Nucleic Acids Research, 2025, 1–7
https://doi.org/10.1093/nar/gkaf334
Web Server issue

antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation

Kai Blin [1,*], Simon Shaw [1], Lisa Vader [1], Judit Szenei [1], Zachary L. Reitz [2], Hannah E. Augustijn [3,4], José D.D. Cediel-Becerra [5], Valérie de Crécy-Lagard [5], Robert A. Koetsier [4], Sam E. Williams [1], Pablo Cruz-Morales [1], Sopida Wongwas [6], Alejandro E. Segurado Luchsinger [7,8], Friederike Biermann [4,9], Aleksandra Korenskaia [10], Mitja M. Zdouc [4], David Meijer [4], Barbara R. Terlouw [4], Justin J.J. van der Hooft [4,11], Nadine Ziemert [10], Eric J.N. Helfrich [9,12], Joleen Masschelein [7,8], Christophe Corre [6], Marc G. Chevrette [5], Gilles P. van Wezel [3], Marnix H. Medema [3,4,*], Tilmann Weber [1,*]

[1]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2950 Kongens Lyngby, Denmark
[2]Department of Ecology, Evolution and Marine Biology, University of California, 1169 Biological Sciences II, Santa Barbara, CA 93106, United States
[3]Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE Leiden, the Netherlands
[4]Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands
[5]Department of Microbiology and Cell Science, University of Florida, 1355 Museum Dr, Gainesville, FL 32603, United States
[6]School of Life Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom
[7]VIB-KU Leuven Center for Microbiology, Kasteelpark Arenberg 31, 3001 Leuven, Belgium
[8]Department of Biology, KU Leuven, Kasteelpark Arenberg 31, 3001 Leuven, Belgium
[9]Institute of Molecular Bio Science, Goethe-University Frankfurt, Max-von-Laue-Straße 9, 60438 Frankfurt am Main, Germany
[10]Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Auf der Morgenstelle 24, 72076 Tübingen, Germany
[11]Department of Biochemistry, University of Johannesburg, C2 Lab Building 224, Kingsway Campus, Cnr University & Kingsway Road, Auckland Park, Johannesburg 2006, South Africa
[12]Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

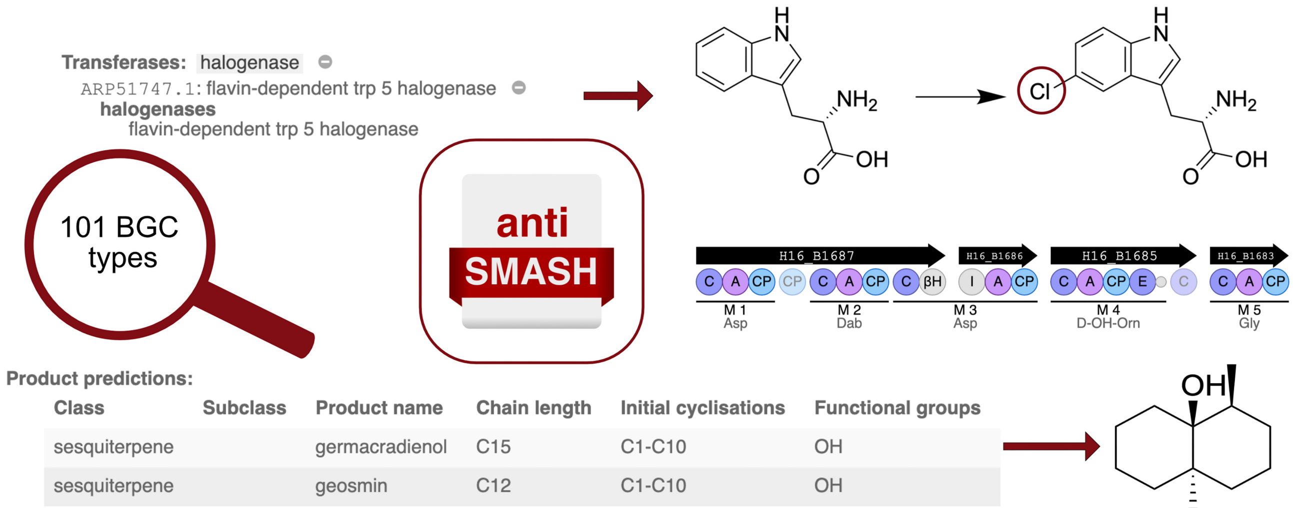
*To whom correspondence should be addressed. Email: kblin@biosustain.dtu.dk
Correspondence may also be addressed to Marnix H. Medema. Email: marnix.medema@wur.nl
Correspondence may also be addressed to Tilmann Weber. Email: tiwe@biosustain.dtu.dk

