC analysis typically begins with genome retrieval and quality assessment. For example, in marine bacterial studies, researchers retrieve 199 bacterial genomes from NCBI, prioritizing complete genomes when available, while including high-quality contig-level assemblies for species without complete genomes [7]. Subsequent BGC prediction employs antiSMASH (antibiotics and Secondary Metabolite Analysis SHell) with default detection settings, enabling KnownClusterBlast, ClusterBlast, SubClusterBlast, and Pfam domain annotation [7] [9]. The results are systematically compiled into structured data formats for downstream analysis.

Advanced Comparative Analysis Methods

Following initial identification, advanced comparative analysis elucidates evolutionary relationships and functional potential. Phylogenetic analysis often employs genetic markers such as the rpoB gene, with sequences aligned using ClustalW multiple alignment tools and phylogenetic trees constructed using maximum likelihood methods with bootstrap replication in MEGA11 software [7]. BGC clustering is performed using BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine) to group clusters into Gene Cluster Families (GCFs) based on domain sequence similarity, with analysis conducted across multiple similarity cutoffs (typically 10% and 30%) [7].

Pangenome analysis represents another powerful methodological approach, implemented through platforms like anvi'o to characterize gene content across multiple strains. This technique enables classification of BGCs into core, accessory, and singleton regions, revealing conservation patterns and taxonomic distribution [9]. For example, application of this method to Xenorhabdus and Photorhabdus bacteria demonstrated that although NRPS BGCs are numerically abundant, their biosynthetic genes scatter predominantly in accessory and singleton genomic regions [9].

Essential Research Reagents and Computational Tools

Successful BGC identification and characterization requires a suite of specialized computational tools and database resources. The table below catalogs essential solutions for conducting comprehensive BGC research.

Table 2: Essential Research Reagent Solutions for BGC Analysis

Tool Category	Specific Tool	Primary Function	Application Context
BGC Prediction	antiSMASH	Identifies BGCs in genomic data	Primary detection of gene clusters across diverse taxa
Comparative Analysis	BiG-SCAPE	Clusters BGCs into families	Evolutionary analysis of BGC relationships
Phylogenetic Analysis	MEGA11	Constructs evolutionary trees	Phylogenetic placement of BGC-harboring organisms
Sequence Alignment	Clustal Omega	Multiple sequence alignment	Core gene comparison within BGC families
Genome Annotation	PGAP	Automated genome annotation	Functional prediction of BGC components
Network Visualization	Cytoscape	Visualizes BGC similarity networks	Representation of gene cluster families
Database Resources	MIBiG	Reference database of known BGCs	Annotation and comparison of putative BGCs
Specialized Prediction	NRPSpredictor	Predicts NRPS substrate specificity	Functional annotation of NRPS BGCs

These research reagents form an integrated ecosystem for BGC discovery and characterization. antiSMASH serves as the cornerstone prediction tool, capable of analyzing DNA sequences or annotated nucleotide files in GenBank or EMBL format, while incorporating additional specialized prediction modules like NRPSpredictor for substrate specificity inference [49]. The continual development of these bioinformatics methods addresses critical challenges in BGC prediction, including the identification of signature enzymes, determination of cluster boundaries, and inference of substrate specificity for non-ribosomal peptide synthetases and polyketide synthases [49].

Integration of Database Ecosystems in Research Applications

Case Study: Marine Bacterial BGC Diversity

The application of integrated database ecosystems has revealed remarkable insights into marine bacterial BGC diversity. Systematic analysis of 199 marine bacterial genomes from 21 species identified 29 distinct BGC types, with non-ribosomal peptide synthetases (NRPS), betalactone, and NI-siderophores representing predominant cluster types [7]. The study employed antiSMASH 7.0 for comprehensive BGC detection, followed by BiG-SCAPE clustering to group NI-siderophore BGCs encoding vibrioferrin into gene cluster families. This approach demonstrated that at 10% similarity, vibrioferrin BGCs formed 12 families, while at 30% similarity, they merged into a single gene cluster family, highlighting the genetic plasticity of these biosynthetic systems [7].

Case Study: Metagenomic BGC Discovery from Pharmaceutical Wastes

Metagenomic approaches leveraging BGC databases have enabled novel compound discovery from complex environmental samples. Research on hospital and pharmaceutical industry wastes employed shotgun metagenomic sequencing followed by antiSMASH analysis to detect multiple BGC classes, including terpenes, bacteriocins, and non-ribosomal peptide synthetases [53]. Functional analysis revealed enrichment of ATP-binding cassette (ABC) transporter protein families and winged-helix protein domains, both linked to antibiotic resistance, metabolite translocation, and antibiotic biosynthesis regulation [53]. This demonstrates how database ecosystems facilitate correlation between genetic potential and functional adaptation in extreme environments.

Evolutionary Analysis of BGC Dynamics

Database ecosystems have also enabled sophisticated evolutionary analyses of BGC dynamics across bacterial taxa. Systematic computational analysis of BGC evolution has revealed that complex metabolite clusters often evolve through successive merger of smaller sub-clusters, which function as independent evolutionary entities [52]. Network analysis of sub-cluster sharing patterns demonstrates mosaic architecture in many BGCs, with >60% of the coding capacity of some clusters (e.g., those encoding vancomycin and rubradirin) composed of individually conserved sub-clusters [52]. This "bricks and mortar" model of gene cluster evolution, wherein gene clusters comprise large, modular "bricks" (sub-clusters) encoding key building blocks and individual genes (the "mortar") encoding tailoring, regulation, and transport functions, provides fundamental insights into natural product evolution [52].

The integration of comprehensive, organism-specific, and metabolite-focused database resources continues to accelerate natural product discovery and engineering. As these ecosystems expand in scope and sophistication, they offer increasingly powerful platforms for connecting genetic potential with chemical diversity, ultimately advancing drug development and biotechnology applications.

The discovery of novel bioactive natural products has entered a transformative phase, shifting from traditional activity-guided isolation to sophisticated, data-driven strategies. A persistent challenge in this field has been the significant gap between the vast biosynthetic potential encoded in microbial genomes and the limited number of characterized metabolites. Genome sequencing routinely reveals that a typical *Streptomyces* strain contains dozens of **biosynthetic gene clusters (BGCs)**, yet only a fraction of these are expressed and detected under standard laboratory conditions [54] [35]. This discrepancy has driven the development of **integrated approaches** that combine genomic mining with sensitive metabolomic profiling to efficiently connect gene clusters to their chemical products, thereby accelerating the discovery of novel compounds with pharmaceutical potential.

This technical guide details the methodology and application of combining genomic data with MS-based molecular networking, a powerful synergy that has become a cornerstone of modern natural product research. By framing these techniques within the context of BGC identification and characterization, this review provides researchers and drug development professionals with a comprehensive framework for implementing these strategies in their discovery pipelines.

Core Methodological Components

Genomic Mining for Biosynthetic Gene Clusters

The initial phase of the integrated approach involves comprehensive genomic analysis to map the biosynthetic potential of a microbial strain.

- **BGC Prediction Tools**: The antiSMASH (antibiotics and Secondary Metabolite Analysis Shell) platform serves as the primary workhorse for BGC identification. The current version 7.0 employs **hidden Markov models (HMMs)** to detect and annotate over 40 distinct types of BGCs, including those for polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS), ribosomally synthesized and post-translationally modified peptides (RiPPs), and hybrids [7] [35]. Complementary tools like DeepBGC utilize **bidirectional long short-term memory networks (BiLSTM)** and Random Forests to identify additional "orphan" clusters in under-explored microbial phyla [35].
- **Phylogenetic Analysis**: Establishing accurate taxonomic classification is crucial for contextualizing BGC diversity. The *rpoB* gene has emerged as a reliable phylogenetic marker due to its conserved nature, providing robust evolutionary relationships among bacterial strains [7]. For *Streptomyces* strains, overall genome-related indices (OGRI) including **Average Nucleotide Identity (ANI)** and **digital DNA-DNA Hybridization (dDDH)** provide definitive taxonomic placement, with values below 95% ANI and 70% dDDH indicating novel species [54].
- **BGC Clustering and Analysis**: Following identification, BGCs are organized into **Gene Cluster Families (GCFs)** using tools like BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine), which groups clusters based on protein domain sequence similarity. This analysis reveals evolutionary relationships and helps prioritize BGCs with novel architectures [7].

Table 1: Representative BGC Distribution in Marine Bacterial Genomes

Bacterial Group	Total Genomes Analyzed	Predicted BGCs	Dominant BGC Types	Noteworthy Findings
Vibrio spp.	58	29 distinct types	NI-siderophores, NRPS, betalactone	Vibrioferrin BGCs showed high genetic variability in accessory genes
Streptomyces sp. B1866	1	42	PKS (17), NRPS (2), PKS-NRPS hybrids (3)	>50% showed <70% similarity to known BGCs
Streptomyces sp. RO-S4	1	19	Angucycline-type PKS	Grincamycin-like BGC with novel modifications

MS-Based Molecular Networking

Molecular networking has revolutionized the visualization and annotation of chemical space in complex metabolomic samples by organizing MS/MS spectra based on structural similarity.

- **Theoretical Foundation**: The core principle of molecular networking is that **structurally similar molecules** produce similar fragmentation patterns in tandem mass spectrometry. The spectral similarity between two molecules is quantified using a **cosine score** (ranging from 0-1), which accounts for matching fragment ions, their relative intensities, and mass accuracy [55]. Spectra are first processed using the MS-Cluster algorithm to merge nearly identical spectra into consensus spectra, reducing redundancy [56].
- **Molecular Networking Variants**: Several specialized networking approaches have been developed to address specific analytical challenges:
  - **Classical Molecular Networking (CLMN)**: The original implementation that clusters MS/MS spectra based on cosine similarity, creating visual maps of chemical relationships [55].
  - **Feature-Based Molecular Networking (FBMN)**: Integrates quantitative chromatographic feature detection from tools like MZmine or XCMS with molecular networking, enabling the incorporation of retention time and peak area data while resolving isomeric compounds [57].
  - **Ion Identity Molecular Networking (IIMN)**: Addresses the challenge of multiple ion species (e.g., [M+H]+, [M+Na]+) from the same compound by incorporating chromatographic peak shape correlation analysis, connecting and collapsing different ion forms of the same molecule to reduce network redundancy [58].
- **Platform Implementation**: The **Global Natural Product Social Molecular Networking (GNPS)** platform serves as the central hub for molecular networking analysis, providing a web-based environment for data processing, library matching, and network visualization [56]. GNPS interfaces with massive spectral libraries, enabling automated annotation of nodes through comparison to reference spectra.

Table 2: Key Parameters for Molecular Networking on GNPS

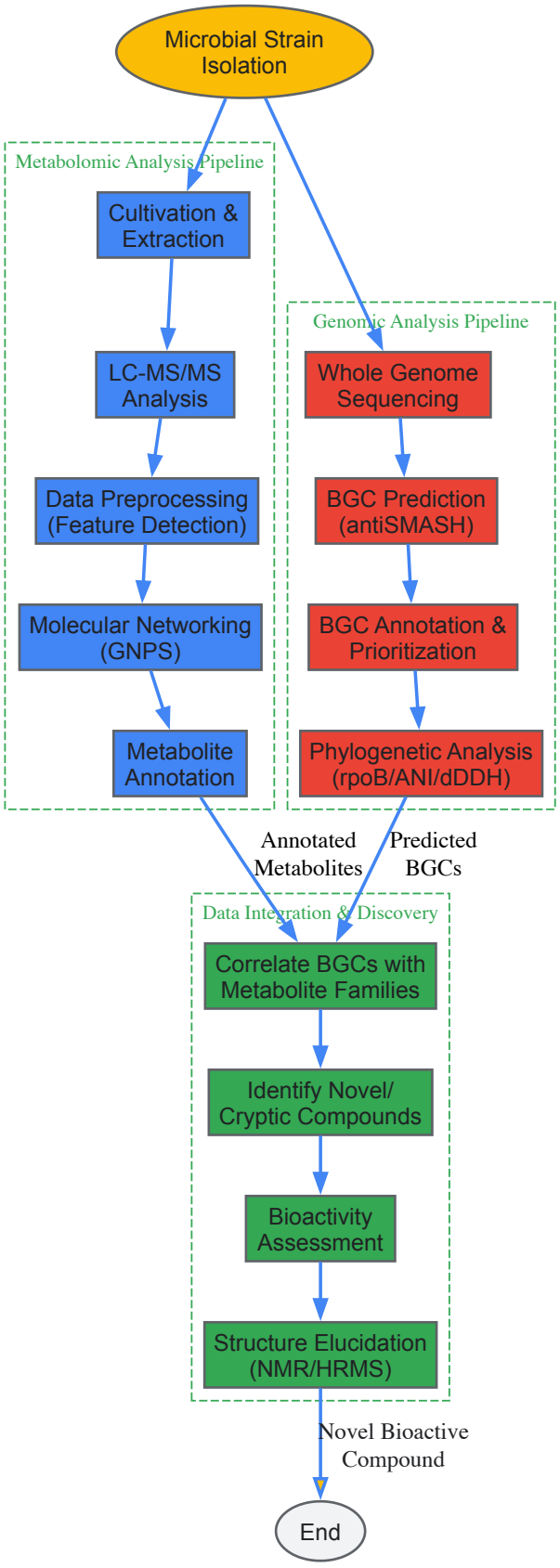
Parameter Category	Specific Parameter	Recommended Setting (High-Resolution MS)	Impact on Network Structure
Basic Options	Precursor Ion Mass Tolerance	± 0.02 Da	Influences MS-Cluster spectrum consolidation
	Fragment Ion Mass Tolerance	± 0.02 Da	Affects spectral similarity calculations
Advanced Network Options	Minimum Cosine Score	0.7	Higher values increase network stringency
	Minimum Matched Fragment Ions	6	Fewer peaks increase connections, reduce specificity
	Maximum Connected Component Size	100 (or 0 for unlimited)	Limits oversized clusters for better visualization
Advanced Library Search	Library Search Min Matched Peaks	6	Balances annotation sensitivity and specificity
	Score Threshold	0.7	Minimum similarity for library-based annotations

Integrated Workflow: From Genes to Metabolites

The power of these approaches emerges from their strategic integration, which creates a bidirectional pipeline connecting genomic potential to metabolomic reality.

Workflow Visualization

The following diagram illustrates the comprehensive integrated workflow for combining genomic data with MS-based molecular networking:



Experimental Protocols

Genomic Sequencing and BGC Identification Protocol

- **DNA Extraction and Sequencing**: Isolate high-quality genomic DNA using standardized kits (e.g., Wizard Genomic DNA Purification Kit). Perform whole-genome sequencing using Illumina platforms (e.g., NovaSeq 6000) to generate 150 bp paired-end reads, achieving sufficient coverage (>50x) for high-quality assembly [59].

- **Genome Assembly and Annotation**: Assemble raw reads using SPAdes (v3.14.0) with "careful" mode enabled. Curate the assembly by removing contigs with low coverage (<500x) and anomalous GC content. Annotate the genome using PROKKA (v1.14.6) with *Streptomyces*-specific databases [59].
- **BGC Prediction and Analysis**: Process the assembled genome through antiSMASH 7.0 with all analysis modules enabled (KnownClusterBlast, ClusterBlast, SubClusterBlast). Manually inspect results and compile BGC predictions in a structured database. Calculate overall genome-related indices (ANI, dDDH) using the EZBiocloud and TYGS servers for taxonomic placement [54] [7] [59].

Metabolomic Profiling and Molecular Networking Protocol

- **Sample Preparation and Chromatography**: Culture strains under various conditions (e.g., different media, time points) to stimulate secondary metabolite production. Extract metabolites with appropriate organic solvents (e.g., ethyl acetate, methanol). Perform UPLC separation using reversed-phase C18 columns (e.g., Acquity UPLC BEH C18, 1.7µm) with water-acetonitrile gradients [54] [57].
- **Mass Spectrometry Analysis**: Acquire data in data-dependent acquisition (DDA) mode on high-resolution mass spectrometers (Q-TOF or Orbitrap instruments). Use positive and negative ionization modes with mass resolution ≥100,000. Set collision energies to stepped values (e.g., 20-40 eV) to generate comprehensive fragmentation patterns [57] [59].
- **Data Preprocessing for FBMN**: Convert raw data to .mzML format using ProteoWizard MSConvert. Process files through MZmine 2 for mass detection, chromatogram building, deconvolution, isotopic grouping, alignment, and gap filling. Export feature lists (.mgf) and quantitative tables (.csv) following GNPS documentation [57] [58].
- **Molecular Networking and Annotation**: Upload processed data to GNPS, selecting appropriate parameters for high-resolution data (precursor mass tolerance 0.02 Da, fragment tolerance 0.02 Da). Utilize FBMN workflow with minimum cosine score of 0.7 and minimum matched peaks of 6. For complex samples, apply IIMN to collapse ion adducts. Annotate nodes through library matching and propagate annotations to structurally related nodes in the network [56] [58].

The Scientist's Toolkit: Essential Research Reagents and Materials

Table 3: Essential Research Reagents and Computational Tools for Integrated Approaches

Category	Specific Tool/Reagent	Function/Purpose	Implementation Example
DNA/Genomics	Wizard Genomic DNA Purification Kit	High-quality DNA extraction for sequencing	*Streptomyces* sp. RO-S4 genome sequencing [59]
	Illumina NovaSeq 6000	High-throughput genome sequencing	Generating 150 bp paired-end reads for assembly
	antiSMASH 7.0	BGC prediction and annotation	Identifying 42 BGCs in *Streptomyces* sp. B1866 [54]
	BiG-SCAPE	BGC clustering into Gene Cluster Families	Analyzing vibrioferrin BGC diversity [7]
Metabolomics	Acquity UPLC BEH C18 Column	Metabolite separation prior to MS analysis	Chromatographic separation of angucyclines [59]
	Q-TOF or Orbitrap Mass Spectrometer	High-resolution MS and MS/MS data acquisition	Metabolic profiling of *Streptomyces* sp. B1866 [54]
	MZmine 2	LC-MS data preprocessing for feature detection	Creating aligned feature lists for FBMN [57]
Data Integration	GNPS Platform	Molecular networking and spectral library matching	Annotating molecular families in RO-S4 extract [59]
	Cytoscape	Advanced network visualization and exploration	Visualizing BGC similarity networks [7]
	MetGem	t-SNE-based network visualization	Complementary visualization to GNPS [55]

Case Studies and Applications

The integrated approach has demonstrated remarkable success in discovering novel bioactive compounds from diverse microbial sources.

Mangrove-Derived Streptomyces Discovery

In a seminal study, *Streptomyces* sp. B1866 isolated from mangrove sediments was identified as a novel species through comprehensive genomic analysis (16S rRNA similarity <97%, ANI 80.1-80.6%, dDDH 24.1-24.7%). Genome mining revealed 42 BGCs, with over half showing low similarity (<70%) to characterized BGCs. Concurrent UPLC-MS/MS-based molecular networking detected several unannotated nodes, guiding the isolation of a previously undescribed benzoxazole compound, streptoxazole A, which exhibited potent anti-inflammatory activity (IC~50~ 38.4 μM against LPS-induced NO production) [54]. This case exemplifies how integration efficiently bridges genomic potential with novel chemical discovery.

Angucycline Discovery from Marine Streptomyces

Analysis of *Streptomyces* sp. RO-S4, isolated from polluted seawater in Algeria, demonstrated the power of integration for antibiotic discovery. The strain showed promising activity against methicillin-resistant *Staphylococcus aureus* (MRSA). Molecular networking annotated the predominant compounds as angucycline-type molecules, most with fridamycin-like aglycones and novel structural features. Genomic sequencing revealed 19 BGCs, including a grincamycin-like BGC responsible for angucycline production. The combined data enabled researchers to propose novel biosynthetic hypotheses for ring-opening and lactonization reactions observed in the metabolites [59].

Siderophore Diversity in Marine Bacteria

A large-scale genomic analysis of 199 marine bacterial genomes identified 29 distinct BGC types, with NRPS, betalactone, and NI-siderophores being predominant. Focused analysis of vibrioferrin BGCs across *Vibrio harveyi*, *Vibrio alginolyticus*, and *Photobacterium damselae* revealed high genetic variability in accessory genes while core biosynthetic genes remained conserved. Clustering analysis showed that vibrioferrin BGCs formed 12 families at 10% similarity but merged into a single gene cluster family at 30% similarity, highlighting the genetic plasticity of these iron-chelating compounds [7].

The strategic integration of genomic mining with MS-based molecular networking represents a paradigm shift in natural product discovery, effectively addressing the historical challenge of connecting biosynthetic potential with chemical output. This synergistic approach enables researchers to prioritize novel BGCs, rapidly annotate metabolite families, and guide the targeted isolation of promising bioactive compounds. As both genomic and metabolomic technologies continue to advance—with improvements in sequencing sensitivity, mass resolution, and computational analytics—this integrated framework will undoubtedly remain central to unlocking the vast untapped chemical diversity encoded in microbial genomes for drug discovery and development.

The discovery of microbial natural products has long been a cornerstone of therapeutic development, yielding clinically essential medicines including penicillin, streptomycin, and doxorubicin [60]. Historically, this process relied on bioactivity-guided isolation, an approach that has increasingly led to the rediscovery of known compounds, creating the impression that nature's medicinal reservoir was largely exhausted [60]. The advent of widespread genome sequencing revealed this to be a fundamental misconception, demonstrating that microbes possess far greater biosynthetic potential than previously recognized [61].

This genomic revolution unveiled that natural product biosynthetic genes are typically organized in **biosynthetic gene clusters** (BGCs) [60]. Sequencing of actinomycetes and filamentous fungi has shown their genomes often contain dozens of these clusters, with the vast majority representing uncharacterized or "orphan" BGCs—those not yet linked to their metabolic products [60] [61]. This disparity between genomic potential and discovered metabolites has catalyzed the development of targeted strategies that prioritize BGC identification and characterization, moving from traditional activity-based screening to genome-informed discovery [4]. This article examines key methodological frameworks and presents case studies demonstrating successful natural product discovery through targeted BGC identification.

Foundational Methodologies in BGC Identification and Analysis

Computational Prediction of Biosynthetic Gene Clusters

The initial critical step in modern natural product discovery is the comprehensive identification of BGCs within genomic data. Computational tools have become indispensable for this purpose, with **antiSMASH** (antibiotics and Secondary Metabolite Analysis SHell) representing the most widely used platform [4] [7]. antiSMASH enables automated identification of BGCs across diverse classes, including polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS), ribosomally synthesized and post-translationally modified peptides (RiPPs), and terpenes [7]. Other notable tools include **PRISM**, which specializes in predicting chemical structures from genomic data, and **ClustScan**, used for detailed analysis of PKS and NRPS assembly lines [4].

A significant limitation of rule-based tools is their reduced effectiveness for novel or "cryptic" BGCs that diverge from known paradigms [4]. This challenge is being addressed by emerging **machine learning and deep learning approaches** (e.g., DeepBGC, NeuRiPP) that can identify BGCs based on statistical patterns rather than predefined rules, offering enhanced potential for discovering novel structural classes [4].

Comparative Genomics and BGC Network Analysis

Following identification, BGCs are analyzed through comparative genomics to map their diversity and evolutionary relationships. The **BiG-SCAPE** (Biosynthetic Gene Similarity Clustering and Prospecting Engine) tool facilitates large-scale analysis by generating sequence similarity networks and

grouping BGCs into **Gene Cluster Families** (GCFs) based on domain architecture and sequence similarity [62] [7]. This clustering operates on the principle that BGCs encoding similar natural products will share significant sequence and organizational homology.

Complementing BiG-SCAPE, **CORASON** (CORe Analysis of Syntenic Orthologues to prioritize Natural product gene clusters) employs a phylogenomic approach to elucidate evolutionary relationships within and across GCFs, providing high-resolution insights into BGC ancestry and diversification [62]. This combined workflow allows researchers to prioritize BGCs based on novelty and phylogenetic distribution, efficiently focusing discovery efforts.

