Lawrence Berkeley National Laboratory
LBL Publications

Title
The secondary metabolism collaboratory: a database and web discussion portal for secondary metabolite biosynthetic gene clusters

Permalink
https://escholarship.org/uc/item/0vc4d53v

Authors
Udwary, Daniel W
Doering, Drew T
Foster, Bryce
et al.

Publication Date
2024-11-14

DOI
10.1093/nar/gkae1060

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Nucleic Acids Research, 2024, 1–7
https://doi.org/10.1093/nar/gkae1060
Database issue

The secondary metabolism collaboratory: a database and web discussion portal for secondary metabolite biosynthetic gene clusters

Daniel W. Udwary *, Drew T. Doering, Bryce Foster, Tatyana Smirnova, Satria A. Kautsar and Nigel J. Mouncey

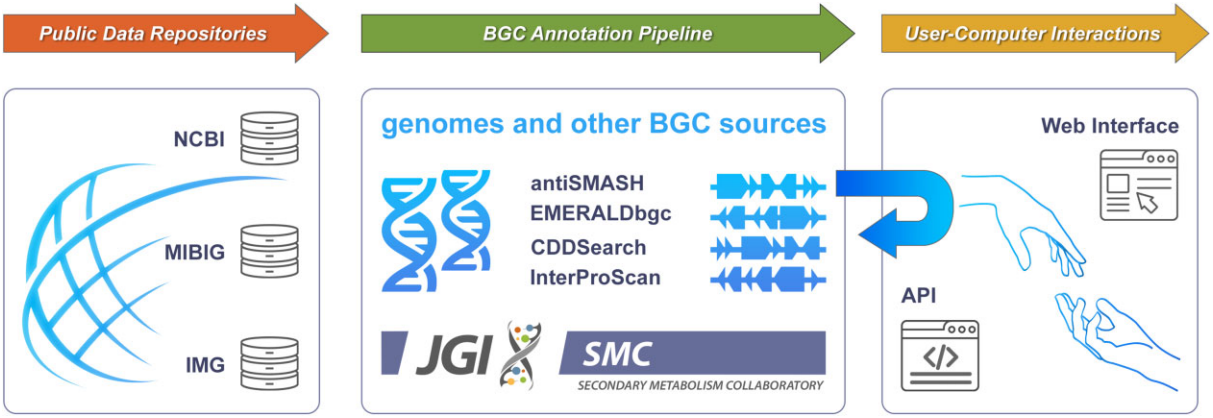
DOE Joint Genome Institute, Lawrence Berkeley National Labs, Berkeley, CA 94720, USA

*To whom correspondence should be addressed. Tel: +1 510 495 8400; Email: dwudwary@lbl.gov

Abstract

Secondary metabolites are small molecules produced by all corners of life, often with specialized bioactive functions with clinical and environmental relevance. Secondary metabolite biosynthetic gene clusters (BGCs) can often be identified within DNA sequences by various sequence similarity tools, but determining the exact functions of genes in the pathway and predicting their chemical products can often only be done by careful, manual comparative analysis. To facilitate this, we report the first release of the secondary metabolism collaboratory (SMC), which aims to provide a comprehensive, tool-agnostic repository of BGC sequence data drawn from all publicly available and user-submitted bacterial and archaeal genome and contig sources. On the website, users are provided a searchable catalog of putative BGCs identified from each source, along with visualizations of gene and domain annotations derived from multiple sequence analysis tools. SMC's data is also available through publicly-accessible application programming interface (API) endpoints to facilitate programmatic access. Users are encouraged to share their findings (and search for others') through comment posts on BGC and source pages. At the time of writing, SMC is the largest repository of BGC information, holding 13.1M BGC regions from 1.3M source sequences and growing, and can be found at https://smc.jgi.doe.gov.

