

The antiSMASH database version 5

Kai Blin^{1,*}, Simon Shaw¹, Marnix H. Medema², Tilmann Weber^{1,*}

¹The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs., Lyngby, Denmark

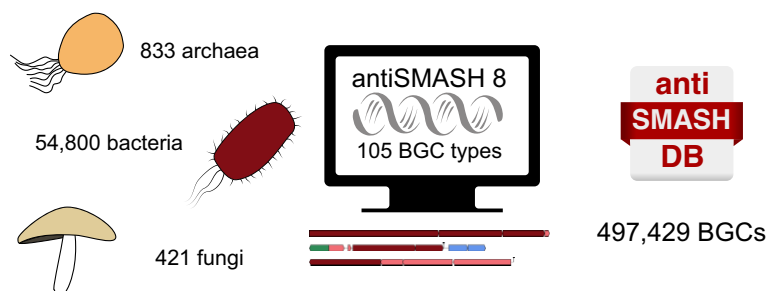
²Bioinformatics Group, Wageningen University, 6708PB, Wageningen, The Netherlands

*To whom correspondence should be addressed. Email: kblin@biosustain.dtu.dk
Correspondence may also be addressed to Tilmann Weber. Email: tiwe@biosustain.dtu.dk

Abstract

Specialized metabolites produced by microorganisms are frequently used in the development of drugs and crop protection agents. Genome mining is a widely used approach to access this potential, and antiSMASH is often the tool of choice for this task. Here, we present version 5 of the antiSMASH database, with biosynthetic gene cluster predictions provided by antiSMASH 8.1 available in an easy-to-use web interface. Version 5 of the database contains 833 archaeal, 54 800 bacterial, and 421 fungal genomes and is available from <https://antismash-db.secondarymetabolites.org/>.

Graphical abstract



Introduction

Specialized or secondary metabolites produced by microorganisms are the main source of bioactive compounds that are in use as antimicrobial and anticancer drugs [1], fungicides, herbicides, pesticides, and biostimulants [2]. In the last decades, the increasing availability of microbial genomes has established genome mining as a very important method for the identification of the biosynthetic gene clusters (BGCs) responsible for the synthesis of these compounds [3]. To assist with genome mining, software tools were soon developed, and the field has seen active development in the past decades (see [4–8] for reviews discussing tools and timelines). While there were a number of tools, only a few databases existed to make BGC data available. An example of the latter development is e.g. ClusterMine360 [9], which was introduced in 2013 but has been discontinued in the meantime.

Initially released in 2011, antiSMASH [10–17] has become the most widely applied tool for genome mining for specialized/secondary metabolites and is considered as the gold standard. antiSMASH uses a rule-based system for identifying genome regions containing BGCs based on conserved biosynthetic enzymes. Currently, it detects 105 different pathway types. For some more well-understood synthesis types, it also performs cluster-specific analyses. antiSMASH also compares

identified regions to similar regions in the MIBiG database [18] of known BGCs and a dataset of antiSMASH results on publicly available, high-quality genomes.

antiSMASH is designed to annotate and analyse a single genome at a time. To enable answering research questions that require comparative analyses across many genomes, we have developed the antiSMASH database [19–22]. This database provides the foundation of antiSMASH's ClusterBlast analysis, with results from ClusterBlast linking directly to the database. In addition, the antiSMASH database website provides BLAST-like searches, taxonomic browsing, and an easy-to-use graphical query builder to dynamically search database contents.

Here, we present the fifth version of this database, covering 833 archaeal, 54 800 bacterial, and 421 fungal genomes.

Materials and methods

Selection of included genomes

Archaeal, bacterial, and fungal genomes were downloaded from the NCBI RefSeq database [23] on 7 January 2025 using the `ncbi-genome-download` tool [24] with 'complete', 'chromosome', and 'scaffold' assembly levels, yielding 1267 archaeal, 182 383 bacterial, and 441 fungal assemblies. The

Received: September 15, 2025. Revised: October 9, 2025. Accepted: October 10, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

