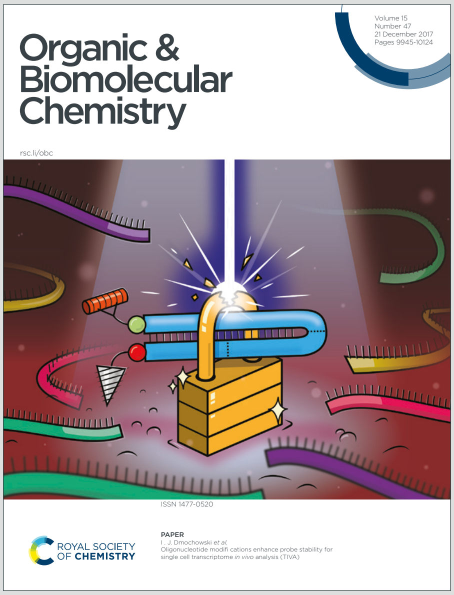
Organic & Biomolecular Chemistry

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: K. L. Steffen, M. Rangel-Grimaldo, T. J. C. Sauters, D. C. Rinker, H. Raja, T. Graf, A. Gumilang, O. L. Riedling, G. Goldman, N. H. Oberlies and A. Rokas, *Org. Biomol. Chem.*, 2025, DOI: 10.1039/D5OB01965F.


Organic & Biomolecular Chemistry
Volume 15
Number 47
21 December 2017
Pages 9945-10124
rsc.li/obc
ISSN 1477-0520
ROYAL SOCIETY OF CHEMISTRY
PAPER
I. J. Dmochowski et al.
Oligonucleotide modifi cations enhance probe stability for single cell transcriptome in vivo analysis (TIVA)

View Article Online
DOI: 10.1039/D5OB01965F

An integrative chemical and genomic similarity approach linking fungal secondary metabolites and biosynthetic gene clusters

Authors

Karin Steffen[1], Manuel Rangel-Grimaldo[2,3], Thomas J. C. Sauters[1], David C. Rinker[1], Huzefa A. Raja[2], Tyler N. Graf[2], Adiyantara Gumilang[1], Olivia L. Riedling[1], Gustavo H. Goldman[4], Nicholas H. Oberlies[2], & Antonis Rokas[1,*]

Addresses

1 Department of Biological Sciences and Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235, USA

2 Department of Chemistry and Biochemistry, University of North Carolina at Greensboro, Greensboro, NC 27402, USA

3 Department of Natural Products, Institute of Chemistry, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico.

4 Faculdade de Ciencias Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto 14040-903, Brazil

Corresponding author: antonis.rokas@vanderbilt.edu

Keywords: Natural product, secondary metabolite (SM), specialized metabolite (SM), biosynthetic gene cluster (BGC), metabologenomics, integrative omics, chemodiversity, secondary metabolite gene cluster (SMGC), extrolites

View Article Online
DOI: 10.1039/D5OB01965F

Abstract

Fungi are well known to biosynthesize structurally complex secondary metabolites (SMs) with diverse bioactivities. These fungal SMs are frequently produced by biosynthetic gene clusters (BGCs). Linking SMs to their BGCs is key to understanding their chemical and biological functions. Reasoning that structural similarity of SMs arises from similarities in the genes involved in their biosynthesis, we developed an integrative approach that leverages known SM-BGC pairs to infer links between detected SMs and genome-predicted BGC regions in fungi. As proof of concept, we structurally characterized 60 metabolites from metabolomic data of 16 strains of the filamentous fungus *Aspergillus fischeri*. Our approach assigned 22 to known SM-BGC pairs and proposed specific links to BGCs and genetic pathways for the remaining 38 metabolites. These results suggest that coupling chemical structure similarity and genomic sequence similarity is a straightforward and high-throughput approach for linking fungal SMs to their BGCs.

Introduction

Secondary or specialized metabolites (SMs), together with allelochemicals, effectors, and extrolites, are all molecules isolated from Nature that are instrumental to fungal ecology (1). Fungal SMs contribute to various functions, including micronutrient acquisition (e.g., siderophores such as ferrichrome (2)), defense (e.g., antibacterials such as penicillin (3)), and pathogenicity (e.g., virulence factors such as gliotoxin (4) and ToxA (5,6)). By virtue of their potent bioactivities, SMs are essential to drug discovery pipelines (7,8) and for medical and agricultural research more broadly (9).

View Article Online
DOI: 10.1039/D5OB01965F

In fungal genomes, the pathways involved in SM biosynthesis typically contain a set of neighboring, co-regulated genes, collectively referred to as biosynthetic gene clusters or BGCs (1). A typical BGC contains genes coding for ‘core’ or ‘backbone-forming’ enzymes responsible for the biosynthesis of the scaffold of the SM, tailoring enzymes that modify the scaffold, and cluster-specific transcription factors and transporters (1). The clustering and content of genes in fungal secondary metabolite pathways led to the development of many different methods to predict BGCs (10). These include CASSIS, a tool for predicting BGCs around a given anchor (or backbone) gene (11); CLOCI, which predicts BGCs based on co-occurring loci and orthologous clusters (12); DeepBGC, a machine learning-based tool trained on distinguishing BGC genomic regions from non-BGC regions in prokaryotic genomes (13); the fai and zol set of tools that employ sequence orthology information for targeted detection of BGCs across genomes (14), and protein domain-based tools like the popular antiSMASH (15) that predict BGC presence using profile hidden Markov models targeting required biosynthetic domains, along with BGC class-specific rules (15,16).

Widespread access of column chromatography coupled with mass spectrometry (i.e., LC-MS and LC-MS/MS or LC-MS^n) has driven the annotation of metabolites from extracts of fungal cultures and even *in situ* from the cultures themselves (17–19). Yet, due to technical limitations, the degree of certainty of an observation of a SM can vary based on the approach used (20). Assigning the chemical identity, and hence structure, of compounds within an extract of an organism can be categorized into four levels of certainty (21): 1) identified compounds for which there are orthogonal supporting structural data, 2) putatively annotated compounds for which there are matches to spectral libraries, 3) putatively characterized compound classes for which

View Article Online
DOI: 10.1039/D5OB01965F

there are matches to the class of compounds, if not the specific compound, and 4) unknown compounds. For the purposes of this report, we focused on the identified compounds (#1 in the list above), where the compounds were isolated and characterized by mass spectrometry and NMR spectroscopy or there were matches to a dereplication database that was built upon fully characterized compounds (22,23).

To date, more than 30,000 fungal metabolites have been characterized (24), and genomic examinations suggest that there are likely millions of predicted BGCs in fungal genomes (25–29). In contrast, there are only about 608 experimentally verified SM-BGC pairs in fungi (27,30–32). This 50-fold discrepancy between identified metabolites and linked BGCs arises largely because SM-BGC pairings are typically established on a case-by-case basis, since confirmation of their pairing requires experimental validation (16,33,34). Thus, the SMs biosynthesized by predicted BGCs in fungal genomes have not yet been discovered, and as such, most of these BGCs are considered “orphans”. Similarly, the biosynthetic pathways responsible for the vast majority of characterized fungal metabolites also remain uncharacterized, hindering efforts to study their biosynthesis.

The very small number of SM-BGC pairs identified to date, coupled with the much larger numbers of fungal metabolites and predicted orphan BGCs in fungal genomes, underscores the need for methods and strategies to predict SM-BGC pairs. To bridge this gap between chemotype and genotype, several general and specific methodologies have been developed to associate SMs and their cognate BGCs (35,16,36,37,34). At the heart of these general approaches lies the independent identification of BGCs via predictions from the genome, and structural

View Article Online
DOI: 10.1039/D5OB01965F

identification of SMs via metabolomics, followed by an algorithm predicting connections. Importantly, many of these algorithms take advantage of the MIBiG database (30,32), a community effort cataloguing BGCs and their SMs, which includes information on the gene/protein sequences of the BGC with their known or putative functions, the organism the SM-BGC pair was identified, and the resulting SM structures and bioactivities.

Strategies have sought to enhance SM-BGC prediction by integrating large metabolomics data. For example, correlation-based approaches statistically associate BGC or gene cluster family (GCF)–SM pairs based on co-occurrence patterns (36), while feature-based approaches rely on specific, searchable attributes (e.g., core enzymes, transcription factors or metabolomic spectral features like fragments and isotopes) to generate “forward” (BGC to SM) or “reverse” (SM to BGC) associations. These approaches have recently uncovered a novel class of BGCs, the isocyanide synthases (37), and linked peptide natural products (e.g., ribosomally synthesized and post-translationally modified peptides (RiPPs) or non-ribosomal peptide synthetases (NRPSs)) to their core genes (35,38,39). Stable isotope labelling has also been used to connect mass spectrometric features (i.e., mass to charge values coupled with chromatographic retention times for metabolites/SMs) to BGCs by tracing the biosynthesis from known BGC substrates (40).

Here, we introduce a new strategy to link the chemical structures of experimentally identified SMs to their cognate BGCs via structural similarity to known SM-BGC pairs. We then applied this strategy to the metabolomes and genomes of 16 strains of the filamentous fungus *Aspergillus fischeri* and the known SM-BGC pairs in the MIBiG database. This enabled us to confidently assign more than one third of detected metabolites to known BGCs that are present in *A. fischeri*

View Article Online
DOI: 10.1039/D5OB01965F

genomes, and generate testable SM-BGC hypotheses in a straightforward, fast and *ab initio* manner for all the remaining SMs. Our results suggest that coupling chemical structure-based similarity with genomic similarity is a powerful approach for linking detected SMs to their BGCs in fungal genomes.