Graphical abstract



Introduction

Clustering of secondary metabolite biosynthetic pathway genes is a phenomenon commonly observed throughout microbial life (1), and occasionally in complex eukaryotes (2,3). A biosynthetic gene cluster (BGC) is a group of genes found adjacent to or in very close proximity to one another in the host organism's genome, and typically encodes all of the enzymes necessary to synthesize a secondary or specialized metabolite (4), including regulatory elements, self-resistance genes, and systems needed for active transport across the membrane. This compact and self-contained organization of genes observed in many BGCs facilitates coordinated regulation of gene expression as well as inheritance by horizontal gene transfer. Understanding BGCs is critical to understanding secondary metabolism as a whole since secondary metabolites play valuable roles in medicine (5,6), chemically-mediated environmental interactions (7), and their genes can be increasingly utilized as building blocks for synthetic or engineered biosynthetic pathways (8).

BGC detection is typically straightforward for known biosynthetic pathways, though may be computationally intensive for specific BGC families, especially when deployed at scale to analyze hundreds or more genomes. Many secondary metabolite biosynthetic gene families are well-established, and once one is located in the genome, a cursory look for other biosynthetic, regulatory, and transport genes surrounding it

Received: August 22, 2024. **Revised:** October 18, 2024. **Editorial Decision:** October 21, 2024. **Accepted:** October 24, 2024
Published by Oxford University Press on behalf of Nucleic Acids Research 2024.
This work is written by (a) US Government employee(s) and is in the public domain in the US.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae1060/7900188 by guest on 13 December 2024

typically unveils the full BGC region. The most commonly used tool for this purpose is undoubtedly antiSMASH, currently in its 7th version (9). antiSMASH detects BGCs using a collection of biosynthetic hidden Markov models coupled with rules-based annotation to categorize BGC classes. Rules-based identification, however, can only identify biosynthetic pathways which have already been recognized as biosynthetic pathways. To help bridge this gap, increasingly sophisticated machine-learning (ML) based tools are also now available, including DeepBGC (10), EMERALDbgc/SanntiS (11), and GECCO (12). Unfortunately, annotation and interpretation of a BGC may still be problematic. While ML tools cannot yet achieve this, likely due to a lack of verified training data, antiSMASH, once again, offers the most sophisticated rules-based analyses of individual catalytic domains, while also providing a suite of predictive tools for, for example, amino acids activated by non-ribosomal peptide synthetase (NRPS) adenylation domains, polyketide synthase (PKS) ketoreductase stereospecificity, and ribosomally processed and post-translationally modified peptide (RiPP) core peptide sequences. Still, accurate automated prediction of the final natural product compound or its biosynthetic pathway from sequence alone remains elusive. Instead, this is typically tackled by careful human-driven comparative domain/gene/pathway analysis, ultimately coupled with laborious chemical isolation and structure elucidation to verify hypotheses. Ultimately, by the inherent nature of secondary metabolism's evolution away from primary metabolic biochemistry, computational analysis of BGCs is critical to identifying and comparing them, but interpreting and elucidating their chemistry is best tacked as a manual human process, for now. This may change rapidly with the advent of new AI-based data tools, but expansive curated databases are going to be critical to providing training sets for those efforts.

Several specialized BGC databases currently exist. The 'Minimum Information about a Biosynthetic Gene Cluster' (MIBiG) database (13) is likely the most-utilized database in natural products biosynthesis. MIBiG is built from community annotation of experimentally validated BGCs, and its v3.0 contains 2502 manually annotated BGC entries. The antismash-db is a database of 232 018 BGCs derived from running current versions of antiSMASH against high-quality complete genomes (14). Previously, the Joint Genome Institute's (JGI) Integrated Microbial Genomics' Atlas of Biosynthetic gene Clusters (IMG-ABC) (15) contained the most comprehensive set of BGCs from IMG microbial sources, also generated with antiSMASH, but discussion with users and the natural products community led us to believe that static, consensus approaches to BGC annotation may not fully capture the nuance needed to interpret BGC data.

So, to further facilitate BGC and biosynthesis research, we present the Secondary Metabolism Collaboratory (SMC) database, with its associated website, containing a comprehensive set of BGCs from bacteria and archaea, pre-annotated with multiple commonly used sequence analysis tools. The database is structured to allow users to add their own source or genome sequences, new BGCs, manual or automated gene annotations, links out to other websites and public databases, tabular data, or images, all shared openly with all other users. At the time of writing, SMC serves as the largest repository of BGC data, and is open to all researchers without access restrictions.