Table 1: Key Bioinformatics Tools for BGC Identification and Analysis

Tool Name	Primary Function	Key Features
antiSMASH	BGC Identification & Prediction	Identifies known BGC classes; predicts core biosynthetic machinery [4] [7]
PRISM	BGC Identification & Chemical Prediction	Predicts chemical structures of secondary metabolites [4]
BiG-SCAPE	BGC Comparison & Networking	Groups BGCs into Gene Cluster Families (GCFs); visualizes relationships [62]
CORASON	Phylogenomic Analysis	Elucidates evolutionary relationships between BGCs [62]
MIBiG	Reference Database	Curated repository of experimentally characterized BGCs [61]

Experimental Activation and Heterologous Expression

A major challenge in BGC-based discovery is that many clusters are "silent" or "cryptic," not expressed under standard laboratory conditions [60]. Multiple strategies have been developed to address this:

- **Heterologous Expression**: BGCs are cloned and expressed in a model host organism engineered for high production yields. This bypasses native regulatory constraints and allows access to cluster-encoded metabolites [60] [27].
- **FAC-MS** (Fungal Artificial Chromosomes and Metabolomic Scoring): This platform involves cloning intact fungal BGCs into an engineered *Aspergillus nidulans* host followed by metabolomic analysis to identify novel metabolites [60].
- **Reporter-Guided Screening**: Engineered reporter strains identify clones containing functional BGCs. For example, a *Streptomyces albus* strain lacking its native phosphopantetheinyl transferase (PPTase) but containing the blue pigment synthase A gene (*bpsA*) produces blue pigment (indigoidine) only when a clone expresses a functional PPTase—a gene commonly found in NRPS and PKS clusters [27].

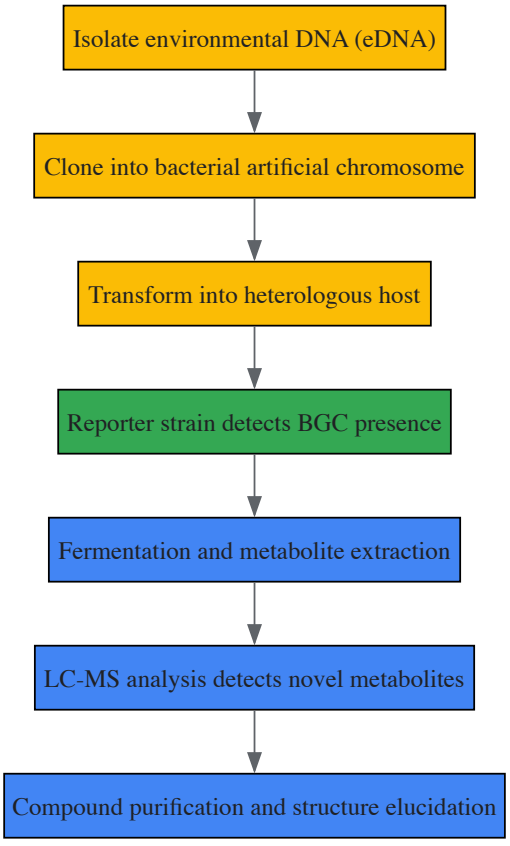


Diagram 1: Functional Metagenomics Workflow for BGC Discovery.

Case Studies in Targeted BGC Discovery

Metabologenomics: Discovery of Novel Analogs from Actinobacteria

Background and Methodology: The Kelleher Research Group, in collaboration with the Metcalf and Thomson labs, developed a **metabologenomics** platform that integrates genomic and metabolomic data from hundreds of actinobacterial strains [60]. This approach involves:

- Sequencing and annotating genomes to identify Biosynthetic Gene Cluster Families (GCFs).
- Analyzing metabolomic profiles using high-resolution LC-MS.
- Calculating correlation scores between GCF presence/absence and metabolite detection across the strain library to link specific natural products to their BGCs [60].

Outcomes and Significance: Metabologenomics enables the simultaneous discovery of novel chemical structures and their associated biosynthetic pathways at a large scale [60]. This data-rich approach allows researchers to mine for new members of known structural classes or identify entirely new structural classes, significantly accelerating the early stages of pharmaceutical discovery [60].

FAC-MS: Accessing Cryptic Fungal Metabolites

Background and Methodology: Filamentous fungi possess extensive biosynthetic potential, with genomes containing 50-80 BGCs, most of which are silent under normal conditions [60]. The FAC-MS platform addresses this challenge through:

- Cloning silent BGCs into Fungal Artificial Chromosomes (FACs).
- Heterologous expression in *Aspergillus nidulans*.
- Metabolomic scoring via LC-MS to identify novel metabolites [60].

Outcomes and Significance: This platform successfully led to the discovery of **valactamide A**, a novel lipopeptide macrolactam, and elucidated the biosynthetic pathway for the orphan benzodiazepine **benzomalvin A/D** [60]. FAC-MS provides a generalizable method to unlock the vast, untapped chemical diversity encoded in fungal genomes.

Reporter-Guided Discovery of Myxochelin from Soil eDNA

Background and Methodology: Functional metagenomics allows access to BGCs from the >99% of environmental bacteria that are unculturable [27]. A key study used a reporter-guided strategy:

- Construction of an environmental DNA (eDNA) cosmid library from Texas soil.
- Cloning into a engineered *Streptomyces albus::bpsA ΔPPTase* reporter strain.
- Screening for clones producing blue pigment (indigoidine), indicating PPTase activity and presence of a functional BGC [27].

Outcomes and Significance: This screen identified clones containing NRPS, PKS, and hybrid BGCs [27]. One NRPS cluster was confirmed to confer production of **myxochelin A**, a known siderophore, demonstrating the method's efficacy in pinpointing functional BGCs within complex metagenomic libraries [27].

BiG-SCAPE/CORASON Workflow: Expansion of the Detoxin/Rimosamide Family

Background and Methodology: The integrated BiG-SCAPE and CORASON workflow was used for large-scale analysis of BGC diversity [62]. This involved:

- Using BiG-SCAPE to generate sequence similarity networks and group BGCs into GCFs.
- Applying CORASON to perform phylogenomic analysis of targeted GCFs, specifically those related to detoxin/rimosamide BGCs.
- Identifying distinct phylogenetic clades representing novel BGC families [62].

Outcomes and Significance: This comprehensive mapping of biosynthetic diversity led to the characterization of **seven novel detoxin/rimosamide analogues** [62]. The study validated the metabologenomic concept by demonstrating strong correlation between GCFs and mass spectrometry features across 363 actinobacterial strains, proving that computational grouping of BGCs effectively predicts shared chemotype [62].

Table 2: Summary of Natural Products Discovered Through Targeted BGC Identification

Case Study	Methodology	Natural Product(s) Discovered	BGC Type	Significance
Actinobacterial Metabologenomics	Genomics-LC-MS Correlation	Novel structural classes & analogs	Various (Actinobacteria)	High-throughput linking of metabolites to BGCs [60]

Case Study	Methodology	Natural Product(s) Discovered	BGC Type	Significance
Fungal FAC-MS	Heterologous Expression in *A. nidulans*	Valactamide A, Benzomalvin A/D pathway	NRPS, Hybrid	Access to silent fungal BGCs [60]
Soil Metagenomics	Reporter-Guided Screening (PPTase)	Myxochelin A	NRPS	Functional mining of uncultured bacteria [27]
Pan-genomic Mining	BiG-SCAPE & CORASON	Seven novel detoxin/rimosamide analogs	Various	Mapping BGC diversity reveals novel analogs [62]

The Scientist's Toolkit: Essential Research Reagents and Solutions

Successful execution of BGC discovery pipelines relies on specialized biological and computational reagents.

Table 3: Key Research Reagent Solutions for BGC Discovery

Reagent / Material	Function in BGC Research	Application Example
Heterologous Host Strains	Engineered model organisms for BGC expression.	*Streptomyces albus* [27], *Aspergillus nidulans* (for FAC-MS) [60]
Reporter Systems	Visual detection of BGC-containing clones.	*bpsA* (blue pigment synthase) in a ΔPPTase background [27]
Fungal Artificial Chromosomes (FACs)	Stable cloning and maintenance of large fungal BGCs.	Heterologous expression of silent fungal gene clusters [60]
BGC Databases (e.g., MIBiG, antiSMASH DB)	Reference repositories for known BGCs and their products.	Annotation, comparison, and prioritization of novel BGCs [4] [61]
Sequence Analysis Tools (e.g., antiSMASH, BiG-SCAPE)	Computational prediction and comparison of BGCs.	Initial genome mining and comparative genomics [4] [62] [7]

Targeted BGC identification has fundamentally reshaped natural product discovery, transforming it from a rediscovery-prone process to a rational, genome-guided endeavor. The case studies detailed herein—spanning metabologenomics, FAC-MS, reporter-guided metagenomics, and computational pan-genomic mining—demonstrate the power of integrating bioinformatics with experimental biology to access novel chemical diversity. These approaches successfully address historical bottlenecks, including the high rediscovery rate of known compounds and the silence of most BGCs in laboratory settings. As computational tools, particularly artificial intelligence, continue to evolve alongside synthetic biology and analytical techniques, the systematic discovery of novel natural products from microbial sources will continue to accelerate, promising a rich pipeline of future therapeutic leads and molecular probes.

Overcoming Challenges: From Silent Clusters to Efficient Heterologous Expression

Microbial secondary metabolites, a prolific source of drugs and antibiotics, are encoded by **Biosynthetic Gene Clusters (BGCs)**. However, genome sequencing has revealed a paradox: the number of BGCs far exceeds the quantity of metabolites detected under standard laboratory conditions [63]. A majority of these BGCs are "cryptic" or "silent," meaning they are not expressed or are poorly expressed, thus representing a vast hidden reservoir of potential new bioactive compounds [64] [65]. The emergence of drug resistance has intensified the critical need to access this untapped chemical diversity for novel drug discovery [63]. This guide details the primary strategies—promoter engineering, regulatory factor manipulation, and ribosome engineering—employed to activate these silent BGCs, providing a technical roadmap for researchers and drug development professionals engaged in genome mining and natural product discovery.

Strategic Approaches to BGC Activation

Promoter Engineering

Promoter engineering is a powerful and direct method for disrupting native transcriptional regulation and activating silent BGCs. This strategy involves replacing the native promoters of biosynthetic genes with constitutive or inducible promoters to ensure strong, controlled expression [65].

Key Experimental Protocols:

- **Construction of Synthetic Promoter Libraries:** A advanced approach involves completely randomizing the regulatory sequences, including both the promoter and ribosomal binding site (RBS) regions. For example, in *Streptomyces albus* J1074, researchers have fixed only

the core -10/-35 regions of the promoter and the Shine-Dalgarno (SD) sequence, randomizing all other nucleotides. The resulting library is then screened using a reporter system, such as an NRPS that produces the blue pigment indigoidine, to isolate regulatory cassettes with a wide range of transcriptional strengths [65].

- **Metagenomic Mining for Universal Promoters:** To move beyond model organisms, natural 5' regulatory sequences can be mined from a phylogenetically diverse set of microbial genomes (e.g., Actinobacteria, Archaea, Bacteroidetes). These putative promoters are cloned into species-specific vectors and systematically quantified using a GFP reporter across different bacterial species and growth conditions to identify orthogonal promoters with broad host ranges [65].
- **CRISPR-Assisted Multiplex Promoter Refactoring:** Techniques like **mCRISTAR (multiplexed CRISPR-based Transformation-Associated Recombination)** allow for the simultaneous replacement of multiple native promoters within a BGC in a single step. This method leverages yeast homologous recombination (YHR) to efficiently swap out up to eight promoters with engineered synthetic counterparts, enabling rapid activation of entire silent pathways [65].

Application Example: The activation of the cryptic **lanthomicin BGC** in *Streptomyces chattanoogensis* L10 was achieved through promoter engineering. Researchers replaced the native promoter with a strong, constitutive kasO*p promoter on a plasmid integrated into the genome, which led to the production and identification of three novel pentangular polyphenols, lanthomicins A-C [63].

Manipulation of Regulatory Factors

Many BGCs are silenced by complex native regulatory networks. Manipulating these networks, either by introducing positive regulators or deleting negative ones, can trigger the expression of silent clusters.

Key Experimental Protocols:

- **Overexpression of Pathway-Specific Regulators:** Identifying and cloning the gene encoding a pathway-specific activator regulator into a multi-copy plasmid under a strong promoter. This plasmid is then introduced into the host strain. The amplified regulator protein can bind to the promoter regions of the BGC, recruiting RNA polymerase and initiating transcription [63].
- **Deletion of Global Repressors:** Using targeted gene knockout systems, such as CRISPR-Cas9 or CRISPR-Cpf1, to disrupt genes encoding global repressor proteins that suppress secondary metabolism. This removes the repression and can lead to the derepression of multiple silent BGCs [63].
- **Exploitation of Interspecies Interactions (Co-culture):** Cultivating the target actinomycete strain in close physical proximity with an "inducer" strain. A highly effective protocol involves using **mycolic acid-containing bacteria (MACB)**, such as *Tsukamurella pulmonis* TP-B0596, as the inducer. The co-culture is established on a solid agar medium, allowing for chemical communication. The MACB produces diffusible signaling molecules (likely mycolic acids) that are sensed by the actinomycete, triggering a regulatory cascade that activates silent BGCs. This method has led to the discovery of over 42 new natural products from 16 different actinobacterial strains [64].

Ribosome Engineering

Ribosome engineering exploits mutations in ribosomal proteins or RNA polymerase to alter cellular physiology and stimulate the production of secondary metabolites. This approach often induces a "relaxed" state, mimicking the stringent response and reallocating resources toward biosynthesis.

Key Experimental Protocols:

- **Selection for Antibiotic Resistance:** The standard protocol involves exposing bacterial cells to sub-inhibitory concentrations of antibiotics that target the protein synthesis machinery, such as **streptomycin** or **rifampicin**. Spontaneous mutants that arise are selected based on their antibiotic resistance. These mutations, often in genes encoding the ribosomal protein S12 (rpsL) or the beta subunit of RNA polymerase (rpoB), can pleiotropically activate silent BGCs [63].
- **Rational Introduction of Known Mutations:** With advanced genetic tools, specific point mutations known to confer resistance (e.g., a K88E mutation in rpsL) can be introduced directly into the genome using CRISPR-based genome editing systems. This allows for precise ribosome engineering without the need for random mutagenesis and screening.

Quantitative Comparison of BGC Activation Strategies

The table below summarizes the core principles, advantages, and challenges of each activation strategy.

Table 1: Comparative Analysis of Strategies for Activating Silent Biosynthetic Gene Clusters

Strategy	Core Principle	Key Advantages	Primary Challenges
Promoter Engineering	Replacing native promoters with synthetic, strong promoters to drive constitutive expression [63] [65].	Direct and predictable; enables fine-tuning of expression levels; highly	Requires prior knowledge of cluster boundaries; can be ineffective for

Strategy	Core Principle	Key Advantages	Primary Challenges
		modular and applicable to heterologous hosts [65].	complex, multi-operon clusters; potential for host toxicity.
Regulatory Factor Manipulation	Altering the expression or function of transcriptional regulators that control BGC expression [63] [64].	Can activate multiple clusters simultaneously (if global regulator); mimics natural activation pathways [64].	Relies on identification of relevant regulators; effects can be highly pleiotropic and strain-specific.
Ribosome Engineering	Introducing mutations in ribosomal proteins or RNA polymerase to globally alter cellular physiology and metabolite production [63].	Simple and low-cost; does not require prior genetic knowledge of the BGC; can be combined with other methods.	Relies on random mutagenesis; can reduce growth rate/fitness; mechanism is indirect and not fully predictable.
Co-culture	Culturing the target organism with another inducing strain to simulate ecological interactions [64].	Culture-based; requires no genetic manipulation; effective at activating clusters responsive to ecological cues [64].	The inducing signal is often unknown; complex metabolite background; results are not always reproducible.

The Scientist's Toolkit: Essential Research Reagents

The following table details key reagents and tools essential for conducting experiments in BGC activation.

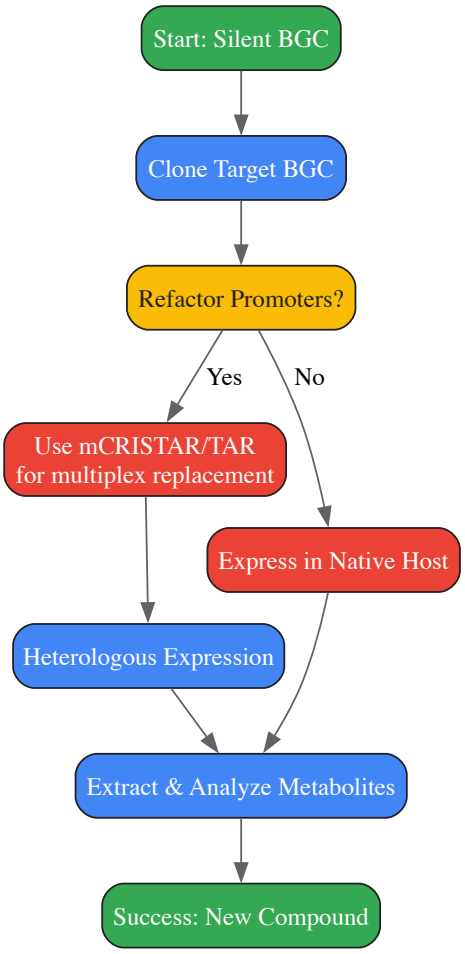
Table 2: Essential Research Reagents for BGC Activation Studies

Reagent / Tool	Function / Application	Examples / Specifications
S. albus J1074	A genetically streamlined and highly popular **heterologous host** for expressing refactored BGCs from diverse origins [65].	Model actinomycete; known for high transformation efficiency and minimal native secondary metabolite background.
CRISPR-Cpf1 / Cas9	**Genome editing systems** used for precise gene knockouts (e.g., repressors), promoter replacements, and introduction of point mutations (ribosome engineering) [63] [65].	Enables targeted genetic modifications without relying on homologous recombination efficiency of the host.
pSET152-spec-kasO*	An example of an **integrative expression vector** for promoter engineering. Carries a spectinomycin resistance marker and a strong constitutive promoter (kasO*p) [63].	Used for chromosomal integration and stable expression of genes or entire BGCs.
antiSMASH	The primary **bioinformatics tool** for in silico identification and annotation of BGCs in genomic data [9] [66].	Rule-based and machine-learning algorithms; essential for genome mining and predicting BGC class and boundaries.
Mycolic Acid-Containing Bacteria (MACB)	**Inducer strains** in co-culture experiments to activate silent BGCs in actinomycetes via interspecies signaling [64].	*Tsukamurella pulmonis* TP-B0596 is a well-characterized and highly effective inducer strain.

Visualizing the Workflows

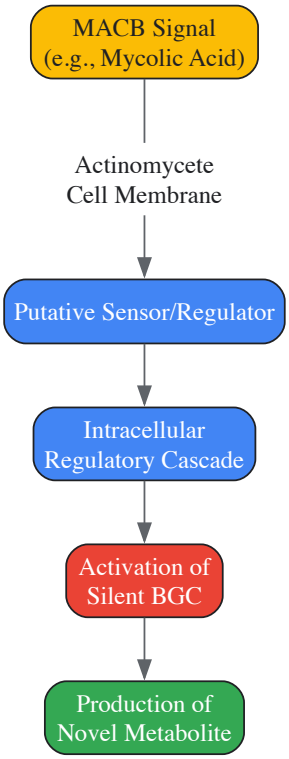
Promoter Engineering and BGC Refactoring Workflow

The following diagram illustrates the key decision points and methodological steps in a promoter engineering pipeline.



Co-culture Activation Signaling Pathway

This diagram outlines the proposed molecular signaling pathway activated during co-culture with mycolic acid-containing bacteria.



The activation of silent biosynthetic gene clusters is a cornerstone of modern natural product research. While promoter engineering, regulatory manipulation, and ribosome engineering each offer distinct paths to unlocking this chemical potential, the future lies in their **strategic integration**.

Combining these methods with co-culture and sophisticated bioinformatics prediction models like **BGC-Prophet** [67] creates a powerful, multi-pronged approach. As these tools continue to evolve, they will undoubtedly accelerate the discovery of novel therapeutic agents from the vast, untapped reservoir of microbial secondary metabolism, directly addressing the growing crisis of antimicrobial resistance.

The discovery of bioactive natural products, which are crucial for developing new antibiotics, immunosuppressants, and anti-cancer agents, hinges on accessing biosynthetic gene clusters (BGCs) within microbial genomes [68] [69]. Large-scale genomic analyses have revealed that microbes harbor a vast reservoir of uncharacterized BGCs; however, a significant majority of these are **silent (or cryptic)** and are not expressed under standard laboratory conditions [68]. This reality has spurred a renaissance in novel drug discovery, shifting the paradigm from traditional cultivation-based screening to **heterologous expression-based genome mining** [68]. In this strategy, target BGCs are directly cloned from their native genomic context and expressed in amenable surrogate hosts, thereby activating their biosynthetic pathways.

The direct cloning of large BGCs, frequently spanning tens to hundreds of kilobases with high GC content, presents a substantial technical challenge and has been the critical rate-limiting step in this discovery pipeline [68] [70] [71]. Conventional methods, such as cosmid library construction, are not only laborious but also restricted by limited cloning capacity, making them poorly suited for targeting specific, large BGCs [70]. This technical bottleneck has driven the development of sophisticated cloning strategies, including Transformation-Associated Recombination (TAR), CRISPR-Cas systems, and other direct capture methods, which are the focus of this technical guide for researchers and drug development professionals.

The core principles of advanced BGC cloning methods involve addressing three fundamental technical issues: 1) preparing high-quality genomic DNA, 2) precisely releasing the target BGC from its chromosomal location, and 3) efficiently assembling the freed fragment into a suitable vector system [68]. The following table summarizes the key characteristics of the primary methods discussed in this guide.

Table 1: Comparison of Advanced Methods for Cloning Large Biosynthetic Gene Clusters

Method	Core Principle	Typical Cloning Capacity	Key Advantages	Primary Limitations
Transformation-Associated Recombination (TAR)	*In vivo* homologous recombination in *S. cerevisiae* [70] [72]	Up to several hundred kb [72]	Cost-effective; highly precise; accessible; captures clusters with high GC content [70]	Low efficiency (0.1-2%) due to vector recircularization; requires intensive clone screening [70]
CRISPR-Cas9-Mediated Assembly	*In vitro* Cas9 cleavage of genomic DNA followed by Gibson assembly [73]	30 - 77 kb (demonstrated) [73]	Simplicity; high fidelity (~100% for <50 kb); does not require specific restriction sites [73]	Efficiency can drop for fragments >50 kb; requires *in vitro* assembly [73]
CAT-FISHING	*In vitro* Cas12a cleavage combined with BAC library principles [71]	Up to 145 kb [71]	Handles very large, high-GC fragments; efficient for complex actinobacterial DNA [71]	Method is technically complex, involving multiple steps [71]
CAPTURE	Cas12a cleavage coupled with *in vivo* Cre-*loxP* recombination in *E. coli* [70] [69]	Not specified in detail	Amenable to full automation; enables high-throughput cloning [69]	Relies on specialized robotic systems for maximum throughput [69]

Detailed Methodologies and Protocols

An Improved Transformation-Associated Recombination (TAR) Cloning Protocol

The TAR cloning method exploits the highly efficient homologous recombination system of the yeast *Saccharomyces cerevisiae* to capture large genomic segments [70] [72]. A common TAR vector contains a yeast centromere and selectable marker for propagation in yeast, a bacterial origin of replication and selectable marker for propagation in *E. coli*, and two short unique targeting sequences ("hooks") homologous to the flanks of the target BGC [72]. Recent improvements have focused on incorporating counterselectable markers to suppress background from empty vectors, thereby reducing the need for laborious screening.

Table 2: Key Research Reagents for Improved TAR Cloning

Reagent / Solution	Function / Explanation
TAR Vector (e.g., pTARa, pCAP01)	Shuttle vector for cloning in yeast and subsequent propagation in *E. coli* [70].

Reagent / Solution	Function / Explanation
Yeast Killer Toxin K1α Subunit	Counterselectable marker that disrupts membrane integrity in yeast, eliminating cells with empty vectors [70].
Homology Hooks (≥60 bp)	Short, unique DNA sequences homologous to the 5' and 3' ends of the target BGC; guide homologous recombination [72].
***S. cerevisiae* BY4742 ΔKu80**	Yeast strain with enhanced recombination efficiency, ideal for TAR cloning [70].
5-Fluoroorotic Acid (5-FOA)	Alternative counterselectable agent for systems using the *URA3* marker [70].