View Article Online
DOI: 10.1039/D5OB01965F

Results

Leveraging chemical and genomic similarity to infer SM-BGC pairs

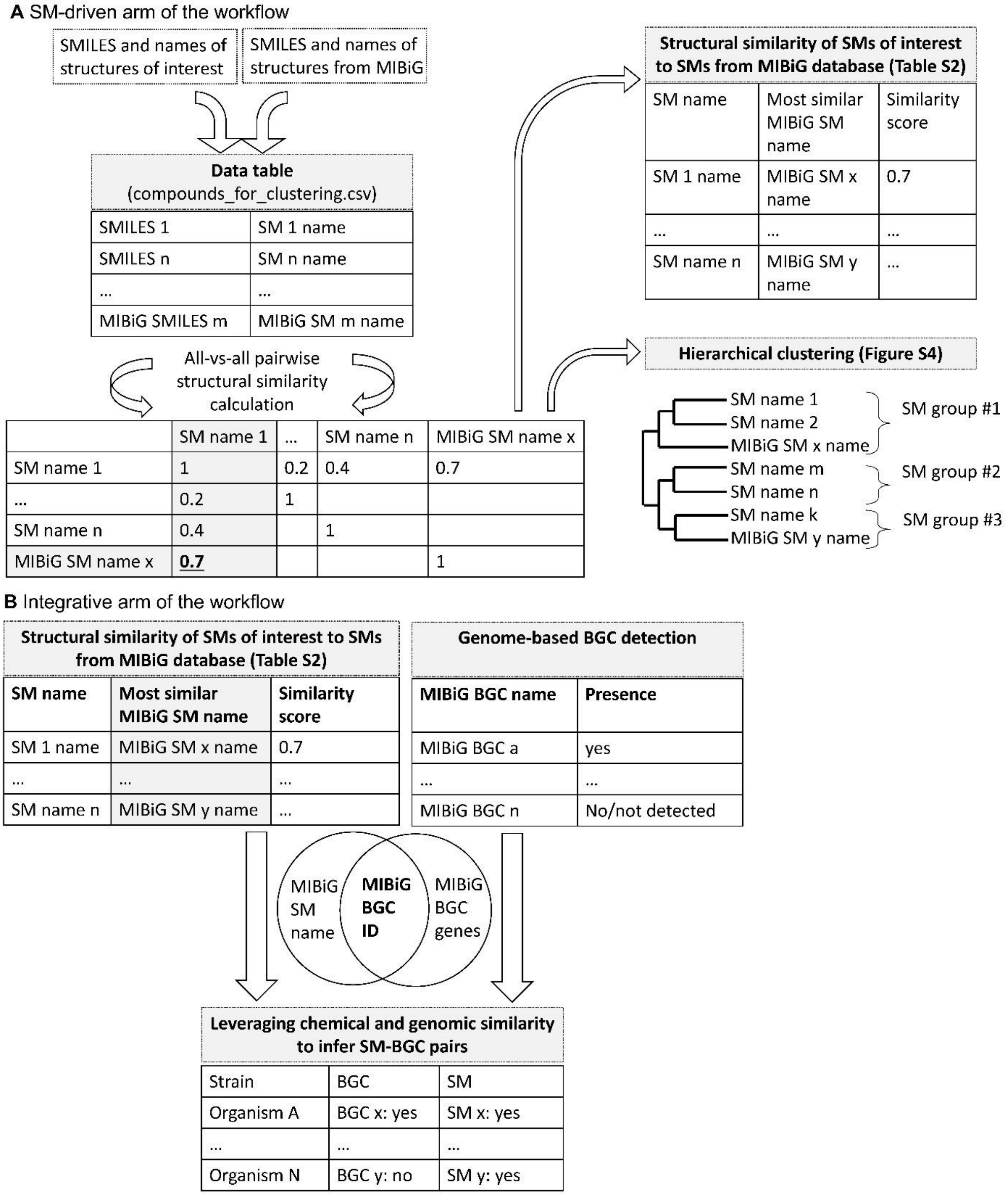

A SM-driven arm of the workflow
SMILES and names of structures of interest
SMILES and names of structures from MIBiG
Data table (compounds_for_clustering.csv)
SMILES 1 | SM 1 name
SMILES n | SM n name
… | …
MIBiG SMILES m | MIBiG SM m name
All-vs-all pairwise structural similarity calculation
| SM name 1 | … | SM name n | MIBiG SM name x
SM name 1 | 1 | 0.2 | 0.4 | 0.7
… | 0.2 | 1
SM name n | 0.4 | | 1
MIBiG SM name x | 0.7 | | | 1
Structural similarity of SMs of interest to SMs from MIBiG database (Table S2)
SM name | Most similar MIBiG SM name | Similarity score
SM 1 name | MIBiG SM x name | 0.7
… | … | …
SM name n | MIBiG SM y name | …
Hierarchical clustering (Figure S4)
SM name 1
SM name 2
MIBiG SM x name
SM group #1
SM name m
SM name n
SM group #2
SM name k
MIBiG SM y name
SM group #3
B Integrative arm of the workflow
Structural similarity of SMs of interest to SMs from MIBiG database (Table S2)
SM name | Most similar MIBiG SM name | Similarity score
SM 1 name | MIBiG SM x name | 0.7
… | … | …
SM name n | MIBiG SM y name | …
Genome-based BGC detection
MIBiG BGC name | Presence
MIBiG BGC a | yes
… | …
MIBiG BGC n | No/not detected
MIBiG SM name
MIBiG BGC ID
MIBiG BGC genes
Leveraging chemical and genomic similarity to infer SM-BGC pairs
Strain | BGC | SM
Organism A | BGC x: yes | SM x: yes
… | … | …
Organism N | BGC y: no | SM y: yes

Figure 1: Schematic of workflow of the SM-BGC co-analyses. A. SM-driven arm of the workflow: All pairwise structural similarities between structures of experimentally

View Article Online
DOI: 10.1039/D5OB01965F

identified SMs and all MIBiG-derived fungal SMs were calculated. From the resulting matrix, the highest structural similarity match between an experimentally identified SM and a MIBiG-derived SM were collected in a table. The matrix was also use to hierarchically cluster structurally similar groups of compounds (i.e., putatively from the same BGC). B. Integrative arm of the workflow: Evaluating the SM-BGC links in the presence of genome-based BGC predictions allowed for orthogonal validation of in silico-predicted BGCs, thereby providing a focused and reliable view of biosynthetic capacities of the fungi.

We developed an integrative approach based on chemical structural similarity to link SMs to BGCs (**Figure 1 A, B**). This approach evaluates structural similarity by matching machine-readable molecular fingerprints from candidate compounds to those stored in the MIBiG database, allowing for the inference of putative SM-BGC relationships. Leveraging the MIBiG database, which contains 3,158 structures from 1,896 bacterial and eukaryotic BGCs, including 692 SMs from 377 fungal BGCs, allows us to connect listed SMs and their BGC genes via their BGC accession IDs (30). Our study demonstrates that metabolomics data from fungal culture extracts can be used to improve the quality and accuracy of genome-based BGC predictions.

Establishing chemical structure similarity

Structural similarity of small molecules can be assessed via digital fingerprints, i.e., a bit vector of each structure generated from SMILES (simplified molecular-input line-entry system, i.e. text abstractions of 2D or 3D structures of molecules) (28,41–43). The similarity between a pair of fingerprints is then expressed using the Tanimoto (Jaccard) index, which is the ratio of the number of shared fingerprint bits (i.e., substructures) to the union of bits in a pairwise comparison. As proposed here, Tanimoto similarity is a heuristic for generating SM-BGC links.

View Article Online
DOI: 10.1039/D5OB01965F

Similarity can be computed between any two given structures and we opted to provide users with the result(s) and leave it up to them to evaluate the quality of the match(es).

The structure similarity linking approach that we employ assumes that SMs from the same BGC are much more similar (as expressed by Tanimoto pairwise similarity) than SMs from different BGCs. To validate this assumption, we calculated the pairwise structural similarity among all SMs in MIBiG (Figure 2). We found that SMs from the same BGC are, on average, significantly more similar than SMs from different BGCs (average Tanimoto pairwise similarity for SMs from the same BGC = 0.568; for SMs from different BGCs = 0.101; permutation test with 1000 permutations gave no permuted statistic as extreme as the observed and a p value ≤ 0.001).

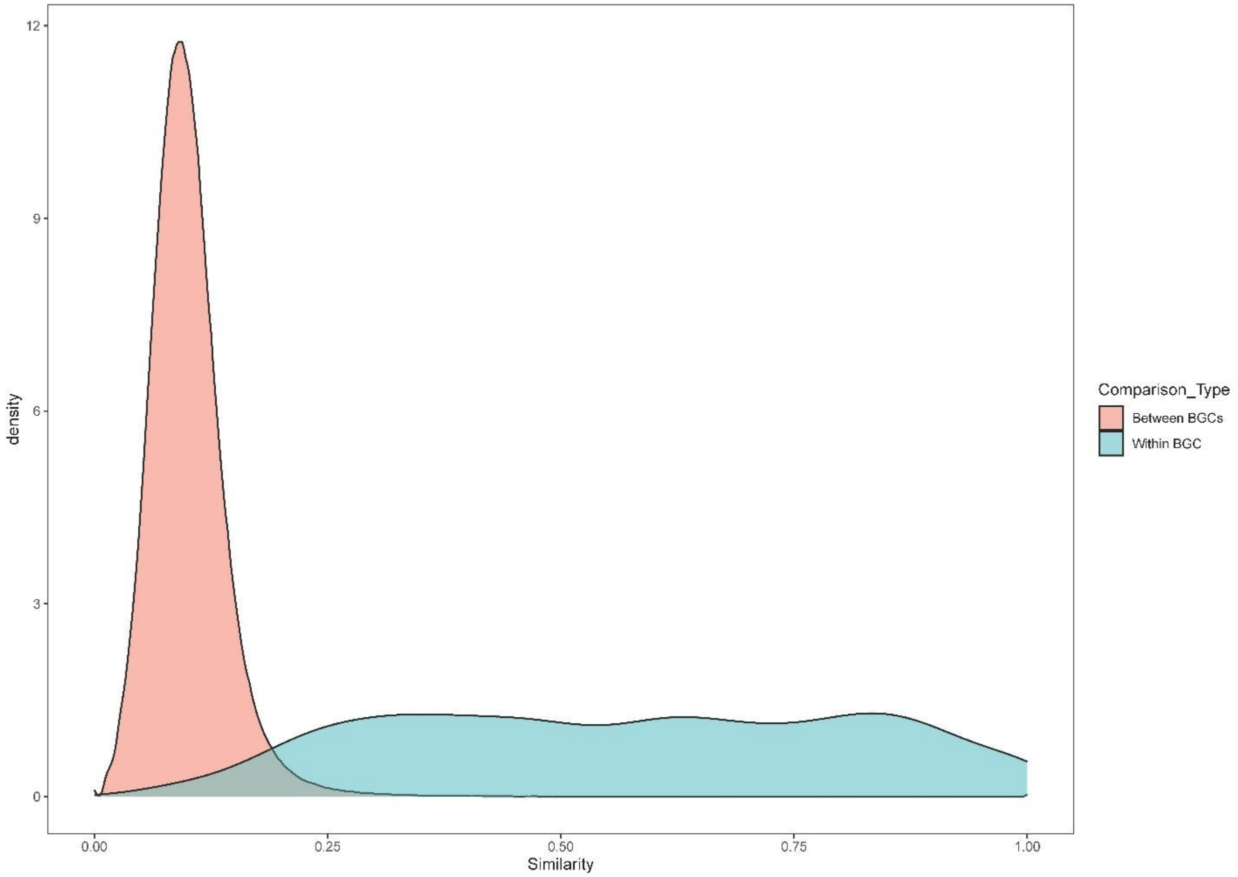


Figure 2: Background distribution of pairwise structure similarities for SMs within the

View Article Online
DOI: 10.1039/D5OB01965F

same BGC (n = 5,889 pairwise comparisons) vs SMs between BGCs (n = 8,586,692 pairwise comparisons). Each density is normalized to integrate to 1; i.e., distributions are shown independent of sample size. *SM structures and their records in MIBiG v4.0 were curated to exclude multiple entries of the same BGCs.

Experimental data: Sixty structurally characterized metabolites from *A. fischeri*

We next applied our approach to a data set containing the metabolomes and genomes of 16 strains of *Aspergillus fischeri*, a filamentous fungus that is gaining attention as a close, non-pathogenic relative of the major human pathogen *Aspergillus fumigatus* (44–46). Using aspects of the “one strain many compounds” approach, the production of SM was evaluated at two temperatures (30°C and 37°C) using UPLC-MS/MS, recently (46,47). In doing so, the number of compounds detected per strain increased,(46) as would have growing them on e.g., different media (44,48). Metabolites were identified based on either a direct match in LC-MS/MS to reference standards, all of which had been fully characterized by NMR, or to a class of fungal metabolites (i.e., via mass defect filtering (22,23)). A total of 60 metabolites were identified at two levels of confidence (**Table 1**, ‘A’ and ‘B’ respectively), subsequently referred to as ‘identified SMs’. Three biological replicates provided insight into the consistency of SM production by the various BGCs and strains (**Figure S1**). Overall, we found the most biosynthetically rich strains across all replicates and temperatures yielded up to three times more SMs than the least-producing strains (e.g., CBS 150748: N=45 vs. CBS 54465 : N=15). Interestingly, strains with the greatest consistency of SM production across all biological replicates produced fewer metabolites (e.g., 18/20 SMs were detected in all replicates of strain

View Article Online
DOI: 10.1039/D5OB01965F

CBS 150750 at 30°C (90%) (**Figure S2**)(46).