Materials and methods

Data generation

All available public and restriction-free bacterial and archaeal genome assemblies, of all assembly quality, were downloaded from NCBI's Genome platform (16) and JGI's IMG platform (17), checked for exact nucleotide sequence redundancy via MD5sum, and files converted to validated GFF3 and FASTA formats for processing by a custom sequence annotation pipeline, which was written to run specifically on JGI's Dori or NERSC's Perlmutter compute clusters. Where possible, taxonomic information for whole genome sequences were assigned according to information provided by NCBI and/or IMG. The annotation pipeline utilizes antiSMASH v7 or later, EMERALDbgc v0.2.4.1, InterProScan 5.60–92.0 (18), and an implementation of NCBI's Conserved Domain Database Search (CDDSearch), which runs rps-blast for 6-frame translations of all BGC regions against NCBI's Conserved Domain Database (19). The full code for this annotation pipeline is available at https://code.jgi.doe.gov/smc/autosmc, but would need adaptation to run outside of DOE computing systems. Because both antiSMASH and EMERALDbgc are sensitive to the accuracy of predicted proteins, we also ran each genome through prodigal (20) to ameliorate the effects of any outdated gene detection methods sometimes encountered in older genomes, and source annotations are retained by the pipeline for later manual comparison by users. Where antiSMASH and EMERALDbgc regions overlapped, the two were combined into a single SMC region. Currently, genomes without identifiable BGC regions (approximately 30% of bacterial genomes and 25% of archaeal genomes) are not stored in SMC, but can be accessed through IMG and/or NCBI.

SMC data releases will not be versioned or scheduled, and any versioning we use will reflect website frontend/backend/API versions. Data updates/additions will be made on a continual rolling basis as new genomes are submitted by users or become available on NCBI and IMG. At the time of writing, SMC data generation has consumed more than 3 million CPU-hours on JGI's Dori compute cluster. Moving forward, SMC has deployed an automated data ingest system, retrieving and processing new genomes and other sequence data from NCBI, IMG, and sequences submitted by SMC users, with the intention to expand upon the current microbial genomic content to include the vast wealth of sequence from metagenomes and metagenome-associated genomes (MAGs), as well as fungal, plant, and other known BGC-containing genomes as BGC detection and annotation technology develops and improves for those taxa.

Results

User-centered design (UCD)

From its inception, we kept users and their needs centered in the design of the SMC database and website using User-Centered Design methods (21). We started with focus group discussions in 2019 to begin to understand what knowledgeable users would want in a BGC database. As development began and proceeded, users were consulted at multiple steps along the way, through early presentations at meetings and seminars, 1-on-1 question and answer sessions, workshops, and focus groups. Direct user engagement at multiple stages of our development and usability testing allowed us to increase

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae1060/7900188 by guest on 13 December 2024

website usability and accessibility. As a result, SMC is structured to be as close as possible to a FAIR (Findable, Accessible, Interoperable, Reproducible) data respository, using open science principles (22). All SMC data is accessible for free via the website or its APIs, and, in addition, users who wish to upload or write data to the database (source sequence uploads, or creating BGCs, annotations, or posts) are invited to register for free using an ORCID login with multifactor authentication (MFA) enabled. SMC source, BGC, gene, etc identifiers are static and hard-linkable, for easy sharing. Data import and export mainly comes through human-readable GFF3 and/or fasta format files.

Data content

SMC's BGC data is derived from BGC 'sources'—input DNA sequences which come primarily from genomes, currently and expanding over the next phase to include metagenomes, metagenome-associated genomes (MAGs), contigs, plasmids, viruses, and artificial constructs. All sources, to the best of our knowledge, are sequences from publicly available repositories with no restrictions on access or utilization, or may come directly from users who have agreed to a policy allowing free access before uploading.

Exact counts of independent, functional BGCs can be challenging because antiSMASH and other tools can typically only identify BGC 'regions', which may contain one or more BGCs in close proximity to one another in the genome, and many genomes may have fragmented BGCs derived from incomplete sequence assembly, made all the more challenging by the fact that modular PKS and NRPS BGCs pose distinct challenges to graph assembly methods using short read sequences (23). For simplicity, we use the term 'BGCs' in SMC to mean 'true BGCs, BGC regions and BGC fragments'. Meaningful derivations of an organism's biosynthetic potential must take these issues into account.