Experimental Workflow:

- **Vector Construction:** A TAR vector is engineered to contain two homology hooks (minimally 60 bp each) specific to the flanking regions of the target BGC. The vector is then linearized between these hooks to expose the recombinogenic ends [72].
- **Genomic DNA Preparation:** High-quality genomic DNA is isolated from the source microbe. To increase recombination efficiency, the DNA can be pre-treated with CRISPR/Cas9 to generate cleaved ends near the BGC boundaries [72].
- **Yeast Transformation:** The linearized TAR vector and the genomic DNA are co-transformed into a suitable yeast strain (e.g., *S. cerevisiae* BY4742 ΔKu80) [70].
- **Selection and Counterselection:** Transformed yeast cells are plated on media that selects for the yeast marker on the vector. The simultaneous expression of the killer toxin K1α subunit in cells that have not incorporated the BGC (i.e., carry an empty vector) leads to their death, thus enriching for correct clones [70].
- **Clone Verification:** Yeast colonies are screened, and the DNA is isolated and transferred to *E. coli* for propagation and subsequent analysis, such as restriction digestion and sequencing, to confirm the correct capture of the BGC [70].

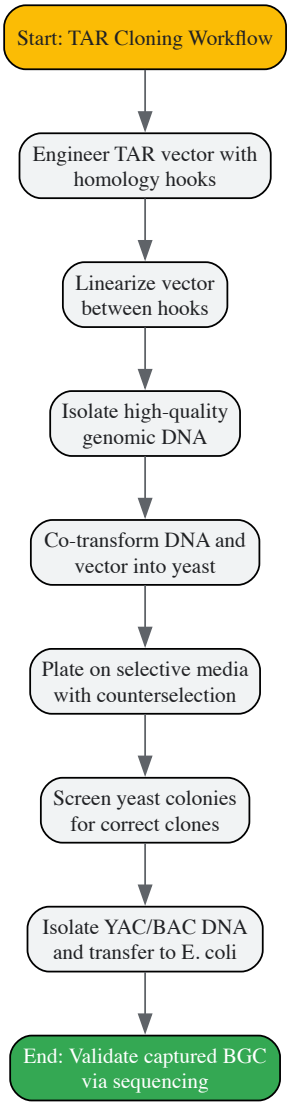


Diagram 1: TAR Cloning Workflow

CRISPR-Cas9-Mediated Large-Fragment Assembly

This method utilizes the programmability of the CRISPR-Cas9 system for the precise *in vitro* cleavage of genomic DNA, followed by the assembly of the target fragment into a vector using Gibson assembly [73].

Experimental Workflow:

- **sgRNA Design and Synthesis:** Two guide RNAs (sgRNAs) are designed to bind sequences immediately flanking the target BGC. The sgRNAs are typically synthesized *in vitro* using a T7 High Yield RNA Transcription Kit and purified [73].
- **Cas9 Protein Purification:** The Cas9 protein is expressed in *E. coli* BL21(DE3) and purified using affinity chromatography (e.g., AKTA system) to ensure high activity and purity [73].
- **Genomic DNA Digestion:** The purified genomic DNA is incubated with Cas9 protein and the two synthesized sgRNAs in an appropriate buffer (e.g., NEB buffer 3.1) at 37°C for several hours to achieve complete cleavage [73].
- **DNA Purification:** The digestion mixture is purified using phenol-chloroform-isoamyl alcohol extraction and ethanol precipitation to isolate the cleaved DNA fragment, which includes the target BGC [73].
- **Gibson Assembly:** The purified BGC fragment is mixed with a linearized vector containing homologous overhangs in a Gibson assembly master mix. This mix contains an exonuclease, a polymerase, and a ligase to seamlessly assemble the insert and vector [73].
- **Transformation and Screening:** The assembly reaction is transformed into competent *E. coli* cells. Transformants are screened via colony PCR or restriction analysis to identify those harboring the desired BGC construct [73].

CRISPR-Cas12a-Mediated Direct Cloning (CAT-FISHING)

CAT-FISHING (CRISPR/Cas12a-mediated fast direct biosynthetic gene cluster cloning) combines the unique cleavage properties of Cas12a with the high cloning capacity of bacterial artificial chromosome (BAC) systems [71]. It is particularly effective for capturing very large BGCs with high GC content from actinomycetes.

Experimental Workflow:

- **Capture Plasmid Construction:** A BAC-based capture plasmid (e.g., pBAC2015) is engineered to contain the *lacZ* gene and two homology arms, each containing a protospacer adjacent motif (PAM) site recognizable by Cas12a. The plasmid is linearized via PCR [71].
- ***In Vitro* Cas12a Digestion:** The linearized capture plasmid and high-molecular-weight genomic DNA, embedded in low-melting-point agarose plugs to protect the large DNA from shear, are separately digested with Cas12a (Cpf1) and crRNAs targeting the PAM sites flanking the BGC. Cas12a creates staggered ends, facilitating subsequent ligation [71].
- **Ligation and Transformation:** The digested genomic DNA and the linearized capture vector are ligated *in vitro* using T4 DNA ligase. The ligation mixture is then transformed directly into *E. coli* [71].
- **Clone Selection and Analysis:** Transformed *E. coli* cells are plated on media containing X-Gal and the appropriate antibiotic. White colonies (indicating successful disruption of *lacZ* by insert DNA) are selected for further analysis. BAC DNA is isolated and validated by restriction analysis and sequencing [71].

Integration with Discovery Pipelines and Concluding Perspectives

The true value of these advanced cloning techniques is realized when they are integrated into a streamlined, automated discovery pipeline. For instance, the **FAST-NPS** platform combines the **CAPTURE** cloning method with bioinformatic prioritization of BGCs using self-resistance genes as markers for bioactivity [69]. This workflow, when executed on an automated foundry like the iBioFAB, can process hundreds of BGCs in parallel, dramatically accelerating the pace of discovery from a previously manual and tedious process [69].

Furthermore, the application of these methods is expanding beyond isolated microbes to the vast genetic reservoir of uncultured organisms via metagenomics. Recent advances in long-read sequencing, coupled with optimized soil DNA extraction that yields megabase-sized assemblies, now allow for the recovery of near-complete bacterial genomes directly from environmental samples [74]. BGCs identified within these large contiguous assemblies can be chemically synthesized *de novo* (a synBNP approach) or captured using the methods described herein, opening up a new frontier for discovering bioactive molecules with rare modes of action [74].

In conclusion, the development of TAR, CRISPR-Cas9, and other direct capture methods has effectively addressed the critical bottleneck of cloning large BGCs. These strategies provide researchers and drug developers with a powerful and versatile toolkit to access the immense chemical diversity encoded in microbial genomes, both cultured and uncultured. The continued refinement of these protocols, particularly through automation and integration with predictive bioinformatics, promises to further accelerate the discovery of novel bioactive natural products for therapeutic and industrial applications.

The identification of biosynthetic gene clusters (BGCs) through genome mining has revealed a vast reservoir of potential natural products (NPs). However, a significant bottleneck persists in translating these genetic blueprints into characterized compounds, as many BGCs remain silent or are poorly expressed in their native hosts under laboratory conditions. Heterologous expression has emerged as a pivotal strategy to overcome these challenges, enabling the activation of cryptic pathways, yield optimization of valuable compounds, and elucidation of biosynthetic mechanisms. This

technical guide provides an in-depth analysis of the three primary microbial chassis systems—*Streptomyces*, *Escherichia coli*, and fungal hosts—framed within the context of BGC identification and characterization research. By synthesizing current methodologies and experimental protocols, this review serves as a comprehensive resource for researchers and drug development professionals engaged in natural product discovery.

Host Organism Profiles and Selection Criteria

1Streptomycesas a Specialized Actinobacterial Chassis

Streptomyces species, Gram-positive bacteria with high GC content genomes, represent the most versatile and widely adopted hosts for heterologous expression of bacterial BGCs, particularly those from actinomycetes [75] [76]. Their natural proficiency in producing diverse secondary metabolites and specialized enzymes makes them ideally suited for expressing complex biosynthetic pathways. A quantitative analysis of over 450 peer-reviewed studies published between 2004 and 2024 confirms *Streptomyces* as the predominant host for heterologous BGC expression, with a clear upward trajectory in application over the past two decades [76].

Advantages: The principal strengths of *Streptomyces* hosts include their genomic compatibility with high-GC BGCs, reducing the need for extensive codon optimization; inherent metabolic capacity for supplying specialized precursors; sophisticated protein secretion systems that facilitate disulfide bond formation and correct folding; presence of native chaperones and post-translational modification enzymes; and self-resistance mechanisms against toxic antimicrobial compounds [75] [77]. The absence of lipopolysaccharides (LPS) simplifies downstream purification processes for therapeutic applications [75]. *Streptomyces lividans* is particularly valued for its low restriction enzyme and proteolytic activities, enabling higher recombinant protein yields [75].

Limitations: Challenges include relatively slow growth rates compared to other bacterial hosts, complex developmental cycles that can complicate fermentation, and less established genetic tools compared to model organisms like *E. coli* [75]. Genetic manipulation often requires specialized techniques such as intergeneric conjugation, and the high endogenous production of proteases in some strains can degrade recombinant proteins [75] [76].

2Escherichia colias a Versatile Model Host

E. coli remains the most extensively studied and utilized prokaryotic host for heterologous expression due to its rapid growth, well-characterized genetics, and extensive molecular biology toolkit [75] [78]. While historically considered suboptimal for expressing large, GC-rich actinobacterial BGCs, recent engineering advances have expanded its utility in natural product research.

Advantages: *E. coli* offers unparalleled advantages in genetic manipulation efficiency, with high transformation rates and numerous available cloning systems. Its fast growth enables rapid strain engineering and screening cycles. The availability of well-characterized inducible expression systems (e.g., T7, lac, araBAD) allows precise temporal control over gene expression [78]. Red recombinase systems enable efficient modification of BGCs using short homology arms (50 bp) in *E. coli*, facilitating pathway refactoring prior to transfer into production hosts [77].

Limitations: The reducing cytoplasmic environment of *E. coli* impedes disulfide bond formation, potentially leading to misfolding of eukaryotic proteins or complex bacterial enzymes [75]. *E. coli* often lacks the specialized precursor pools, post-translational modification machinery (e.g., phosphopantetheinyl transferases for activating non-ribosomal peptide synthetases), and self-resistance mechanisms needed for producing bioactive natural products [75] [78]. Expression of large, GC-rich BGCs from actinomycetes frequently requires extensive codon optimization and refactoring [75].

Fungal Systems for Eukaryotic BGC Expression

Fungal hosts, particularly *Aspergillus nidulans* and *Saccharomyces cerevisiae*, provide specialized platforms for expressing eukaryotic BGCs, offering cellular environments and machinery more compatible with complex eukaryotic biosynthesis [79]. *S. cerevisiae* is commonly used due to established genetic tools, recombinant DNA stability, and capacity for post-translational modifications, while *A. nidulans* serves as a preferred host for filamentous fungal BGCs [75] [79].

Advantages: Fungal systems enable correct folding, modification, and compartmentalization of eukaryotic proteins, with native cytochrome P450 systems for oxidative transformations [79]. *S. cerevisiae* has been successfully employed in high-throughput platforms (HEx) for expressing diverse fungal BGCs [75]. The compartmentalized cellular architecture supports potential enzyme co-compartmentalization, as demonstrated in the biosynthesis of methylene-bridged depsides where protein-protein interactions appear crucial for intermediate channeling [79].

Limitations: Genetic manipulation can be more challenging and time-consuming compared to prokaryotic systems. Growth rates are generally slower than bacterial hosts, and transformation efficiency is typically lower. Fungal hosts may possess competing secondary metabolic pathways that divert precursors or produce interfering compounds [79].

Table 1: Comparative Analysis of Heterologous Expression Hosts

Feature	Streptomyces	E. coli	Fungal Systems
Genomic GC Compatibility	High (natural fit for actinobacterial BGCs)	Low (often requires codon optimization)	Variable
Genetic Tool Availability	Moderate (specialized techniques needed)	Extensive (well-established protocols)	Moderate to good
Growth Rate	Slow (complex life cycle)	Fast (rapid biomass accumulation)	Moderate
Post-translational Modifications	Extensive (native PTMs for bacterial enzymes)	Limited (reducing cytoplasm)	Eukaryotic-specific PTMs
Precursor Availability	Broad (native SM precursor pools)	Limited for specialized metabolites	Specialized for eukaryotic metabolism
Secretion Capacity	High (efficient protein secretion)	Limited (periplasmic accumulation)	Moderate to high
Toxicity Tolerance	High (native resistance mechanisms)	Variable (often low for novel compounds)	Variable
Typical Applications	Actinobacterial PKS/NRPS clusters, secreted proteins	Soluble enzymes, refactored pathways, initial cloning	Fungal BGCs, eukaryotic proteins, P450 transformations

Table 2: Quantitative Performance Metrics for Heterologous Hosts

Parameter	Streptomyces	E. coli	Fungal Systems
Transformation Efficiency	10^4-10^6 CFU/μg (conjugation)	10^7-10^{10} CFU/μg (electroporation)	10^3-10^5 CFU/μg
Typical Fermentation Duration	5-7 days	1-2 days	3-5 days
Maximum BGC Size Demonstrated	>100 kb	~50 kb (refactored)	~80 kb
Representative Success Rates for Actinobacterial BGCs	~70% [76]	~30% (with refactoring)	Limited data
Protein Secretion Yield	100 mg/L - 1 g/L [75]	<100 mg/L (periplasmic)	Variable

Engineering Strategies for Optimized Chassis Strains

1StreptomycesGenome Engineering

Advanced engineering approaches have been developed to enhance *Streptomyces* as heterologous hosts. The Micro-HEP (microbial heterologous expression platform) exemplifies modern chassis development, featuring *S. coelicolor* A3(2)-2023 with four deleted endogenous BGCs to reduce metabolic competition and background metabolites [77]. This platform incorporates multiple recombinase-mediated cassette exchange (RMCE) sites (Cre-lox, Vika-vox, Dre-rox, and phiBT1-attP) for orthogonal integration of heterologous BGCs, avoiding plasmid backbone integration and enabling stable multi-copy integration [77].

Precursor Enhancement: Engineering precursor supply has proven effective for increasing titers of specific natural product classes. Strategies include overexpressing key enzymes in primary metabolic pathways, deleting competing pathways, and introducing heterologous enzymes to enhance cofactor availability [75] [76].

Secretory Pathway Optimization: To enhance protein secretion, engineers have manipulated signal peptides, deleted extracellular proteases, and overexpressed chaperones to improve folding efficiency [75].

2E. coliStrain Engineering

E. coli chassis have been engineered to address limitations in expressing complex natural product pathways. Key modifications include:

Cofactor and Precursor Supplementation: Introduction of heterologous pathways for malonyl-CoA, methylmalonyl-CoA, and other specialized precursors enables polyketide production [78]. Co-expression of phosphopantetheinyl transferases (e.g., Sfp from *B. subtilis*) activates carrier protein domains in non-ribosomal peptide synthetases [78].

Folding and Solubility Enhancement: Expression of molecular chaperones, use of fusion tags, and cytoplasmic redox engineering (e.g., expression of sulfhydryl oxidases or disruption of reductase pathways) improve folding of complex proteins [75] [78].

Toxicity Mitigation: Engineering efflux systems and inducible resistance genes protects hosts from toxic pathway intermediates and products [78].

Fungal Host Engineering

Fungal chassis have been optimized through promoter engineering, deletion of competing pathways, and enhancement of precursor pathways. In *A. nidulans*, deletion of the sterigmatocystin gene cluster reduces background metabolites [79]. Promoter replacement strategies using strong constitutive or inducible promoters help activate silent BGCs [79] [78].

Experimental Protocols and Workflows

BGC Capture and Refactoring

Transformation-Associated Recombination (TAR) Cloning: This yeast-based method captures large DNA fragments (up to 100+ kb) directly from genomic DNA [77] [76]. The protocol involves: (1) designing capture vectors with homology arms targeting BGC flanking regions, (2) co-transforming genomic DNA and linearized vector into yeast, (3) selecting for recombinant clones, and (4) verifying captured BGCs by restriction analysis and sequencing [76].

Red Recombineering in *E. coli*: The λ phage Redα/Redβ system enables precise BGC modification in *E. coli* strains such as GB2005/DH5G [77]. The two-step process involves: (1) rhamnose-induced recombinase expression to replace target sequences with selectable markers, and (2) counterselection to excise markers, achieving markerless modifications [77]. Redα provides 5'→3' exonuclease activity generating 3' overhangs, while Redβ facilitates homologous recombination [77].

Exonuclease Combined with RecET Recombination (ExoCET): This method combines RecET recombination with exonuclease treatment for direct cloning of large BGCs from genomic DNA, useful for capturing GC-rich regions that challenge other methods [77].

Intergeneric Conjugation fromE. colitoStreptomyces

Efficient BGC transfer from *E. coli* to *Streptomyces* employs conjugation protocols improved beyond traditional ET12567(pUZ8002) systems [77]. The enhanced method utilizes *E. coli* strains with stabilized repeat sequences and superior transfer efficiency [77].

Protocol: (1) Introduce BGC-containing plasmid (with oriT) into donor *E. coli*, (2) grow *E. coli* and *Streptomyces* recipient to appropriate densities, (3) mix cells on appropriate medium, (4) after conjugation, overlay with selective antibiotics, (5) incubate until exconjugants appear, (6) validate by PCR and antibiotic resistance [77].

Multi-Chassis Expression Platform

Recent advances enable simultaneous BGC expression across multiple hosts using broad-host-range vectors [78]. The pMSV series vectors utilize RSF1010 origin with constitutive (PJ23119, Ptrc, Ptac) and inducible (Prham) promoters functional in *E. coli*, *B. subtilis*, and cyanobacteria [78].

Workflow: (1) Clone BGC into pMSV vectors, (2) transform/transfer into multiple hosts, (3) evaluate expression under standardized conditions, (4) identify optimal host for scale-up [78].

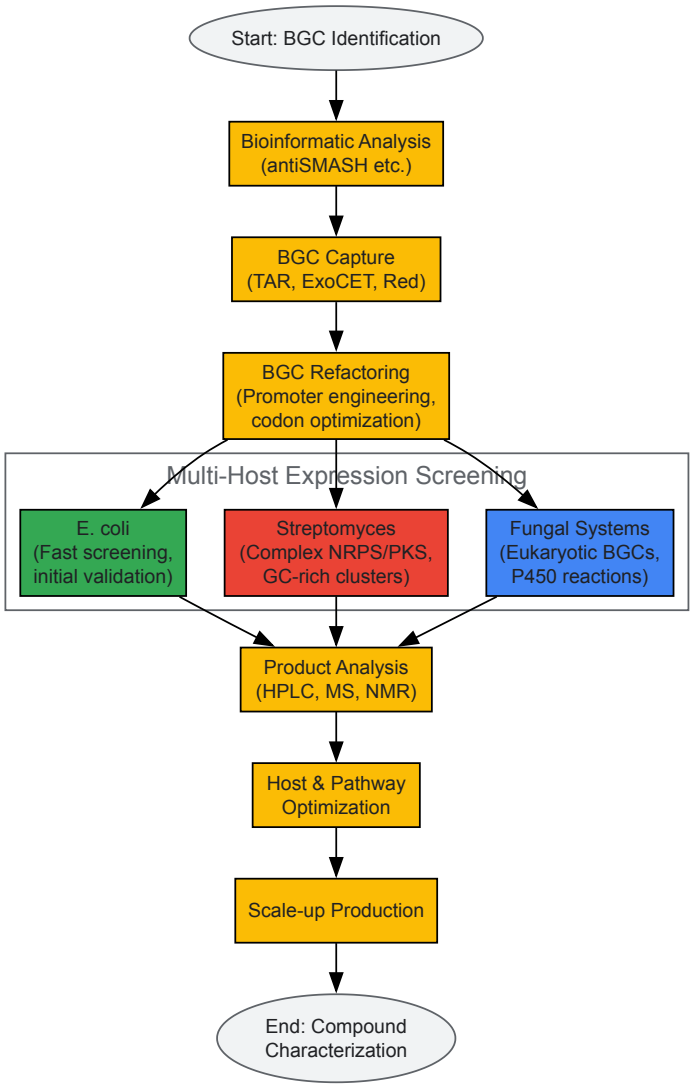


Diagram 1: Heterologous Expression Workflow for BGC Characterization. This workflow outlines the comprehensive process from BGC identification to scaled-up production, emphasizing multi-host screening strategies.

Essential Research Reagents and Tools

Table 3: Key Research Reagent Solutions for Heterologous Expression

Reagent/Tool	Function	Application Examples
pMSV Vector Series	Broad-host-range expression vectors with constitutive/inducible promoters	Multi-chassis BGC expression in *E. coli*, *B. subtilis*, cyanobacteria [78]
Micro-HEP System	Engineered *E. coli* strains for BGC modification and *S. coelicolor* chassis with RMCE sites	Efficient BGC transfer and multi-copy integration in *Streptomyces* [77]
Redα/Redβ/Redγ System	λ phage recombinases enabling efficient genetic manipulation with short homology arms	BGC refactoring and modification in *E. coli* [77]
RMCE Cassettes (Cre-lox, Vika-vox, Dre-rox)	Orthogonal recombination systems for precise genomic integration	Markerless BGC integration in *Streptomyces* [77]
Broad-Host-Range oriT Systems	Origin of transfer for conjugative plasmid mobilization from *E. coli* to actinomycetes	BGC transfer to *Streptomyces* and other actinobacteria [77]
Constitutive Promoters (ermEp, kasOp)	Strong, constitutive transcription initiation in *Streptomyces*	Driving expression of heterologous BGC genes [76]
Inducible Systems (TetR, TipA, Prham)	Chemically regulated gene expression	Controlled expression of toxic genes or pathway tuning [78] [76]

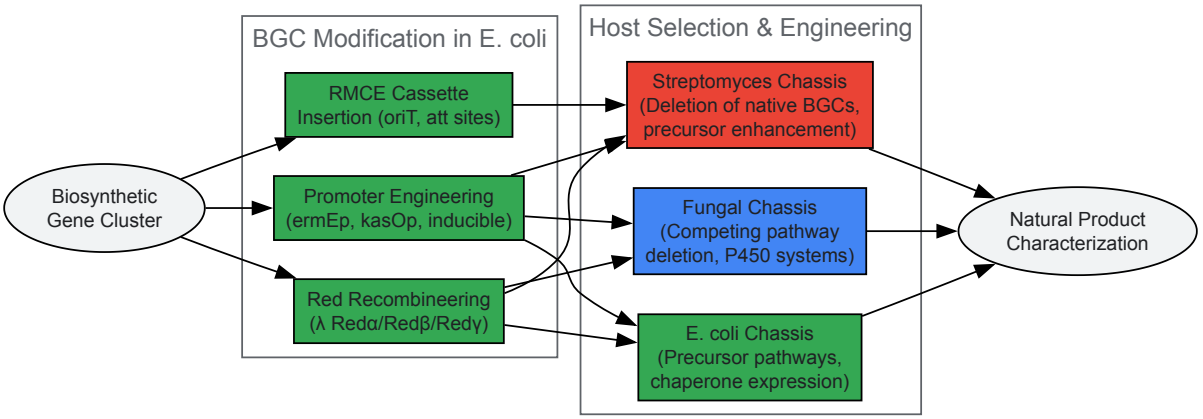


Diagram 2: BGC Engineering and Host Selection Strategy. This diagram illustrates the parallel engineering approaches for BGC modification and host optimization to enable successful heterologous expression.

Troubleshooting and Optimization Approaches

Even with carefully selected hosts and engineered BGCs, heterologous expression may fail due to various biological barriers. Systematic troubleshooting is essential for successful pathway activation.

Transcriptional/Translational Barriers: If BGC transcription is inefficient, verify promoter functionality in the chosen host and consider promoter replacement or refactoring. For translation issues, analyze codon adaptation indices and implement codon optimization for problematic genes, particularly when expressing high-GC content BGCs in low-GC hosts like *E. coli* [75] [78].

Enzyme Solubility and Function: For insoluble enzymes, consider fusion tags (MBP, GST), co-expression of chaperones (GroEL/GroES), lower expression temperatures, or targeted mutagenesis of hydrophobic regions. If enzymes are expressed but inactive, verify cofactor requirements, potential need for post-translational modifications (e.g., phosphopantetheinylation), and subcellular localization [75] [78].

Precursor Limitations: When pathway intermediates accumulate, analyze precursor availability through metabolomics or enzyme assays. Consider precursor feeding, overexpression of bottleneck enzymes, or engineering of precursor supply pathways [75] [76].