Predicting the BGCs linked to experimentally identified SMs

To generate hypotheses about the biosynthetic origin of SMs from *A. fischeri*, we calculated pairwise Tanimoto similarities for all 60 experimentally identified chemical structures from *A. fischeri* and all known SMs from the MIBiG database. We next used the all-versus-all structural similarity matrix to perform hierarchical clustering and generate groups of highly similar SMs (**Table 1**, **Figure 3; Figure S3**). The highest match between an identified SM and an SM (or a set of SMs) from MIBiG, which is already linked to a BGC, enabled us to assign the identified SM to that corresponding BGC; we refer to these assignments as hypothetical SM-BGC links (**Table 1**).

Identified SMs were thus linked to putative BGCs via their highest structural similarity to SMs from MIBiG. In doing so, we generated BGC hypotheses for all 60 identified metabolites from *A. fischeri*. Of these, 22 *A. fischeri* metabolites were identical to SMs in the MIBiG database, i.e., representing known SM-BGC links, and 37 metabolites were structurally similar, but not identical, to SMs in MIBiG (**Table S1** ‘confidence’ column: ‘reported’ and ‘predicted,’ respectively). The sole remaining metabolite is a sterol, which was not linked to a BGC, as sterol biosynthesis is part of primary metabolism (49,50). Structural similarity between identified SMs and SMs in MIBiG varied substantially for the 37 metabolites examined (**Figure 3**). For example, the experimentally identified SM acetylaszonalenin produces an exact match with the acetylaszonalenin SM present in the MIBiG database (**Figure 3A**). Two additional

View Article Online
DOI: 10.1039/D5OB01965F

experimentally identified SMs have a high similarity with acetylaszonalenin: aszonalenin and 4-hyrdroxyaszonalenin. Upon further investigation, the link of aszonalenin with the acetylaszonalenin BGC is confirmed by literature (but not recorded in MIBiG), while the link between 4-hyrdroxyaszonalenin and the acetylaszonalenin BGC remains a hypothetical connection not yet experimentally confirmed. In other cases, such the breviamides (**Figure 3B**), the similarity score between experimentally identified SMs and MIBiG SMs is lower, which suggests that these metabolites may be biosynthesized by a BGC not currently present in the MIBiG database. All SM groups and hypotheses are described in **Table 1**, and are subsequently evaluated in depth.

View Article Online
DOI: 10.1039/D5OB01965F

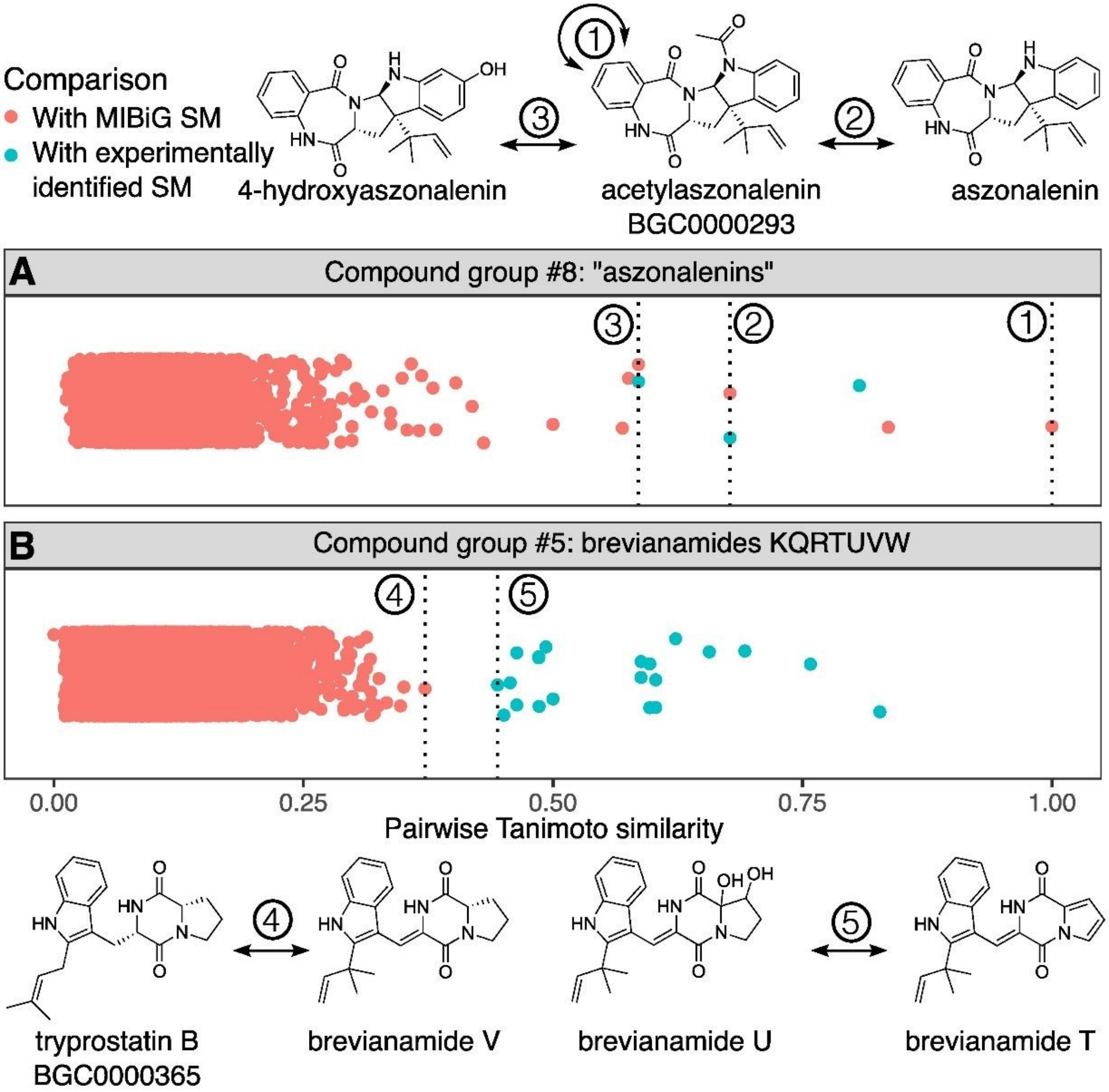


Figure 3. Pairwise structural similarities among secondary metabolites (SMs) within an SM group (blue) and between experimentally identified SMs and all MIBiG metabolites (red). Each dot represents a Tanimoto similarity between two structures. A. Matching an SM group to a known MIBiG metabolites. Three experimentally identified aszonalenin analogs show high mutual similarity (0.59–0.81) and were therefore grouped. Their similarities to MIBiG metabolites are shown in red; dashed lines mark each metabolite's highest MIBiG match. Acetylaszonalenin, which is also present in MIBiG (BGC0000293), matches itself with a similarity of 1 and also shows the highest similarity to aszonalenin and 4-hydroxyaszonalenin. This group was therefore linked to

View Article Online
DOI: 10.1039/D5OB01965F

BGC0000293. B. SM group without a MIBiG counterpart. Seven brevianamides show high within-group similarity (0.44–0.83) and were grouped accordingly. Their similarities to MIBiG SMs (red) show that the closest MIBiG match (tryprostatin B to brevianamide V) falls below the lowest within-group similarity. Although all share the same *L*-Trp/*L*-Pro diketopiperazine core, they differ in prenylation and other modifications. Thus, this group was not linked to any known MIBiG metabolite.

Table 1: The 60 metabolites identified from 16 strains of *A. fischeri* were hierarchically clustered into 25 SM groups based on structural similarity. Each group was assigned an arbitrary identifier (i.e., 1 to 25). The superscript after the SM name indicates the level of experimental support: $^{\mathbf{A}}$ MS/MS and NMR or MS/MS and dereplication with in-house database/standard; $^{\mathbf{B}}$ MS/MS only. For each SM, the BGC(s) linked by structural similarity clustering are indicated, with the underlined BGCs yielding the highest Tanimoto similarity match. Hypothetical links that were confirmed post-hoc based on experimental data (e.g., identical SM structures, evidence from the literature) are denoted as ‘reported’, and all newly generated hypotheses without additional evidence are denoted as ‘predicted’. For SMs of known BGCs, all generated hypotheses were accurate. For an overview of all structurally similar metabolites from *A. fischeri* together with their top SM hits in MIBiG database, where available, see **Figure S5**.

SM group #	SM	BGC link	BGC present	confidence	Reference
1	Ilicicolin E ^{B}	BGC0001923, BGC0001924 (New BGC 1)	no	predicted	(51)
2	(3β,22E)-Ergosta-4,6,8(14),22-tetraene-3-ol ^{A}	(primary metabolism)	–	–	(49)
3	Fumagillol ^{B}	BGC0001067	yes	reported	(52)
4	Brevianamide A/B ^{B}	BGC0001084, BGC0000816 (New BGC 2)	no	predicted	
4	Brevianamide C/D ^{B}	BGC0001084, BGC0000816 (New BGC 2)	no	predicted	
5	Brevianamide Q ^{B}	BGC0000442 (New BGC 3)	no	predicted	
5	Brevianamide R ^{B}	BGC0000442 (New BGC 3)	no	predicted	
5	Brevianamide T ^{B}	BGC0000442 (New BGC 3)	no	predicted	
5	Brevianamide U ^{B}	BGC0000442 (New BGC 3)	no	predicted	
5	Brevianamide V/W ^{B}	BGC0000356 (New BGC 3)	no	predicted	
5	Brevianamide K ^{B}	BGC0000442 (New BGC 3)	no	predicted	
6	Cottoquinazoline E ^{A}	BGC0000355	putative	predicted	(53,54)
6	Cottoquinazoline F ^{A}	BGC0000355	putative	predicted	(53,54)
6	Cottoquinazoline G ^{A}	BGC0000355	putative	predicted	(53,54)
7	Fumitremorgin F ^{B}	BGC0001142, BGC0000355 (New BGC 4)	no	predicted	(55)
7	Fumitremorgin G/L ^{B}	BGC0001142, BGC0000355 (New BGC 4)	no	predicted	(55)