At the time of writing, SMC holds over 13M BGCs, derived from more than 1.3M source sequences. There are an average of 4–5 annotation tracks for each BGC, which comprise a total of 2.1 billion individual gene and domain annotations. We compared each BGC against the most recently available MIBiG database (v3.0) and found only 24 439 SMC BGCs showed high similarity (95% or greater sequence identity over 80% of the MIBiG entry), indicating that approximately 99.82% of BGCs in public sequence cannot unambiguously be assigned to an experimentally-characterized natural product biosynthetic pathway. In Table 1, we list the top 20 BGC-containing phyla and observe that the Pseudomonadota led with the largest number of BGCs, but over very large number of genomes. The leading phylum in BGC density was the Myxococcota, with an average of 24.4 BGCs per genome, across only 1028 genomes, followed closely by the well-established secondary metabolite producers, the Actinomycetota, with 20 BGCs per genome. However, by another metric of density, the Actinomycetota led, averaging 4.8 BGCs/megabase. While it is tempting to draw deeper conclusions about trends in secondary metabolism across taxa, it is important to note that the public sequence databases are not necessarily representative of the true diversity of nature, and care should be taken to draw conclusions about the natural occurrence of BGCs without a good understanding of what has entered the public sphere and why.

Visualizations

The main route for exploring BGC content on SMC's website is through its Source page and BGC region visualizations. SMC's web frontend provides easy access to SMC's database via its Source and BGC pages.

The SMC web frontend generates a Source Page for each BGC-containing DNA source sequence. Source pages contain basic information about the source, including its name, taxonomic information where available, the public sequence database or user from which it was acquired, and some other basic statistics (Figure 1A). Below that, there are links to download the source files, typically the fasta nucleic acid sequence and a GFF3 table with all available annotations (Figure 1B). Most of the remainder of the page provides information about BGC regions and their locations in the source sequence, including a visual map of each BGC's location on each contig, color-coded by rough BGC class (Figure 1C), and the same information is found below that in a tabular format (Figure 1D). At the bottom of the page (not depicted) one can find a button to add a BGC to the source, and the 'Comments' section, where a user may view or post any additional information relevant to the source sequence or its BGC content.

By clicking through to a BGC region entry from a Source page, SMC generates a BGC Page. The top of the page provides basic information about the BGC region: compound name, if known, BGC class(es), a link back to its source page, and location information with a link to downloadable nucleotide sequence of that region (Figure 2A). Below that, one finds a scrollable visualization of the gene and domain content of each annotation split out into separated tracks (Figure 2B). The exact contents of each annotation track will depend upon the output of the tool that generated the annotation. For example, NCBI source annotations generally depict only gene content, while antiSMASH tracks show genes, domains, motifs, candidate clusters, protoclusters, etc, and will be color-coded per antiSMASH's default color scheme, while EMERALDbgc tracks generally only depict the location of its predicted BGC region. Each feature is clickable and will depict any information provided by the annotation tool, frequently with links back to any outside websites or databases where possible. Below that (not depicted) one will find another BGC-specific comments section, functionally identical to the Source page comments section.

Community tools

Beyond the Postgres database itself, the SMC platform provides functionality in the form of its web interface, and its APIs. On the website, users may search or browse for BGCs based on a compound name, by taxonomy, the original source accession ID (i.e. NCBI or IMG accession IDs), or by BLAST (24). From the homepage, users may also upload their own sequences for inclusion in SMC, if they are willing to agree that the sequence is available for unrestricted use by the scientific community, at which point the user-provided sequence will go through SMC's annotation pipeline and make its way to the database and be viewable by all users. All functions of SMC are available through its APIs, which are suitable for simple programmatic queries or smaller data retrievals, with full documentation accessible from the homepage. For more complex or global data queries and downloads, users can contact us via a 'Contact Us' form available from the home page to discuss

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae1060/7900188 by guest on 13 December 2024