Abstract

Microorganisms synthesize small bioactive compounds through their secondary or specialized metabolism. Those compounds play an important role in microbial interactions and soil health, but are also crucial for the development of pharmaceuticals or agrochemicals. Over the past decades, advancements in genome sequencing have enabled the identification of large numbers of biosynthetic gene clusters directly from microbial genomes. Since its inception in 2011, antiSMASH (https://antismash.secondarymetabolites.org/), has become the leading tool for detecting and characterizing these gene clusters in bacteria and fungi. This paper introduces version 8 of antiSMASH, which has increased the number of detectable cluster types from 81 to 101, and has improved analysis support for terpenoids and tailoring enzymes, as well as improvements in the analysis of modular enzymes like polyketide synthases and nonribosomal peptide synthetases. These modifications keep antiSMASH up-to-date with developments in the field and extend its overall predictive capabilities for natural product genome mining.

Received: February 27, 2025. **Revised:** April 4, 2025. **Editorial Decision:** April 8, 2025. **Accepted:** April 15, 2025

Graphical abstract



Introduction

Small, bioactive molecules produced by microorganisms are an important source for drugs [1] and agrochemicals [2], and play important roles in microbial interactions and soil health. Over the past 20–25 years, the abundance of available genome data has allowed enhancing the traditional workflows of isolating microbial strains, extracting compounds, and then screening for desired activities by searching for biosynthetic gene clusters (BGCs) encoding for the biosynthesis of these molecules in microbial genomes [3]. Software tools for searching genomes for these secondary/specialized metabolite-producing BGCs have existed for over a decade [4–7].

Since its initial release in 2011, antiSMASH [8–14] has become the leading tool in the field. Around antiSMASH, a wide ecosystem of related tools that incorporate or rely on antiSMASH predictions has evolved. Examples include the resistance-based mining tool ARTS [15], the mass-spectrometry-guided Seq2PKS [16], the genome-engineering tool StreptoCAD [17], the BGC networking and clustering tool BiG-SCAPE [18], and the paired omics analysis tool NPLinker [19]. In turn, antiSMASH can also incorporate BGC predictions from other tools [13]. Originally built for DeepBGC [20], other machine-learning-based tools like GECCO [21] also provide their results in the required format. antiSMASH BGC predictions are also used in many genomic and BGC-related databases, such as the Joint Genome Institute's Secondary Metabolite Collaboratory [22], the MicroScope platform for genome annotation and analysis [23], the MIBiG database of manually curated BGCs [24], the BGC family database BiG-FAM [25], the chemical diversity metagenome database BGC Atlas [26], and the antiSMASH database [27]. Furthermore, antiSMASH is part of several systematic workflows for large-scale analyses of genomic data, e.g. BGCFlow [28] or the MicroOrganisms Pipelines Service [29] of the European Food Safety Authority (EFSA).

Here, we present version 8 of antiSMASH. This release increases the number of detectable biosynthetic pathway types from 81 to 101, adds an analysis module for terpenoid BGCs, and provides in-depth analysis of tailoring enzymes. Additionally, the KnownClusterBlast and ClusterCompare datasets were updated to reflect the data from MIBiG release 4, we added proper support for BGCs spanning the origin of replication in circular genomes, and transcription factor binding site predictions were extended with datasets from the CollecTF database [30]. In BGCs containing nonribosomal peptide synthetases (NRPSs) or type I polyketide synthases (PKSs), more biosynthetic domains are detected and analyzed.