View Article Online
DOI: 10.1039/D5OB01965F

8	4-Hydroxyaszonalenin [B]	BGC0000293, (BGC0002272)	yes	predicted	(56)
8	Acetylaszonalenin [A]	BGC0000293, (BGC0002272)	yes	reported	(56)
8	Aszonalenin [A]	BGC0000293, (BGC0002272)	yes	reported	(56,57)
9	Isoroquefortine C [B]	BGC0000420	yes	reported	(58,59)
9	Roquefortine C [B]	BGC0000420	yes	reported	(58,59)
10	Brevianamide E [B]	BGC0002272, BGC0002617	no	predicted	
11	13-O-prenylfumitremorgin B [A]	BGC0000356	yes	predicted	(60,61)
11	Brevianamide F [B]	BGC0000356	yes	reported	(60,61)
11	Deoxybrevianamide E [B]	BGC0000356	yes	predicted	(60,61)
11	Fumitremorgin A [A]	BGC0000356	yes	reported	(62)
11	Fumitremorgin B [A]	BGC0000356	yes	reported	(60,61)
11	Fumitremorgin C [A]	BGC0000356	yes	reported	(60,61)
11	spiro[5H,10H-dipyrrolo-[1,2-a:1′,2′-d]pyrazine-2-(3H),2′-[2H]-indole]-3′,5,10(1′H)trione [A]	BGC0000356	yes	predicted	(63)
11	Tryprostatin B [B]	BGC0000356	yes	reported	(60,61)
11	Tryprostatin C/D [B]	BGC0000356	yes	predicted	(60,61)
11	Verruculogen [B]	BGC0000356	yes	reported	(60,61)
12	hexadehydroastechrome (monomer) [B]	BGC0000372	yes	reported	(64)
12	Trihistatin [A]	BGC0000420, BGC0000372	yes	predicted	(64)
13	16-O-deacetyl helvolic acid 21,16-lactone [B]	BGC0000686	yes	predicted	(65)
13	Helvolic acid [A]	BGC0000686	yes	reported	(65)
14	Pyripyropene F [B]	BGC0000129, BGC0001068	yes	predicted	(66)
14	Pyripyropene H [B]	BGC0000129, BGC0001068	yes	predicted	(66)
14	Pyripyropene I [B]	BGC0000129, BGC0001068	yes	predicted	(66)
14	Pyripyropene O [B]	BGC0000129, BGC0001068	yes	predicted	(66)
15	Azonapyrone A [A]	BGC0002604	yes	predicted	(67)
15	Sartorypyrone A [A]	BGC0002604	yes	reported	(67)
16	Circumdatin C [A]	BGC0000355, BGC0001652, BGC0000448, BGC0000409, BGC0000303 (New BGC 5)	no	predicted	
16	Dimetoxycircumdatin C [A]	BGC0000355, BGC0001652, BGC0000448,	no	predicted	

View Article Online
DOI: 10.1039/D5OB01965F

		BGC0000409, BGC0000303 (New BGC 5)			
17	Betaenone E ^{B}	BGC0002165, BGC0001264 (New BGC 6)	no	predicted	(68)
17	Betaenone G/I/J ^{B}	BGC0002165, BGC0001264 (New BGC 6)	no	predicted	(68)
17	Betaenone H ^{B}	BGC0002165, BGC0001264 (New BGC 6)	no	predicted	(68)
18	Clavaric acid ^{B}	BGC0001248	yes	reported	(69,70)
19	Chaetoglobosin 542 ^{B}	BGC0002539, BGC0000968, BGC0001182	yes	predicted	(71)
20	Neosartoricin ^{B}	BGC0001144	yes	reported	(72,73)
20	Neosartoricin C ^{B}	BGC0001144	yes	reported	(72,73)
20	Neosartoricin D ^{B}	BGC0001144	yes	reported	(72,73)
21	Brevianamide L ^{B}	BGC0002208, BGC0002242 (New BGC 7)	no	predicted	(74)
21	Brevianamide O ^{B}	BGC0002208, BGC0002242 (New BGC 7)	no	predicted	(74)
21	Brevianamide P ^{B}	BGC0002208, BGC0002242 (New BGC 7)	no	predicted	(74)
22	Secalonic acids (A/ B/ C/ D/ E/ F/ F1/ G; 4,4'-Secalonic acid E) ^{B}	BGC0002063, BGC0001886, BGC0001988	yes	reported	(75)
23	Nidiascin C ^{A}	BGC0002275, BGC0002171 (New BGC 8)	no	predicted	
24	Neosartorin ^{A}	BGC0001988	yes	reported	(76)
25	Bisdethiobis(methylthio)-gliotoxin ^{A}	BGC0000361	yes	reported	(4)
25	Gliotoxin ^{A}	BGC0000361	yes	reported	(4)

Assigning BGCs to identical pairs of structures

There were 13 *A. fischeri* SMs that had an identical SM structure included in the MIBiG database (**Table 1**). While unsurprising and seemingly trivial, the ability of our approach to quickly assign BGCs for experimentally identified SMs also present in the MIBiG database offers considerable practical utility, since the natural products literature does not dictate a consistent nomenclature process for SMs, which makes lookups by name futile. Lack of well-catalogued data further complicates fast identification (43).

View Article Online
DOI: 10.1039/D5OB01965F

The 13 SMs identified from *A. fischeri* with an identical SM match in the MIBiG database are: acetylaszonalenin, brevianamide F, clavaric acid, fumagilol, fumitremorgin B and C, helvolic acid, hexadehydroastechrom, neosartorin, roquefortine C, sartorypyrone A, tryprostatin B, and verruculogen. Notably, three known SM-BGC pairs were missed by our structure similarity approach due to database limitations. These were bisdethiobis(methylthio)gliotoxin and gliotoxin, both produced by gliotoxin BGC0000361, which was retired in MIBiG v3.1, and secalonic acid(s) for which the SM(s) were not structurally identified in our study nor when describing the BGC (and hence neither in the corresponding MIBiG BGC0001886 entry).

Uncovering BGCs for SMs not present in the MIBiG database

Not all known biosynthetic intermediates, shunt products or possible SMs are deposited in the MIBiG database. Thus, six additional SM-BGC links were confirmed based on primary literature. These are aszonalenin in BGC0000293 (56), fumitremorgin A in BGC0000356 (62), isoroquefortine C in BGC0000420(58), and neosartoricin, neosartoricin C, and neosartoricin D in BGC0001144 (72). Notably, isoroquefortine C is an artifact produced by the isomerization of roquefortine C caused by pH or light (58). Similarly, neosartoricin C and D might be artifacts related to the production of neosartoricin B (72). Indeed, artifacts – compounds that were isolated but whose structure slightly differs from the true SM, possibly due to extraction solvents or sample handling – are a well-known challenge in the natural products literature (77). Finally, fumitremorgin A is technically not considered a product of the verruculogen BGC (BGC0000356), as the gene encoding the FtmPT3 protein responsible for converting verruculogen to fumitremorgin A is not part of the BGC (62). However, this variation in the degree of biosynthetic gene clustering is not unusual (78). Given that fumitremorgin A is

View Article Online
DOI: 10.1039/D5OB01965F

produced from verruculogen, an SM of this BGC, it is reasonable to include it in the set of SMs attributed to BGC0000356. In summary, our approach directly assigned BGCs for 22 of the 60 experimentally identified SMs (36%).

The remaining 38 *A. fischeri* identified metabolites did not have identical matches to SM structures included in the MIBiG database or biosynthetic information in the literature. Thus, we augmented the SM-BGC hypotheses for each of these metabolites based on structural similarity by examining whether *A. fischeri* genomes contained the SM-linked BGCs (for details on BGC prediction/detection, see Methods and the next section). Our predictions can be broadly grouped into three level-of-confidence categories: (i) attributing the metabolite to a known BGC that is present in the respective *A. fischeri* strain genome(s) (e.g., 4-hydroxyaszonalenin – BGC0000293, **Figure 3A**), (ii) linking the SM to a BGC not present in the respective *A. fischeri* genome(s), and (iii) ascribing the SM (or SM group) as a novel metabolite(s) likely encoded by an unknown BGC (i.e., no similar SMs are present in the MIBiG database, **Figure 3B**). Given the dearth of fungal BGCs in MIBiG (i.e., only 377), we were pleased that our approach predicted 13 SMs in category (i), 11 SMs in category (ii), and 13 SMs in category (iii) (see extended **Table S1** for more details, and **Table S2** for all Tanimoto similarities). Notably, we found that the BGCs associated with 38 of the 59 predicted SM-BGC links (64%) are in the curated list of *A. fischeri* BGCs (**Table S1**). For the remaining "undetected" BGCs, hypotheses to explain this pattern are consistent with the presence of a homologous but divergent/convergent BGC or with genome incompleteness. The latter possibility is less likely because the estimated genome completeness is very high (46).

View Article Online
DOI: 10.1039/D5OB01965F

Genomic characterization of *A. fischeri* BGCs

To evaluate the hypotheses generated for the 38 remaining metabolites without known BGCs, we next examined the BGC content of the *A. fischeri* genomes. We first analyzed the 16 genomes using antiSMASH v7, which predicts 'BGC regions' –i.e., continuous stretches in the genome containing BGC(s) and other genes (**Figure 4**). For traceability, we also identified and grouped homologous BGCs across the individual genomes. Across all 16 genomes, antiSMASH predicted 44 BGC regions that corresponded to 42 unique BGCs (BGC0001248 and BGC0002710 were each detected in two regions of the genomes), as well as 20 candidate BGC regions ('unnamed' or 'orphan' putative BGCs). The mean number of BGCs per strain was 53.3 (range 51–56), a number consistent with previous reports (79). Note that we refer to these predicted BGCs by the accession numbers of their reference BGCs in the MIBiG database.

antiSMASH-predicted BGCs were classified as 'present' in *A. fischeri* when they contained all the genes present in the reference MIBiG entry, or when they were incomplete but supported by evidence from structurally identified SMs. Additionally, BGCs were classified as 'putative' in *A. fischeri* when they were incomplete with at least half of the genes detected but without evidence from the metabolomics study. Otherwise, BGCs were classified as 'absent' (i.e., fewer than half of the genes were found and no evidence from metabolomics was present). Examining each of the 44 antiSMASH-predicted BGC regions across *A. fischeri* genomes, we classified 20 BGCs as 'present', 9 as 'putative', and 15 as 'absent' (**Table S3**). We also specifically searched for the protein sequences of each MIBiG BGC in the RNAseq-based gene annotations(46), allowing us

View Article Online
DOI: 10.1039/D5OB01965F

to manually curate and revise the antiSMASH predictions. These additional analyses enabled us to classify 7 additional BGCs as ‘present’ and 4 BGCs as ‘putative’. A full list of BGCs is given in **Table S3**, with information on sequence identity with known BGC genes and genomic location in **Table S4**. Subsequently, we detail issues and difficulties in faithfully assessing the number and identity of BCGs present in genomes.

Among all 40 BGCs classified as ‘present’ or ‘putative’, five artifacts reduce the total BGC count. These primarily stem from the current cataloging approach for BGCs and the scientific community’s limited understanding of them. MIBiG defines each BGC in the genome it is reported in, sometimes listing the 'same' (i.e., homologous) BGC multiple times from different organisms. Similarly, one BGC that biosynthesizes one SM can be nested within another, larger BGC that biosynthesizes a different SM. These situations can lead to ‘collisions’, i.e. the assignment of the same set of genes to multiple BGCs.

There were collisions in two pairs of BGCs where the same set of proteins in *A. fischeri* is classified as two different BGCs due to similarity of the MIBiG reference sequences (BGC0000361 gliotoxin / BGC0001609 gliovirin, and BGC0001144 neosartoricin B / BGC0002646 hancockinone A), reducing the number of unique BGCs by two. The BGC for biotin is listed twice in the MIBiG database (BGC0001238 and BGC0001239) but was counted only once, as it matches the same set of genes. Similarly, there are two slightly different BGCs matching a congruent set of genes for the metabolite ilicicolin H (BGC0002035 and BGC0002093), which were counted as one BGC, further decreasing the total count by two. Additionally, the BGC for clavaric acid (BGC0001248), which is composed of a single gene,

View Article Online
DOI: 10.1039/D5OB01965F

was found twice (**Table S4**). However, only one of the two homologs identified (homolog ID 221721_1) contains the sequence motif VSDCISE, which was previously found in *Fusarium graminearum* to be involved in clavaric acid production (70).

In total, we infer that *A. fischeri* contains 35 'present' and 'putative' BGCs (**Figure 4**, **Table S3**). Overall, BGC content was largely conserved and consistent across the 16 strains, with most BGCs (82%; 29/35 total 'present' and 'putative' BGCs) detected in all strains.