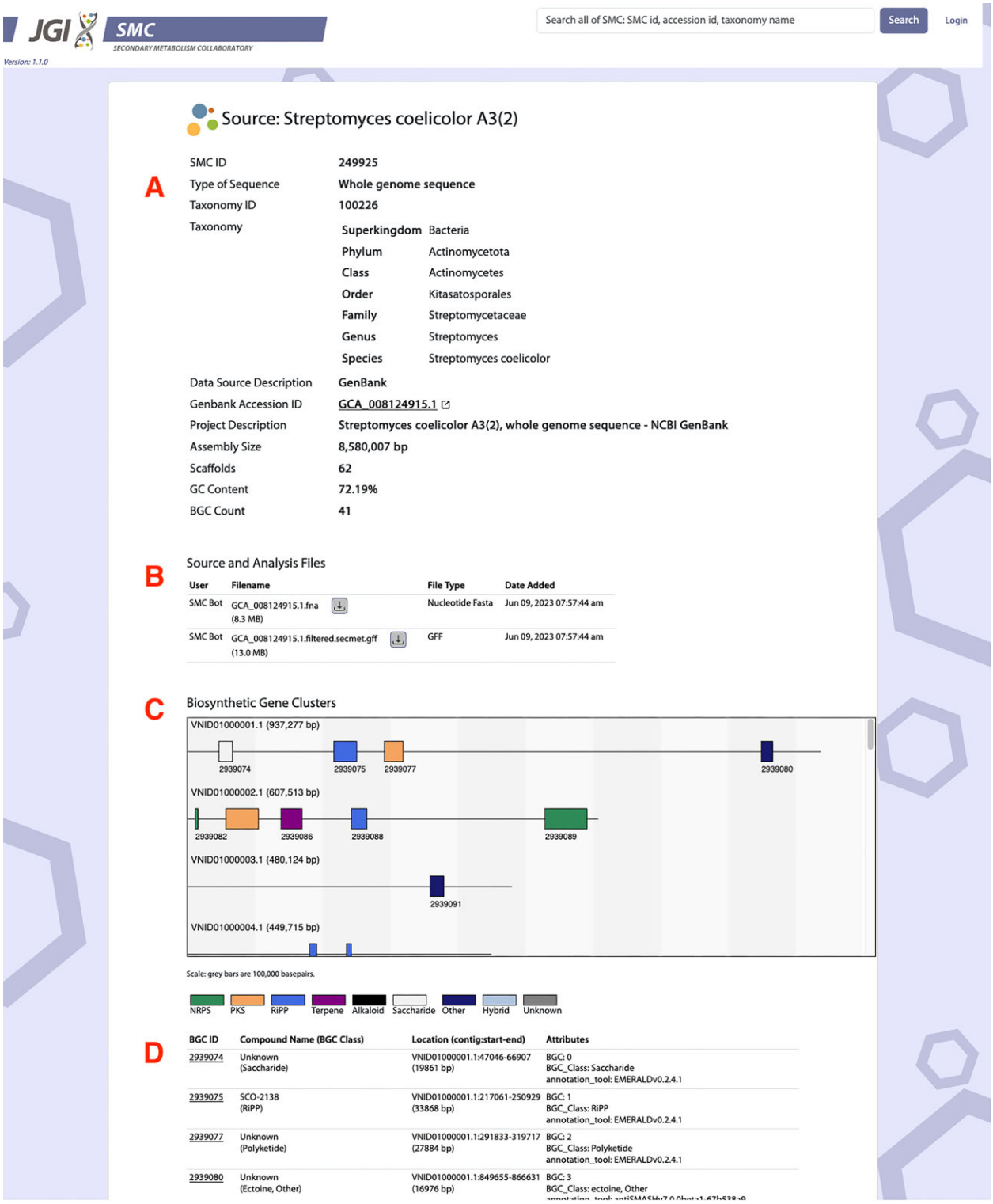


Figure 1. Example of an SMC Source page, for *Streptomyces coelicolor* A3(2). (**A**) Basic information and statistics on the source. (**B**) Files available for download. (**C**) A visual representation of locations and general BGC class for each contig within the source sequence. (**D**) A tabular representation of BGC regions within the source sequence.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae1060/7900188 by guest on 13 December 2024