New features and updates

BGC detection updates

antiSMASH uses manually curated rules to define what biosynthetic functions need to exist in a genomic region in order to define a BGC. To identify these biosynthetic functions, antiSMASH makes use of both profile hidden Markov models (pHMMs) and, to a lesser extent, dynamic profiles specified in Python code files. The pHMMs are sourced from public datasets such as PFAM [31], TIGRFAMS [32], SMART [33], BAGEL [34], Yadav *et al.* [35], or created specifically for use in antiSMASH. antiSMASH 7 contained 81 of such BGC rules [14], this number has increased to 101 in this release, with a number of existing rules having been refined. Due to the large overlap in biosynthetic enzymes, it is hard to differentiate between linear azole-containing peptides and the thio-linked circularized thiopeptides. In the current antiSMASH database [27], ∼16% of thiopeptide and ∼29% of linear azole-containing peptide BGC calls overlap, and it is unclear how many of the remaining assignments are correct. To address this uncertainty and avoid potential confusion, the detection rules for both types of BGCs were merged into a new "azole-containing RiPPs rule". The detection rules for terpenes, mycosporines, NRPS-independent siderophores, *trans*-AT PKSs, NRPSs, NRPS-like clusters, and fatty acids were also updated. Additionally, new rules for archaeal ribosomally synthesized and post-translationally modified peptides (RiPPs), atropopeptides, nitropropanoic acid, azoxy-containing compounds, polyynes, deazapurines, polyhalogenated pyrroles, hydroxytropolones, hydrogen-cyanides, darobactins, isocyanides, bacterial and fungal cyclic dipeptides, triceptides, highly reducing type II PKSs, and fungal NRPS-like lysine biosynthesis were added.

To cover cases where contiguous sections of some core genes were too far away from other core genes, a new “EXTENDS” condition was added to the rule definition to ensure that all core genes are properly marked. The *trans*-AT PKS rules use this to ensure that all genes containing modules are detected correctly, even when the single *trans*-acting acyltransferase (AT) domain is far away.

Terpene analysis

While antiSMASH has been detecting terpene gene clusters since version 1 [8], a more detailed analysis of the terpene synthases/cyclases was missing. antiSMASH version 4 [11] added an initial analysis module, but due to limitations in performance of the phylogenetic placement algorithm and maintainability challenges, this was dropped again in antiSMASH 5 [12]. For version 8, we see a return of the terpene analysis based on carefully curated pHMMs (see Fig. 1A). For every region containing a terpene BGC, a list of potential product types is shown. Every prediction includes the terpenoid class (e.g. diterpene, sesquiterpene, etc.) and chain length of the predicted product. For more well-understood terpene synthase subfamilies, the prediction can also contain the terpenoid subclass (e.g. indole diterpenoid), initial cyclizations (e.g. C1–C15), and product name.

Gene function analysis

During the big community push to better annotate tailoring enzymes in MIBiG entries during last year’s MIBiG annotathons (see [24] for a description), we noticed that a more user-friendly interface to access tailoring enzyme reactions was needed in antiSMASH. Using the collection of tailoring enzymes in the MITE database (https://mite.bioinformatics.nl/) [36] annotated during those annotathons, as well as our existing smCOG annotations, we now present tailoring enzyme information in a dedicated “tailoring” tab (see Fig. 1B). This also gave us an opportunity to report the substrate specificities of many flavin-dependent halogenases, also added in this release. Using custom pHMMs and conserved motif signatures, predictions range in detail from “halogenated pyrrole” to regioselectivity of e.g. tryptophan halogenases, depending on data availability.