Product Toxicity: If host growth is impaired upon pathway induction, implement inducible expression systems to delay production until sufficient biomass accumulates, engineer efflux systems, or introduce resistance genes. For intracellular toxicity, consider enzyme engineering to modify substrate specificity or product structure [78].

The strategic selection and engineering of heterologous expression hosts is fundamental to advancing biosynthetic gene cluster research. *Streptomyces* platforms offer unparalleled compatibility for actinobacterial BGCs, *E. coli* provides rapid prototyping capabilities, and fungal systems enable expression of eukaryotic pathways. The emerging trend toward multi-chassis screening maximizes the probability of successful BGC expression and natural product discovery.

Future developments will likely focus on creating increasingly specialized chassis with expanded precursor pools, enhanced folding capacity, and improved stress tolerance. Machine learning approaches promise to improve prediction of optimal host-BGC pairings, while synthetic biology tools will enable more sophisticated pathway control strategies. As these technologies mature, heterologous expression will continue to be a cornerstone methodology for unlocking the vast chemical diversity encoded in microbial genomes, fueling drug discovery and biotechnology innovation.

Biosynthetic Gene Clusters (BGCs) represent genomic repositories encoding the synthetic machinery for secondary metabolites, which have been a foundational source of therapeutic agents. Traditional methods for BGC characterization face significant throughput limitations, requiring individual handling of strains and clusters that renders comprehensive discovery campaigns prohibitively resource-intensive [80]. The emerging paradigm of multiplexed platforms addresses these constraints by leveraging parallel processing, automation, and computational integration to systematically access cryptic biosynthetic diversity. This technical guide examines integrated experimental frameworks that transform BGC discovery from a piecemeal process to a high-throughput pipeline, enabling researchers to navigate the vast landscape of microbial secondary metabolism with unprecedented efficiency.

Core Multiplexed Platform Architectures

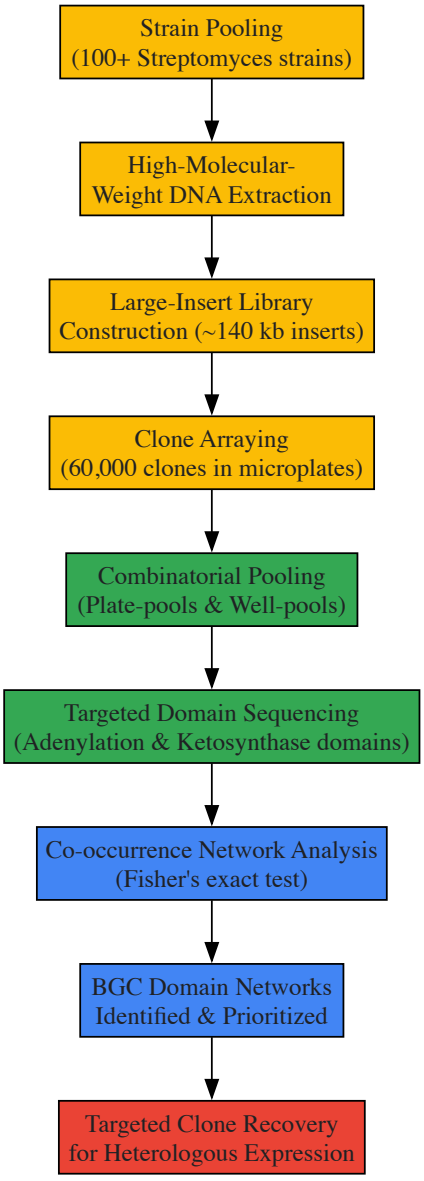
Parallelized BGC Capture and Localization

The CONKAT-seq (co-occurrence network analysis of targeted sequences) platform establishes an efficient workflow for capturing, identifying, and prioritizing numerous BGCs simultaneously from microbial strain collections. This approach transforms the laborious process of individual BGC cloning into a scalable sequencing problem [80].

Table 1: Key Components of the CONKAT-seq Platform

Component	Description	Function
Multi-Genome Library	Single large-insert library from pooled strain genomes (~140 kb average insert)	Simultaneously captures BGCs from hundreds of strains in one cloning effort
PAC Shuttle Vector	*E. coli*-Streptomyces* shuttle vector with integration elements	Enables cloning in *E. coli* and transfer to actinobacterial expression hosts
Pooling Strategy	Plate-pools and well-pools across microtiter plates	Compresses library for efficient screening via combinatorial indexing
Domain-Targeted Amplification	Degenerate primers for conserved adenylation (A) and ketosynthase (KS) domains	Amplifies diagnostic regions from NRPS and PKS BGCs for sequencing
Co-occurrence Analysis	Statistical detection (Fisher's exact test) of domain co-localization	Identifies domains belonging to the same BGC through distribution patterns

The methodology begins with pooling mycelia from hundreds of *Streptomyces* strains, typically 100 or more, followed by extraction of high-molecular-weight DNA and construction of a large-insert library in a shuttle vector [80]. The resulting library, consisting of tens of thousands of clones, is arrayed in microtiter plates. To localize BGCs within this complex library, CONKAT-seq employs a pooling strategy where clones are combined into plate-pools (all clones from the same plate) and well-pools (clones from the same well position across different plates) [80]. This creates a barcoded system that allows triangulation of specific clone positions through targeted sequencing of conserved biosynthetic domains.



High-Throughput Expression Screening Platforms

Once BGCs are captured and prioritized, multiplexed expression screening platforms enable the systematic activation and analysis of their metabolic products. Two complementary approaches have emerged: multi-host heterologous expression and high-throughput elicitor screening.

Table 2: Multiplexed Expression Screening Platforms

Platform	Core Principle	Throughput Advantage	Detection Method
Multi-Host Heterologous Expression	Parallel expression in multiple optimized hosts	24% of cryptic BGCs produced detectable compounds across hosts	Liquid chromatography-mass spectrometry (LC-MS)
High-Throughput Elicitor Screening (HiTES)	Chemical induction using library of 320+ elicitors	5+ novel metabolites identified from "drained" strains	UPLC-Qtof-MS with MetEx analysis software
Agar-Based HiTES	Solid-phase cultivation mimicking natural habitats	12-15-fold induction of cryptic metabolites	3D metabolite mapping with differential analysis

The multi-host heterologous expression approach involves transferring mobilized BGCs into well-characterized expression hosts such as *Streptomyces albus* J1074 and *Streptomyces lividans* RedStrep 1.7, which have demonstrated superior capabilities for activating cryptic clusters [80]. Following fermentation, cultures are extracted and analyzed by liquid chromatography-mass spectrometry, with BGC-specific features identified by comparing the chemical profile of each strain against all others in the series [80]. This comparative analysis readily highlights unique mass features attributable to the introduced BGC.

The HiTES platform employs a different strategy, challenging bacteria with hundreds to thousands of exogenous elicitors in a high-throughput format and monitoring natural product synthesis through mass spectrometry [81]. The recent adaptation of HiTES to agar-based cultivation more accurately recapitulates native microbial habitats and has proven particularly effective for accessing metabolites not produced in liquid culture [81]. The platform utilizes robotic liquid handling to dispense media into microtiter plates, followed by addition of elicitor libraries and bacterial inoculation in low-percentage agar. After incubation, the entire well content is extracted with methanol and analyzed by UPLC-Qtof-MS coupled with the MetEx software, which generates a three-dimensional map of the secondary metabolome as a function of the elicitor library [81].

Quantitative Performance Metrics

Multiplexed BGC platforms demonstrate significant advantages in throughput and efficiency compared to traditional approaches. Quantitative assessments reveal their capacity to substantially accelerate natural product discovery.

Table 3: Performance Metrics of Multiplexed BGC Platforms

Metric	CONKAT-seq Performance	HiTES Performance	Traditional Methods
BGC Recovery Rate	72% of NRPS/PKS BGCs detected from source collection	N/A (strain-specific)	Typically individual BGCs
Expression Success	24% of uncharacterized BGCs produced detectable compounds	5+ novel metabolites from pre-studied strains	Variable, often <5% for cryptic BGCs
Host Dependence	14 BGCs unique to S. albus, 2 unique to S. lividans, 9 in both	Compound production exclusively in solid-phase	Host optimization requires sequential testing
Novel Compound Rate	Multiple uncharacterized structural families (prolinolexin, cinnamexin, conkatamycin)	Burkethyl A and B with unusual m-ethylbenzoyl functionality	High rediscovery rates
Platform Scalability	70 BGCs interrogated in parallel	320+ elicitors tested simultaneously	Limited by individual processing

The CONKAT-seq platform achieved approximately 72% recovery of nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) BGCs from a source strain collection when validated against completely sequenced genomes [80]. In expression experiments, 24% of previously uncharacterized BGCs produced differential mass spectral features in at least one heterologous host, with notable host-dependent effects: 14 BGCs were activated exclusively in *S. albus*, two exclusively in *S. lividans*, and nine in both hosts [80]. This host-specific expression pattern underscores the value of parallel expression in multiple genetic backgrounds.

HiTES demonstrates unique capabilities for activating cryptic BGCs in their native hosts. In application to extensively studied *Burkholderia* species, the platform revealed five novel cryptic metabolites, including burkethyl A and B which feature an unusual *m*-ethylbenzoyl moiety [81]. Critically, these compounds were exclusively produced in agar-based culture and undetectable in liquid fermentation, highlighting the importance of cultivation format [81]. Dose-response optimization showed 12-15-fold induction of target compounds at optimal elicitor concentrations (90 μM), enabling sufficient production for structural characterization (2 mg burkethyl A and 1 mg burkethyl B from 150 agar plates) [81].

Experimental Protocols and Methodologies

CONKAT-seq Library Construction and Screening Protocol

Multi-Genome Library Construction:

- **Strain Pooling:** Combine equal mycelial amounts from 100+ *Streptomyces* strains into a single pooled biomass sample [80].
- **DNA Extraction:** Isolate high-molecular-weight DNA using standard actinobacterial protocols, minimizing shear to preserve large fragments.
- **Vector Preparation:** Linearize PAC shuttle vector (containing *E. coli* replication origin, Streptomyces integration elements, and selection markers) with appropriate restriction enzymes.
- **Library Construction:** Perform in vitro packaging and transformation into *E. coli* host, plating on selective media. The resulting library should contain ~60,000 clones with average insert sizes of ~140 kb to ensure adequate coverage of large BGCs [80].
- **Arraying and Storage:** Pick individual colonies into 384-well microtiter plates containing growth medium with appropriate antibiotics. Store library as separate clones across 150+ microplates at -80°C with glycerol cryopreservation.

CONKAT-seq BGC Localization:

- **Library Pooling:** Create two types of pools from the arrayed library:
  - **Plate-pools:** Combine aliquots from all wells of individual plates
  - **Well-pools:** Combine aliquots from the same well position across all plates [80]
- **Targeted Amplification:** Perform PCR on all pools using barcoded degenerate primers targeting conserved biosynthetic domains (e.g., adenylation domains for NRPS, ketosynthase domains for PKS) [80].
- **Sequencing and Analysis:** Sequence amplicons using high-throughput platforms. Process data to identify domain sequences present in each pool.
- **Co-occurrence Network Construction:** Apply Fisher's exact test to identify domain pairs that significantly co-occur across pools, indicating they originate from the same physical clone and thus the same BGC [80]. Construct domain networks where connected components represent candidate BGCs.
- **Clone Recovery:** Identify well positions containing clones with complete domain networks and retrieve corresponding clones from the original library array for heterologous expression.

Agar-Based HiTES Screening Protocol

Preparation and Inoculation:

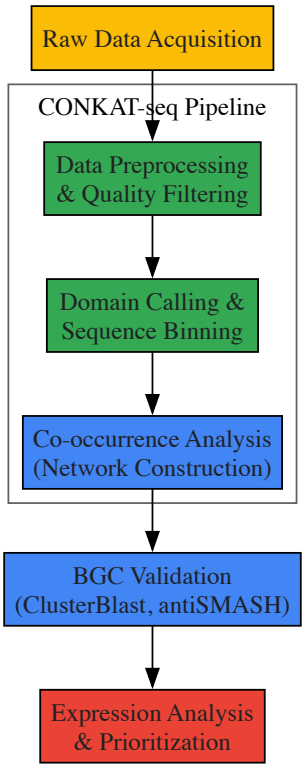
- **Media Dispensing:** Using robotic liquid handlers, dispense appropriate liquid media into all wells of 96-well microtiter plates.
- **Elicitor Library Addition:** Add 320+ candidate elicitors from compound libraries (e.g., FDA drug library) to individual wells, with DMSO-only controls [81].
- **Inoculation:** Prepare bacterial suspension in 1% agar maintained at 45°C (liquid but ready to solidify). Rapidly dispense into all wells using robotic systems and allow to solidify at room temperature, creating an even lawn of bacterial growth within and on the agar surface [81].

Incubation and Metabolite Analysis:

- **Incubation:** Incubate plates at optimal growth temperature (e.g., 30°C for *Burkholderia*) for 3-7 days to allow growth and metabolite production [81].
- **Extraction:** Add methanol to each well to extract metabolites, then filter to remove cellular debris and agar particulates.
- **Metabolite Profiling:** Analyze extracts by UPLC-Qtof-MS with automated injection. Process data using MetEx software, which:
  - Bins detected ions above a selected abundance threshold
  - Subtracts twice the average value for each bin in vehicle-treated controls
  - Displays positive values in a 3D plot showing m/z, intensity, and elicitor condition [81]
- **Hit Validation:** Select features of interest based on novelty and induction level. Scale up production using larger agar plates (e.g., 10-20 mL media) with optimized elicitor concentrations determined through dose-response testing (typically 15-120 μM range) [81].

Integrated Data Analysis Workflows

The massive datasets generated by multiplexed BGC platforms require specialized computational workflows for effective interpretation. CONKAT-seq produces complex amplicon sequencing data that must be decoded to map BGCs to their physical clone locations, while HiTES generates multidimensional metabolite profiles that demand sophisticated differential analysis.



For CONKAT-seq, the computational workflow begins with demultiplexing pooled amplicon sequencing data and assigning reads to their original plate-pool and well-pool origins [80]. Domain sequences are then identified through translation and alignment to conserved domain databases. The core analysis applies statistical tests (typically Fisher's exact test) to identify domain pairs that co-occur across pools significantly more often than expected by chance, indicating they likely reside on the same physical DNA fragment [80]. These co-occurrence patterns are used to construct BGC-specific domain networks, with each connected network representing a candidate complete BGC. Finally, networks are prioritized based on biosynthetic novelty assessed through comparison to reference databases like MIBiG, with clusters containing domains showing <80% amino acid identity to characterized proteins flagged as high-priority targets [80].

For HiTES data analysis, the MetEx software performs automated extraction of ion features from UPLC-Qtof-MS data across all experimental conditions [81]. The algorithm normalizes abundance values and identifies significantly induced features by comparing elicitor-treated samples to vehicle controls, typically applying a threshold of twice the average control abundance for hit detection [81]. The resulting three-dimensional visualization maps m/z values against elicitor conditions, enabling rapid identification of candidate cryptic metabolites specifically induced by particular elicitors. These candidates are then prioritized for purification and structure elucidation based on induction fold-change, novelty of mass, and abundance.

Essential Research Reagent Solutions

Implementation of multiplexed BGC platforms requires specific reagent systems and tools optimized for high-throughput applications.

Table 4: Essential Research Reagents for Multiplexed BGC Platforms

Reagent/Tool	Specifications	Application in BGC Platforms
PAC Shuttle Vector	~30 kb, *E. coli* and *Streptomyces* replicons, ΦC31/int/attP integration	Large-insert library construction, heterologous expression in actinomycetes [80]
Degenerate Primers	Target conserved adenylation (A) and ketosynthase (KS) domains	Amplification of diagnostic regions from NRPS and PKS BGCs for CONKAT-seq [80]
Elicitor Libraries	320+ compounds (e.g., FDA drug library), structurally diverse bioactives	Chemical induction of cryptic BGCs in HiTES screening [81]
Mass Spectrometry	UPLC-Qtof-MS systems with automated liquid handling	High-resolution metabolite profiling of thousands of samples [81]
Bioinformatics Tools	antiSMASH, MIBiG, BiG-SCAPE, MetEx software	BGC annotation, comparative analysis, and metabolomics data processing [81]

The PAC shuttle vector system is particularly critical for CONKAT-seq implementation, as it must accommodate large DNA inserts (>100 kb) while maintaining compatibility with both *E. coli* (for library construction and maintenance) and *Streptomyces* (for heterologous expression) [80]. Vectors should contain necessary elements for replication in both hosts, selection markers, and site-specific integration components for stable chromosomal insertion in expression hosts.

For HiTES implementations, the composition of the elicitor library significantly influences success rates. The FDA-approved drug library has proven particularly effective, likely because it contains structurally diverse compounds with known bioactivities that may interact with cellular regulatory networks [81]. Additionally, the low-percentage agar used in agar-based HiTES must be carefully calibrated to maintain porosity and air permeability while supporting robust microbial growth.

Multiplexed platforms for BGC mobilization and screening represent a transformative advancement in natural product discovery, fundamentally shifting the paradigm from individual cluster characterization to systematic interrogation of biosynthetic diversity. The integrated workflows of CONKAT-seq and HiTES demonstrate complementary strengths: CONKAT-seq enables extensive mining of biosynthetic diversity across strain collections, while HiTES activates cryptic pathways through environmental and chemical simulation. Together, these approaches provide researchers with powerful toolkits to navigate the vast untapped reservoir of microbial secondary metabolism, offering new avenues for discovering therapeutic agents in an era of escalating antibiotic resistance and diminishing discovery returns. As these platforms continue to evolve through integration with artificial intelligence and automation technologies, they promise to further accelerate the pace of natural product discovery and development.

The identification of biosynthetic gene clusters (BGCs) through genome mining has revealed a vast untapped reservoir of natural products with potential therapeutic applications. However, a significant bottleneck lies in translating genetic potential into detectable and quantifiable compounds. Many BGCs are poorly expressed in their native hosts under laboratory conditions, a challenge addressed through two complementary synthetic biology strategies: **biosynthetic pathway refactoring** and **chassis strain development**. Pathway refactoring involves the systematic redesign of BGCs for optimized expression and functionality, while chassis development focuses on engineering microbial hosts that provide an optimal physiological environment for heterologous expression. This technical guide examines current methodologies and platforms that integrate these approaches to facilitate efficient natural product discovery and overproduction, with particular emphasis on their application within the broader context of BGC research.

Chassis Strain Development: Building Specialized Host Platforms

Host Selection Criteria and Comparative Performance

Selecting an appropriate chassis organism is a foundational decision that significantly influences the success of heterologous expression efforts. The ideal chassis combines genetic tractability with physiological compatibility to support the biosynthesis of target compounds. While model organisms like *Escherichia coli* and *Saccharomyces cerevisiae* offer well-established genetic tools, **actinomycetes—particularly *Streptomyces* species—have emerged as preferred hosts** for expressing complex bacterial BGCs due to their native capacity for secondary metabolism, appropriate post-translational modification systems, and natural tolerance to many bioactive compounds [76].

Quantitative comparisons between potential chassis strains reveal substantial differences in performance. Recent studies directly comparing conventional *Streptomyces* chassis with specialized engineered variants demonstrate the dramatic impact of host selection on production titers (Table 1).

Table 1: Comparative Performance of Streptomyces Chassis Strains for Type II Polyketide Production

Chassis Strain	Modifications	Target Compound	Production Level	Reference
S. aureofaciens Chassis2.0	Deletion of two endogenous T2PKs gene clusters	Oxytetracycline	370% increase vs. commercial strains	[82]
S. aureofaciens Chassis2.0	Native chassis optimization	Actinorhodin	High efficiency production	[82]
S. aureofaciens Chassis2.0	Native chassis optimization	TLN-1 (pentangular T2PK)	Direct activation and production	[82]
S. coelicolor A3(2)-2023	Deletion of four endogenous BGCs + RMCE sites	Xiamenmycin	Copy number-dependent yield increase	[77]
S. albus J1074	Unmodified model strain	Oxytetracycline	No detectable production	[82]
S. lividans TK24	Unmodified model strain	Oxytetracycline	No detectable production	[82]

Industrial antibiotic producers have recently been explored as chassis candidates, leveraging their inherent metabolic capabilities. For instance, *Streptomyces aureofaciens* J1-022, a high-yield chlortetracycline producer, was engineered into Chassis2.0 through in-frame deletion of two endogenous T2PKs gene clusters, creating a pigmented-faded host with minimal precursor competition [82]. This chassis demonstrated remarkable

versatility, successfully producing tetra-ring antibiotics (oxytetracycline), tri-ring pigments (actinorhodin), and even enabling the discovery of novel pentangular polyketides (TLN-1) through direct activation of a previously unidentified BGC [82].

Genome Reduction and Metabolic Streamlining

A primary strategy in chassis development involves the elimination of endogenous BGCs to reduce metabolic burden and prevent the formation of competing products. The *Streptomyces coelicolor* A3(2)-2023 chassis exemplifies this approach, with **four native BGCs deleted** to create a clean metabolic background optimized for heterologous expression [77]. This genome reduction serves multiple purposes: it redirects metabolic flux toward heterologous pathways, simplifies the analytical process by eliminating native metabolites, and can enhance genetic stability.

Additional modifications to optimize cellular physiology include:

- **Precursor pathway enhancement**: Overexpression of key enzymes in central metabolism to increase cofactor and building block supply
- **Regulatory system engineering**: Deletion of global regulators that repress secondary metabolism or introduction of inducible regulatory elements
- **Secretion system optimization**: Engineering transport mechanisms to facilitate product excretion and reduce feedback inhibition

Biosynthetic Pathway Refactoring Strategies

Genetic Optimization and Modular Design

Pathway refactoring involves the systematic redesign of BGCs to improve their performance in heterologous hosts. This process typically includes **codon optimization** to match the host's tRNA pool, replacement of native regulatory elements with well-characterized synthetic promoters and ribosome binding sites, and reorganization of gene order to minimize potential transcriptional conflicts [76]. Refactored pathways are often designed with modularity in mind, allowing for straightforward manipulation of individual components and rapid prototyping of different configurations.

Advanced refactoring approaches include:

- **Regulatory element standardization**: Implementation of synthetic promoter/RBS libraries for fine-tuning expression levels
- **Orthogonal control systems**: Incorporation of inducible expression systems responsive to exogenous inducers not present in the production host
- **Synthetic operon design**: Rational grouping of genes based on functional relationships and expression requirements

Multi-Copy Integration and Pathway Amplification

Chromosomal integration of multiple BGC copies represents an effective strategy for enhancing production titers. The **recombinase-mediated cassette exchange (RMCE) system** enables precise integration of heterologous DNA at designated chromosomal loci without incorporating plasmid backbone sequences [77]. This technology leverages orthogonal recombination systems (Cre-lox, Vika-vox, Dre-rox, and phiBT1-attP) to facilitate simultaneous integration at multiple sites.

Recent studies demonstrate a direct correlation between BGC copy number and product yield. When the xiamenmycin BGC was integrated as two to four copies in an engineered *S. coelicolor* chassis via RMCE, researchers observed a **clear copy number-dependent increase in xiamenmycin production** [77]. This gene dosage effect highlights the importance of chromosomal integration strategies in maximizing pathway flux.

Integrated Workflows: From BGC Capture to Compound Production

The Micro-HEP Platform Workflow

Comprehensive heterologous expression platforms integrate multiple technologies into streamlined workflows. The Microbial Heterologous Expression Platform (Micro-HEP) exemplifies this integrated approach, combining efficient BGC capture, modification, and conjugation systems with optimized chassis strains (Figure 1).

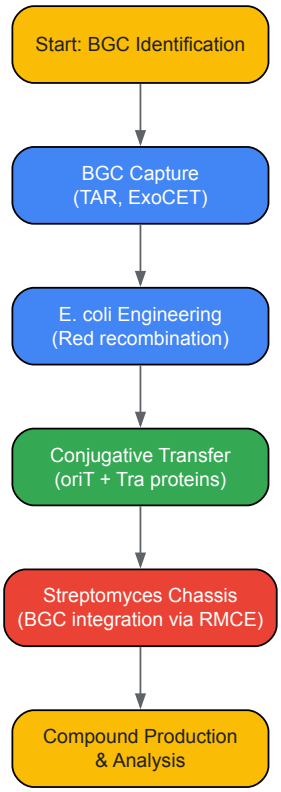


Figure 1: Integrated workflow for heterologous expression of biosynthetic gene clusters, illustrating the pathway from BGC identification to compound production.