Table 1. Top 20 Phyla by BGC count

Phylum	Genome count	BGC count	Genome_length_sum	BGCs/genome	BGCs/Mbase
Pseudomonadota	785 326	8 652 588	3.81151E + 12	11.01 782 954	2.27 012 141
Bacillota	312 079	2 225 994	9.24346E + 11	7.132 790 095	2.40 818 213
Actinomycetota	61 566	1 237 623	2.58589E + 11	20.10 237 794	4.78 605 898
Campylobacterota	87 448	438 750	1.52174E + 11	5.017 267 405	2.88 320 451
Bacteroidota	39 675	220 624	1.39811E + 11	5.560 781 348	1.57 801 716
Cyanobacteriota	4438	56 166	17 116 670 286	12.65 570 077	3.2 813 625
Unclassified	8316	38 765	19 657 015 687	4.661 495 911	1.97 206 944
Planctomycetota	3786	28 940	15 687 512 179	7.6 439 514	1.84 477 944
Acidobacteriota	2957	27 868	12 154 984 307	9.424 416 638	2.29 272 201
Verrucomicrobiota	4723	25 184	14 018 005 982	5.332 204 108	1.79 654 653
Myxococcota	1028	25 136	6 936 079 277	24.45 136 187	3.62 394 935
Chloroflexota	4437	23 860	13 559 856 083	5.377 507 325	1.7 596 057
Thermodesulfobacteriota	2554	17 800	8 192 601 631	6.969 459 671	2.172 692
Spirochaetota	4358	15 003	13 786 072 543	3.442 634 236	1.08 827 224
Euryarchaeota	1699	8611	5 164 062 363	5.068 275 456	1.66 748 567
Nitrospirota	1247	8405	3 080 362 104	6.740 176 423	2.72 857 532
Gemmatimonadota	994	6173	3 358 385 200	6.210 261 569	1.83 808 576
Candidatus Omnitrophota	1025	4278	1 724 194 171	4.173 658 537	2.48 115 907
Bdellovibrionota	710	4256	2 061 312 869	5.994 366 197	2.06 470 355
Ignavibacteriota	762	3906	2 417 780 950	5.125 984 252	1.61 553 097

Figure 2. Example of a BGC page, for the *Streptomyces coelicolor* A3(2) coelichelin BGC region. (**A**) Basic information about the BGC region. (**B**) Visualized annotation tracks for BGC regions.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae1060/7900188 by guest on 13 December 2024

how best to conduct them using computing resources available outside of the web server.

SMC is intended to be, first, a biosynthetic gene cluster sequence repository, and second, a 'collaboratory' where researchers can work together and share information to improve the data itself. Therefore, we have included capabilities that allow each user to upload a new annotation (or edit their own existing annotation) for any BGC. Individual annotation tracks may be downloaded as GFF3 format tables from BGC pages, and new annotations are accepted in GFF3 format. We also allow users to leave comments on sources or their BGCs, in the form of text, hyperlinks to other sites, images, PDFs, or tables. Comments may also be made using SMC's APIs, as we hope that may facilitate the use of automated tools for information sharing. In addition, users may create 'collections' of sources or BGCs, which can then be linked for easy access or sharing with other users. Users must be logged into the website (where you can also retrieve an access token for API usage) to write or upload or collect data, and must adhere to a code of conduct, which we hope will foster accountability, create a record of discovery, and facilitate discussion and collaboration between researchers worldwide.

Discussion

SMC and its contents are designed to be tool-independent, unlike, for example, antismash-db, which is centered and built on antiSMASH results. While the pipeline used to generate SMC's initial contents is composed of antiSMASH, EMERALDbgc, InterProScan and NCBI CDDSearch, new annotations may be added by users, and new sequence annotation tools not yet invented may be easily added to the pipeline in the future, or annotations from new tools may simply be added to existing BGC regions as new annotation tracks, perhaps uploaded programmatically via the API. We know and expect that different tools and different user interpretations of the data may lead to conflicting annotations, and so our intention is that SMC's discussion tools will facilitate deeper, collaborative scientific investigation of specific BGCs and BGC families. Over time, we intend to introduce additional visualization methods to clarify the 'best' or community-established annotations and reduce complexity of visuals. Our hope is that, if fully utilized by the community, SMC can be a living document of the latest information for a given BGC or genome, illustrating a history of the scientific community's progress toward understanding a given biosynthetic pathway or the producing organisms.

