

*antiSMASH 5*  
*antiSMASH database*  
**MANUAL**

Tilmann Weber

Kai Blin

Simon Shaw

The Novo Nordisk Foundation Center for Biosustainability

Technical University of Denmark

Kemitorvet bygning 220

2800 Kgs. Lyngby

Denmark

Email: [tiwe@biosustain.dtu.dk](mailto:tiwe@biosustain.dtu.dk)

[kblin@biosustain.dtu.dk](mailto:kblin@biosustain.dtu.dk)

[sisha@biosustain.dtu.dk](mailto:sisha@biosustain.dtu.dk)



The Novo Nordisk Foundation  
Center for Biosustainability

Technical University  
of Denmark



## Table of Content

<b>1</b>	<b>Introduction.....</b>	<b>3</b>
<b>2</b>	<b>Access to antiSMASH .....</b>	<b>4</b>
2.1	Web-access to antiSMASH .....	4
2.2	Web-access to the antiSMASH database.....	4
2.3	Local installation of standalone antiSMASH.....	4
<b>3</b>	<b>Suggested reading and further information.....</b>	<b>5</b>
3.1	The Secondary Metabolite Bioinformatics Portal.....	5
3.2	The antiSMASH publications .....	5
3.3	Tools integrated into antiSMASH .....	5
3.4	The antiSMASH database publications.....	6
3.5	Genome Mining related reviews .....	6
<b>4</b>	<b>Using the antiSMASH webserver .....</b>	<b>8</b>
4.1	Job submission .....	8
4.1.1	Job submission with standard parameters (recommended).....	8
4.1.2	Extended parameters .....	10
4.2	antiSMASH results .....	12
4.2.1	Cluster overview table .....	12
4.2.2	The antiSMASH 5 region concept .....	13
4.2.3	Detailed results view .....	15
4.2.4	Gene cluster details: Click on gene arrow: .....	16
4.4	Downloading results.....	20
4.5	Obtaining antiSMASH .....	21
<b>6</b>	<b>The antiSMASH database .....</b>	<b>22</b>
6.1	Browsing the antiSMASH database .....	22
6.2	Querying the antiSMASH database .....	22
6.2.1	Simple search.....	23
6.2.2	Complex searches.....	23
<b>7</b>	<b>References .....</b>	<b>25</b>

## 1 Introduction

Many microbial genomes contain several (up to 30-40) gene clusters encoding the biosynthesis of secondary metabolites. Thus mining genetic data has become a very important method in modern screening approaches for bioactive compounds like antibiotics or chemotherapeutics (1).

The antibiotics and secondary metabolites analysis shell antiSMASH is a comprehensive pipeline for the automated mining of finished or draft genome data for the presence of secondary metabolite biosynthetic gene clusters (2–8).

antiSMASH is an Open Source software written in Python. The development is coordinated by Tilmann Weber/Kai Blin (DTU) and Marnix Medema (Wageningen University). There are also many international contributors to the software.

## 2 Access to antiSMASH

### 2.1 Web-access to antiSMASH

#### Stable versions:

For bacterial sequences: <https://antismash.secondarymetabolites.org/>

For fungal sequences: <https://fungismash.secondarymetabolites.org>

For plant sequences: <https://plantismash.secondarymetabolites.org>

#### Latest development versions (use on own risk):

For bacterial sequences: <https://dev.antismash.secondarymetabolites.org/>

For fungal sequences: <https://dev.fungismash.secondarymetabolites.org/>

### 2.2 Web-access to the antiSMASH database

The antiSMASH database is located at <https://antismash-db.secondarymetabolites.org/>

### 2.3 Local installation of standalone antiSMASH

antiSMASH can be installed locally on Linux computers or is available as Docker container or via bioconda (9)(also for Mac).

### 3 Suggested reading and further information

#### 3.1 The Secondary Metabolite Bioinformatics Portal

A web portal describing all software and databases relevant for Natural Products research can be found at <https://www.secondarymetabolites.org/>.