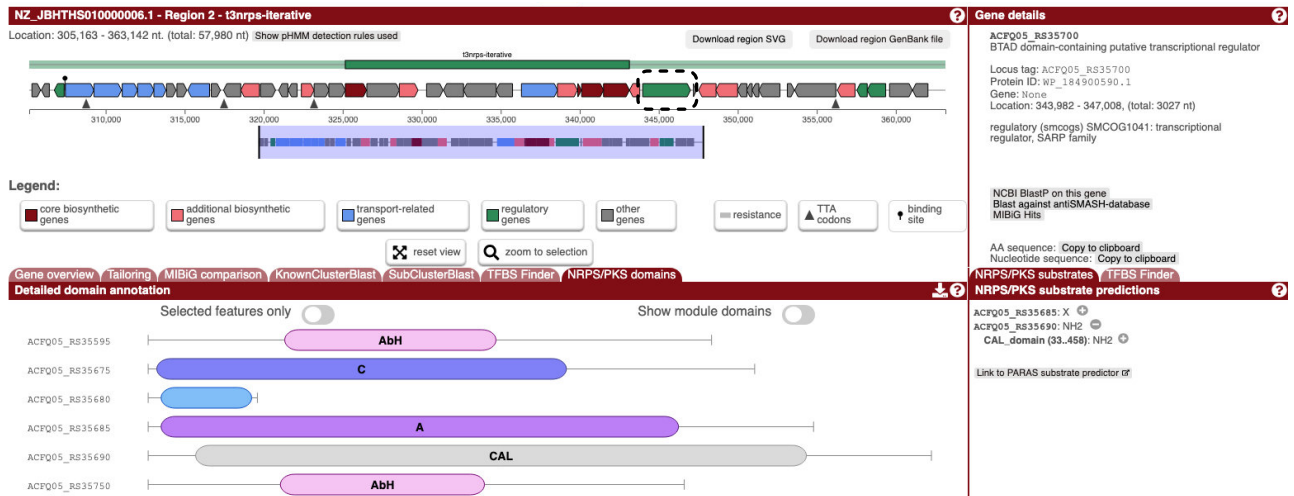


Figure 1. An iterative type III NRPS cluster from *Amycolatopsis umgeniensis* (NCBI assembly ID GCF_042676565.1). The green opaque bar above the gene arrows spans all core genes. A nearby SARP regulator, highlighted by a striped box, strongly suggests that this region indeed contains a BGC.

quality filtering deviated from previous versions that used the contig count (see [22] for a description). We decided to instead use the $L50$ value, which is the minimal number of contigs needed to account for half the total assembly length. This decision was made in order to not penalize assemblies that had one or two large contigs but also had a number of small contigs due to repeats, contamination, or leftover adapter sequences. The limits were set to an $L50$ value of at most 10 for archaea and bacteria and 60 for fungi. As described before [22], duplicated genomes were assigned to a similarity cluster when they had a Mash [25] distance of 0.04 or less. Another difference from previous versions is that we selected the representative genome of a similarity cluster using a slightly more complex calculation. Each assembly was assigned a score S with $S = w1 \times Ln - w2 \times L50n$, where Ln is the normalized length of all the assemblies in the similarity cluster, $L50n$ is the normalized $L50$ value of the similarity cluster, $w1$ is a length bonus, and $w2$ is a fragmentation penalty. The assembly in the similarity cluster with the highest score was selected as the representative assembly. For our selection, we used a length bonus and fragmentation penalty of 0.3. This method was chosen because our old contig-count minimization could pick assemblies of genome-minimized, plasmid-free mutants (e.g. *Streptomyces collinus* str. SQ GCF_021496465.1) over their corresponding wild types (e.g. *Streptomyces collinus* str. Tü365 GCF_000444875.1) because the plasmids increased the contig count. After redundancy filtering, 843 archaeal, 55 680 bacterial, and 431 fungal assemblies remained. Of the fungal assemblies, 36 contain genes that carry splice variants outside their coding sequences, leading to CDS entries that have identical locations, locus tags, and translations while having differing protein IDs in the GenBank files. As antiSMASH requires all CDSs to be uniquely identifiable by location, translation, and ID, we removed the duplicate CDSs on those assemblies.

antiSMASH annotations and data import

With the filtered, non-redundant dataset, we used GNU parallel [26] to run antiSMASH 8 with the options ‘--cb-knownclusters --cb-subclusters --cc-mibig --rre --asf’. antiSMASH successfully processed all archaeal, bacterial, and fungal sequences. Using the data from this first run, we extracted

all predicted ribosomally synthesized and post-translationally modified peptide (RiPP) precursors and all predicted regions. These were used to build new CompaRiPPson and ClusterBlast datasets, respectively. With these new datasets, antiSMASH 8 was re-run on the first round’s results using the options ‘--reuse --cb-general --clusterhmmmer --tigrfam --pfam2go’ to allow cross-references to the other results in the antiSMASH database.