Moving forward, we intend to continue SMC development, making improvements to its user-facing APIs, advancing its text and sequence search methods, integrating comparative analysis tools like BiG-SLICE (25) and GATOR-GC (https://github.com/chevrettelab/gator-gc), introducing more community engagement features, enhancing SMC's ability to recognize more data formats for inputs and exports (including MIBiG compatible data formats), and refining SMC's visualizations of BGCs and associated chemistry. We will also continue to explore and intend to integrate new BGC prediction and secondary metabolism-focused gene annotation tools, potentially including GECCO (12) and other new machine-learning tools as they become widely adopted and trusted by the community. Additional information about SMC's development roadmap is linked from the homepage ('About SMC').

SMC is now the largest repository of natural product BGCs, and our results in generating data for SMC show that more than 99% of publicly available BGC sequences cannot currently be unambiguously associated with experimentally characterized BGC sequences. This clearly demonstrates that there is much work to be done in the natural products community to elucidate the chemical endpoints and biosynthetic mechanisms of many more BGCs. The community frequently comes together for periodic MIBiG 'annotathons', where new MIBiG entries are created or improved, and, now, with SMC we hope to give users the similar ability to make their expertise heard on non-experimentally characterized BGC data at any time, through naming of clusters, annotation of gene/domain functions, redefining cluster boundaries, adding links to publications or associated chemical structures, and making connections to external data sources (eg NPAtlas (26), GNPS (27), MiBIG (13)). We expect that the best BGC data will be 'living' data, constantly updated and refined by active users who care about it. SMC is the first platform available for this kind of data and activity, and we hope that the community will embrace it, making it the invaluable community resource we envision.

Data availability

The database and all data underlying this article are available in the Secondary Metabolism Collaboratory (SMC), at https://smc.jgi.doe.gov.

Acknowledgements

Many thanks to all the anonymous scientists who gave their time to participate in focus groups, interviews, and who tested all versions of SMC and provided feedback. Kandasoft Inc (Boston, MA) provided an early prototype under contract. We thank David Gilbert for helpful suggestions, including a name for the project.

Funding

U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy [DE-AC02-05CH11231]. Funding for open access charge: U.S. Department of Energy [DE-AC02-05CH11231].

Conflict of interest statement

None declared.

References

1. Bauman,K.D., Butler,K.S., Moore,B.S. and Chekan,J.R. (2021) Genome mining methods to discover bioactive natural products. *Nat. Prod. Rep.*, **38**, 2100–2129.
2. Osbourn,A. (2010) Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends Genet.*, **26**, 449–457.
3. Scesa,P.D., Lin,Z. and Schmidt,E.W. (2022) Ancient defensive terpene biosynthetic gene clusters in the soft corals. *Nat. Chem. Biol.*, **18**, 659–663.
4. Barona-Gomez,F., Chevrette,M.G. and Hoskisson,P.A. (2023) On the evolution of natural product biosynthesis. *Adv. Microb. Physiol.*, **83**, 309–349.
5. Katz,L. and Baltz,R.H. (2016) Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.*, **43**, 155–176.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae1060/7900188 by guest on 13 December 2024