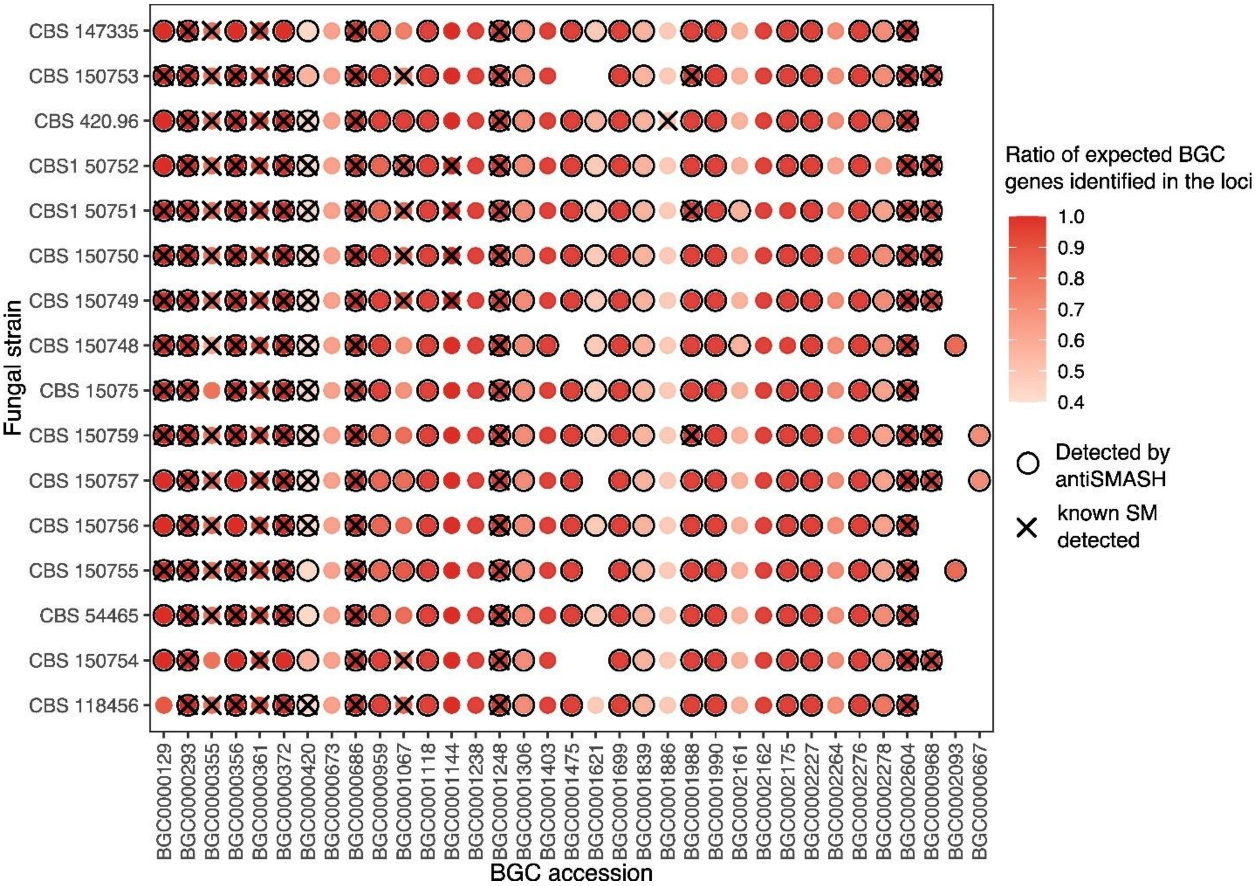


Figure 4: Map of BGC and SM presence in *A. fischeri*. BGCs across the 16 strains of *Aspergillus fischeri*. The black circle around a given data point indicates the BGC was detected by antiSMASH in the respective genome. The fill indicates the BGC completeness (ratio of expected to verified genes). The x denotes instances in which a known SM for a given BGC was identified in the respective strain. The five artifactual BGCs, antiSMASH-predicted BGCs found 'absent', and unnamed BGC regions were not included. For a complete evaluation of all antiSMASH-predicted BGCs, see **Table S3**. Empirically, we find that the lowest number of verified genes for an active BGC (SM

View Article Online
DOI: 10.1039/D5OB01965F

identified), is in BGC0000420, where we detect 3 or 4 out of 7 expected genes. This aligns with our threshold of 50% of genes present for 'putatively present' BGCs.

Caveats for using antiSMASH as tool for accurate BGC surveys

We chose the comparison and validation of antiSMASH, since it is a widely used (and very useful) tool for BGC prediction in fungal genomes (34). antiSMASH is designed to discover regions containing known or novel BGCs (15). In practice, the tool is frequently also used to discover BGCs (rather than regions containing BGCs) in fungal genomes, with the results being taken at face value without further scrutiny. While examining the correspondence between *A. fischeri* BGCs identified by antiSMASH and their inferred references in the MIBiG database, we noted five sources of error associated with the common practice of conflating the BGC regions identified by antiSMASH with individual BGCs.

First, in most known BGCs, the locus predicted by antiSMASH to contain a BGC was much larger (up to approximately three times the number of genes) than the actual BGC. This is by design, as BGC boundaries are difficult to define (formerly possible with CASSIS (11)) and hence the more relaxed/inclusive 'region' concept in antiSMASH (**Figure 4, 5A**). Second, as BGCs are known to co-localize, particularly in telomeric or low complexity regions of genomes (80,81), their physical proximity on chromosomes, in combination with this 'regions' concept, can lead to BGCs masking each other (**Figure 5B**). This masking occurred in the proximal BGCs BGC0001403 for trypacidin and BGC0001988 for neosartorin, and with BGC0000356 for verruculogen and BGC0001067 for fumagillin. Examination of 16 strains of *A. fischeri* revealed some instances where the same homologous genes were predicted as part of different BGCs in different strains (**Figure 5C**). Third, at low identities, the BGC predictions from the module '--

View Article Online
DOI: 10.1039/D5OB01965F

cb-knownclusters' may be misleading/arbitrary as we found several instances of the same orthologous region being labeled as different BGCs. A fourth source of inaccuracies stems from version differences of the MIBiG database used for the BGC prediction. Curation processes continually expand the knowledge base (32), but sometimes, valid BGCs are removed or lacking, thus leading to missed predictions (e.g., the extensively studied gliotoxin BGC, which was retired, i.e. removed in MIBiG v3.1/v4.0) (**Figure 4**). Finally, we noted instances of BGCs missed by antiSMASH (but detected by protein sequence searches) for reasons that are not apparent (**Figure 4**).

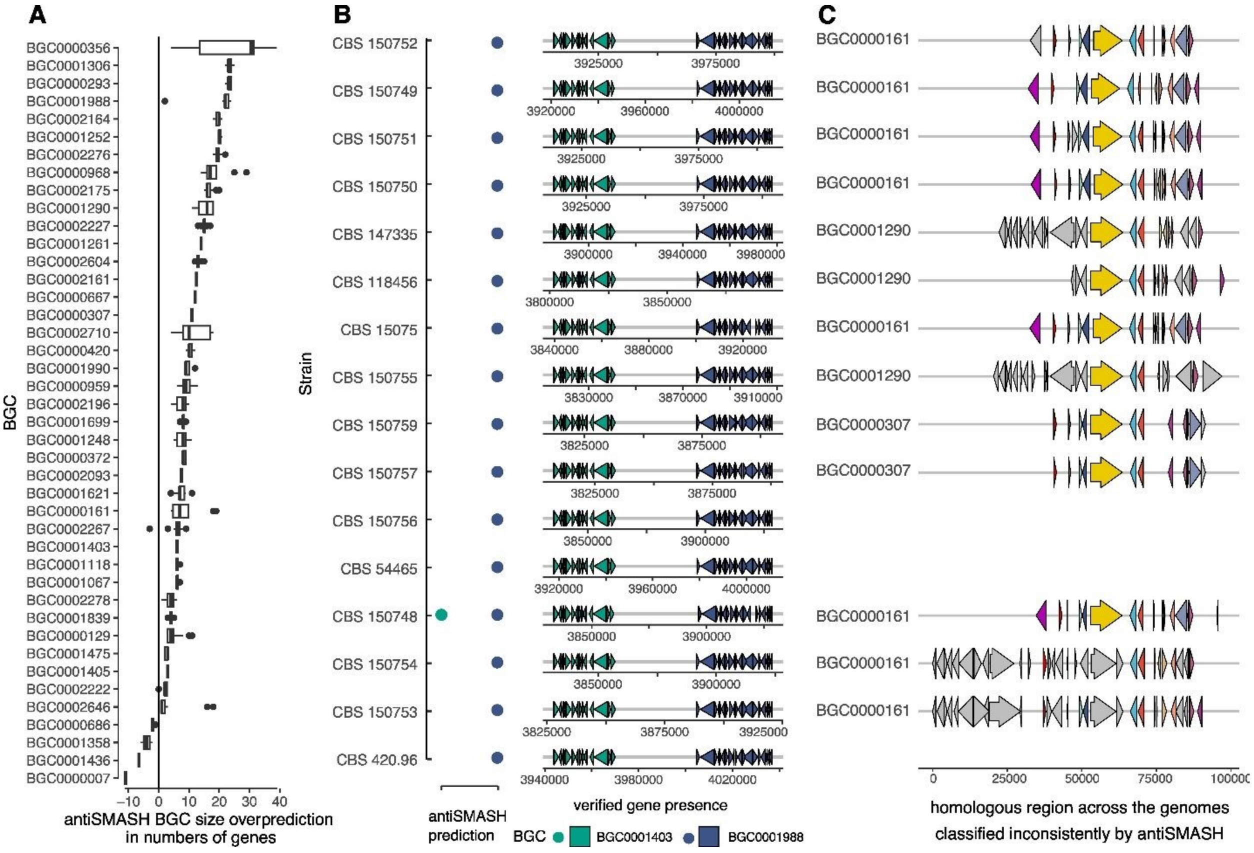


Figure 5. antiSMASH-predicted BGC regions in fungal genomes do not correspond to predicted BGCs. There are five reasons for this lack of correspondence, including (A) region overprediction, (B) merging/masking, and (C) inconsistent BGC assignments. **A. Region overprediction:** Overprediction is defined as the difference between the

View Article Online
DOI: 10.1039/D5OB01965F

number of genes included in the antiSMASH "region" and the true BGC boundaries. Predicted regions frequently extend into neighboring BGCs, which can promote artificial merging of adjacent clusters (see B). **B. Merging and masking:** Example of BGC0001403 and BGC0001988, two adjacent clusters in *A. fischeri*. Their proximity leads antiSMASH to merge them in all but one genome, causing BGC0001988 to mask BGC0001403. Although both BGCs occur in all strains, antiSMASH failed to list BGC0001403 in 15 of 16 cases. **C. Inconsistent prediction across strains:** Using the same strains as in panel B, ortholog tracking shows that an identical genomic region was assigned to different BGC accessions (BGC0000161, BGC0000307, BGC0001290). Plotting all orthologs attributed to BGC0001290 illustrates that the same gene set was labeled as three different BGCs across strains.

Discussion

There are at least 30,000 reported fungal metabolites (24,82–84) and millions of BGCs predicted in fungal genomes (26–28) but only a few hundred SM-BGC pairs (32), suggesting that linking SMs and BGCs remains challenging. To address this challenge, we developed an SM-BGC linking approach based on chemical similarity, that requires a minimum of input data (e.g., a single SM) and can be performed using experimental data or data retrieved from natural product databases (82–85). Across 16 strains of a single fungal species, our approach recovered 22 known SM-BGC pairs and generated hypotheses for 37 more, including 11 that could be SMs attributed to BGCs present in MIBiG. Thus, our approach efficiently automated SM-driven linking of SM and BGCs, and faithfully recovered known connections, additional links not included in MIBiG, and new hypotheses. This approach offers two advantages: (1) it can provide orthogonal BGC validation (in case of known links i.e., Tanimoto similarity =1), and (2) it can generate hypotheses for SMs whose biosynthetic pathways are not known (i.e., Tanimoto similarity <1).