In the tailoring tab, tailoring enzymes are organized by Enzyme Commission category, i.e. oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. Only categories with hits in the region are shown. Clicking the plus icons expands the category to show the genes with relevant hits and a short summary of the most detailed prediction of the tailoring function possible. Clicking on the plus icon of a gene of interest expands all information antiSMASH provides about a tailoring enzyme. If a tailoring enzyme shows at least 60% amino acid sequence identity to any entry in the MITE database, a cross-link to that MITE entry is provided.

NRPS and PKS improvements

antiSMASH provides a detailed analysis of protein domains encoded in NRPS/PKS BGCs. To provide a more comprehensive overview of the domains present, we added profiles for siderophore-associated β-hydroxylases and interface domains (see [37] for a detailed discussion) and a more generic α/β-hydrolase profile, as some enzymes with presumed proofreading functions were missed by the existing specific thioesterase profile.

CoA-ligase (CAL) domains are often involved in loading fatty acid-derived starter units in lipopeptide BGCs, but were not previously considered starting modules in the antiSMASH module detection. Together with the aforementioned β-hydroxylases and interface domains they can now be detected as part of modules.

Following the information collected in [38], the active sites of NRPS condensation (C) and epimerization (E) domains are now checked for the presence of catalytic residues and flagged as inactive when those residues are missing. To complement the NRPS adenylation (A) domain substrate specificity predictions already performed in antiSMASH, we now also provide a link to the external PARAS substrate specificity predictor [39] to provide researchers with even more analysis options.

Miscellaneous changes

Many other smaller changes have been included in version 8. The overview page view showing the most similar known clusters has been simplified to address some sources of user confusion. Instead of directly showing the cluster similarity in “percent of the genes having a sequence similarity of at least 30%”, we now show the similarity in three confidence levels: “high” for a cluster similarity of larger or equal to 75%, “medium” for a cluster similarity between 75% and 50%, and “low” for a cluster similarity between 50% and 15%. Cluster similarities of <15% are no longer considered to be similar enough and are no longer shown in the overview. KnownClusterBlast, ClusterCompare, and CompaRiPPson have been updated to the data provided in the MIBiG 4.0 release [24]. ClusterBlast SVG generation was moved into JavaScript to reduce file sizes and improve the user experience of viewing antiSMASH results opened locally. Historically, antiSMASH has struggled with BGCs spanning the origin in circular genomes. At best, those BGCs were split in two, at worst one or both of the parts on the opposite sides of the origin could be missed entirely. In antiSMASH 8.0, we have completely overhauled our coordinate handling for circular genomes to be detected and reported properly.

Historically, antiSMASH has relied on GlimmerHMM [40] for fungal gene calling, as it was able to run without manually selecting the right gene model. Unfortunately, with more and more fungal genomes becoming available, it became evident that this gene model selection is crucial for high-quality gene calling. As we cannot easily select the right gene model for uploaded genomes of unknown taxa, we have decided to remove fungal gene calling functionality from antiSMASH. We recommend users to run a dedicated gene-calling tool like AUGUSTUS [41] and then provide antiSMASH with the gene annotations. Due to software incompatibilities in modern systems, we have also deprecated our support for the MEME suite of tools, specifically MEME [42] and FIMO [43]. Users of the standalone version of antiSMASH can still provide those binaries themselves, but they are no longer part of the containers we provide or the web service. This effectively disables the CASSIS [44] fungal BGC border detection and might affect the RODEO [45] score for lanthipeptide, sactipeptide, lasso peptide, and azole-containing RiPP precursors.