6. Baltz,R.H. (2008) Renaissance in antibacterial discovery from actinomycetes. *Curr. Opin. Pharmacol.*, **8**, 557–563.
7. Traxler,M.F. and Kolter,R. (2015) Natural products in soil microbe interactions and evolution. *Nat. Prod. Rep.*, **32**, 956–970.
8. Nava,A.A., Roberts,J., Haushalter,R.W., Wang,Z. and Keasling,J.D. (2023) Module-based polyketide synthase engineering for de novo polyketide biosynthesis. *ACS Synth. Biol.*, **12**, 3148–3155.
9. Blin,K., Shaw,S., Augustijn,H.E., Reitz,Z.L., Biermann,F., Alanjary,M., Fetter,A., Terlouw,B.R., Metcalf,W.W., Helfrich,E.J.N., *et al.* (2023) antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.*, **51**, W46–W50.
10. Hannigan,G.D., Prihoda,D., Palicka,A., Soukup,J., Klempir,O., Rampula,L., Durcak,J., Wurst,M., Kotowski,J., Chang,D., *et al.* (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.*, **47**, e110.
11. Sanchez,S., Rogers,J.D., Rogers,A.B., Nassar,M., McEntyre,J., Welch,M., Hollfelder,F. and Finn,R.D. (2023) Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. bioRxiv doi: https://doi.org/10.1101/2023.05.23.540769, 02 June 2023, preprint: not peer reviewed.
12. Carroll,L.M., Larralde,M., Fleck,J.S., Ponnudurai,R., Milanese,A., Cappio,E. and Zeller,G. (2021) Accurate *de novo* identification of biosynthetic gene clusters with GECCO. bioRxiv doi: https://doi.org/10.1101/2021.05.03.442509, 04 May 2021, preprint: not peer reviewed.
13. Terlouw,B.R., Blin,K., Navarro-Munoz,J.C., Avalon,N.E., Chevrette,M.G., Egbert,S., Lee,S., Meijer,D., Recchia,M.J.J., Reitz,Z.L., *et al.* (2023) MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.*, **51**, D603–D610.
14. Blin,K., Shaw,S., Medema,M.H. and Weber,T. (2024) The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more. *Nucleic Acids Res.*, **52**, D586–D589.
15. Palaniappan,K., Chen,I.A., Chu,K., Ratner,A., Seshadri,R., Kyrpides,N.C., Ivanova,N.N. and Mouncey,N.J. (2020) IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.*, **48**, D422–D430.
16. O'Leary,N.A., Cox,E., Holmes,J.B., Anderson,W.R., Falk,R., Hem,V., Tsuchiya,M.T.N., Schuler,G.D., Zhang,X., Torcivia,J., *et al.* (2024) Exploring and retrieving sequence and metadata for species across the tree of life with NCBI datasets. *Sci. Data*, **11**, 732.
17. Chen,I.A., Chu,K., Palaniappan,K., Ratner,A., Huang,J., Huntemann,M., Hajek,P., Ritter,S.J., Webb,C., Wu,D., *et al.* (2023) The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.*, **51**, D723–D732.
18. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
19. Wang,J., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R., Gwadz,M., Lu,S., Marchler,G.H., Song,J.S., Thanki,N., Yamashita,R.A., *et al.* (2023) The conserved domain database in 2023. *Nucleic Acids Res.*, **51**, D384–D388.
20. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.*, **11**, 119.
21. Norman,D.A. and Draper,S.W. (1986) In: User Centered System Design : New Perspectives on human-computer Interaction. L. Erlbaum Associates, Hillsdale, NJ.
22. Parsons,S., Azevedo,F., Elsherif,M.M., Guay,S., Shahim,O.N., Govaart,G.H., Norris,E., O'Mahony,A., Parker,A.J., Todorovic,A., *et al.* (2022) A community-sourced glossary of open scholarship terms. *Nat. Hum. Behav.*, **6**, 312–318.
23. Goldstein,S., Beka,L., Graf,J. and Klassen,J.L. (2019) Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *Bmc Genomics [Electronic Resource]*, **20**, 23.
24. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
25. Epstein,S.C., Charkoudian,L.K. and Medema,M.H. (2018) A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences. *Stand. Genomic Sci.*, **13**, 16.
26. van Santen,J.A., Poynton,E.F., Iskakova,D., McMann,E., Alsup,T.A., Clark,T.N., Fergusson,C.H., Fewer,D.P., Hughes,A.H., McCadden,C.A., *et al.* (2022) The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res.*, **50**, D1317–D1323.
27. Wang,M., Carver,J.J., Phelan,V.V., Sanchez,L.M., Garg,N., Peng,Y., Nguyen,D.D., Watrous,J., Kapono,C.A., Luzzatto-Knaan,T., *et al.* (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.*, **34**, 828–837.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae1060/7900188 by guest on 13 December 2024

Received: August 22, 2024. **Revised:** October 18, 2024. **Editorial Decision:** October 21, 2024. **Accepted:** October 24, 2024
Published by Oxford University Press on behalf of Nucleic Acids Research 2024.
This work is written by (a) US Government employee(s) and is in the public domain in the US.