Method development in BGC detection from genomic data has produced many tools (e.g., SMURF (86), antiSMASH (15), BiG-SCAPE (87), cblaster (88), DeepBGC (13), BGCFlow

View Article Online
DOI: 10.1039/D5OB01965F

(89), CLOCI (12), zol and fai (14)). Additionally, SM-BGC links can be established via correlation analyses, an approach termed 'metabologenomics', or specific experiments such as 'IsoAnalyst' (40). Metabologenomics can yield *de novo* SM-BGC links, but requires large datasets (> 100 species) from extensive experimental data as well as sophisticated fine-tuning of scoring functions and parameters, dependent on BGC class (NP Linker)(36,90). These genome- or BGC-driven innovations stand in contrast to the number of integrative tools for linking SMs (SANDPUMA (91), GNP (92), PARAS (93)), which typically are limited to specific taxa or classes of SMs. Furthermore, only two tools are available that can link SMs to BGCs: RIPPminer (94) and Prism (95,96), which again are limited to specific classes of SMs or taxa. The method outlined here fills a gap, where the strategy of connecting SMs to BGC is agnostic to chemical structural class, organism, data size or specificity.

As a consequence of the aforementioned challenges in SM-BGC linking, existing strategies for straightforward orthogonal validation of BGCs are lacking. SM-BGC links are typically validated via gene knock-out studies (e.g., (65,97,98)). In contrast, *in silico* tools linking SMs to BGCs deliver unvalidated predictions or connections. The wealth of BGC prediction tools with various strategies, focused on specific BGC classes or more general tools (10) poses a challenge because presence/absence or identity of a predicted BGC are frequently a function of arbitrary cutoffs (**Figure 5**). BGCs can be interpreted with some fluidity, e.g., many genes in described BGCs are of unknown function and may not be essential to the BGC, synteny conservation is sometimes low, and with increasing phylogenetic/evolutionary distance, gene and protein sequences naturally diverge. Using metabolomics as orthogonal validation can be a means to

View Article Online
DOI: 10.1039/D5OB01965F

avoid arbitrary thresholds confirming the presence of a BGC with the unambiguous presence of its SM product(s).

In our structural similarity analyses, we refrained from setting a similarity threshold, due to the known patchiness of SMs present in the MIBiG database (i.e., only 692 fungal SMs out of >30,000 reported in the literature), general limitations in SM-BGC pairing knowledge, and previously documented challenges with threshold-based approaches (36). Moreover, some SMs (particularly biosynthetic intermediates) are not necessarily unique to any single BGC. For example, the diketopiperazine brevianamide F (cyclo-L-Trp-L-Pro) is the first product of the biosynthesis of verruculogen by BGC000356 in *A. fumigatus* (60) as well as of the biosynthesis of notoamide A by BGC0000818 in *Aspergillus versicolor* (99) and brevianamide A by the *bvn* gene cluster (currently not listed in MIBiG) in *Penicillium brevicompactum* NRRL 864 (100).

Our chemical structure similarity approach also has caveats. Our results are based on the examination of strains of an *Aspergillus* species, one of the most well studied fungal genera in terms of prior knowledge of SM-BGC pairs. Studies of less-studied organisms may be more challenging, especially if their chemodiversity differs from the SMs currently represented in the MIBiG database, resulting in hypotheses (SM-BGC links and SM groups) that may be a poor fit. As we have shown, some published SM-BGC pairs are not currently included in the MIBiG database. Yet, as databases grow, so does the utility of this approach. This methodology could further be expanded to work on partial structures or *m/z* fingerprints in the same manner as using SMILES as input. Additionally, chemical conversions that alter the backbone or skeleton of a SM sufficiently could mask a better clustering fit. Furthermore, when SMs are produced by

View Article Online
DOI: 10.1039/D5OB01965F

multiple non-homologous BGCs (e.g., brevianamide F), genomic evidence is necessary to determine which BGC it is produced by. Such instances of convergently evolved SMs would only be detected in this strategy when finding the SM and not the BGC (but this inference would be based on the absence of evidence). Putative examples of this in our data are chaetoglobosin 542 and ilicicolin E. Chaetoglobosin 542 is structurally very similar to chaetoglobosin A produced by BGC0000968 (101), which is similar to two different *A. fischeri* BGCs. Interestingly, the presence-absence patterns of the BGC and the SM match only for one of the BGCs. In the second case, ilicicolin E differs from ascochlorin of BGC0001923 (51) only by the presence of an α,β-unsaturated ketone instead of the aliphatic ketone, respectively, in the 6-membered ring. However, *A. fischeri* genomes do not contain any related BGCs, suggesting that the observed structural similarity of the two SMs may result from convergent evolution. Of course, another possibility is that the BGC is present in the genome but not part of the genome assembly (e.g., because it resides in an otherwise highly repetitive region).

These caveats notwithstanding, our approach successfully inferred SM-BGC pairs for nearly one third of the fungal metabolites identified and predicted SM-BGC pair hypotheses for nearly all the rest. Ultimately, our approach is a hypothesis-generating strategy and must be validated experimentally (e.g., by modifying putative BGCs in the native host or through heterologous expression of the putative BGC)(33,102). The approach applied in this work leveraged similarity among known SM structures and BGCs to bidirectionally link SMs and BGCs via the MIBiG database and thereby successfully generated testable biosynthetic hypotheses in a high-throughput fashion and validated the presence of predicted BGCs. This increases the fidelity of the biological conclusions drawn based on the BGCs and their implications for the chemotype

View Article Online
DOI: 10.1039/D5OB01965F

(i.e., SM profile), lifestyle, and niche of an organism. While our approach is hypothesis-generating and requires further validation, it can augment the fidelity of stand-alone tools that operate solely either on metabolomic or genomic data.

Methods

All genomic and metabolomic data were taken from Rinker *et al*.(46) and are available via FigShare (https://doi.org/10.6084/m9.figshare.25316452).

Chemical fingerprinting and clustering

Structures (SMILES, simplified molecular-input line-entry system; a text string representing the molecule) for all SMs identified from untargeted metabolomics were collected via ChemDraw v23.1.1 (Revvity) and combined with structures from known BGCs deposited in the MIBiG database (32). Chemoinformatic analyses were carried out in Jupyter notebook(103) using RDKit and PubChemPy(104). To facilitate the subsequent search for the detected metabolites, we prefixed the names of structures from MIBiG with the BGC accession ID, and those of metabolites found in extracts with 'chem_'. For comparing structural similarity and clustering the metabolites, we calculated the Morgan fingerprint for each metabolite with GetMorganFingerprintAsBitVect() using chirality with a radius of 2, and 2048 bits, and converted the fingerprints to binary strings using ToBitString(). We calculated Tanimoto similarity (Jaccard index, the intersection of set bitflags divided by the union) between all pairwise comparisons using calculate_tanimoto() resulting in a symmetric similarity matrix of all-vs-all comparisons. With linkage(method='average', metric='euclidean') and dendrogram()

View Article Online
DOI: 10.1039/D5OB01965F

from scipy, we performed hierarchical clustering of the metabolites based on the distance matrix and used matplotlib to plot and save the resulting figure (**Figure S3**). SM groups were initially delineated by searching the dendrogram for the tag "chem_" and grouping similar structures. Subsequently, for every identified SM, the highest pairwise similarity scores with a SM (or a set of SMs) in the MIBiG database was extracted from the similarity matrix, thereby generating biosynthetic hypotheses for each SM. To test validity, we performed bootstrap resampling (1000 replicates) yielding a 95% confidence interval of [0.475, 0.498] for the median difference, confirming the robustness of the result.

BGC predictions

BGCs were predicted using antiSMASH v7.1.0(15) and DIAMOND v2.1.6.160 blastp searches(105) of the MIBiG database v3.1(30). All subsequent analyses were performed in R v4.4.0(106). Conventionally, BGCs are defined in a specific genome. However, in this manuscript, we refer to the predicted candidate BGCs by their MIBiG accession number for convenience.

After the antiSMASH prediction (--fullhmmer --rre --cc-MIBiG --cb-knownclusters --cb-subclusters --cb-general, using the corresponding gff3 annotation file), we aggregated results from individual runs into a single file. Across the different strains, known and unknown BGCs were aggregated by comparing gene content. This approach yielded meaningful clusters as evidenced by the correct grouping of known BGCs (with their MIBiG BGC accession ID). This clustering revealed instances in which the same genes were attributed to different BGCs (both known and "anonymous" candidate clusters) by antiSMASH.

View Article Online
DOI: 10.1039/D5OB01965F

For the amino acid sequence search, the 16 genomes were queried with all sequences in MIBiG v3.1 (diamond blastp -f6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qcovhsp qlen slen full_sseq) and the results concatenated.

To validate the antiSMASH BGC predictions, the genes in each predicted region were searched using DIAMOND blastp. Additionally, the hits were filtered for high identity (pident >80%, minimum 50% query coverage), as well as for runs of hits against the same BGC in proximity (low identity clustering of putative, diverged BGCs). By using DIAMOND blastp to confirm *A. fischeri* BGC genes based on known BGCs, we tagged every BGC gene with a BGC ID from MIBiG hence allowing for interoperability of biological and chemical data.

BGCs were classified as present if all genes were found in proximity, regardless of whether a corresponding SM was detected, or if they were recovered partially, i.e. incomplete but with evidence from SMs. BGCs were classified as putative if more than half of the genes were present but there was no evidence for their presence based on metabolomics. BGCs were classified as absent if fewer than half of the genes were found and no evidence from metabolomics was present.

Additional data (https://figshare.com/s/27b1a13ca534c1e646f4) and analysis code (https://figshare.com/s/a2c267ec94e82e062bdd) for this study can be found on FigShare.

View Article Online
DOI: 10.1039/D5OB01965F

Funding information

KS was supported by a PostDoc stipend of the Swedish Pharmaceutical Society. Research in the AR lab is supported by the National Science Foundation (DEB-2110404) and the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R01 AI153356). OLR was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 2444112. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. GHG thanks the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação Coordenação de Aperfeiçoamento do Pessoal do Ensino Superior (CAPES) grant number 405934/2022-0 (The National Institute of Science and Technology INCT Funvir), and CNPq 301058/2019-9 from Brazil.

Competing Interest Statement

AR is a scientific consultant for LifeMine Therapeutics, Inc. NHO has ownership interests in Ionic Pharmaceuticals, LLC and is a member of the Scientific Advisory Board of Mycosynthetix, Inc. HAR, TNG, and N.H.O. are members of the Scientific Advisory Board of Clue Genetics, Inc. KS is a data scientist at Olink part of Thermo Fisher Scientific.

View Article Online
DOI: 10.1039/D5OB01965F

References

1. Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. Nat Rev Microbiol. 2019 Mar;17(3):167–80.

2. Wiemann P, Lechner BE, Baccile JA, Velk TA, Yin WB, Bok JW, et al. Perturbations in small molecule synthesis uncovers an iron-responsive secondary metabolite network in *Aspergillus fumigatus*. Front Microbiol [Internet]. 2014 Oct 24 [cited 2024 Nov 14];5. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2014.00530/abstract

3. Fleming A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. Br J Exp Pathol. 1929;10(3):226–36.

4. Dolan SK, O'Keeffe G, Jones GW, Doyle S. Resistance is not futile: gliotoxin biosynthesis, functionality and utility. Trends Microbiol. 2015 July;23(7):419–28.