#### 3.2 The antiSMASH publications

1. Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E. and Breitling, R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**, W339-W346.  
<http://www.ncbi.nlm.nih.gov/pubmed/21672958>  
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125804/pdf/gkr466.pdf>
2. Blin, K., Medema, M.H., Kazempour, D., Fischbach, M., Breitling, R., Takano, E. and Weber, T. (2013) antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.*, **41**, W204-W212.  
<http://www.ncbi.nlm.nih.gov/pubmed/23737449>  
<http://nar.oxfordjournals.org/content/early/2013/06/03/nar.gkt449.full.pdf>
3. Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Brucolieri, R., Lee, S.Y., Fischbach, M.A., Müller, R., Wohlleben, W. et al. (2015) antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.*, **43**, W237-W243.  
<http://www.ncbi.nlm.nih.gov/pubmed/25948579>  
<http://nar.oxfordjournals.org/content/43/W1/W237.full.pdf>
4. Blin, K., Wolf, T., Chevrette, M.G., Lu, X., Schwalen, C.J., Kautsar, S.A., Suarez Duran, H.G., de los Santos, E.L.C., Kim, H.U., Nave, M. et al. (2017) antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**, W36-W41.  
<https://www.ncbi.nlm.nih.gov/pubmed/28460038>  
<https://academic.oup.com/nar/article/45/W1/W36/3778252>

#### 3.3 Tools integrated into antiSMASH

1. Villebro, R., Shaw, S., Blin, K. and Weber, T. (2019) Sequence-based classification of type II polyketide synthase biosynthetic gene clusters for antiSMASH. *J. Ind. Microbiol. Biotechnol.*, doi: 10.1007/s10295-10018-02131-10299.  
<https://www.ncbi.nlm.nih.gov/pubmed/30610412>  
<https://link.springer.com/article/10.1007%2Fs10295-018-02131-9>
2. Tietz, J.I., Schwalen, C.J., Patel, P.S., Maxson, T., Blair, P.M., Tai, H.C., Zakai, U.I. and Mitchell, D.A. (2017) A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.*, **13**, 470-478.  
<https://www.ncbi.nlm.nih.gov/pubmed/28244986>  
<http://www.nature.com/nchembio/journal/v13/n5/pdf/nchembio.2319.pdf>
3. Chevrette, M.G., Aicheler, F., Kohlbacher, O., Currie, C.R. and Medema, M.H. (2017) SANDPUMA: Ensemble Predictions of Nonribosomal Peptide Chemistry Reveals Biosynthetic Diversity across

Actinobacteria. *Bioinformatics*, **33**, 3202-3210.

<https://www.ncbi.nlm.nih.gov/pubmed/28633438>

4. Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., de Bruijn, I., Chooi, Y.H., Claesen, J., Coates, R.C. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625-631.  
<http://www.ncbi.nlm.nih.gov/pubmed/26284661>  
<http://www.nature.com/nchembio/journal/v11/n9/pdf/nchembio.1890.pdf>
5. Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J. *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, **158**, 412-421.  
<http://www.ncbi.nlm.nih.gov/pubmed/25036635>
6. Blin, K., Kazempour, D., Wohlleben, W. and Weber, T. (2014) Improved lanthipeptide detection and prediction for antiSMASH. *PLoS ONE*, **9**, e89420.  
<http://www.ncbi.nlm.nih.gov/pubmed/24586765>  
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3930743/pdf/pone.0089420.pdf>

### 3.4 The antiSMASH database publications

1. Blin, K., Medema, M.H., Kottmann, R., Lee, S.Y. and Weber, T. (2017) The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **45**, D555-D559.  
<http://nar.oxfordjournals.org/content/early/2016/10/24/nar.gkw960.full.pdf>  
<https://www.ncbi.nlm.nih.gov/pubmed/27924032>
2. Blin, K., Pascal Andreu, V., de los Santos, E.L., Del Carratore, F., Lee, S.Y., Medema, M.H. and Weber, T. (2019) The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **47**, D625-D630.  
<https://www.ncbi.nlm.nih.gov/pubmed/30395294>