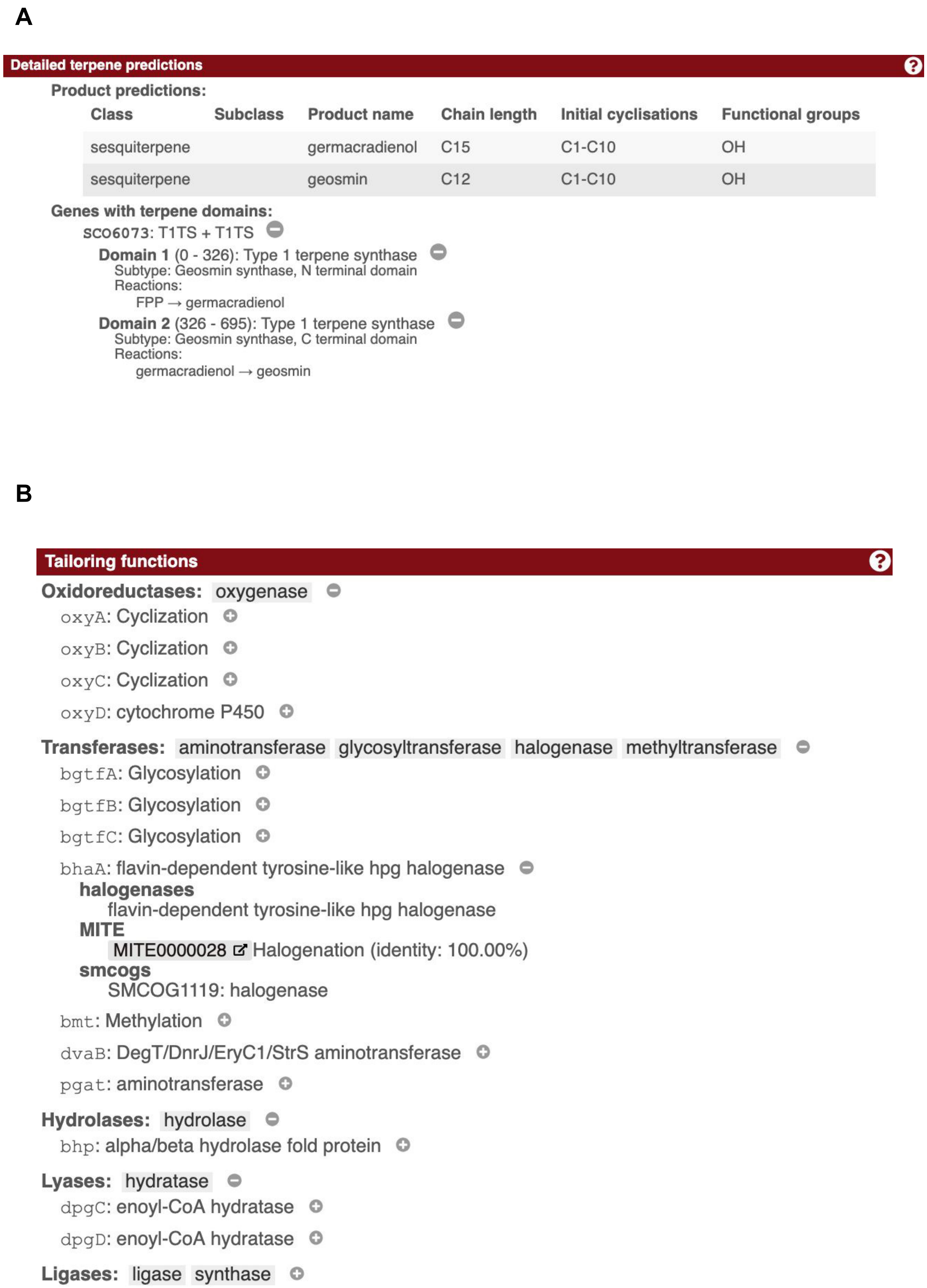


Figure 1. (**A**) The "terpene" tab of the geosmin BGC of *Streptomyces coelicolor* A3(2) (NCBI ID NC_003888.3:6656219–6678399) with all details expanded. The product predictions cover both the geosmin precursor germacradienol and geosmin itself. (**B**) Tailoring functions identified in the balhimycin BGC of *Amycolatopsis balhimycina* (NCBI ID Y16952.3). The oxidoreductase category has been expanded to show that all four P450 monooxygenases were identified. OxyA–C are correctly annotated as having a cyclization function due to their similarity to the MITE entries of the corresponding cyclization enzymes from the vancomycin biosynthesis. In the transferases category, the bhaA halogenase entry has been expanded further to show the predictions based on the built-in halogenase prediction, MITE similarity, and smCoG-based gene function annotations.