The Micro-HEP platform employs specialized *E. coli* strains capable of both modifying and conjugatively transferring large DNA constructs. A key innovation is the implementation of a **rhamnose-inducible redαβγ recombination system** that facilitates precise insertion of RMCE cassettes into BGC-containing plasmids [77]. These cassettes incorporate the transfer origin site (oriT), integrase genes, and corresponding recombination target sites, enabling efficient conjugation and subsequent genomic integration.

Conjugative Transfer Optimization

Traditional conjugation systems using *E. coli* ET12567 (pUZ8002) face limitations including low transformation efficiency and instability with repetitive sequences. Improved conjugation systems address these challenges through:

- **Dedicated transfer strains**: Engineered *E. coli* strains with stable genomic integration of conjugation machinery
- **Optimized selection schemes**: Implementation of antibiotic resistance markers compatible with *Streptomyces* genetics
- **Enhanced transfer efficiency**: Refinement of mating conditions to improve exconjugant formation rates, particularly for large BGCs

Computational Tools Supporting Refactoring and Chassis Selection

Bioinformatics and Pathway Prediction

Computational methods have become indispensable for BGC analysis and refactoring design. **Genome mining tools** such as antiSMASH enable researchers to identify BGCs in genomic data and predict their structural outputs [4]. For poorly characterized BGCs, **retrobiosynthesis tools** like BioNavi-NP can propose plausible biosynthetic pathways from simple building blocks, achieving 72.8% accuracy in recovering reported building blocks for test compounds [83].

These computational approaches leverage:

- **Rule-based systems**: Matching BGC components to known enzymatic reactions and pathways
- **Machine learning algorithms**: Predicting novel BGC functions based on training datasets of characterized clusters
- **Comparative genomics**: Identifying conserved regulatory elements and potential bottlenecks across related pathways

Table 2: Computational Resources for Biosynthetic Pathway Design and Analysis

Resource Type	Examples	Primary Function	Application in Refactoring
BGC Databases	MIBiG, antiSMASH DB	Curated repository of known BGCs	Identification of regulatory elements and common architectures
Pathway Prediction	BioNavi-NP, RetroPathRL	Retrobiosynthetic pathway design	Proposal of optimal biosynthetic routes
Enzyme Databases	BRENDA, UniProt	Enzyme function and kinetics	Selection of optimal enzyme variants
Compound Databases	PubChem, NPAtlas	Chemical structure information	Product identification and characterization
Metabolic Models	GEMs, FBA models	Host metabolic network analysis	Prediction of metabolic bottlenecks

Design-Build-Test-Learn Cycles

The iterative DBTL framework represents the state-of-the-art approach for optimizing heterologous expression systems. This engineering paradigm involves:

- **Design**: Computational prediction of optimal refactoring strategies based on available data
- **Build**: High-throughput DNA assembly and strain construction using automated platforms
- **Test**: Parallelized small-scale fermentation and analytical screening
- **Learn**: Data analysis to inform the next design iteration, potentially incorporating machine learning

Advanced analytical techniques, including **LC-MS/MS and NMR spectroscopy**, provide critical feedback on pathway performance and compound structural validation.

Case Studies in Successful Expression Optimization

Type II Polyketide Production in Engineered Chassis

The development of *Streptomyces aureofaciens* Chassis2.0 demonstrates the power of integrated chassis development and pathway refactoring. By deleting two endogenous T2PKs gene clusters to mitigate precursor competition, researchers created a platform capable of efficiently producing diverse polyketide structures [82]. This chassis achieved a **370% increase in oxytetracycline production** compared to commercial strains, without additional metabolic engineering [82]. Furthermore, the chassis successfully produced tri-ring T2PKs like actinorhodin and flavokermesic acid, and directly activated an unidentified pentangular T2PK BGC, leading to the discovery of structurally distinct TLN-1 [82].

Multi-Copy Integration for Enhanced Titers

The application of the Micro-HEP platform to xiamenmycin production illustrates the efficacy of multi-copy integration strategies. Through RMCE-based integration of two to four copies of the xim BGC into the *S. coelicolor* A3(2)-2023 chassis, researchers demonstrated a **direct correlation between copy number and product yield** [77]. Similarly, expression of the griseorhodin (grh) BGC in this system enabled production of the architecturally complex polyketide and identification of a new derivative, griseorhodin H [77].

The Scientist's Toolkit: Essential Research Reagents

Table 3: Key Research Reagents for Biosynthetic Pathway Refactoring and Chassis Development

Reagent/System	Function	Application Example	Reference
ExoCET	Cloning of large BGCs	Direct cloning of oxytetracycline BGC	[82]
Redαβγ recombination	DNA modification in *E. coli*	Insertion of RMCE cassettes into BGC plasmids	[77]
RMCE systems (Cre-lox, Vika-vox, Dre-rox, phiBT1-attP)	Site-specific genomic integration	Multi-copy BGC integration in *Streptomyces*	[77]
antiSMASH	BGC identification and analysis	Genome mining for novel BGCs	[4]
BioNavi-NP	Retrobiosynthetic pathway prediction	Designing heterologous expression routes	[83]
Engineered *E. coli* conjugation strains	BGC transfer to actinomycetes	Improved conjugation efficiency for large BGCs	[77]

Reagent/System	Function	Application Example	Reference
S. coelicolor A3(2)-2023	Deletion chassis for heterologous expression	Expression of xiamenmycin and griseorhodin BGCs	[77]
S. aureofaciens Chassis2.0	High-yield T2PK production platform	Overproduction of diverse type II polyketides	[82]

The continued integration of synthetic biology, metabolic engineering, and computational design promises to further enhance our ability to access and optimize natural product biosynthesis. Emerging areas include:

- **Automated strain engineering**: High-throughput robotic platforms for rapid prototyping of refactored pathways
- **Machine learning-guided design**: Algorithms trained on multi-omics data to predict optimal refactoring strategies
- **Orthogonal chassis development**: Hosts engineered with minimal genomes and customized metabolic networks
- **Dynamic regulation systems**: Circuits that automatically adjust pathway expression in response to metabolic status

In conclusion, the synergistic application of biosynthetic pathway refactoring and chassis strain development has dramatically advanced our capacity to exploit the genetic potential encoded in BGCs. These technologies not only facilitate the production of known compounds at industrially relevant levels but also enable the discovery of novel chemical entities through the activation of cryptic metabolic pathways. As these platforms continue to mature, they will play an increasingly vital role in natural product-based drug discovery and development.

Connecting Genotype to Phenotype: Validation, Dereplication, and Bioactivity Assessment

The systematic linking of **biosynthetic gene clusters (BGCs)** to the chemical structures they encode represents a critical bottleneck in natural product discovery. This technical guide comprehensively outlines contemporary analytical chemistry and spectroscopic validation methodologies that bridge this gap, enabling researchers to accelerate the identification of novel bioactive compounds. By integrating genomic data with advanced mass spectrometry techniques, including **feature-based** and **correlation-based** approaches, scientists can now efficiently connect biosynthetic potential to chemical reality. This whitepaper details experimental protocols for paired omics analysis, spectroscopic validation, and computational workflows, providing a structured framework for BGC functional characterization within the broader context of biosynthetic gene cluster research for drug development professionals.

The discovery that microbial natural products are encoded by grouped biosynthetic genes—**biosynthetic gene clusters**—has revolutionized natural product research [4]. Traditional natural product discovery approaches relied on bioactivity-guided fractionation and were limited in terms of dereplication [84]. While genome sequencing has revealed a vast reservoir of BGCs, connecting these genetic blueprints to their corresponding chemical products remains a fundamental challenge [85]. This linkage is essential for understanding chemical and biological functions and represents a crucial step in modern drug discovery pipelines [86].

The disconnect between biosynthetic potential and characterized metabolites is particularly striking in human microbiome studies, where thousands of BGCs have been identified but only a fraction have been linked to their chemical products [87]. Similarly, eukaryotic algae represent a largely untapped resource for natural product discovery, with 2,762 putative BGCs recently identified across 212 genomes but few experimentally validated [11]. This guide addresses the critical analytical chemistry and spectroscopic validation strategies needed to bridge this gap, providing researchers with comprehensive methodologies for confidently linking BGCs to chemical structures.

Computational Approaches for BGC-Chemical Structure Linking

Integrative Genomic and Chemical Similarity

Reasoning that structural similarity of secondary metabolites arises from similarities in their biosynthetic genes, researchers have developed integrative approaches that leverage known BGC-secondary metabolite pairs to predict global links across compounds and BGCs in fungi and bacteria [86]. This methodology systematically interrogates metabolomes and genomes across multiple strains, detecting metabolites and proposing specific hypotheses for uncharacterized compounds [86]. The core principle involves coupling **genomic similarity** with **chemical structure-based similarity** to enable high-throughput linking of metabolites to their BGCs.

Large-Scale BGC Analysis and Classification

For large-scale analyses involving thousands of BGCs, tools such as **BiG-SCAPE** (Biosynthetic Gene Similarity Clustering and Prospecting Engine) facilitate the grouping of BGCs into **Gene Cluster Families (GCFs)** based on domain sequence similarity [62]. These GCFs can then be correlated with **Molecular Families (MFs)** identified from mass spectrometry data, enabling pattern-based genome mining and metabologenomics approaches [84] [62]. The CORASON (CORe Analysis of Syntenic Orthologues to prioritize Natural product gene clusters) tool further elucidates phylogenetic relationships within and across these families, providing high-resolution multi-locus phylogenies of BGCs [62].

Table 1: Computational Tools for BGC Analysis and Chemical Linking

Tool Name	Primary Function	Application in Chemical Linking	Reference
BiG-SCAPE	BGC similarity networks and GCF classification	Groups BGCs into families for correlation with mass spectral molecular families	[62]
CORASON	Phylogenomic analysis of BGCs	Elucidates evolutionary relationships within and across GCFs	[62]
antiSMASH	BGC identification and annotation	Predicts BGCs from genomic data for subsequent chemical correlation	[4]
GNPS	Mass spectral data analysis	Molecular networking and metabolomic analysis for chemical family identification	[84]

Mass Spectrometry-Based Linking Methodologies

Feature-Based Approaches: Peptidogenomics and Glycogenomics

Peptidogenomics leverages specific mass shifts related to peptide fragmentation into amino acids, generating sequence tags that can be linked to BGCs in the producer's genome [84] [85]. For ribosomally synthesized and post-translationally modified peptides (RiPPs), the sequence tag corresponds to an encoded precursor peptide, while for nonribosomal peptides (NRPs), it relates to modules with predictable adenylation domain specificities [84]. Automated tools like **RiPP-Quest**, **NRP-Quest**, and **MetaMiner** have been developed to streamline this process, enabling discovery of novel peptides such as informatipeptin and seven previously unknown RiPPs from GNPS datasets [84].

Glycogenomics utilizes diagnostic mass shifts or fragments from sugar moieties, particularly modified deoxysugars, to connect metabolites to their BGCs [84]. This approach successfully enabled discovery of arenimycin B from Salinispora arenicola CNB527, which was found to be more bioactive than the previously isolated arenimycin A, showing a twofold or greater increase in activity against clinically relevant, multidrug-resistant strains of *Staphylococcus aureus* [84].

Correlation-Based Approaches: Metabologenomics

Metabologenomics correlates the presence of BGCs/GCFs with mass spectral molecular families across multiple bacterial strains, operating on the principle that strains sharing a GCF should produce similar metabolites [84] [85]. This approach was validated by correlating GCFs to metabolomic data across 363 actinobacterial strains [62]. In practice, paired datasets comprising MS/MS data of culture extracts and genome sequences of their producers are analyzed to identify co-occurrence patterns, enabling simultaneous establishment of multiple compound-BGC linkages in a high-throughput workflow [84].

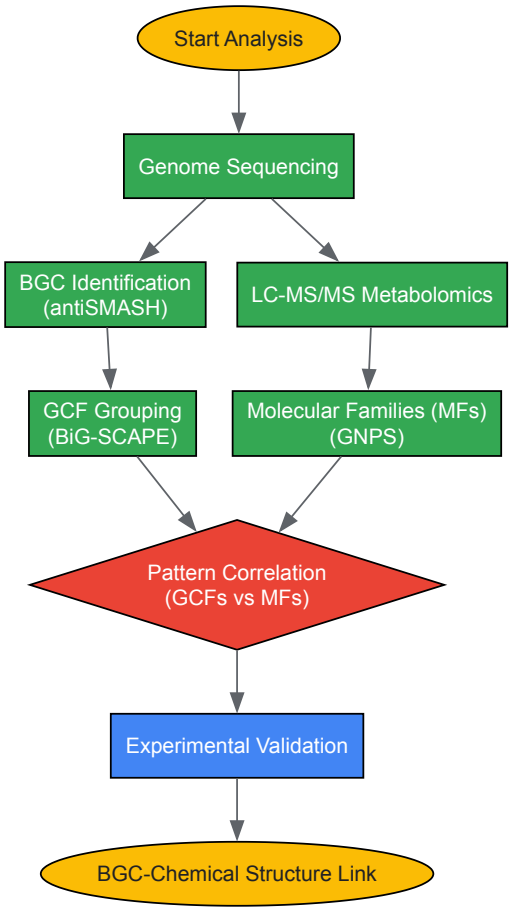


Figure 1: Integrated Workflow for Correlation-Based BGC-Chemical Structure Linking

Experimental Protocols for BGC-Chemical Structure Validation

Paired Genomic and Metabolomic Data Acquisition

Genome Sequencing and BGC Identification

- **DNA Extraction**: Use high-molecular-weight genomic DNA extraction protocols suitable for sequencing platforms [88].
- **Sequencing**: Perform whole-genome sequencing using Illumina NovaSeq or comparable platforms to generate high-quality assemblies [88] [87].
- **BGC Prediction**: Process genome sequences through antiSMASH (v7.0 or higher) with default detection settings, enabling KnownClusterBlast, ClusterBlast, SubClusterBlast, and Pfam domain annotation [7] [88].
- **GCF Classification**: Input antiSMASH results into BiG-SCAPE for sequence similarity network analysis and GCF grouping at appropriate similarity cutoffs (typically 10-30%) [62].

Metabolomic Profiling and Molecular Family Analysis

- **Culture Extraction**: Grow strains under appropriate conditions (consider OSMAC approach to elicit cryptic BGC expression) and extract metabolites using organic solvents [85] [89].
- **LC-MS/MS Analysis**: Employ reversed-phase C18 or HILIC chromatography coupled to high-resolution tandem mass spectrometry using data-dependent acquisition (DDA) or data-independent acquisition (DIA) methods [89].
- **Molecular Networking**: Process MS/MS data through Global Natural Products Social Molecular Networking (GNPS) to generate molecular families based on spectral similarity [84].

Spectroscopic Validation Techniques

Structural Elucidation of Prioritized Compounds

- **Compound Isolation**: Use bioactivity-guided or MS-guided fractionation to isolate target compounds from complex extracts [84].
- **NMR Spectroscopy**: Apply 1D and 2D NMR techniques (including COSY, HSQC, HMBC) for definitive structural characterization [89].
- **Tandem MS Analysis**: Employ CID, HCD, UVPD, or ECD fragmentation to obtain structural information, particularly for labile modifications [89].

Genetic Validation of BGC-Chemical Links

- **Gene Inactivation**: Perform targeted gene knockouts or CRISPR-Cas9-mediated disruption of core biosynthetic genes to abolish metabolite production [85].
- **Heterologous Expression**: Clone complete BGCs into suitable expression hosts and confirm compound production in the heterologous system [84].
- **Isotope Labeling**: Use 13C or 15N labeled precursors to confirm biosynthetic pathways through tracking incorporation into final metabolites [89].

The Scientist's Toolkit: Essential Research Reagents and Solutions

Table 2: Key Research Reagent Solutions for BGC-Chemical Structure Linking

Reagent/Resource	Function	Application Notes
antiSMASH	BGC identification from genomic data	Essential for initial BGC detection; regularly updated with new profile HMMs [4]
BiG-SCAPE	BGC similarity analysis and GCF grouping	Critical for large-scale BGC comparisons; uses glocal alignment for fragmented BGCs [62]
GNPS Platform	Mass spectral analysis and molecular networking	Enables MF creation and spectral library matching [84]
MIBiG Database	Repository of experimentally characterized BGCs	Reference for known BGC-chemical structure pairs [7]
High-Resolution Mass Spectrometer	Metabolite detection and characterization	Q-TOF or Orbitrap instruments provide required mass accuracy and resolution [89]

Case Studies and Applications

Fungal Secondary Metabolite Discovery

An integrative genomic and chemical similarity approach successfully linked fungal secondary metabolites to their BGCs in *Aspergillus fischeri* [86]. By systematically interrogating metabolomes and genomes of 16 strains, researchers detected 60 metabolites, assigned 22 to known BGC pairs, and proposed specific hypotheses for the remaining 38 metabolites [86]. This demonstrated that coupling genomic similarity with chemical structure-based similarity provides a straightforward and high-throughput approach for linking fungal metabolites to their BGCs.

Marine Bacterial Natural Products

In marine bacteria, coupling mass spectral and genomic information enabled the discovery of arenimycin B from *Salinispora arenicola* CNB527 through glycogenomics [84]. Additionally, pattern-based genome mining combined with comparative metabolomics across 35 *Salinispora* strains linked an uncharacterized NRPS BGC to retimycin A, a new quinomycin-like depsipeptide [85]. These examples highlight how integrated approaches can successfully connect BGCs to novel chemical structures with potent bioactivities.

Human Microbiome Metabolomics

A comprehensive analysis of 4,744 human gut microbial genomes revealed extensive BGC diversity, with *Paenibacillus* identified as a dominant genus with significant biosynthetic capacity, including potential for leinamycin synthesis [87]. This large-scale study highlighted the gut microbiome as a rich, untapped resource for novel drug discovery, particularly when combining BGC identification with metabolomic validation to establish ecological and therapeutic relevance.

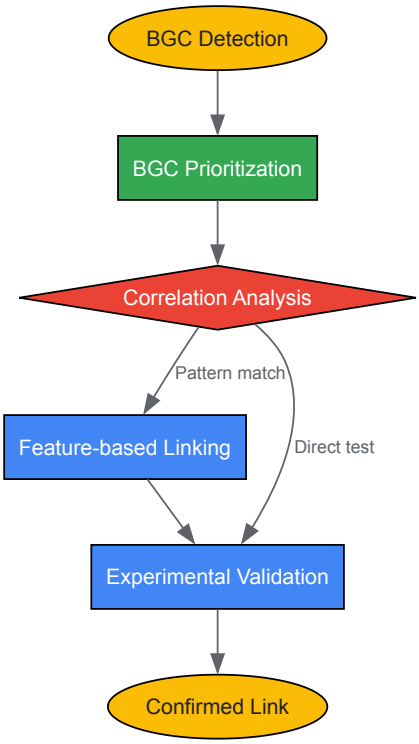


Figure 2: Decision Workflow for BGC-Chemical Structure Validation Strategies

The integration of genomic and metabolomic data through feature-based and correlation-based approaches has dramatically accelerated the process of linking BGCs to their chemical products. As computational methods evolve, particularly with the incorporation of machine learning and artificial intelligence, the efficiency and accuracy of these linkages will continue to improve [4]. Future developments in database curation, analytical instrumentation, and genetic manipulation techniques will further streamline the validation process, enabling comprehensive mapping of biosynthetic diversity to chemical space. For researchers and drug development professionals, mastering these integrated approaches is essential for unlocking the vast potential of microbial natural products in therapeutic development.

Within the genomic architecture of microorganisms lies a treasure trove of biosynthetic potential encoded by **Biosynthetic Gene Clusters (BGCs)**. These clusters are responsible for producing a vast array of **secondary metabolites**, many of which form the foundation of clinically essential antibiotics, anticancer agents, and other therapeutics [90] [91]. The central challenge in modern natural product discovery has shifted from genome sequencing to the functional interpretation of this encoded potential, particularly as genomic data reveals that traditionally cultivated microorganisms possess far more BGCs than previously known metabolites [90] [92].

This technical guide details a sophisticated comparative genomics framework designed to systematically prioritize BGCs for experimental characterization. By integrating **phylogenetic distribution analysis** with **sequence similarity networking**, this approach provides researchers with a powerful methodology to navigate the immense diversity of BGCs and focus discovery efforts on those gene clusters most likely to encode novel chemical scaffolds. The protocols outlined herein enable the transition from raw genomic data to confidently prioritized targets, addressing a critical bottleneck in the pipeline from gene sequence to new therapeutic compound [90] [46].

Core Concepts and Rationale

The Prioritization Challenge in Natural Product Discovery

The development of efficient genome mining tools has revealed a surprising disparity: microbial genomes harbor a remarkable abundance of BGCs, far exceeding the number of secondary metabolites detected under standard laboratory conditions [90]. For example, actinobacterial genomes typically contain **20–29 BGCs on average**, with certain phylogenetic lineages exhibiting even greater potential [91]. A specific clade of *Streptomyces* characterized by rugose-ornamented spores was found to possess an average of **50 BGCs per genome**, with genomes averaging 11.5 Mb in size [90].

This discrepancy creates a fundamental resource allocation problem in discovery pipelines. As activating silent BGCs requires substantial investment in culture optimization, genetic manipulation, or analytical chemistry, strategic prioritization becomes essential [90] [92]. The framework described in this work addresses this challenge by leveraging evolutionary principles and genetic relationships to identify BGCs with high novelty potential before committing to laborious experimental characterization.

Conceptual Foundation: Phylogeny and BGC Distribution

The theoretical underpinning of this approach rests on the established relationship between phylogenetic lineage and biosynthetic potential. Comparative genomics analyses have demonstrated that BGC distribution across microbial taxa is not random but follows **discernible phylogenetic patterns** [90] [93]. Certain monophyletic lineages exhibit significantly enriched biosynthetic capabilities, representing promising targets for metabolite characterization studies [90].

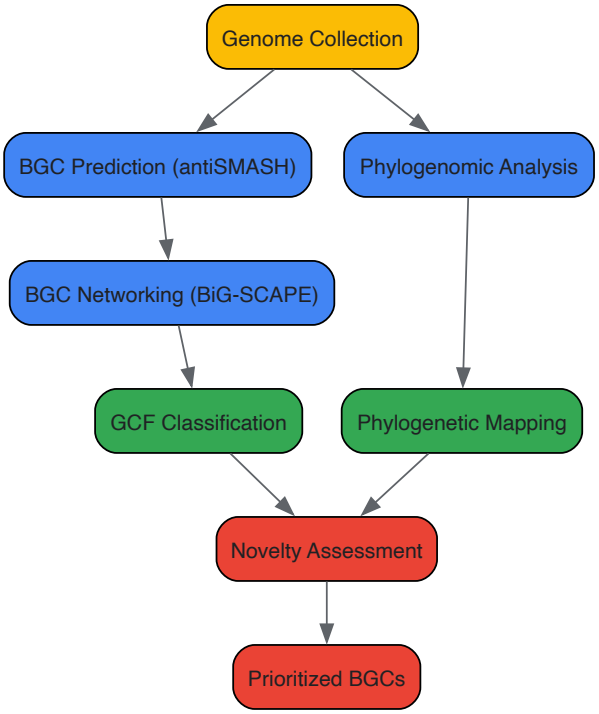
Two primary evolutionary mechanisms shape BGC distribution:

- **Vertical Inheritance**: BGCs conserved within specific phylogenetic lineages, indicating possible functional importance to that clade [91].
- **Horizontal Gene Transfer**: BGCs distributed sporadically across taxa, often located in genomic regions with high plasticity [91].

Understanding these distribution patterns enables researchers to make informed decisions about which phylogenetic groups and which types of BGCs warrant further investigation based on their novelty and taxonomic prevalence.

Computational Framework and Workflow

The following workflow diagram illustrates the integrated computational pipeline for BGC classification and phylogenetic analysis, from genome collection to final prioritization:



Stage 1: Genomic Data Acquisition and Curation

Protocol 1.1: Genome Selection and Quality Control

Initiate the workflow with comprehensive genome collection representing the taxonomic group of interest. For a robust analysis, include:

- **Type strains** with validated taxonomic assignments
- **Environmental isolates** from diverse ecological niches
- **Reference genomes** with previously characterized metabolites

Quality assessment is critical at this stage. Filter genomes based on:

- **Completeness**: >95% completeness as determined by CheckM [93]
- **Contamination**: <5% contamination threshold
- **Contiguity**: Prefer genomes with <100 contigs to minimize BGC fragmentation [90]

For phylogenetic analysis, ensure balanced representation across taxonomic groups and environments. The resulting genome set should capture the diversity of the clade while maintaining sufficient quality for comparative analysis.