5. Tomas A. Purification of a Cultivar-Specific Toxin from *Pyrenophora tritici-repentis,* Causal Agent of Tan Spot of Wheat. Mol Plant Microbe Interact. 1990;3(4):221.

6. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, et al. Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet. 2006 Aug;38(8):953–6.

7. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod. 2020 Mar 27;83(3):770–803.

8. the International Natural Product Sciences Taskforce, Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov. 2021 Mar;20(3):200–16.

9. Niego AGT, Lambert C, Mortimer P, Thongklang N, Rapior S, Grosse M, et al. The contribution of fungi to the global economy. Fungal Divers. 2023 July;121(1):95–137.

10. Weber T, Kim HU. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. Synth Syst Biotechnol. 2016 June;1(2):69–79.

11. Wolf T, Shelest V, Nath N, Shelest E. CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Bioinformatics. 2016 Apr 15;32(8):1138–43.

View Article Online
DOI: 10.1039/D5OB01965F

12. Konkel Z, Kubatko L, Slot JC. CLOCI: unveiling cryptic fungal gene clusters with generalized detection. Nucleic Acids Res. 2024 July 17;gkae625.

13. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 2019 Oct 10;47(18):e110–e110.

14. Salamzade R, Tran PQ, Martin C, Manson AL, Gilmore MS, Earl AM, et al. zol and fai: large-scale targeted detection and evolutionary investigation of gene clusters. Nucleic Acids Res. 2025 Jan 24;53(3):gkaf045.

15. Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res. 2023 July 5;51(W1):W46–50.

16. Van Der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH. Linking genomics and metabolomics to chart specialized metabolic diversity. Chem Soc Rev. 2020;49(11):3297–314.

17. Oberlies NH, Knowles SL, Amrine CSM, Kao D, Kertesz V, Raja HA. Droplet probe: coupling chromatography to the *in situ* evaluation of the chemistry of nature. Nat Prod Rep. 2019;36(7):944–59.

18. Jarmusch SA, Van Der Hooft JJJ, Dorrestein PC, Jarmusch AK. Advancements in capturing and mining mass spectrometry data are transforming natural products research. Nat Prod Rep. 2021;38(11):2066–82.

19. Dong Y, Aharoni A. Image to insight: exploring natural products through mass spectrometry imaging. Nat Prod Rep. 2022;39(7):1510–30.

20. Alseekh S, Aharoni A, Brotman Y, Contrepois K, D'Auria J, Ewald J, et al. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Nat Methods. 2021 July;18(7):747–56.

21. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics. 2007 Sept 19;3(3):211–21.

22. El-Elimat T, Figueroa M, Ehrmann BM, Cech NB, Pearce CJ, Oberlies NH. High-Resolution MS, MS/MS, and UV Database of Fungal Secondary Metabolites as a Dereplication Protocol for Bioactive Natural Products. J Nat Prod. 2013 Sept 27;76(9):1709–16.

23. Paguigan ND, El-Elimat T, Kao D, Raja HA, Pearce CJ, Oberlies NH. Enhanced dereplication of fungal cultures via use of mass defect filtering. J Antibiot (Tokyo). 2017 May;70(5):553–61.

View Article Online
DOI: 10.1039/D5OB01965F

24. Bérdy J. Thoughts and facts about antibiotics: Where we are now and where we are heading. J Antibiot (Tokyo). 2012 Aug;65(8):385–95.

25. Amos GCA, Awakawa T, Tuttle RN, Letzel AC, Kim MC, Kudo Y, et al. Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. Proc Natl Acad Sci [Internet]. 2017 Dec 26 [cited 2024 Nov 15];114(52). Available from: https://pnas.org/doi/full/10.1073/pnas.1714381115

26. Kautsar SA, Blin K, Shaw S, Weber T, Medema MH. BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res. 2021 Jan 8;49(D1):D490–7.

27. Palaniappan K, Chen IMA, Chu K, Ratner A, Seshadri R, Kyrpides NC, et al. IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. Nucleic Acids Res. 2019 Oct 29;gkz932.

28. Zhang S, Shi G, Xu X, Guo X, Li S, Li Z, et al. Global Analysis of Natural Products Biosynthetic Diversity Encoded in Fungal Genomes. J Fungi. 2024 Sept 13;10(9):653.

29. Riedling OL, Rokas A. mGem: How many fungal secondary metabolites are produced by filamentous fungi? Conservatively, at least 1.4 million. Rodrigues M, editor. mBio. 2025 Oct 8;16(10):e01381-25.

30. Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. 2023 Jan 6;51(D1):D603–10.

31. Riedling O, Walker AS, Rokas A. Predicting fungal secondary metabolite activity from biosynthetic gene cluster data using machine learning. Anderson MZ, editor. Microbiol Spectr. 2024 Feb 6;12(2):e03400-23.

32. Zdouc MM, Blin K, Louwen NLL, Navarro J, Loureiro C, Bader CD, et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Res. 2024 Dec 9;gkae1115.

33. Kjærbølling I, Mortensen UH, Vesth T, Andersen MR. Strategies to establish the link between biosynthetic gene clusters and secondary metabolites. Fungal Genet Biol. 2019 Sept;130:107–21.

34. Lv HW, Tang JG, Wei B, Zhu MD, Zhang HW, Zhou ZB, et al. Bioinformatics assisted construction of the link between biosynthetic gene clusters and secondary metabolites in fungi. Biotechnol Adv. 2025 July;81:108547.

35. Dejong CA, Chen GM, Li H, Johnston CW, Edwards MR, Rees PN, et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nat Chem Biol. 2016 Dec;12(12):1007–14.

View Article Online
DOI: 10.1039/D5OB01965F

36. Caesar LK, Butun FA, Robey MT, Ayon NJ, Gupta R, Dainko D, et al. Correlative metabologenomics of 110 fungi reveals metabolite–gene cluster pairs. Nat Chem Biol. 2023 July;19(7):846–54.

37. Nickles GR, Oestereicher B, Keller NP, Drott MT. Mining for a new class of fungal natural products: the evolution, diversity, and distribution of isocyanide synthase biosynthetic gene clusters. Nucleic Acids Res. 2023 Aug 11;51(14):7220–35.

38. Kersten RD, Yang YL, Xu Y, Cimermancic P, Nam SJ, Fenical W, et al. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. Nat Chem Biol. 2011 Nov;7(11):794–802.

39. Behsaz B, Bode E, Gurevich A, Shi YN, Grundmann F, Acharya D, et al. Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery. Nat Commun. 2021 May 28;12(1):3225.

40. McCaughey CS, Van Santen JA, Van Der Hooft JJJ, Medema MH, Linington RG. An isotopic labeling approach linking natural products with biosynthetic gene clusters. Nat Chem Biol. 2022 Mar;18(3):295–304.

41. Voser TM, Campbell MD, Carroll AR. How different are marine microbial natural products compared to their terrestrial counterparts? Nat Prod Rep. 2022;39(1):7–19.

42. Morgan HL. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. J Chem Doc. 1965 May 1;5(2):107–13.

43. Steffen K, Oberlies NH, Rokas A. Machine-Readable Structural Information Is Essential for Natural Products Research. J Nat Prod. 2025 Nov 28;88(11):2815–21.

44. Mead ME, Knowles SL, Raja HA, Beattie SR, Kowalski CH, Steenwyk JL, et al. Characterizing the Pathogenic, Genomic, and Chemical Traits of *Aspergillus fischeri* , a Close Relative of the Major Human Fungal Pathogen *Aspergillus fumigatus*. Mitchell AP, editor. mSphere. 2019 Feb 27;4(1):e00018-19.

45. Rokas A. Evolution of the human pathogenic lifestyle in fungi. Nat Microbiol. 2022 May 4;7(5):607–19.

46. Rinker DC, Sauters TJC, Steffen K, Gumilang A, Raja HA, Rangel-Grimaldo M, et al. Strain heterogeneity in a non-pathogenic *Aspergillus fungus* highlights factors associated with virulence. Commun Biol. 2024 Sept 4;7(1):1082.

47. Bode HB, Bethe B, Höfs R, Zeeck A. Big Effects from Small Changes: Possible Ways to Explore Nature's Chemical Diversity. ChemBioChem. 2002 July 3;3(7):619.

48. VanderMolen KM, Raja HA, El-Elimat T, Oberlies NH. Evaluation of culture media for the production of secondary metabolites in a natural products screening program. AMB Express. 2013 Dec;3(1):71.

View Article Online
DOI: 10.1039/D5OB01965F

49. Desmond E, Gribaldo S. Phylogenomics of Sterol Synthesis: Insights into the Origin, Evolution, and Diversity of a Key Eukaryotic Feature. Genome Biol Evol. 2009 Jan 1;1:364–81.

50. Dhingra S, Cramer RA. Regulation of Sterol Biosynthesis in the Human Fungal Pathogen *Aspergillus fumigatus*: Opportunities for Therapeutic Development. Front Microbiol [Internet]. 2017 Feb 1 [cited 2025 Feb 6];8. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2017.00092/full

51. Araki Y, Awakawa T, Matsuzaki M, Cho R, Matsuda Y, Hoshino S, et al. Complete biosynthetic pathways of ascofuranone and ascochlorin in *Acremonium egyptiacum*. Proc Natl Acad Sci. 2019 Apr 23;116(17):8269–74.

52. Lin HC, Chooi YH, Dhingra S, Xu W, Calvo AM, Tang Y. The Fumagillin Biosynthetic Gene Cluster in *Aspergillus fumigatus* Encodes a Cryptic Terpene Cyclase Involved in the Formation of β- *trans* -Bergamotene. J Am Chem Soc. 2013 Mar 27;135(12):4616–9.

53. O'Hanlon KA, Gallagher L, Schrettl M, Jöchl C, Kavanagh K, Larsen TO, et al. Nonribosomal Peptide Synthetase Genes *pesL* and *pes1* Are Essential for Fumigaclavine C Production in Aspergillus fumigatus. Appl Environ Microbiol. 2012 May;78(9):3166–76.

54. Ames BD, Haynes SW, Gao X, Evans BS, Kelleher NL, Tang Y, et al. Complexity Generation in Fungal Peptidyl Alkaloid Biosynthesis: Oxidation of Fumiquinazoline A to the Heptacyclic Hemiaminal Fumiquinazoline C by the Flavoenzyme Af12070 from *Aspergillus fumigatus*. Biochemistry. 2011 Oct 11;50(40):8756–69.

55. Gao X, Chooi YH, Ames BD, Wang P, Walsh CT, Tang Y. Fungal Indole Alkaloid Biosynthesis: Genetic and Biochemical Investigation of the Tryptoquialanine Pathway in *Penicillium aethiopicum*. J Am Chem Soc. 2011 Mar 2;133(8):2729–41.

56. Yin WB, Grundmann A, Cheng J, Li SM. Acetylaszonalenin Biosynthesis in *Neosartorya fischeri*. J Biol Chem. 2009 Jan;284(1):100–9.

57. Wakana D, Hosoe T, Itabashi T, Nozawa K, Kawai K ichi, Okada K, et al. Isolation of Isoterrein from Neosartorya fischeri. Mycotoxins. 2006;56(1):3–6.