Conclusion and future perspective

Genome mining technologies like antiSMASH constitute an important piece of the natural product discovery puzzle. Over the past 14 years, antiSMASH has seen continuous updates and improvements, making sure it stays at the forefront of natural product genome mining tools. As evident by the numerous contributions from outside of the core development team, the antiSMASH project's Open Source and Open Science model remains successful. It is an important tool that many other, more specialized, tools, workflows, and databases rely on. It serves as the technology platform for a number of other genome mining tools currently in development. In future updates, we will continue our work on new algorithms and analysis modules to decipher the full biosynthesis pathways of detected clusters. We will build on the foundation of the tailoring enzyme prediction improvements in this version and cover even more enzyme families. Similarly, our work on regulator and regulator binding site detection will continue. As always, we will also integrate, or integrate with, new tools developing in the ecosystem.

Acknowledgements

Author contributions: Kai Blin (Conceptualization [lead], Data curation [lead], Funding acquisition [equal], Methodology [lead], Project administration [lead], Software [equal], Supervision [equal], Writing—original draft [lead], Writing—review & editing [lead]), Simon Shaw (Data curation [equal], Methodology [equal], Software [lead], Visualization [lead]), Lisa Vader (Data curation [supporting], Software [supporting], Visualization [supporting]), Judit Szenei (Data curation [supporting], Software [supporting], Visualization [supporting]), Zachary L. Reitz (Data curation [supporting], Software [supporting]), Hannah E. Augustijn (Data curation [supporting], Software [supporting]), José D.D. Cediel-Becerra (Data curation [supporting], Software [supporting]), Valérie de Crécy-Lagard (Data curation [supporting]), Robert A. Koetsier (Data curation [supporting], Software [supporting]), Sam E. Williams (Data curation [supporting]), Pablo Cruz-Morales (Data curation [supporting]), Sopida Wongwas (Data curation [supporting]), Alejandro E. Segurado Luchsinger (Data curation [supporting]), Friederike Biermann (Data curation [supporting]), Aleksandra Korenskaia (Data curation [supporting]), Mitja M. Zdouc (Data curation [supporting], Writing—original draft [supporting]), David Meijer (Data curation [supporting]), Barbara R. Terlouw (Data curation [supporting]), Justin J.J. van der Hooft (Funding acquisition [equal], Writing—original draft [supporting]), Nadine Ziemert (Funding acquisition [equal], Supervision [equal]), Eric J.N. Helfrich (Funding acquisition [equal], Supervision [equal], Writing—original draft [supporting]), Joleen Masschelein (Data curation [supporting], Funding acquisition [equal], Supervision [equal]), Christophe Corre (Funding acquisition [equal], Supervision [equal]), Marc G. Chevrette (Funding acquisition [equal], Supervision [equal]), Gilles P. van Wezel (Funding acquisition [equal], Supervision [equal]), Marnix H. Medema (Conceptualization [supporting], Funding acquisition [equal], Supervision [equal], Writing—original draft [supporting]), and Tilmann Weber (Conceptualization [supporting], Funding acquisition [equal], Supervision [equal], Validation [supporting], Writing—original draft [supporting])

Conflict of interest

J.J.J.vdH. is member of the Scientific Advisory Board of NAICONS Srl, Milano, Italy and consults for Corteva Agriscience, Indianapolis, IN, USA. M.H.M. is a member of the Scientific Advisory Board of Hexagon Bio. All other authors declare to have no conflicts of interest.