### 3.5 Genome Mining related reviews

1. Kim, H.U., Blin, K., Lee, S.Y. and Weber, T. (2017) Recent development of computational resources for new antibiotics discovery. *Curr. Opin. Microbiol.*, **39**, 113-120.  
<https://www.ncbi.nlm.nih.gov/pubmed/29156309>
2. Blin, K., Kim, H.U., Medema, M.H. and Weber, T. (2017) Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief. Bioinform.*, doi: 10.1093/bib/bbx1146.
3. Ziemert, N., Alanjary, M. and Weber, T. (2016) The evolution of genome mining in microbes – a review. *Nat. Prod. Rep.*, **33**, 988-1005.  
<http://pubs.rsc.org/en/content/articlepdf/2016/np/c6np00025h>  
<https://www.ncbi.nlm.nih.gov/pubmed/27272205>
4. Leclère, V., Weber, T., Jacques, P. and Pupin, M. (2016) In Evans, B. S. (ed.), *Nonribosomal peptide and polyketide biosynthesis: methods and protocols*. Springer Science and Business Media, New York, Vol. 1401, pp. 209-232.
5. Zhao, H. and Medema, M.H. (2016) Standardization for natural product synthetic biology. *Nat. Prod. Rep.*, **33**, 920-924.  
<http://www.ncbi.nlm.nih.gov/pubmed/27313083>  
<http://pubs.rsc.org/en/content/articlepdf/2016/np/c6np00030d>

6. Medema, M.H. and Osbourn, A. (2016) Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Nat. Prod. Rep.*, **33**, 951-962.  
<https://www.ncbi.nlm.nih.gov/pubmed/27321668>  
<http://pubs.rsc.org/en/content/articlepdf/2016/np/c6np00035e>

## 4 Using the antiSMASH webserver

### 4.1 Job submission

#### 4.1.1 Job submission with standard parameters (recommended)

antiSMASH can work with different formats of data. Data can be uploaded in Genbank (recommended), EMBL, or plain FASTA format. If the user uploads FASTA files, which do not contain any annotation or annotated coding sequences, putative genes are identified with Prodigal (10) (for bacterial sequences) or GlimmerHMM (11) (for fungal sequences). If the NCBI accession number is known, antiSMASH can also automatically retrieve the data from NCBI.

If you work with draft genome sequences, it is preferable to use scaffolded sequences containing “N” characters for the gaps, as gene clusters can only be identified if positional information is available.

#### Note:

*The quality of the prediction by antiSMASH is highly dependent on the quality of the input data. If you analyze poor quality draft genomes with many (thousands) of small contigs and low N<sub>50</sub>, antiSMASH may not be able to detect gene clusters if they are scattered over multiple contigs.*

*Please be also aware, that many of the specificity determination algorithms in antiSMASH depend on the presence/absence of conserved amino acids at specific positions in the enzyme sequence. If the sequence quality is low and thus also the amino acid of the translated proteins cannot be fully trusted all predictions have to be taken with precautions.*

*For more detailed descriptions about caveats in genome mining, please see (7)*

## Job submission:

1. Open antiSMASH homepage <https://antismash.secondarymetabolites.org> for bacterial sequences, <https://fungismash.secondarymetabolites.org> for fungal or <http://plantismash.secondarymetabolites.org/> for plant sequences

Screenshot of the bacterial antiSMASH job submission page:

2. To start an analysis, enter your email address (optional, but highly recommended → you get an email when your results have been processed) and either
  - Enter Genbank or RefSeq accession number to directly download sequence from NCBI
  - Upload your sequence by using the “Upload file” button and selecting the sequence file (Fasta or GenBank format) to upload
3. Press “Submit” button at the end of the page. If you did not provide an email address, please bookmark the link to this page – otherwise you will not be able to access the results.
4. Wait...