Protocol 1.2: Genome Annotation and Functional Prediction

Process curated genomes through standardized annotation pipelines:

- Utilize **RAST** or **PROKKA** for consistent gene calling [94]
- Employ **antiSMASH** for initial BGC identification with default detection settings [7] [93]
- Enable all analysis modules including **KnownClusterBlast**, **ClusterBlast**, and **SubClusterBlast** to facilitate later comparative analysis [7]

Stage 2: BGC Prediction and Classification

Protocol 2.1: Comprehensive BGC Detection

Execute antiSMASH analysis uniformly across all genomes:

Key parameters to enable:

- **ClusterBlast** for comparative analysis against known clusters
- **SubClusterBlast** for detecting conserved subregions
- **Pfam2GO** for functional domain annotation [93]

Process results to compile all predicted BGCs into a standardized table format, recording:

- BGC type (PKS, NRPS, terpene, RiPP, etc.)
- Contig location and coordinates
- Domain architecture and key biosynthetic genes
- Similarity to known MIBiG reference clusters [94]

Protocol 2.2: Gene Cluster Family (GCF) Analysis

Group BGCs into Gene Cluster Families using **BiG-SCAPE**:

- Use default **0.3 similarity cutoff** for initial clustering [93]
- Employ **10% cutoff** for fine-scale family resolution when needed [7]
- Reference **MIBiG database** to identify known GCFs [94]

BiG-SCAPE calculates pairwise distances between BGCs based on domain sequence similarity and organization, generating network files that can be visualized in Cytoscape [7]. This classification places each BGC into context with known and unknown families, enabling novelty assessment.

Stage 3: Phylogenomic Reconstruction

Protocol 3.1: Robust Phylogeny Construction

For reliable phylogenetic placement, employ **Multi-Locus Sequence Analysis (MLSA)** with conserved single-copy housekeeping genes:

- Select 5-7 core genes (*atpD, clpB, gapA, gyrB, nuoD, pyrH, rpoB*) [91]
- Extract sequences using **HMMER** with Pfam domain models [94]
- Align concatenated sequences with **MAFFT** or **ClustalW**
- Construct maximum likelihood tree with **IQ-TREE** or **RAxML** with 1000 bootstrap replicates [91]

For broader taxonomic comparisons, the **rpoB** gene alone can provide reliable phylogenetic signal [7]. However, for species-level resolution within a genus, MLSA provides superior discriminatory power compared to 16S rRNA alone [91].

Protocol 3.2: Phylogenetic Tree Annotation and Visualization

Annotate the resulting phylogenetic tree with key metadata:

- Species/strain identifiers and isolation sources
- Genome size and BGC abundance
- Presence of specific GCFs or BGC classes
- Ecological origins or specific phenotypes

Use **iTOL** for interactive tree visualization and annotation [7], enabling exploration of the relationship between phylogeny and biosynthetic potential.

Stage 4: Integrated Analysis and Prioritization

Protocol 4.1: Cross-Ranking BGCs by Phylogeny and Similarity

The core analytical stage integrates phylogenetic and GCF data to identify high-priority BGCs. Implement a dual-axis ranking system that considers:

- **Phylogenetic Distribution Patterns**:

- **Strain-specific BGCs**: Present in only one strain, indicating recent acquisition [91]
- **Clade-specific BGCs**: Restricted to a monophyletic subgroup [90]
- **Widely distributed BGCs**: Found across the phylogeny, suggesting conserved function

- **Sequence Similarity Assessment**:
  - **Novel GCFs**: No significant similarity to known BGCs in MIBiG [91]
  - **Divergent GCFs**: Distant relatives of known clusters
  - **Known GCFs**: High similarity to characterized BGCs

Protocol 4.2: Prioritization Heuristics

Apply the following decision matrix to identify high-priority targets:

Phylogenetic Distribution	Sequence Similarity	Priority Level	Rationale
Strain-specific	Novel GCF	Very High	Recent acquisition + novel genetics
Clade-specific	Novel GCF	High	Evolutionary conservation + novelty
Strain-specific	Divergent GCF	Medium-High	Recent acquisition + potential structural variation
Clade-specific	Divergent GCF	Medium	Possible ecological specialization
Widely distributed	Known GCF	Low	Likely conserved function, lower novelty

This integrated approach enables systematic identification of BGCs that represent truly novel biosynthetic potential rather than rediscovering known compounds.

Key Analytical Outputs and Interpretation

Quantitative Assessment of Biosynthetic Potential

Comparative genomics analyses across multiple bacterial genera have revealed striking patterns in BGC distribution and diversity. The table below summarizes key quantitative findings from published studies:

Table 1: BGC Distribution Patterns Across Bacterial Genera

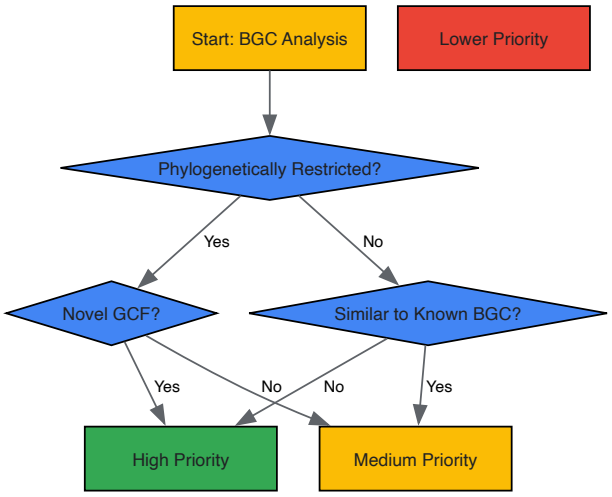
Genus	Avg. Genome Size	Avg. BGCs per Genome	Notable Phylogenetic Patterns	Study
Streptomyces	8.5 Mb (average across genus)	33 (average across genus)	Rugose-ornamented spore group: 11.5 Mb genome, 50 BGCs	[90]
Amycolatopsis	8.5-10.5 Mb	20-29	Four major phylogenetic lineages with distinct BGC content	[91]
Marine Bacteria (Proteobacteria)	Varies by species	29 BGC types identified across 199 strains	NRPS, betalactone, NI-siderophores most predominant	[7]
Plant Microbiomes	Not specified	Terpene and aryl polyene BGCs show strongest phylogenetic conservation	Conservation strength varies by BGC class and habitat	[93]

Interpretation of Phylogenetic-BGC Relationships

The relationship between phylogeny and BGC distribution provides critical insights for prioritization:

- **Strong phylogenetic conservation** of specific BGC classes (e.g., terpenes in plant-associated bacteria) suggests important ecological functions and enables targeted discovery in related taxa [93].
- **Patchy phylogenetic distribution** indicates horizontal transfer and recent acquisition, often associated with adaptation to specific environments [91].
- The presence of **monophyletic groups with enriched BGC content** (e.g., the *Streptomyces* rugose-ornamented spore group) identifies lineages with exceptional biosynthetic potential worthy of focused investigation [90].

The following diagram illustrates the decision process for BGC prioritization based on phylogenetic distribution and sequence novelty:



Computational Toolkit for BGC Analysis

Table 2: Essential Computational Resources for BGC Classification and Phylogenetic Analysis

Tool/Resource	Primary Function	Key Features	Application in Workflow
antiSMASH [7] [93]	BGC Prediction	Identifies BGCs based on known biosynthetic rules; integrates with MIBiG	Initial BGC detection and annotation
BiG-SCAPE [7] [93]	BGC Networking	Clusters BGCs into Gene Cluster Families (GCFs) based on sequence similarity	GCF classification and novelty assessment
MIBiG [95] [94]	Reference Database	Curated repository of experimentally characterized BGCs	Reference for known BGCs and their products
Phylogenetic Software (IQ-TREE, RAxML) [91]	Tree Building	Constructs robust phylogenetic trees from sequence alignments	Phylogenomic analysis and evolutionary inference
Cytoscape [7]	Network Visualization	Visualizes BiG-SCAPE output networks	Exploration of GCF relationships
iTOL [7]	Tree Visualization	Annotates and displays phylogenetic trees with metadata	Integration of phylogenetic and BGC data

Experimental Validation Strategies

Following computational prioritization, selected BGCs require experimental validation:

- **Heterologous Expression**: Clone prioritized BGCs into suitable production hosts (e.g., *Streptomyces coelicolor*) with optimized promoters [92].
- **Culture Optimization**: Employ OSMAC (One Strain Many Compounds) approaches with varied media composition and cultivation parameters [92].
- **Metabolite Analysis**: Utilize LC-MS/MS and NMR to characterize novel compounds, comparing to known metabolites in databases.
- **Regulatory Manipulation**: Target pathway-specific regulators or epigenetic controls (e.g., histone deacetylase inhibitors) to activate silent clusters [92].

The integrated framework of BGC family classification and phylogenetic analysis represents a powerful strategy for navigating the complex landscape of microbial biosynthetic potential. By applying this systematic approach, researchers can transition from indiscriminate BGC screening to targeted investigation of phylogenetically informed, genetically novel gene clusters with the highest potential for yielding new chemical entities.

This methodology directly addresses the critical bottleneck in natural product discovery—the prioritization of targets from thousands of predicted BGCs—by leveraging evolutionary principles and comparative genomics. As computational methods continue to advance, particularly through machine learning approaches [46] [4], the integration of phylogenetic distribution patterns with BGC classification will remain fundamental to rational discovery pipelines aimed at unlocking nature's chemical diversity for therapeutic development.

Functional characterization of biosynthetic gene clusters (BGCs) is a critical pathway for converting genomic data into novel therapeutic agents. Within the broader thesis of BGC identification research, determining the chemical structure and biological activity of the metabolites these clusters encode represents the ultimate step from in silico prediction to tangible discovery. Many BGCs are **silent or cryptic**, meaning they are not expressed under standard laboratory conditions, and their potential products remain unknown [16]. Furthermore, the native producers of intriguing BGCs are often **uncultivable** or genetically intractable, creating a significant bottleneck in natural product discovery [16].

To overcome these challenges, researchers employ a two-pronged technical approach: heterologous expression and advanced compound isolation. **Heterologous expression** involves transferring a target BGC into a genetically amenable host chassis, thereby activating its biosynthetic potential in a controlled environment [96]. Following successful expression, sophisticated **chromatographic isolation techniques** are required to separate, purify, and characterize the target compound from a complex broth [97]. This guide details the current methodologies and best practices for this end-to-end process, providing a technical roadmap for scientists and drug development professionals.

Heterologous Expression of Biosynthetic Gene Clusters

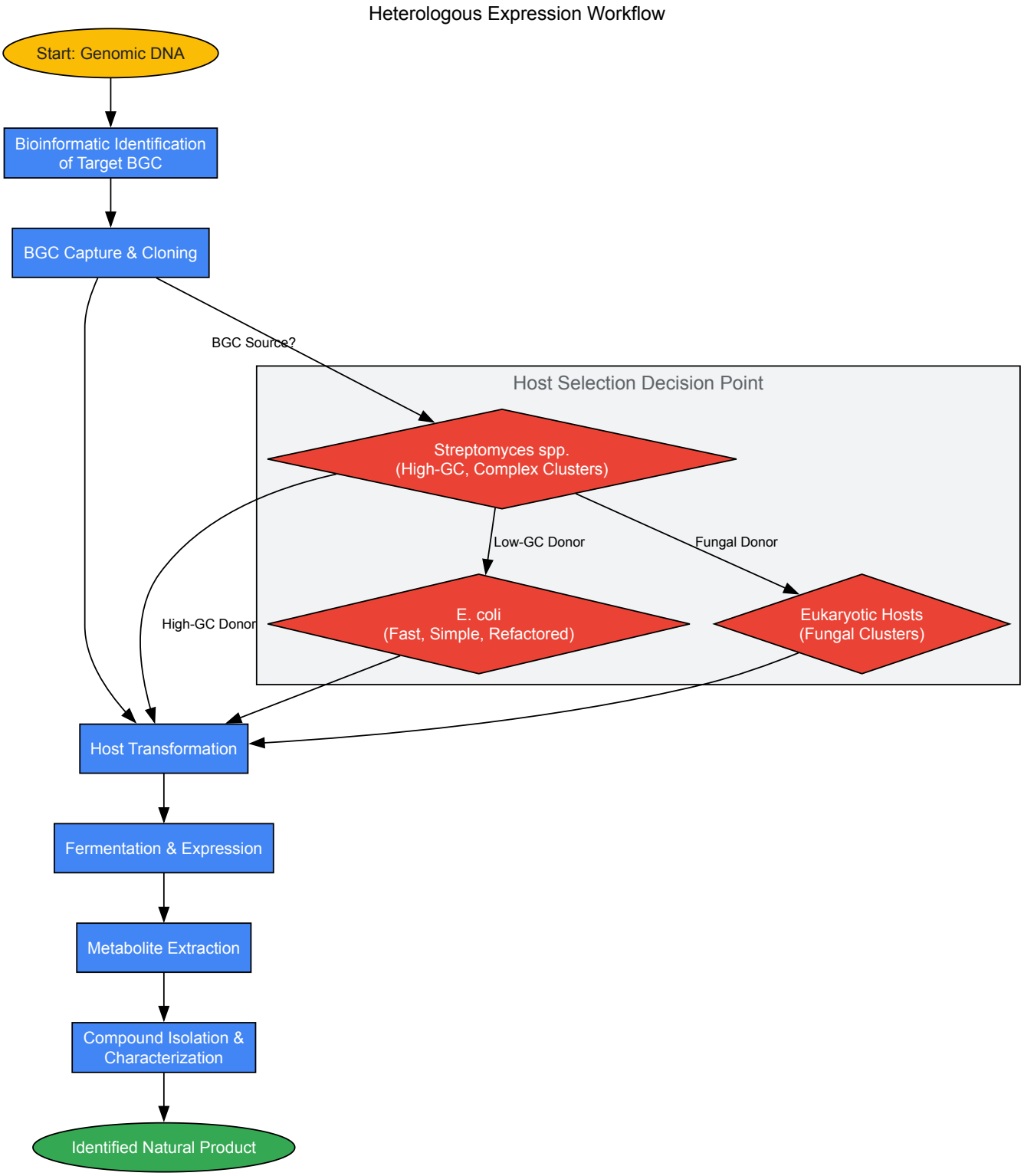
Fundamental Workflow and Host Selection

The typical workflow for heterologous expression is a sequence of four continuous steps: 1) bioinformatics-driven identification of a target BGC; 2) capture and genetic refactoring of the BGC; 3) selection of a suitable heterologous host; and 4) isolation and identification of the synthesized metabolite [96]. The success of this process is heavily influenced by the initial choice of a heterologous host, which must provide the necessary enzymatic precursors, co-factors, and cellular environment for the BGC to function.

Table 1: Common Heterologous Hosts for BGC Expression

Host Organism	Key Features and Advantages	Ideal for BGCs from	Common Applications
Streptomyces spp.	High GC-content compatibility; native capacity for producing complex polyketides and non-ribosomal peptides; well-established fermentation [76]	Actinobacteria; other high-GC bacteria [76]	Expression of large, complex clusters (PKS, NRPS) [76]
Escherichia coli	Fast growth; highly tractable genetics; extensive toolkit for protein expression [98]	Low-GC bacteria; refactored clusters [98]	Production of single proteins or refactored pathways [98]
***Aspergillus* spp.**	Advanced eukaryotic protein processing and secretion [96]	Fungal kingdoms [96]	Expression of fungal BGCs requiring eukaryotic machinery [96]
Saccharomyces cerevisiae	Eukaryotic host; strong homologous recombination for DNA assembly [99]	Fungi and refactored bacterial clusters [99]	Pathway assembly and expression of eukaryotic clusters [99]

As a general rule, the closer the phylogenetic relationship between the donor and the host, the more likely the expression will be successful due to shared codon usage patterns and regulatory factors [16]. Analysis of over 450 studies confirms **Streptomyces species** as the most versatile and widely used chassis, particularly for expressing BGCs from other actinobacteria [76]. Their intrinsic advantages include genomic compatibility (high GC content), a native capacity for producing complex secondary metabolites, and advanced, scalable fermentation processes [76].



BGC Capture and DNA Assembly Methods

Capturing an intact BGC, which can range from 10 to over 100 kb in size, is a technically demanding step. A variety of both in vitro and in vivo DNA assembly tools have been developed to accomplish this [99]. The choice of method depends on the size and nature of the cluster, as well as the desired throughput.

Table 2: DNA Assembly Methods for BGC Reconstruction

Method	Principle	Maximum Assembled Size (Number of Fragments)	Key Application
Transformation-Associated\nRecombination (TAR)	In vivo homologous recombination in *S. cerevisiae* [99]	~67 kb [99]	Direct capture of large BGCs from genomic DNA [99]
Modular Cloning (MoClo)	Type IIs restriction enzyme-based seamless assembly in vitro [99]	50 kb (68 fragments) [99]	High-throughput, automated assembly of refactored clusters [99]
DNA Assembler	In vivo homologous recombination in *S. cerevisiae* [99]	50 kb [99]	Assembly of pathways from individual genes or modules [99]
Cas9-Assisted Targeting (CATCH)	Uses Cas9 nuclease to linearize and capture specific genomic regions [76]	> 100 kb [76]	Precise isolation of BGCs directly from chromosome [76]

For clusters that are silent in their native state, **refactoring** is often necessary. This process involves replacing native regulatory elements (promoters, ribosome binding sites) with well-characterized, synthetic parts to ensure robust expression in the heterologous host [99]. This strategy was successfully used to activate the silent spectinabilin BGC from *Streptomyces orinoci* [99].

Experimental Protocol: BGC Expression in aStreptomycesHost

The following provides a detailed methodology for the heterologous expression of a BGC in a *Streptomyces* host, a common and powerful approach in the field [96] [76] [99].

- **Bioinformatic Identification and Analysis**: Identify the target BGC from a sequenced genome using tools like **antiSMASH** [16] [96]. Analyze the cluster for key features like GC content, codon usage, and potential regulatory genes.
- **BGC Capture**:
  - For large clusters (>30 kb), use direct capture methods like **TAR** or **CATCH** [76]. Design specific primers or guide RNAs to flanks the BGC precisely.
  - For smaller clusters or refactored pathways, use in vitro assembly methods like **MoClo** or Gibson Assembly. Clone the assembled cluster into an appropriate *Streptomyces* integration vector (e.g., containing the φC31 or φBT1 *attP* site) [99].
- **Host Preparation and Transformation**:
  - Culture the selected *Streptomyces* host strain (e.g., *S. albus* or *S. coelicolor*) in a suitable liquid medium to prepare protoplasts or competent cells.
  - Introduce the constructed vector into the host via polyethylene glycol (PEG)-mediated protoplast transformation or intergeneric conjugation from *E. coli*.
  - Select for exconjugants or transformants using the appropriate antibiotic(s) present on the vector.
- **Screening and Fermentation**:
  - Genotypically verify correct integrants by PCR across the integration junctions.
  - Inoculate positive clones into multiple production media (e.g., SFM, R5) to trigger secondary metabolism. Incubate with shaking for 3–7 days.
- **Metabolite Analysis**:
  - Extract the culture broth and mycelia with an equal volume of organic solvent (e.g., ethyl acetate or butanol).
  - Concentrate the organic extract under reduced pressure and resuspend in methanol for analysis by liquid chromatography-mass spectrometry (LC-MS).
  - Compare the metabolic profiles of the expression strain to the empty vector control to identify new peaks specific to the BGC.

Chromatographic Isolation of Bioactive Compounds

Once heterologous expression is confirmed, the target compound must be isolated from the complex fermentation broth in sufficient quantity and purity for structural elucidation and biological testing.

Core Chromatographic Techniques

The choice of chromatographic technique depends on the physicochemical properties of the target compound.

- **Liquid Chromatography (LC)**: The workhorse for natural product isolation. In preparative LC, the goal shifts from separating all components to maximizing the resolution around the peak of interest to enhance loading capacity, purity, and recovery [100].

- **Ion-Exchange Chromatography (IEX)**: Separates proteins and other charged molecules based on electrostatic interactions with a solid support matrix. Molecules are eluted by changing the pH or ionic strength of the buffer [97].
- **Hydrophobic Interaction Chromatography (HIC)**: Separates molecules based on their hydrophobicity. It is particularly useful for separating proteins with minimal denaturation [97].
- **Affinity Chromatography**: A highly specific technique where a ligand that binds the target protein is immobilized on the matrix. The target is retained while impurities pass through, and is later eluted under specific conditions [97].
- **Gel-Permeation Chromatography (GPC) / Size Exclusion**: Separates molecules based on their size and shape. It is often used for desalting protein solutions or determining molecular weights [97].

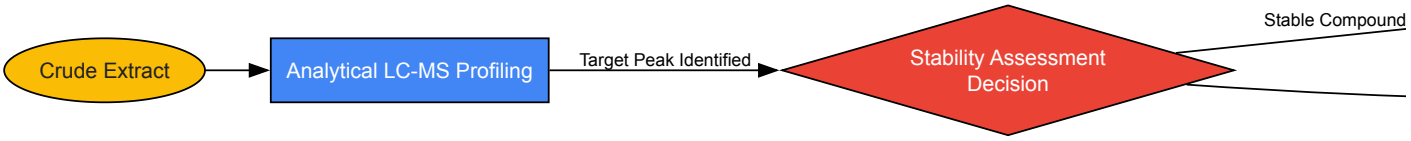
Managing Compound Instability During Isolation

A major challenge in isolating novel natural products is their potential instability under standard purification conditions. Degradation can occur during the separation or the subsequent solvent removal steps, leading to poor recovery or incorrect structural assignment [100].

Experimental Protocol: Isolation of Sensitive Compounds

- **Method Development with Stability in Mind**: When developing the LC method, systematically test different columns, mobile phases (using volatile buffers like ammonium bicarbonate or TFA), and, critically, **temperatures**. A method that gives optimal peak shape at high temperature (e.g., 55°C) may cause degradation of the target compound [100].
- **Low-Temperature Operation**: If instability is suspected, lower the temperature of the entire system (autosampler, column compartment, and fraction collector) to 5-10°C. While this may cause some peak broadening, it dramatically reduces degradation and improves the purity of the collected fraction [100].
- **Gentle Fraction Drying**: For aqueous fractions containing sensitive compounds, avoid harsh drying methods.
  - Use a **rotary evaporator** at low pressure (e.g., 5 mbar) and with the water bath temperature maintained at a low level (e.g., 6°C) [100].
  - To facilitate the removal of the last amounts of water, add a small amount of a solvent like acetonitrile to form an **azeotrope**, which evaporates more readily [100].
  - Lyophilization is another option but may not be suitable for all compounds [100].

Cc



The Scientist's Toolkit: Essential Reagents and Materials

Successful functional characterization relies on a suite of specialized reagents and tools. The following table details key solutions used in the workflows described above.

Table 3: Key Research Reagent Solutions for Functional Characterization

Reagent / Material	Function and Application	Specific Examples / Notes
antiSMASH Software	The primary bioinformatics tool for the identification and annotation of BGCs in genomic data [16] [96].	Enables detailed analysis of cluster architecture and predicts core biosynthetic machinery [16].
Bacterial Artificial Chromosomes (BACs)	Vectors used to clone and maintain very large DNA inserts (100-200 kb) for BGC library construction [76].	Essential for storing and manipulating large, complex BGCs prior to heterologous expression [76].
Expression Vectors (φC31-based)	Integration vectors for stable introduction of BGCs into the chromosome of *Streptomyces* hosts [99].	Provides stable, single-copy integration, avoiding plasmid instability issues during fermentation [99].
CyDisCo System	A strain/co-expression system enabling the production of proteins with multiple disulfide bonds in the cytoplasm of *E. coli* [98].	Crucial for functional expression of complex eukaryotic proteins in a prokaryotic host [98].