Funding

Novo Nordisk Foundation [NNF20CC0035580 to T.W., K.B., and P.C.-M., NNF22OC0079021 to S.E.W.]; Center for Microbial Secondary Metabolites (CeMiSt), Danish National Research Foundation [DNRF137 to T.W.]; EU Horizon Europe Programme MAGic-MOLFUN [MSC 101072485 to T.W., K.B., M.H.M., N.Z., A.K., and J.J.J.vdH.]; EU Horizon 2020 MARBLES [101000392 to M.M.Z., J.J.J.vdH., M.H.M., and G.P.vW.]; NWO [KICH1.LWV04.21.013 to M.M.Z., J.J.J.vdH., and M.H.M.]; EU ERC Community [101055020 to G.P.vW.]; NWO-XL [OCENW.XL21.XL21.088 to R.A.K.]; National Institutes of Health [RM1GM145426 to V.dC.-L.]; German Center for Infection Research (DZIF) [TTU09.716 to N.Z.]; EU Horizon 2020 DECIPHER [948770 to D.M. and M.H.M.]; The Royal Thai Government [PhD funding to S.W.]; Emmy Noether Program of the German Research Foundation [504947087 to E.J.N.H.], Hessian Ministry for Science and the Arts [LOEWE Center for Translational Biodiversity Genomics to E.J.N.H.]; FWO [I008520N and G061821N to J.M.]. Funding to pay the Open Access publication charges for this article was provided by the EU Horizon Europe Program MAGic-MOLFUN.

Data availability

The bacterial and fungal versions of antiSMASH 8.0 can be freely accessed at https://antismash.secondarymetabolites.org and https://fungismash.secondarymetabolites.org, respectively.

The antiSMASH documentation is available at https://docs.antismash.secondarymetabolites.org/.

The antiSMASH source code is licensed under the GNU Affero General Public License (AGPL) version 3.0. antiSMASH is also available via Docker. See the documentation website for details on how to download and install antiSMASH.

References

1. Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *BMC Genomics* 2020;**83**:1138–43. https://doi.org/10.1021/acs.jnatprod.9b01285
2. 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
3. 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
4. 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
5. Medema MH, Fischbach MA. Computational approaches to natural product discovery. *Nat Chem Biol* 2015;**11**:639–48. https://doi.org/10.1038/nchembio.1884
6. 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