The SQL schema for the database and the importer were updated to support antiSMASH 8 results. During the import process, assemblies without any antiSMASH hits were dropped. This results in a final count of 833 archaeal, 54 800 bacterial, and 421 fungal assemblies in the database.

Results and discussion

The NCBI RefSeq database is a valuable resource of a lot of microbial genomic diversity. It does also contain a large number of very similar genomes, especially for common pathogens like *Staphylococcus aureus*, *Escherichia coli*, or *Salmonella enterica*. It also contains a large number of very fragmented genomes. Fragmentation has been shown to negatively impact BGC detection [27]. To ensure that the antiSMASH database covers the whole microbial tree of life as well as possibly without overly biasing towards the most frequently sequenced organisms, and to ensure the BGC predictions are of the best possible quality, the data were rigorously filtered for quality and sequence dissimilarity. After the filtering and processing, 833 archaeal, 54 800 bacterial, and 421 fungal assemblies are present in the database. Annotations were performed using antiSMASH 8.1, which, on top of the rules described in [17], adds two new detection rules, one for adenosine derivatives like aureonuclemycin and another for iterative NRPS type III clusters (so NRPSs that are thiotemplated but do not use a condensation domain as the peptide-bond forming enzyme, see [28] for a detailed description and Fig. 1 for an example), for a total of 105 supported BGC types.

Support for new antiSMASH 8 annotations has been added to the query builder. A common user request for the BLAST-style sequence searches has been to add an option to download the result table. We have now added the option to download

A

Job 34e357a3-0b49-480f-b394-2e969f00e9e7 comparippson

Query	Hit ID	% Identity	Hit Record	Type
lanA	BW151_RS03795_lanthipeptide	100.00	NZ_MTJP01000006.1	Lanthipeptides
lanA	CV702_RS02990_lanthipeptide	100.00	NZ_CP024954.1	Lanthipeptides
lanA	LLUC08_RS03255_lanthipeptide	100.00	NZ_CP015903.1	Lanthipeptides
lanA	D6108_RS11395_lanthipeptide	100.00	NZ_RAGL01000022.1	Lanthipeptides
lanA	LSG16_RS07195_lanthipeptide	100.00	NZ_JAJONJ01000005.1	Lanthipeptides
lanA	BW154_RS09850_lanthipeptide	100.00	NZ_MTJS01000002.1	Lanthipeptides
lanA	PGV42_RS12330_lanthipeptide	100.00	NZ_JAQEOU01000022.1	Lanthipeptides
lanA	LG36_RS03420_lanthipeptide	100.00	NZ_CP009472.1	Lanthipeptides
lanA	PGW09_RS10765_lanthipeptide	100.00	NZ_JAQEOV01000015.1	Lanthipeptides
lanA	LILO_RS03015_lanthipeptide	100.00	NC_020450.1	Lanthipeptides
lanA	PGU01_RS08180_lanthipeptide	100.00	NZ_JAQDWM01000007.1	Lanthipeptides
lanA	LacL0098_RS03205_lanthipeptide	100.00	NZ_CP066300.1	Lanthipeptides
lanA	PGT80_RS11850_lanthipeptide	100.00	NZ_JAQDXV01000018.1	Lanthipeptides
lanA	LLKF_RS13210_lanthipeptide	100.00	NC_013656.1	Lanthipeptides

Download TSV

B

Query

Your search gave 25818 results, showing 1 to 100

Species	Region	Type	From	To	Edge	Similarity confidence	Most similar MIBIG cluster	In strain collection
Streptomyces Unknown MP131-18	1.1	Hybrid region: Terpene & Type II polyketide	0	60587	Yes	Low	prejadomycin/abelomycin/gaudimycin C/gaudimycin D/UWM6/gaudimycin A (BGC0000262.5)	DSM 42172
Streptomyces Unknown MP131-18	1.2	Hybrid region: Non-ribosomal peptide metallophores & non-ribosomal peptide synthase	177104	237638	No	High	coelibaactin (BGC0000324.5)	DSM 42172
Streptomyces Unknown MP131-18	1.3	Hybrid region: Catches NRPS-like fragments that are not detected by the NRPS rule & hydrogen cyanide & Type I polyketide & Type III polyketide	396157	445235	No	High	lagunapyrone A/lagunapyrone B/lagunapyrone C (BGC0001647.3)	DSM 42172
Streptomyces Unknown MP131-18	1.4	Hybrid region: Butyrolactone & Type I polyketide	535955	607675	No	Low	camporidine A/camporidine B (BGC0002308.2)	DSM 42172
Streptomyces Unknown MP131-18	1.5	Hybrid region: (Thio) azol (in) e-containing peptides , linear and macrocyclic & RRE-element containing cluster	618298	652572	No	Low	dehydroxynocardamine (BGC0002073.3)	DSM 42172
Streptomyces Unknown MP131-18	1.6	Type III polyketide	750512	791616	No	High	naringenin (BGC0001310.5)	DSM 42172
Streptomyces Unknown MP131-18	1.7	Terpene	808537	832852	No	Low	hopene (BGC0000663.5)	DSM 42172
Streptomyces Unknown MP131-18	1.8	Type I polyketide	891781	945787	No	Low	aculeximycin (BGC0000002.5)	DSM 42172
Streptomyces Unknown MP131-18	1.9	Type I polyketide	946079	992517	No	Low	tirandamycin (BGC0001052.5)	DSM 42172
Streptomyces Unknown MP131-18	1.10	deazapurine containing secondary metabolites	1149709	1175375	No	High	hulimycin (BGC0002354.4)	DSM 42172
Streptomyces Unknown MP131-18	1.11	Crocagin-like cluster	1341354	1365867	No	Low	spicamycin (BGC0001774.5)	DSM 42172