Reagent / Material	Function and Application	Specific Examples / Notes
Cibacron Blue Dye	A dye-ligand used in affinity chromatography for purifying nucleotide-binding proteins like dehydrogenases and kinases [97].	Mimics the structure of NAD, allowing for one-step purification of various enzymes [97].
Volatile Buffers	Mobile phase additives for preparative LC that are easily removed during the drying of collected fractions [100].	Trifluoroacetic acid (TFA, 0.05-0.1%), ammonium formate, ammonium bicarbonate [100].

The integrated strategies of heterologous expression and sophisticated compound isolation form a powerful pipeline for transforming the vast amount of data from BGC identification research into novel chemical entities. The continuous development of more efficient DNA assembly techniques, a growing portfolio of engineered host chassis, and sensitive analytical methods is steadily overcoming previous technical barriers. As these tools mature, the systematic functional characterization of the countless silent BGCs predicted by genomics will undoubtedly accelerate, unlocking new bioactive compounds with potential applications in drug development and other biotechnology sectors. This structured, technical guide provides a framework for researchers to navigate this complex but rewarding process, from genome to compound.

Biosynthetic gene clusters (BGCs) are genomic arrangements of co-localized genes that encode the production of microbial secondary metabolites, which represent a primary source of bioactive compounds with therapeutic potential [101]. These metabolites, also known as natural products, have historically constituted over 60% of approved pharmaceuticals, including essential antibiotics, anticancer agents, and immunosuppressants [11] [102]. The traditional discovery pipeline for these compounds relied heavily on bioactivity-guided fractionation, which is often labor-intensive, time-consuming, and prone to rediscovering known compounds [101] [103]. Modern genome mining approaches have revealed a striking disparity between genetic potential and characterized metabolites, with computational analyses identifying tens of thousands of putative BGCs across microbial genomes, suggesting a vast untapped resource for drug discovery [101] [37]. This guide examines contemporary methodologies for evaluating the therapeutic potential of BGC-encoded metabolites, integrating computational prediction with experimental validation to accelerate natural product-based drug development.

Computational Prediction and Prioritization of BGCs

The initial identification and prioritization of BGCs from genomic data relies on sophisticated bioinformatics tools that detect genetic signatures of secondary metabolite biosynthesis. These computational approaches have become indispensable for navigating the extensive genomic landscape and focusing experimental efforts on the most promising targets.

Table 1: Computational Tools for BGC Detection and Analysis

Tool Name	Primary Function	Target Organisms	Key Features
antiSMASH [7] [101]	BGC detection & annotation	Bacteria, Fungi, Plants	Rule-based detection using profile HMMs; most widely used tool
PRISM 4 [37]	Chemical structure prediction	Bacteria	Predicts complete chemical structures of encoded metabolites
BAGEL4 [104]	Bacteriocin & RiPP detection	Bacteria	Specialized for ribosomally synthesized and post-translationally modified peptides
BiG-SCAPE [7]	BGC clustering & network analysis	Bacteria, Fungi	Groups BGCs into gene cluster families based on sequence similarity
ARTS [101]	BGC prioritization	Bacteria	Identifies promising BGCs based on resistance genes and genomic context

The effectiveness of these tools was demonstrated in a 2025 study that analyzed 199 marine bacterial genomes, identifying 29 distinct BGC types using antiSMASH 7.0, with non-ribosomal peptide synthetases (NRPS), betalactone, and NI-siderophores being most prevalent [7]. The integration of multiple tools creates a powerful pipeline for BGC characterization, from initial detection through structural prediction and prioritization.

BGC Prioritization Strategies

With the ability to detect numerous BGCs in a single genome, prioritization becomes essential. Biological hypotheses provide a framework for identifying BGCs with heightened potential for novel bioactivity. Three key principles guide this process:

- **Self-resistance Mechanisms**: Bacteria protect themselves from their own antibiotics through resistance genes, which are often encoded within or near the BGC. The ARTS tool exploits this by identifying BGCs with associated self-resistance genes, indicating bioactivity [101].
- **Gene Duplication Events**: The presence of duplicated biosynthetic genes suggests evolutionary selection for enhanced production of valuable metabolites [101].

- **Horizontal Gene Transfer**: BGCs with evidence of horizontal transfer may provide competitive advantages to their hosts and are more likely to encode bioactive compounds [101].

These computational prioritization strategies enable researchers to focus experimental validation on BGCs most likely to yield novel therapeutics with potent biological activities.

Experimental Screening Methodologies

Once promising BGCs are computationally identified and prioritized, experimental screening methodologies are essential for confirming their bioactivity and therapeutic potential. These approaches can be broadly categorized into affinity-based and cell-based screening platforms.

Affinity-Based Screening Platforms

Affinity-based screening methods operate on the principle of molecular recognition, where target biomolecules selectively bind to bioactive compounds from complex mixtures. These techniques are uniquely advantageous as they can identify ligands without prior separation of individual components [105].

Table 2: Affinity-Based Screening Platforms for Bioactive Compound Identification

Method	Principle	Applications	Key Advantages
Cell Membrane Chromatography (CMC) [105] [103]	Cell membranes immobilized on silica carriers retain ligands binding to specific receptors	Screening receptor-targeting compounds from natural extracts; extensively used for Traditional Chinese Medicines	Maintains native receptor conformation; can screen complex mixtures directly
Affinity Ultrafiltration (UF) [103]	Macromolecular targets incubated with extracts; unbound compounds removed by ultrafiltration	Enzyme inhibition screening (e.g., acetylcholinesterase, xanthine oxidase)	Suitable for targets without established immobilization protocols
Magnetic Separation [105] [103]	Targets immobilized on magnetic nanoparticles; ligand-target complexes separated magnetically	High-throughput screening of enzyme inhibitors and receptor ligands	Rapid separation; easily scalable; compatible with automation
Surface Plasmon Resonance (SPR) [105]	Real-time monitoring of molecular interactions on sensor chip surfaces	Kinetic analysis of ligand-target interactions; determination of binding affinity	Label-free; provides kinetic parameters (association/dissociation rates)

The experimental workflow for CMC exemplifies the integration of screening with analytical characterization. In a typical setup, a CMC column is connected to an HPLC-MS/MS system through a multi-port switching valve. Natural product extracts are first applied to the CMC dimension, where compounds with affinity for the immobilized receptors are retained. The retained fractions are then transferred to the HPLC-MS/MS for separation and identification, enabling simultaneous activity screening and compound characterization [105].

Cell-Based Screening Assays

Cell-based screening methods utilize whole living cells to evaluate bioactivity, maintaining the physiological context of molecular targets including receptors, enzymes, and signaling pathways. These approaches can identify compounds with complex mechanisms of action that might be missed in target-based screens [103].

Advanced cell-based screening platforms now incorporate high-content imaging and omics technologies to provide deeper insights into mechanism of action. For example, image-based high-throughput screening in 384-well formats has been developed to discover biofilm inhibitors against pathogenic bacteria like *Pseudomonas aeruginosa*, using constitutively expressed fluorescent proteins for detection [102]. These systems can monitor phenotypic changes in real-time, providing information on cytotoxicity and therapeutic windows simultaneously.

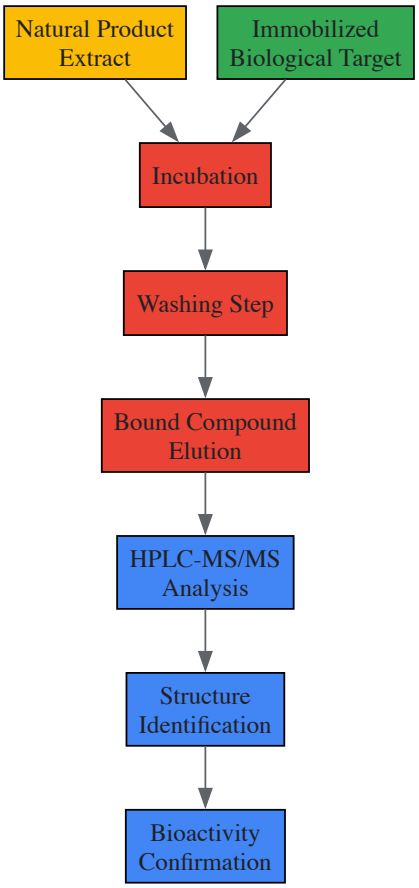


Figure 1: Affinity-Based Screening Workflow. This diagram illustrates the key steps in affinity-based screening approaches, from incubation of natural product extracts with immobilized targets through to bioactivity confirmation.

Integrating Genomic and Experimental Data

The most powerful approaches for BGC-encoded metabolite discovery integrate genomic information with experimental screening data, creating a bidirectional pipeline that connects genetic potential with chemical output and biological activity.

Metabolomics-Driven Genome Mining

Metabolomics-guided strategies leverage analytical chemistry data to direct genomic analyses, creating a target-oriented discovery process. Key approaches include:

- **Molecular Networking**: Mass spectrometry-based molecular networking clusters compounds by structural similarity, allowing identification of novel analogs of known bioactive molecules and connection to their BGCs [102].
- **Dereplication Strategies**: Advanced dereplication combines hyphenated techniques (LC-MS, GC-MS, LC-NMR) with database searching to rapidly identify known compounds, focusing efforts on novel chemical entities [102].

Heterologous Expression and Structure Elucidation

For BGCs from unculturable organisms or those with silent expression, heterologous expression in model hosts such as *E. coli* or *S. cerevisiae* enables production and characterization of encoded metabolites [101]. Following production, structure elucidation represents a critical step in the pipeline, with particular challenges in determining stereochemistry. Advanced methods include:

- **Quantum chemical calculations** for NMR chemical shift prediction and DP4 probability analysis [102]
- **Electronic and vibrational circular dichroism** for absolute configuration determination [102]
- **Computer-assisted structure elucidation (CASE)** systems that integrate experimental spectroscopic data with computational predictions [102]

The Scientist's Toolkit: Essential Research Reagents and Materials

Successful bioactivity screening of BGC-encoded metabolites requires specialized reagents and materials tailored to both computational and experimental workflows.

Table 3: Essential Research Reagent Solutions for BGC Bioactivity Screening

Reagent/Material	Function	Application Examples
antiSMASH Database [7]	Reference database of known BGCs for comparative analysis	Annotating putative BGCs; identifying novel gene clusters
MIBiG Reference Dataset [11]	Curated repository of experimentally characterized BGCs	Training and validation of prediction algorithms; dereplication
Cell Membrane Stationary Phase (CMSP) [105]	Immobilized cell membranes for CMC screening	Identifying ligands for specific transmembrane receptors
Enzyme-Magnetic Nanoparticle Conjugates [103]	Target immobilization for magnetic separation screening	High-throughput fishing of enzyme inhibitors from complex mixtures
Biosynthetic Domain Architectures [11]	Vectorized representations of biosynthetic domains	Comparative analysis of BGCs across phylogenetically diverse organisms

The field of BGC-encoded metabolite discovery has evolved from traditional activity-guided fractionation to sophisticated integrated approaches that connect genomic potential with therapeutic activity. The convergence of computational prediction, advanced affinity-based screening, and metabolomic profiling has created a powerful pipeline for identifying novel bioactive compounds. Current challenges include improving the accuracy of in silico structure prediction, particularly for stereochemical elements, and developing efficient heterologous expression systems for silent BGCs. Future directions will likely see increased incorporation of machine learning algorithms throughout the discovery pipeline, from BGC boundary prediction to bioactivity forecasting, further accelerating the identification of therapeutic candidates from the vast untapped reservoir of microbial biosynthetic diversity.

Within natural product drug discovery, **dereplication** represents the critical process of rapidly identifying known compounds in complex biological extracts to prioritize novel leads early in the discovery pipeline [106] [107]. This process has evolved from simple analytical comparisons to sophisticated integrated frameworks that combine high-throughput technologies, bioinformatics, and genomic data. When framed within the context of **biosynthetic gene cluster (BGC) research**, dereplication transforms from merely recognizing known chemical structures to understanding and predicting the genetic potential of microbial producers [108] [109]. The convergence of analytical chemistry, genomics, and computational biology has created powerful dereplication frameworks capable of distinguishing novel compounds from known natural products with unprecedented efficiency, thereby accelerating the discovery of new therapeutic agents from natural sources [106] [108].

The fundamental challenge in natural product discovery lies in the extensive chemical complexity of biological extracts coupled with the high probability of compound rediscovery [106] [110]. Historically, bioassay-guided fractionation served as the primary discovery approach, but this method is time-consuming, resource-intensive, and often leads to the repeated isolation of known compounds [111]. Modern dereplication frameworks address these limitations by integrating multiple data streams to enable informed decisions about which extracts warrant further investigation [111] [107]. Within BGC research, this involves connecting chemical profiles to genetic blueprints, allowing researchers to prioritize silent or cryptic gene clusters that may encode novel compounds with desirable bioactivities [108] [109].

The Evolution of Dereplication in Natural Product Discovery

Dereplication has undergone a significant transformation from its initial conception as a simple chromatographic comparison technique to its current state as a multidisciplinary framework integrating advanced analytical technologies and bioinformatics [106] [107]. The term "dereplication" itself encompasses the entire process of analyzing extracts from microbial fermentation broths or plant samples to identify known compounds early in the screening process [107]. This evolution has been driven by several key factors, including the increasing availability of genomic data, advancements in mass spectrometry, and the development of specialized databases and computational tools [106].

In the context of BGC research, dereplication has taken on an additional dimension that extends beyond chemical identification to encompass genetic potential [108] [109]. The revelation that microbial genomes contain a vast reservoir of silent or cryptic BGCs – genetic elements not expressed under standard laboratory conditions – has created new opportunities and challenges for natural product discovery [108] [109]. Modern dereplication frameworks must therefore address both the chemical diversity present in extracts and the genetic potential encoded within microbial genomes, creating a comprehensive approach that connects genotype to chemotype [109] [112].

The integration of dereplication with BGC research has revealed the staggering untapped potential of microbial natural products. Studies of mangrove swamp microbiomes, for instance, have identified 3,740 BGCs across 809 reconstructed genomes, with 86% showing no similarity to known clusters in the MIBiG repository [109]. Similarly, genome mining of bacterial species has revealed that microbial strains often contain numerous BGCs that remain silent under standard laboratory conditions [108]. This expanded understanding has transformed dereplication from a defensive strategy against rediscovery to an offensive approach for unlocking nature's hidden chemical diversity.

Core Components of Modern Dereplication Frameworks

Analytical Instrumentation and Chromatographic Techniques

The foundation of any dereplication framework lies in its analytical capabilities, with liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) serving as the workhorse technology [113] [114] [110]. Modern approaches utilize ultra-high performance liquid chromatography (UHPLC) systems equipped with high-resolution mass spectrometers, providing exceptional separation efficiency and mass accuracy for analyzing complex natural product mixtures [106] [107]. The specific configuration and parameters of these systems significantly impact the quality and depth of dereplication data.

Different mass spectrometry acquisition strategies offer complementary advantages for dereplication. Data-dependent acquisition (DDA) generates cleaner MS/MS spectra suitable for database matching, while data-independent acquisition (DIA) approaches like SWATH (Sequential Window Acquisition of All Theoretical Fragment-Ion Spectra) provide comprehensive fragmentation data for all detectable analytes [114]. The integration of both approaches creates a powerful dereplication strategy that leverages their respective strengths [114]. For example, in a study of Sophora flavescens, the combination of DIA and DDA enabled the annotation of 51 compounds, demonstrating how these complementary approaches enhance dereplication comprehensiveness [114].

The analytical workflow typically involves separation on reversed-phase C18 columns using gradient elution with water-acetonitrile or water-methanol mobile phases modified with acids or volatile buffers to enhance ionization [113] [114]. High-resolution mass analyzers such as Q-TOF (Quadrupole-Time of Flight) or Orbitrap instruments provide accurate mass measurements essential for elemental composition determination, with mass accuracy thresholds typically set at <5 ppm for confident annotations [113] [110]. These analytical parameters create the foundational data upon which subsequent dereplication strategies are built.

Bioinformatics and Data Analysis Platforms

Bioinformatics platforms serve as the computational engine of modern dereplication frameworks, transforming raw analytical data into actionable chemical insights [106] [114]. The Global Natural Products Social Molecular Networking (GNPS) platform stands as a cornerstone technology, enabling the organization and visualization of untargeted MS/MS data based on spectral similarity [106] [114] [110]. Molecular networking within GNPS clusters mass spectra of related molecules, creating visual maps that reveal structural relationships within complex mixtures and facilitate the annotation of both known and novel compounds [114].

Complementary bioinformatics tools expand dereplication capabilities through different approaches. The SIRIUS platform utilizes database-independent structure predictions, leveraging computational mass spectrometry to annotate compounds beyond the scope of spectral libraries [110]. antiSMASH (antibiotics & Secondary Metabolite Analysis Shell) enables the identification and analysis of BGCs in genomic data, creating critical connections between genetic potential and chemical output [108] [109] [112]. Specialized algorithms like NRPSpredictor2 predict substrate specificity for non-ribosomal peptide synthetase adenylation domains, enabling in silico prediction of chemical structures from genetic sequences [112].

The integration of these bioinformatics platforms creates a powerful ecosystem for dereplication. For instance, LC-MS/MS spectral data can be simultaneously analyzed using GNPS for spectral library matching and SIRIUS for database-independent structure elucidation, significantly expanding annotation capabilities [110]. Similarly, the combination of genomic analysis with antiSMASH and metabolomic data through GNPS enables the connection of BGCs to their metabolic products, bridging the gap between genotype and chemotype [109] [112].

Databases and Spectral Libraries

Comprehensive databases and spectral libraries provide the reference standards essential for compound identification in dereplication workflows [106] [107]. These resources span chemical, genetic, and taxonomic domains, creating a multidimensional knowledge base for natural product discovery. The Dictionary of Natural Products represents one of the most comprehensive chemical databases, while the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) repository serves as a curated database of experimentally characterized BGCs [111] [109].

The development of specialized mass spectral libraries significantly enhances dereplication capabilities [113] [107]. For example, researchers have constructed in-house MS/MS libraries containing spectral data for 31 commonly occurring natural products from different classes, enabling rapid dereplication of these compounds in plant and food extracts [113]. Such libraries capture not only precursor ion masses but also fragmentation patterns at different collision energies, providing robust fingerprints for compound identification [113]. Public spectral libraries within GNPS contain approximately 600,000 molecule-annotated spectra, while structure databases like PubChem and ChemSpider provide reference data for over 110 million unique structures, dramatically expanding the scope of dereplication [110].

The continuous expansion and curation of these databases remains critical for effective dereplication. As new compounds are discovered and characterized, their integration into reference databases enhances future dereplication efforts. Similarly, the development of taxon-specific databases or specialized collections focused on particular compound classes addresses the unique challenges of dereplication in targeted discovery programs [113] [107].

Integrated Dereplication Strategies and Workflows

LC-MS/MS-Based Molecular Networking

LC-MS/MS-based molecular networking has emerged as a powerful strategy for dereplication and novel compound discovery [114]. This approach leverages the organizational power of molecular networks to visualize chemical relationships within complex mixtures, enabling simultaneous dereplication of known compounds and identification of structural variants or novel analogues [114]. The workflow typically begins with LC-MS/MS analysis using either DDA or DIA modes, followed by data conversion and processing using tools like MS-DIAL or MZmine [114]. The processed data is then uploaded to GNPS for molecular network construction and analysis.

A representative application of this strategy enabled the annotation of 51 compounds from Sophora flavescens root extracts [114]. The analytical pipeline incorporated both DIA and DDA LC-MS/MS data, with DIA data processed to construct molecular networks and DDA data used for direct database matching [114]. This complementary approach overcame the limitations of individual methods, with molecular networking particularly effective for identifying trace compounds that challenged direct database matching [114]. The integration of orthogonal data sources and analysis methods created a robust dereplication framework with enhanced annotation capabilities.

Molecular networking also facilitates the discovery of structural variants within compound families. In a study of Thermoactinomyces vulgaris, researchers integrated genome mining with LC-HRMS/MS molecular networking to identify 10 structural variants of the bioactive cyclohexapeptide thermoactinoamide A, five of which were new compounds [112]. The molecular network revealed clusters of related peptides, enabling targeted isolation and characterization of novel analogues. This approach demonstrates how molecular networking transforms dereplication from merely identifying known compounds to exploring chemical diversity around promising structural scaffolds.

Ligand Affinity-Based Screening Systems

Ligand affinity-based screening systems represent a targeted dereplication approach that integrates biological activity with chemical analysis [111]. These systems utilize molecular targets as affinity capture agents to selectively isolate bioactive compounds from complex mixtures, directly linking chemical identification to mechanism of action. The Lickety-Split Ligand-Affinity-Based Molecular Angling System (LLAMAS) exemplifies this approach, employing an ultrafiltration-based LC-PDA-MS/MS-guided DNA-binding assay to identify DNA-interactive compounds from natural product extracts [111].

The LLAMAS workflow involves four interconnected phases: (1) incubation of natural product extracts with DNA targets to facilitate binding; (2) ultrafiltration to separate ligand-bound DNA complexes from unbound small molecules; (3) untargeted hyphenated mass spectrometric analysis of filtrates to detect candidate DNA-binding molecules; and (4) employment of natural product data resources for dereplication and identification [111]. By comparing filtrates from extracts incubated with DNA versus control samples processed without DNA, compounds bound to DNA are revealed through their differential abundances [111]. This approach condenses multiple rounds of purification and bioassays into a single step while providing mechanism-based classification of bioactive compounds.

In practice, LLAMAS successfully identified seven DNA-binding compounds from a library of 332 plant samples used in traditional Chinese medicine, including berberine, palmatine, coptisine, fangchinoline, tetrandrine, daurisoline, and dauricine [111]. The system was validated using eight known DNA-binding compounds representing different interaction mechanisms (intercalation, groove binding, and covalent binding), demonstrating its ability to detect diverse structural classes regardless of their solubility or detection characteristics [111]. This targeted dereplication approach efficiently links chemical identification to biological function, providing valuable mechanistic insights early in the discovery process.

Genomics-Integrated Dereplication

Genomics-integrated dereplication represents the cutting edge of natural product discovery, connecting the genetic potential encoded in BGCs with the chemical output observed in metabolomic profiles [108] [109] [112]. This approach leverages the fundamental premise that genes encoding natural product biosynthetic pathways are clustered in microbial genomes, enabling prediction of chemical potential from genetic sequences [108] [109]. The integration of genomic and metabolomic data creates a powerful framework for targeted discovery of novel compounds, particularly those from silent or cryptic gene clusters that are not expressed under standard laboratory conditions [108].

The genomics-integrated dereplication workflow typically begins with genome sequencing and analysis using tools like antiSMASH to identify BGCs [108] [109] [112]. For cultivated microorganisms, this may involve sequencing individual strains, while for complex microbial communities, metagenomic approaches reconstruct genomes from environmental samples [109]. The identified BGCs are then prioritized based on novelty, structural features, or potential bioactivities. Gene expression under various cultivation conditions is monitored using metatranscriptomic approaches, connecting genetic potential to actual compound production [109]. Finally, metabolomic profiling using LC-MS/MS and molecular networking links the expressed metabolites to their genetic origins, completing the connection from gene cluster to natural product [109] [112].

A compelling application of this approach revealed the extensive biosynthetic potential of mangrove swamp microbiomes [109]. Through analysis of 809 metagenome-assembled genomes, researchers identified 3,740 BGCs, with 86% showing no similarity to known clusters in the MIBiG repository [109]. Metatranscriptomic analysis confirmed that most identified gene clusters were active in environmental samples, while untargeted metabolomics revealed that 98% of the mass spectra generated were unrecognizable, further supporting the novelty of these BGCs [109]. This study demonstrates how genomics-integrated dereplication can access the vast hidden chemical diversity encoded in environmental microbiomes.

Table 1: Key Bioinformatic Tools for Dereplication and BGC Analysis

Tool Name	Primary Function	Application in Dereplication	Reference
GNPS	Molecular networking & spectral library matching	Organizes MS/MS data based on spectral similarity; enables compound annotation & discovery	[106] [114]
antiSMASH	BGC identification & analysis	Predicts secondary metabolite potential from genomic data; prioritizes novel BGCs	[108] [109]
SIRIUS	Database-independent structure elucidation	Annotates compounds beyond spectral libraries; expands dereplication scope	[110]
NRPSpredictor2	Substrate specificity prediction	Predicts amino acid incorporation in NRPS; enables in silico structure prediction	[112]
PRISM	BGC identification & structural prediction	Predicts chemical structures from genetic sequences; guides compound isolation	[109] [112]

Experimental Protocols for Dereplication

Protocol 1: LC-MS/MS Analysis for Dereplication

Principle: This protocol provides a standardized approach for LC-MS/MS analysis of natural product extracts, generating high-quality data for subsequent dereplication through molecular networking and database matching [113] [114].