58. Shangguan N, Hehre WJ, Ohlinger WS, Beavers MP, Joullié MM. The Total Synthesis of Roquefortine C and a Rationale for the Thermodynamic Stability of Isoroquefortine C over Roquefortine C. J Am Chem Soc. 2008 May 1;130(19):6281–7.

59. García-Estrada C, Ullán RV, Albillos SM, Fernández-Bodega MÁ, Durek P, von Döhren H, et al. A Single Cluster of Coregulated Genes Encodes the Biosynthesis of the Mycotoxins Roquefortine C and Meleagrin in Penicillium chrysogenum. Chem Biol. 2011 Nov;18(11):1499–512.

60. Maiya S, Grundmann A, Li S, Turner G. The Fumitremorgin Gene Cluster of *Aspergillus fumigatus* : Identification of a Gene Encoding Brevianamide F Synthetase. ChemBioChem. 2006 July 3;7(7):1062–9.

View Article Online
DOI: 10.1039/D5OB01965F

61. Grundmann A, Kuznetsova T, Afiyatullov SSh, Li S. FtmPT2, an *N* -Prenyltransferase from *Aspergillus fumigatus* , Catalyses the Last Step in the Biosynthesis of Fumitremorgin B. ChemBioChem. 2008 Sept;9(13):2059–63.

62. Mundt K, Wollinsky B, Ruan H, Zhu T, Li S. Identification of the Verruculogen Prenyltransferase FtmPT3 by a Combination of Chemical, Bioinformatic and Biochemical Approaches. ChemBioChem. 2012 Nov 26;13(17):2583–92.

63. Tsunematsu Y, Ishikawa N, Wakana D, Goda Y, Noguchi H, Moriya H, et al. Distinct mechanisms for spiro-carbon formation reveal biosynthetic pathway crosstalk. Nat Chem Biol. 2013 Dec;9(12):818–25.

64. Yin WB, Baccile JA, Bok JW, Chen Y, Keller NP, Schroeder FC. A Nonribosomal Peptide Synthetase-Derived Iron(III) Complex from the Pathogenic Fungus *Aspergillus fumigatus*. J Am Chem Soc. 2013 Feb 13;135(6):2064–7.

65. Lv JM, Hu D, Gao H, Kushiro T, Awakawa T, Chen GD, et al. Biosynthesis of helvolic acid and identification of an unusual C-4-demethylation process distinct from sterol biosynthesis. Nat Commun. 2017 Nov 21;8(1):1644.

66. Itoh T, Tokunaga K, Matsuda Y, Fujii I, Abe I, Ebizuka Y, et al. Reconstitution of a fungal meroterpenoid biosynthesis reveals the involvement of a novel family of terpene cyclases. Nat Chem. 2010 Oct;2(10):858–64.

67. Wang WG, Du LQ, Sheng SL, Li A, Li YP, Cheng GG, et al. Genome mining for fungal polyketide-diterpenoid hybrids: discovery of key terpene cyclases and multifunctional P450s for structural diversification. Org Chem Front. 2019;6(5):571–8.

68. Li H, Hu J, Wei H, Solomon PS, Stubbs KA, Chooi Y. Biosynthesis of a Tricyclo[6.2.2.0 ^{2,7}]dodecane System by a Berberine Bridge Enzyme-Like Aldolase. Chem – Eur J. 2019 Nov 27;25(66):15062–6.

69. Godio RP, Fouces R, Martín JF. A Squalene Epoxidase Is Involved in Biosynthesis of Both the Antitumor Compound Clavaric Acid and Sterols in the Basidiomycete H. sublateritium. Chem Biol. 2007 Dec;14(12):1334–46.

70. Godio RP, Martín JF. Modified oxidosqualene cyclases in the formation of bioactive secondary metabolites: Biosynthesis of the antitumor clavaric acid. Fungal Genet Biol. 2009 Mar;46(3):232–42.

71. Perlatti B, Nichols CB, Lan N, Wiemann P, Harvey CJB, Alspaugh JA, et al. Identification of the Antifungal Metabolite Chaetoglobosin P From Discosia rubi Using a Cryptococcus neoformans Inhibition Assay: Insights Into Mode of Action and Biosynthesis. Front Microbiol. 2020 July 28;11:1766.

72. Yin WB, Chooi YH, Smith AR, Cacho RA, Hu Y, White TC, et al. Discovery of Cryptic Polyketide Metabolites from Dermatophytes Using Heterologous Expression in *Aspergillus nidulans*. ACS Synth Biol. 2013 Nov 15;2(11):629–34.

View Article Online
DOI: 10.1039/D5OB01965F

73. Chooi YH, Fang J, Liu H, Filler SG, Wang P, Tang Y. Genome Mining of a Prenylated and Immunosuppressive Polyketide from Pathogenic Fungi. Org Lett. 2013 Feb 15;15(4):780–3.

74. Zheng L, Wang H, Ludwig-Radtke L, Li SM. Oxepin Formation in Fungi Implies Specific and Stereoselective Ring Expansion. Org Lett. 2021 Mar 19;23(6):2024–8.

75. Neubauer L, Dopstadt J, Humpf HU, Tudzynski P. Identification and characterization of the ergochrome gene cluster in the plant pathogenic fungus Claviceps purpurea. Fungal Biol Biotechnol. 2016 Dec;3(1):2.

76. Matsuda Y, Gotfredsen CH, Larsen TO. Genetic Characterization of Neosartorin Biosynthesis Provides Insight into Heterodimeric Natural Product Generation. Org Lett. 2018 Nov 16;20(22):7197–200.

77. Capon RJ. Extracting value: mechanistic insights into the formation of natural product artifacts – case studies in marine natural products. Nat Prod Rep. 2020;37(1):55–79.

78. Rokas A, Wisecaver JH, Lind AL. The birth, evolution and death of metabolic gene clusters in fungi. Nat Rev Microbiol. 2018 Dec;16(12):731–44.

79. Steenwyk JL, Mead ME, Knowles SL, Raja HA, Roberts CD, Bader O, et al. Variation Among Biosynthetic Gene Clusters, Secondary Metabolite Profiles, and Cards of Virulence Across *Aspergillus* Species. Genetics. 2020 Oct 1;216(2):481–97.

80. Keller NP, Turner G, Bennett JW. Fungal secondary metabolism — from biochemistry to genomics. Nat Rev Microbiol. 2005 Dec;3(12):937–47.

81. Zhang X, Leahy I, Collemare J, Seidl MF. Secondary metabolite biosynthetic gene clusters and their genomic localization in the fungal genus *Aspergillus* [Internet]. 2024 [cited 2024 June 21]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.02.20.581327

82. Van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Cent Sci. 2019 Nov 27;5(11):1824–33.

83. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, et al. The LOTUS initiative for open knowledge management in natural products research. eLife. 2022 May 26;11:e70780.

84. Chemnetbase Dictionary of Natural Products 33.2 [Internet]. 2024. Available from: https://dnp.chemnetbase.com/chemical/ChemicalSearch.xhtml?dswid=918

85. Chandrasekhar V, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, et al. COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res. 2025 Jan 6;53(D1):D634–43.

View Article Online
DOI: 10.1039/D5OB01965F

86. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010 Sept;47(9):736–41.

87. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2020 Jan;16(1):60–8.

88. Gilchrist CLM, Booth TJ, Van Wersch B, Van Grieken L, Medema MH, Chooi YH. cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. Ouangraoua A, editor. Bioinforma Adv. 2021 June 9;1(1):vbab016.

89. Nuhamunada M, Mohite OS, Phaneuf PV, Palsson BO, Weber T. BGCFlow: systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. Nucleic Acids Res. 2024 June 10;52(10):5478–95.

90. Hjörleifsson Eldjárn G, Ramsay A, Van Der Hooft JJJ, Duncan KR, Soldatou S, Rousu J, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. Nagarajan N, editor. PLOS Comput Biol. 2021 May 4;17(5):e1008920.

91. Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across *Actinobacteria*. Birol I, editor. Bioinformatics. 2017 Oct 15;33(20):3202–10.

92. Johnston CW, Skinnider MA, Wyatt MA, Li X, Ranieri MRM, Yang L, et al. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. Nat Commun. 2015 Sept 28;6(1):8421.

93. Terlouw BR, Huang C, Meijer D, Cediel-Becerra JDD, He R, Rothe ML, et al. PARAS: high-accuracy machine-learning of substrate specificities in nonribosomal peptide synthetases [Internet]. Bioinformatics; 2025 [cited 2025 Dec 6]. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.01.08.631717

94. Agrawal P, Khater S, Gupta M, Sain N, Mohanty D. RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. Nucleic Acids Res. 2017 July 3;45(W1):W80–8.

95. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res. 2017 July 3;45(W1):W49–54.

96. Spencer NR, Gunabalasingam M, Dial K, Di X, Malcolm T, Magarvey NA. An integrated AI knowledge graph framework of bacterial enzymology and metabolism. Proc Natl Acad Sci. 2025 Apr 15;122(15):e2425048122.

View Article Online
DOI: 10.1039/D5OB01965F

97. Ma K, Zhang P, Tao Q, Keller NP, Yang Y, Yin WB, et al. Characterization and Biosynthesis of a Rare Fungal Hopane-Type Triterpenoid Glycoside Involved in the Antistress Property of *Aspergillus fumigatus*. Org Lett. 2019 May 3;21(9):3252–6.

98. Heard SC, Wu G, Winter JM. Discovery and characterization of a cytochalasan biosynthetic cluster from the marine-derived fungus Aspergillus flavipes CNL-338. J Antibiot (Tokyo). 2020 Nov;73(11):803–7.

99. Li S, Srinivasan K, Tran H, Yu F, Finefield JM, Sunderhaus JD, et al. Comparative analysis of the biosynthetic systems for fungal bicyclo[2.2.2]diazaoctane indole alkaloids: the (+)/(−)-notoamide, paraherquamide and malbrancheamide pathways. MedChemComm. 2012;3(8):987.

100. Ye Y, Du L, Zhang X, Newmister SA, McCauley M, Alegre-Requena JV, et al. Fungal-derived brevianamide assembly by a stereoselective semipinacolase. Nat Catal. 2020 May 18;3(6):497–506.

101. Schümann J, Hertweck C. Molecular basis of cytochalasan biosynthesis in fungi: gene cluster analysis and evidence for the involvement of a PKS-NRPS hybrid synthase by RNA silencing. J Am Chem Soc. 2007 Aug 8;129(31):9564–5.

102. Caesar LK, Robey MT, Swyers M, Islam MN, Ye R, Vagadia PP, et al. Heterologous Expression of the Unusual Terreazepine Biosynthetic Gene Cluster Reveals a Promising Approach for Identifying New Chemical Scaffolds. Davies JE, editor. mBio. 2020 Aug 25;11(4):e01691-20.

103. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks -- a publishing format for reproducible computational workflows. In: Positioning and Power in Academic Publishing: Players, Agents and Agendas. 2016. p. 87–90.

104. Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, et al. rdkit/rdkit: 2024_09_3 (Q3 2024) Release [Internet]. Zenodo; 2024 [cited 2024 Dec 2]. Available from: https://zenodo.org/doi/10.5281/zenodo.591637

105. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59–60.

106. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/

View Article Online
DOI: 10.1039/D5OB01965F

Data availability statement

Analysis code is deposited together with supplementary files in Figshare (https://figshare.com/s/a2c267ec94e82e062bdd, private link for reviewing). The genomic and metabolomic data for the 16 *A. fischeri* strains used in this study were published by Rinker et al. 43 and can be accessed in GenBank via BioProject accession number PRJNA1129834, and in the corresponding Figshare repository (https://doi.org/10.6084/m9.figshare.25316452).