Figure 2. A A screenshot of the RiPP precursor (CompaRiPPson) search results on a nisin-like sequence. The new 'Download TSV' button is highlighted with a striped box. **(B)** A screenshot of a 'genus *Streptomyces* AND available in strain collection' search. The strain collection identifier column is highlighted with a striped box.

the sequence search results table (Fig. 2A). To make it easier for researchers to identify strains they can obtain from strain collections, we have worked with colleagues from the German Collection of Microorganisms and Cell Cultures (DSMZ) and the DTU Biosustain NBC strain collection to flag strains that are available to order (Fig. 2B). We are in contact with other strain collections about getting their strains included in a future release.

Compared to version 4 with 36 554 assemblies containing 231 534 BGC regions not touching a contig edge, version 5 raises these numbers to 56 054 assemblies (a 53% increase) and 497 429 BGC regions (a 107% increase). The CompaRiPPson dataset of predicted RiPP precursor peptides has grown from 16 533 to 34 401 sequences (a 108% increase).

Conclusion

Genome mining continues to be a valuable tool for assessing microbial biosynthetic potential. In the past 15 years, antiSMASH has aided these efforts. The antiSMASH database provides a user-friendly web interface to compare, contextualize, and cross-reference findings across genomes. With 497 429 BGC regions across archaea, bacteria, and fungi, the antiSMASH database version 5 is a comprehensive collection of specialized/secondary metabolite biosynthetic gene clusters with high-quality, up-to-date antiSMASH annotations for the whole natural product research community.

Acknowledgements

The authors would like to acknowledge Prof. Yvonne Mast, Dr. Isabel Schober, and Dr. Lorenz Reimer of DSMZ for providing the dataset of RefSeq assembly ID to DSMZ ID mappings.

Author contributions: Kai Blin (Conceptualization [lead], Data curation [lead], Methodology [lead], Project administration [lead], Software [equal], Validation [lead], Writing—original draft [lead], Writing—review & editing [lead]), Simon Shaw (Data curation [supporting], Software [equal], Writing—original draft [supporting]), Marnix H. Medema (Conceptualization [equal], Writing—original draft [equal]), and Tilmann Weber (Conceptualization [equal], Supervision [equal], Writing—original draft [equal])

Conflict of interest

M.H.M. is a member of the Scientific Advisory Boards of Hexagon Bio and Hothouse Therapeutics Ltd. All other authors declare to have no competing interests.

Funding

This work was funded by grants of the Novo Nordisk Foundation (grant number NNF20CC0035580 to T.W., K.B., S.S.) and an ERC Starting Grant (948770-DECIPHER to M.H.M.). Funding to pay the Open Access publication charges for this article was provided by Novo Nordisk Foundation.

Data availability

The antiSMASH database is available at <https://antismash-db.secondarymetabolites.org/>. There are no access restric-

tions for academic or commercial use of the web server. The source code components and SQL schema for the antiSMASH database are available on GitHub (<https://github.com/antismash>) under an OSI-approved Open Source license. The complete set of antiSMASH results, antiSMASH JSON files, and an SQL dump of the database can be downloaded from the antiSMASH download server (<https://dl.secondarymetabolites.org/database/5.0>).