Materials and Reagents:

- Natural product extract (e.g., 50 mg powder extracted with 10 mL methanol/water/formic acid, 49:49:2 v/v/v)
- Chromatographic grade methanol and acetonitrile
- Formic acid (purity >98.0%) or ammonium acetate (purity >98.0%)
- Purified water (Milli-Q or equivalent)
- Reference standards for system calibration and retention time alignment

Instrumentation:

- UPLC system equipped with binary pump, autosampler, and column oven
- High-resolution mass spectrometer (Q-TOF or Orbitrap)
- Reversed-phase C18 column (e.g., 2.1 × 150 mm, 1.8 μm)

Procedure:

- **Sample Preparation:**
  - Extract natural material using appropriate solvent system (e.g., methanol/water/formic acid, 49:49:2 v/v/v)
  - Centrifuge at high speed (e.g., 13,000 × g) to remove particulate matter
  - Filter supernatant through 0.22 μm membrane
  - Adjust final concentration to approximately 10 mg/mL for LC-MS analysis [114]
- **LC Conditions:**
  - Mobile Phase A: 8.0 mmol/L ammonium acetate in water or water with 0.1% formic acid
  - Mobile Phase B: acetonitrile
  - Flow rate: 0.300 mL/min
  - Column temperature: 40°C
  - Injection volume: 2.0 μL
  - Gradient program: 3-5% B (0-3 min), 5-5% B (3-5 min), 5-15% B (5-8 min), 15-60% B (8-12 min), 60-98% B (12-20 min), 98-98% B (20-21 min) [114]
- **MS Conditions (Positive Ion Mode):**
  - Ionization voltage: +5.5 kV
  - Nebulizing gas: 55 psi
  - Auxiliary gas: 55 psi
  - Curtain gas: 35 psi
  - Source temperature: 550°C
  - TOF scan range: m/z 100-2000

- For DDA: Top 4 ions selected for CID, collision energy: 50 eV, CE spread: 10 eV
- For DIA (SWATH): 50 Da mass windows across 100-1000 Da, collision energy: 50 eV [114]

- **Data Processing:**
  - Convert raw data to mzML format using MSConvert (ProteoWizard)
  - Process with MS-DIAL (DIA) or MZmine (DDA) for feature detection and alignment
  - Export results for GNPS molecular networking and database matching

Protocol 2: Genome Mining for BGC Identification

Principle: This protocol outlines the bioinformatic workflow for identifying BGCs in microbial genomes, enabling genomics-integrated dereplication and targeted discovery of novel natural products [108] [109] [112].

Materials and Data Requirements:

- Microbial genome sequence data (assembled contigs or complete genome)
- High-performance computing resources
- Reference databases: MIBiG, GenBank, antiSMASH database

Software Tools:

- antiSMASH for BGC identification
- BiG-SCAPE for gene cluster family analysis
- PRISM for structural prediction
- NRPSpredictor2 for substrate specificity prediction (for NRPS clusters)

Procedure:

- **Genome Assembly and Quality Assessment:**
  - Assemble sequencing reads into contigs using appropriate assembler (e.g., SPAdes)
  - Assess genome completeness and contamination using CheckM or similar tool
  - Annotate genome using Prokka or RAST

- **BGC Identification with antiSMASH:**
  - Access antiSMASH web server or install local version
  - Input genome sequence in FASTA format
  - Set detection strictness to "relaxed" for comprehensive analysis
  - Enable all extra features: KnownClusterBlast, ClusterBlast, SubClusterBlast, ActiveSiteFinder, Cluster Pfam analysis [112]
  - Execute analysis and retrieve results

- **BGC Analysis and Prioritization:**
  - Assess novelty through comparison with MIBiG database
  - Identify core biosynthetic genes (PKS, NRPS, hybrid, RiPP, terpene, etc.)
  - Evaluate cluster organization and domain architecture
  - Prioritize based on novelty, structural features, or potential bioactivity

- **Substrate Specificity Prediction (for NRPS clusters):**
  - Extract adenylation domain sequences from identified NRPS clusters
  - Input sequences into NRPSpredictor2
  - Retrieve substrate predictions with confidence scores
  - Compile predicted building blocks for structural prediction [112]

- **Gene Cluster Family Analysis:**
  - Input identified BGCs into BiG-SCAPE
  - Define cutoff parameters for gene cluster family formation
  - Analyze phylogenetic relationships between clusters
  - Identify unique gene cluster families lacking characterized representatives [109]

- **Integration with Metabolomic Data:**

- Correlate BGC predictions with LC-MS/MS metabolomic profiles
- Use molecular networking to connect BGCs to metabolic products
- Identify expressed compounds potentially encoded by target BGCs

Table 2: Experimental Parameters for LC-MS/MS Dereplication

Parameter	Specification	Purpose	Reference
Column Type	Reversed-phase C18 (2.1 × 150 mm, 1.8 µm)	Optimal separation of diverse natural products	[114]
Mass Accuracy	<5 ppm	Confident elemental composition assignment	[113]
Collision Energy	10-62 eV (range); 50 eV (standard)	Comprehensive fragmentation data	[113] [114]
Acquisition Mode	DDA & DIA (SWATH)	Complementary spectral information	[114]
Ionization Mode	Positive ESI with formic acid modification	Enhanced detection of diverse compound classes	[113] [114]

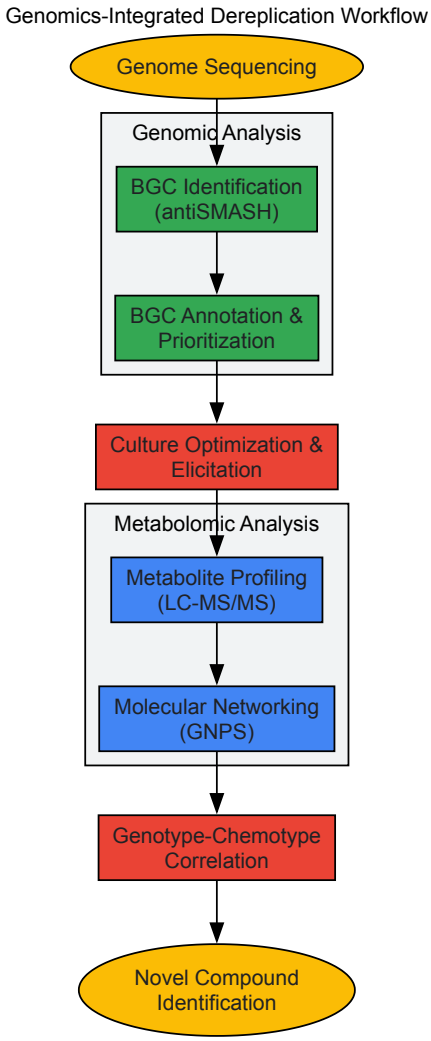
The Scientist's Toolkit: Essential Research Reagents and Materials

Table 3: Essential Research Reagents and Materials for Dereplication Studies

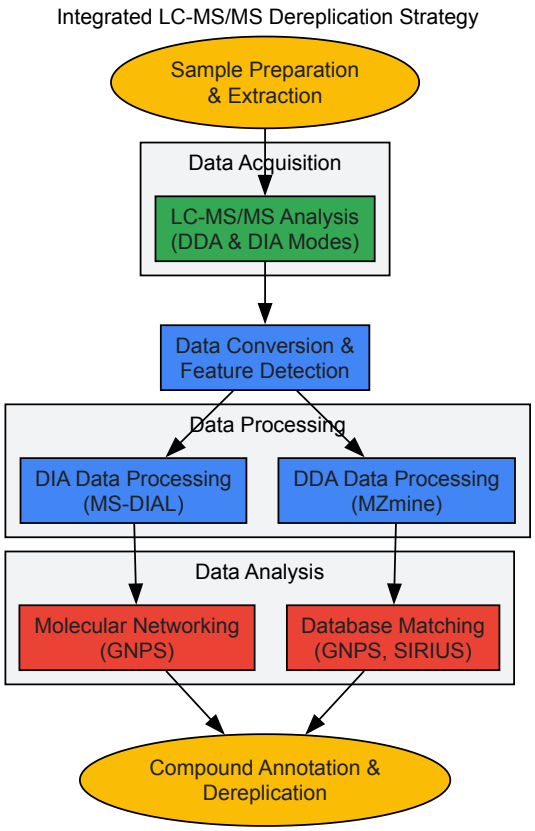
Category	Specific Items	Function/Application	Technical Notes	
Chromatography	UPLC system with C18 column (1.8 µm), methanol, acetonitrile, formic acid, ammonium acetate	Separation of complex natural product mixtures	Gradient elution with volatile modifiers enhances separation and ionization	[113] [114]
Mass Spectrometry	High-resolution mass spectrometer (Q-TOF or Orbitrap), calibration solutions	Accurate mass measurement and fragmentation data	Mass accuracy <5 ppm enables confident formula assignment	[113] [110]
Bioinformatics	GNPS, antiSMASH, SIRIUS, MZmine, MS-DIAL	Data processing, molecular networking, genome mining	Integrated workflow from raw data to compound annotation	[106] [108] [114]
Genomic Analysis	DNA extraction kits, sequencing services, genome assembly software	BGC identification and analysis	Essential for genomics-integrated dereplication	[108] [109]
Reference Standards	Compound standards for relevant chemical classes	Retention time alignment and spectral validation	In-house library development improves dereplication accuracy	[113] [114]
Sample Preparation	Solvents (methanol, chloroform, water), sonicator, centrifuge, filtration units	Extraction of metabolites from biological material	Standardized protocols ensure reproducibility	[112] [114]

Visualization of Integrated Dereplication Workflows

Genomics-Integrated Dereplication Workflow



Integrated LC-MS/MS Dereplication Strategy



Dereplication frameworks have evolved from simple chemical screening approaches to sophisticated integrated systems that combine advanced analytical technologies, bioinformatics, and genomic data [106] [108]. Within the context of BGC research, dereplication has expanded beyond merely identifying known compounds to encompass the comprehensive analysis of an organism's genetic potential for natural product synthesis [109] [112]. The integration of LC-MS/MS-based molecular networking, genomics-guided discovery, and ligand affinity-based screening creates a powerful multidimensional approach that significantly accelerates the identification of novel bioactive compounds while minimizing resource expenditure on rediscovered molecules [111] [114] [110].

The future of dereplication lies in the continued integration of complementary technologies and data streams. Advances in instrumental analysis, particularly in mass spectrometry resolution and sensitivity, will enhance our ability to detect and characterize minor compounds in complex mixtures [106] [107]. The expansion of spectral and genomic databases will improve annotation accuracy, while machine learning approaches will enable more sophisticated prediction of chemical structures from both spectral and genomic data [106] [110]. As these technologies mature, dereplication will increasingly shift from primarily identifying known compounds to proactively predicting and prioritizing novel chemical entities, ultimately transforming natural product discovery from a screening process to a targeted exploration of nature's chemical diversity.

Conclusion

The identification and characterization of biosynthetic gene clusters has transformed natural product discovery, bridging genomic potential with chemical reality through integrated computational and experimental approaches. The field has progressed from foundational BGC recognition to sophisticated activation and expression systems that unlock previously inaccessible chemical diversity. Current methodologies now enable researchers to efficiently navigate the vast genomic landscape, prioritize promising targets, overcome expression barriers, and validate connections between genetic architecture and bioactive molecules. Future directions will likely focus on leveraging artificial intelligence for more accurate BGC prediction, developing universal heterologous expression platforms, and creating high-throughput workflows that further accelerate discovery. As these technologies mature, systematic exploration of BGCs promises to yield novel therapeutic compounds addressing urgent medical needs, particularly in combating antibiotic-resistant pathogens and treating complex diseases. The continued convergence of bioinformatics, synthetic biology, and analytical chemistry will ensure BGC identification remains a cornerstone of drug discovery pipelines, tapping into nature's largely unexplored biochemical repertoire for the next generation of medicines.

References

- 1. Metabolic gene cluster - Wikipedia [https://en.wikipedia.org/wiki/Metabolic_gene_cluster]
- 2. New Approaches to Detect Biosynthetic Gene Clusters in ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC6473659/]
- 3. Insights into secondary metabolism from a global analysis ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC4123684/]
- 4. Computational advances in biosynthetic gene cluster ... [https://www.sciencedirect.com/science/article/abs/pii/S0734975025000187]
- 5. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity - PMC [https://pmc.ncbi.nlm.nih.gov/articles/PMC7849407/]
- 6. A systematic analysis of biosynthetic gene clusters in the ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC4164201/]
- 7. Genomic insights into biosynthetic gene cluster diversity ... [https://www.nature.com/articles/s41598-025-21523-3]
- 8. Diverse secondary metabolites are expressed in particle ... [https://www.nature.com/articles/s41467-023-36026-w]
- 9. Global analysis of biosynthetic gene clusters reveals ... [https://www.nature.com/articles/s41557-022-00923-2]
- 10. A Machine Learning Bioinformatics Method to Predict ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC8243324/]
- 11. Global characterization of biosynthetic gene clusters in non ... [https://www.nature.com/articles/s41598-023-50095-3]
- 12. An interpreted atlas of biosynthetic gene clusters from ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC8126772/]
- 13. Global Analysis of Natural Products Biosynthetic Diversity ... [https://www.mdpi.com/2309-608X/10/9/653]
- 14. Genome mining reveals novel biosynthetic gene clusters in ... [https://www.nature.com/articles/s41598-023-47121-9]
- 15. Mining Natural Product Biosynthesis in Eukaryotic Algae - PMC [https://pmc.ncbi.nlm.nih.gov/articles/PMC7073580/]
- 16. Recent Advances in the Heterologous Expression of ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC9225448/]
- 17. Biopharma Trends 2025 [https://www.bcg.com/publications/2025/biopharma-trends]
- 18. The Minimum Information about a Biosynthetic Gene ... [https://globalplantcouncil.org/the-minimum-information-about-a-biosynthetic-gene-cluster-mibig-database/]
- 19. MIBiG 2.0: a repository for biosynthetic gene clusters of known ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC7145714/]
- 20. MIBiG 3.0: a community-driven effort to annotate ... [https://www.jcvi.org/publications/mibig-30-community-driven-effort-annotate-experimentally-validated-biosynthetic-gene]
- 21. A standardized workflow for submitting data to the Minimum ... [https://environmentalmicrobiome.biomedcentral.com/articles/10.1186/s40793-018-0318-y]
- 22. Plant biosynthetic gene clusters in the context of metabolic ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC9298681/]
- 23. "Biosynthetic Gene Clusters, Microbiomes, and Secondary ... [https://digitalcommons.usf.edu/etd/9524/]
- 24. Biosynthetic Gene Clusters in Sequenced Genomes of Four ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC10269915/]
- 25. Minimum Information about a Biosynthetic Gene cluster | Nature Chemical Biology [https://www.nature.com/articles/nchembio.1890]
- 26. BGC identification and clustering using automated genome ... [https://bio-protocol.org/exchange/minidetail?id=19123911&type=30]

- 27. Identification of biosynthetic gene clusters from ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC5827622/]
- 28. Genome mining reveals the distribution of biosynthetic gene ... [https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-025-11754-z]
- 29. Genome Mining as an Alternative Way for Screening the ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC9394291/]
- 30. Genome mining-driven discovery of enzymes catalyzing ... [https://www.nature.com/articles/s41429-025-00881-0]
- 31. Genome mining for drug discovery: progress at the front end [https://pmc.ncbi.nlm.nih.gov/articles/PMC8788784/]
- 32. Genome mining as a biotechnological tool for the discovery ... [https://www.frontiersin.org/journals/fungal-biology/articles/10.3389/ffunb.2022.993171/full]
- 33. Microbial unusual gene clusters without prominent core ... [https://pubmed.ncbi.nlm.nih.gov/41243002/]
- 34. Genome mining of tailoring enzymes from biosynthetic ... [https://www.sciencedirect.com/science/article/pii/S109671762500120X]
- 35. The Deep Mining Era: Genomic, Metabolomic, and Integrative ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC12299652/]
- 36. Mini review: Genome mining approaches for the ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC7327026/]
- 37. Comprehensive prediction of secondary metabolite ... [https://www.nature.com/articles/s41467-020-19986-1]
- 38. Sequence modeling tools to decode the biosynthetic diversity ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC12282076/]
- 39. antiSMASH 2.0—a versatile platform for genome mining of ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC3692088/]
- 40. BGCFlow: systematic pangenome workflow for the analysis ... [https://pubmed.ncbi.nlm.nih.gov/38686794/]
- 41. Cluster mining tools - SMBP - SecondaryMetabolites.org [https://www.secondarymetabolites.org/mining/]
- 42. Machine learning-enabled genome mining and bioactivity ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC10615616/]
- 43. Deep self-supervised learning for biosynthetic gene cluster ... [https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011162]
- 44. RFBGCpred: A random forest based tool for prediction of ... [https://www.sciencedirect.com/science/article/abs/pii/S0304416525001047]
- 45. A deep learning genome-mining strategy for biosynthetic ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC6765103/]
- 46. Computational advances in biosynthetic gene cluster ... [https://pubmed.ncbi.nlm.nih.gov/39924008/]
- 47. pmobio/Deep-BGCpred: DeepBGCpred - Biosynthetic ... [https://github.com/pmobio/Deep-BGCpred]
- 48. A Case Study of the Biome-BGC Model [https://www.mdpi.com/1999-4907/15/9/1609]
- 49. Bioinformatics tools for the identification of gene clusters that ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC6171489/]
- 50. The PLSDB 2025 update: enhanced annotations and ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC11701622/]
- 51. Identification of Secondary Metabolite Gene Clusters in the ... [https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.01494/full]
- 52. A Systematic Computational Analysis of Biosynthetic Gene ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC4256081/]
- 53. Secondary metabolite biosynthetic gene clusters and ... [https://www.nature.com/articles/s41598-025-03467-w]
- 54. Integrated genome mining and molecular networking uncover ... [https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-025-11966-3]
- 55. MS/MS-Based Molecular Networking: An Efficient ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC9822519/]
- 56. Molecular Networking - GNPS Documentation - GitHub Pages [https://ccms-ucsd.github.io/GNPSDocumentation/networking/]
- 57. Combining Feature-Based Molecular Networking and ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC9610267/]
- 58. Ion identity molecular networking for mass spectrometry ... [https://www.nature.com/articles/s41467-021-23953-9]
- 59. Integrated Metabolomic, Molecular Networking, and ... [https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.906161/full]
- 60. Natural Products Discovery - Kelleher Research Group [https://www.kelleher.northwestern.edu/research/natural-products-discovery/]
- 61. Natural Products and the Gene Cluster Revolution - PMC [https://pmc.ncbi.nlm.nih.gov/articles/PMC5123934/]
- 62. A computational framework to explore large-scale biosynthetic ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC6917865/]
- 63. Activation and Characterization of Lanthomicins A–C by ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC9127795/]
- 64. unlocking silent secondary metabolism with mycolic acid ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC12187902/]
- 65. Refactoring biosynthetic gene clusters for heterologous ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC8238852/]
- 66. Automated genome mining predicts structural diversity and ... [https://elifesciences.org/reviewed-preprints/109154]
- 67. Deciphering the biosynthetic potential of microbial genomes ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC11995264/]
- 68. Recent advances in the direct cloning of large natural ... [https://www.sciencedirect.com/science/article/pii/S2667370323000176]
- 69. New automated method increases the efficiency of bioactive ... [https://chbe.illinois.edu/news/stories/New-automated-method-increases-the-efficiency-bioactive-natural-product-discovery]
- 70. An Improved Transformation-Associated Recombination ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC11626649/]
- 71. Activating cryptic biosynthetic gene cluster through a CRISPR ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC8989516/]
- 72. Transformation-associated recombination (TAR) cloning ... [https://www.oncotarget.com/article/28546/text/]
- 73. A CRISPR-Cas9-Mediated Large-Fragment Assembly ... [https://www.mdpi.com/2076-2607/12/7/1462]
- 74. Bioactive molecules unearthed by terabase-scale long- ... [https://www.nature.com/articles/s41587-025-02810-w]
- 75. Streptomyces as Microbial Chassis for Heterologous ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC8724576/]
- 76. Streptomyces as a versatile host platform for heterologous ... [https://pubs.rsc.org/en/content/articlehtml/2025/np/d5np00036j]

- 77. A highly efficient heterologous expression platform to facilitate ... [https://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-025-02722-z]
- 78. Multi-chassis expression of cyanobacterial and other ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC12416320/]
- 79. Heterologous expression of an in planta-upregulated gene ... [https://pubs.rsc.org/en/Content/ArticleLanding/2025/SC/D5SC05442G]
- 80. Multiplexed mobilization and expression of biosynthetic ... [https://www.nature.com/articles/s41467-022-32858-0]
- 81. Discovery of Cryptic Natural Products Using High-Throughput ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC12303134/]
- 82. Development of a versatile chassis for the efficient ... [https://www.nature.com/articles/s41467-025-62659-0]
- 83. Deep learning driven biosynthetic pathways navigation for ... [https://www.nature.com/articles/s41467-022-30970-9]
- 84. Coupling Mass Spectral and Genomic Information to ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC7998270/]
- 85. Linking biosynthetic and chemical space to accelerate ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC6697067/]
- 86. An integrative genomic and chemical similarity approach ... [https://pubmed.ncbi.nlm.nih.gov/40654994/]
- 87. A comprehensive analysis of human gut microbial ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC12282132/]
- 88. Whole genome analysis and biosynthetic gene cluster ... [https://www.sciencedirect.com/science/article/pii/S071734582500034X]
- 89. Mass spectrometry-based metabolomics approaches to ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC12169106/]
- 90. Comparative Genomics Reveals a Remarkable ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC8407293/]
- 91. Comparative genomics reveals phylogenetic distribution ... [https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4809-4]
- 92. Recent advances in awakening silent biosynthetic gene ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC3117463/]
- 93. Global analyses of biosynthetic gene clusters in ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC10469690/]
- 94. Phylogenetic classification of natural product biosynthetic ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC10663231/]
- 95. Phylogenetic classification of natural product biosynthetic ... [https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1290473/full]
- 96. BGC heteroexpression strategy for production of novel ... [https://www.sciencedirect.com/science/article/abs/pii/S1096717625000540]
- 97. Separation techniques: Chromatography - PMC [https://pmc.ncbi.nlm.nih.gov/articles/PMC5206469/]
- 98. Heterologous Expression of Difficult to Produce Proteins in ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC10815505/]
- 99. New tools for reconstruction and heterologous expression of ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC4742407/]
- 100. Chromatographic Isolation of Sensitive Compounds [https://www.chromatographyonline.com/view/chromatographic-isolation-of-sensitive-compounds-challenges-and-solutions]
- 101. Detecting and prioritizing biosynthetic gene clusters for ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC6449301/]
- 102. Advanced Methods for Natural Products Discovery [https://www.mdpi.com/1660-3397/21/5/308]
- 103. Screening techniques for the identification of bioactive ... [https://www.sciencedirect.com/science/article/pii/S0731708518301948]
- 104. New investigation of encoding secondary metabolites gene ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC11015633/]
- 105. Recent advances in screening active components from natural ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC7606101/]
- 106. Advanced Methods for Natural Products Discovery: Bioactivity ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC10222211/]
- 107. Dereplication of microbial extracts and related analytical ... [https://www.nature.com/articles/ja201412]
- 108. High-Throughput Mining of Novel Compounds from Known ... [https://www.mdpi.com/1420-3049/29/13/3237]
- 109. Novel Gene Clusters for Natural Product Synthesis Are ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC10304795/]
- 110. Dereplication of Natural Product Antifungals via Liquid ... [https://www.mdpi.com/1420-3049/30/1/77]
- 111. An Integrated Strategy for the Detection, Dereplication, and ... [https://pmc.ncbi.nlm.nih.gov/articles/PMC9229839/]
- 112. Identification of the Biosynthetic Gene Cluster ... [https://www.frontiersin.org/journals/chemistry/articles/10.3389/fchem.2020.00397/full]
- 113. Rapid Dereplication of Bioactive Compounds in Plant and ... [https://pubmed.ncbi.nlm.nih.gov/40873593/]
- 114. Dereplication of secondary metabolites from Sophora ... [https://www.nature.com/articles/s41598-025-94958-3]