7. 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
8. Medema MH, Blin K, Cimermancic 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
9. 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
10. 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
11. 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
12. 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
13. 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
14. 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
15. Mungan MD, Alanjary M, Blin K *et al.* ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Res* 2020;**48**:W546–52. https://doi.org/10.1093/nar/gkaa374
16. Yan D, Zhou M, Adduri A *et al.* Discovering type I cis-AT polyketides through computational mass spectrometry and genome mining with Seq2PKS. *Nat Commun* 2024;**15**:5356. https://doi.org/10.1038/s41467-024-49587-1
17. Levassor L, Whitford CM, Petersen SD *et al.* StreptoCAD: an open-source software toolbox automating genome engineering workflows in streptomycetes. bioRxiv, https://doi.org/10.1101/2024.12.19.629370, 20 December 2024, preprint: not peer reviewed.
18. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 2020;**16**:60–8. https://doi.org/10.1038/s41589-019-0400-9
19. Eldjárn GH, Ramsay A, Hooft JJJvd *et al.* Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLOS Comput Biol* 2021;**17**:e1008920. https://doi.org/10.1371/journal.pcbi.1008920
20. Hannigan GD, Prihoda D, Palicka A *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* 2019;**47**:e110. https://doi.org/10.1093/nar/gkz654
21. Carroll LM, Larralde M, Fleck JS *et al.* Accurate de novo identification of biosynthetic gene clusters with GECCO. bioRxiv, https://doi.org/10.1101/2021.05.03.442509, 4 May 2021, preprint: not peer reviewed.
22. Udwary DW, Doering DT, Foster B *et al.* The secondary metabolism collaboratory: a database and web discussion portal for secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* 2025;**53**:D717–23. https://doi.org/10.1093/nar/gkae1060
23. Vallenet D, Calteau A, Dubois M *et al.* MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Res* 2020;**48**:D579–89. https://doi.org/10.1093/nar/gkz926
24. 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
25. Kautsar SA, Blin K, Shaw S *et al.* BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res* 2021;**49**:D490–7. https://doi.org/10.1093/nar/gkaa812
26. Bağcı C, Nuhamunada M, Goyat H *et al.* BGC Atlas: a web resource for exploring the global chemical diversity encoded in bacterial genomes. *Nucleic Acids Res* 2025;**53**:D618–24. https://doi.org/10.1093/nar/gkae953
27. 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
28. Nuhamunada M, Mohite OS, Phaneuf PV *et al.* BGCFlow: systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. *Nucleic Acids Res* 2024;**52**:5478–95. https://doi.org/10.1093/nar/gkae314
29. Yolanda GC. Microorganisms pipelines service (MOPS): pipelines documentation. Technical report.
30. Kılıç S, White ER, Sagitova DM *et al.* CollecTF: a database of experimentally validated transcription factor-binding sites in bacteria. *Nucleic Acids Res* 2014;**42**:D156–60. https://doi.org/10.1093/nar/gkt1123
31. Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9. https://doi.org/10.1093/nar/gkaa913
32. Haft DH, Selengut JD, Richter RA *et al.* TIGRFAMs and genome properties in 2013. *Nucleic Acids Res* 2013;**41**:D387–95. https://doi.org/10.1093/nar/gks1234
33. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res* 2021;**49**:D458–60. https://doi.org/10.1093/nar/gkaa937
34. van Heel AJ, de Jong A, Song C *et al.* BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res* 2018;**46**:W278–81. https://doi.org/10.1093/nar/gky383
35. Yadav G, Gokhale RS, Mohanty D. Towards prediction of metabolic products of polyketide synthases: an *in silico* analysis. *PLOS Comput Biol* 2009;**5**:e1000351. https://doi.org/10.1371/journal.pcbi.1000351
36. Zdouc MM, Meijer D, Biermann F *et al.* The minimum information about a tailoring enzyme/maturase data standard for capturing natural product biosynthesis. ChemRxiv, https://doi.org/10.26434/chemrxiv-2024-78mtl, 9 April 2024, preprint: not peer reviewed.
37. Reitz ZL, Hardy CD, Suk J *et al.* Genomic analysis of siderophore β-hydroxylases reveals divergent stereocontrol and expands the condensation domain family. *Proc Natl Acad Sci USA* 2019;**116**:19805–14. https://doi.org/10.1073/pnas.1903161116
38. Bloudoff K, Schmeing TM. Structural and functional aspects of the nonribosomal peptide synthetase condensation domain superfamily: discovery, dissection and diversity. *Biochim Biophys Acta Proteins Proteom* 2017;**1865**:1587–604. https://doi.org/10.1016/j.bbapap.2017.05.010
39. Terlouw BR, Huang C, Meijer D *et al.* PARAS: high-accuracy machine-learning of substrate specificities in nonribosomal peptide synthetases. bioRxiv, https://doi.org/10.1101/2025.01.08.631717, 10 January 2025, preprint: not peer reviewed.
40. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* 2004;**20**:2878–9. https://doi.org/10.1093/bioinformatics/bth315
41. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 2005;**33**:W465–7. https://doi.org/10.1093/nar/gki458

42. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994;**2**:28–36.
43. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;**27**:1017–8. https://doi.org/10.1093/bioinformatics/btr064
44. Wolf T, Shelest V, Nath N *et al.* CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics* 2016;**32**:1138–43. https://doi.org/10.1093/bioinformatics/btv713
45. Walker MC, Eslami SM, Hetrick KJ *et al.* Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *BMC Genomics* 2020;**21**:387. https://doi.org/10.1186/s12864-020-06785-7

Received: February 27, 2025. **Revised:** April 4, 2025. **Editorial Decision:** April 8, 2025. **Accepted:** April 15, 2025