References

- 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;83:770–803. <https://doi.org/10.1021/acs.jnatprod.9b01285>
- Sparks TC, Bryant RJ. Impact of natural products on discovery of, and innovation in, crop protection compounds. *Pest Manag Sci* 2022;78:399–408. <https://doi.org/10.1002/ps.6653>
- Ziemert N, Alanjary M, Weber T. The evolution of genome mining in microbes—a review. *Nat Prod Rep* 2016;33:988–1005. <https://doi.org/10.1039/C6NP00025H>
- Weber T. *In silico* tools for the analysis of antibiotic biosynthetic pathways. *Int J Med Microbiol* 2014;304:230–5. <https://doi.org/10.1016/j.ijmm.2014.02.001>
- Medema MH, Fischbach MA. Computational approaches to natural product discovery. *Nat Chem Biol* 2015;11:639–48. <https://doi.org/10.1038/nchembio.1884>
- Weber T, Kim HU. The secondary metabolite bioinformatics portal: computational tools to facilitate synthetic biology of secondary metabolite production. *Synth Syst Biotechnol* 2016;1:69–79. <https://doi.org/10.1016/j.synbio.2015.12.002>
- Baltz RH. Natural product drug discovery in the genomic era: realities, misconceptions, and opportunities. *J Ind Microbiol Biotechnol* 2019;46:281–99. <https://doi.org/10.1007/s10295-018-2115-4>
- Blin K, Kim HU, Medema MH *et al.* Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform* 2019;20:1103–13. <https://doi.org/10.1093/bib/bbx146>
- Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Res* 2013;41:D402–7. <https://doi.org/10.1093/nar/gks993>
- Medema MH, Blin K, Cimercanic P *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 2011;39:W339–46. <https://doi.org/10.1093/nar/gkr466>
- Blin K, Medema MH, Kazempour D *et al.* antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res* 2013;41:W204–12. <https://doi.org/10.1093/nar/gkt449>
- Weber T, Blin K, Duddela S *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 2015;43:W237–43. <https://doi.org/10.1093/nar/gkv437>
- Blin K, Wolf T, Chevrette MG *et al.* antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* 2017;45:W36–41. <https://doi.org/10.1093/nar/gkx319>
- Blin K, Shaw S, Steinke K *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;47:W81–7. <https://doi.org/10.1093/nar/gkz310>
- Blin K, Shaw S, Kloosterman AM *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 2021;49:W29–35. <https://doi.org/10.1093/nar/gkab335>
- Blin K, Shaw S, Augustijn HE *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* 2023;51:W46–50. <https://doi.org/10.1093/nar/gkad344>

17. Blin K, Shaw S, Vader L *et al.* antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Res* 2025;53:W32–8. <https://doi.org/10.1093/nar/gkaf334>
18. Zdouc MM, Blin K, Louwen NLL *et al.* MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res* 2025;53:D678–90. <https://doi.org/10.1093/nar/gkae1115>
19. Blin K, Medema MH, Kottmann R *et al.* The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* 2017;45:D555–9. <https://doi.org/10.1093/nar/gkw960>
20. Blin K, Pascal Andreu V, de los Santos ELC *et al.* The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* 2019;47:D625–30. <https://doi.org/10.1093/nar/gky1060>
21. Blin K, Shaw S, Kautsar SA *et al.* The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res* 2021;49:D639–43. <https://doi.org/10.1093/nar/gkaa978>
22. Blin K, Shaw S, Medema MH *et al.* The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more. *Nucleic Acids Res* 2024;52:D586–9. <https://doi.org/10.1093/nar/gkad984>
23. Goldfarb T, Kodali VK, Pujar S *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res* 2025;53:D243–57. <https://doi.org/10.1093/nar/gkae1038>
24. Blin K. ncbi-genome-download. Zenodo. <https://doi.org/10.5281/zenodo.8192486>. 28 July 2023.
25. Ondov BD, Treangen TJ, Melsted P *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132. <https://doi.org/10.1186/s13059-016-0997-x>
26. Tange O. GNU Parallel 20221122 (Херсо́н'). Zenodo. <https://doi.org/10.5281/zenodo.7347980>. 22 November 2022.
27. Sánchez-Navarro R, Nuhamunada M, Mohite OS *et al.* Long-read metagenome-assembled genomes improve identification of novel complete biosynthetic gene clusters in a complex microbial activated sludge ecosystem. *mSystems* 2022;7:e0063222. <https://doi.org/10.1128/msystems.00632-22>
28. Dell M, Dunbar KL, Hertweck C. Ribosome-independent peptide biosynthesis: the challenge of a unifying nomenclature. *Nat Prod Rep* 2022;39:453–9. <https://doi.org/10.1039/D1NP00019E>