⇒ ***For a typical bacterial genome computing time is ~0.5-2 hours under normal server load; if you select the long runtime options, a typical analysis will take several days.***

#### 4.1.2 Extended parameters

If you want to analyze bacterial sequences, the standard antiSMASH parameters just work fine. Still, the antiSMASH run can be influenced by optional parameters:

*For bacterial sequences (<https://antismash.secondarymetabolites.org/>)*

The screenshot shows the antiSMASH web interface for bacterial sequences. At the top, there are buttons for 'Extra features' (All off, All on) and a note about being considerate in job usage. Below this, several analysis options are listed with checkboxes:

- KnownClusterBlast
- ClusterBlast
- SubClusterBlast
- ActiveSiteFinder
- Cluster Pfam analysis
- Pfam-based GO term annotation

A large 'Submit' button is centered at the bottom of the form.

Please be considerate in your use of antiSMASH. Help us keep antiSMASH available for everybody by limiting yourself to 5 concurrent jobs. Need to run more? See the [antiSMASH install guide](#) for instructions for getting your own antiSMASH installation.

**KnownClusterBlast analysis** (default: on): If this option is enabled, the identified clusters are searched against the MIBiG repository (13). MIBiG is a hand curated data collection on gene clusters, which have been experimentally characterized.

**ClusterBlast analysis** (default: off): If this option is enabled, the identified clusters are searched against a comprehensive gene cluster database and similar clusters are identified. (Selecting this option increases runtime significantly). The algorithm used is similar to MultiGeneBlast (14)

**Subcluster Blast analysis** (default: on): If this option is enabled, the identified clusters are searched against a database containing operons involved in the biosynthesis of common secondary metabolite building blocks (e.g. the biosynthesis of non-proteinogenic amino acids)

**smCOG analysis for functional prediction and phylogenetic analysis of genes** (always on): If this option is enabled, each gene of the cluster is compared to a database of clusters of orthologous groups of proteins involved in secondary metabolism. This information then is used to provide an annotation of the putative function of the gene products.

**Active site finder** (default: on): If this option is enabled, active sites of several highly conserved biosynthetic enzymes are detected and variations of the active sites are reported.

**Detect TTA codons** (always on for high GC genomes): Some high-GC bacteria, for example streptomycetes, use the rare Leu-codon “TTA” as a mean for post-transcriptional regulation by limiting/controlling the amount of TTA-tRNA in the cell. This type of regulation is commonly found in secondary metabolite BGCs. Selecting this option will annotate such TTA codons in the identified BGCs

**Cluster PFAM analysis** (default: off): If this option is enabled, each gene product encoded in the detected BGCs is analyzed against the PFAM database (15). Hits are annotated in the final Genbank/EMBL files that can be downloaded after the analysis is finished. Note: These

results are not displayed on the antiSMASH HTML results page. Selecting this option normally increases the runtime.

**NOTE:**

*Due to very high runtime requirements and our observation that results generated with the ClusterFinder option were often misinterpreted, we no longer support ClusterFinder via the antiSMASH web interface. If you are interested in ClusterFinder (and are aware on how to interpret the data), it still is included in the download version of antiSMASH.*

**ClusterFinder:** ClusterFinder is an alternative method to detect gene clusters is used. By default, clusters are detected based on a manually curated set of rules (e.g. NRPS cluster is assigned genes encoding enzymes which have condensation, adenylation and PCP domains are found). The alternative ClusterFinder approach (12) uses a statistical method to identify regions, which accumulate genes encoding for enzymes that are commonly found in the context of secondary metabolite biosynthesis. Using this method, it is possible to identify cluster types, which are not included in the ruleset – but it will give many false positive predictions.

**Use ClusterFinder algorithm for BGC border prediction:** The information obtained by ClusterFinder can be used to predict potential borders of the BGCs. When selecting this option, these border predictions are included in the output.

For fungal sequences (<https://fungismash.secondarymetabolites.org>)

Extra features    All off    All on

---

<input checked="" type="checkbox"/> KnownClusterBlast	<input type="checkbox"/> ClusterBlast	<input checked="" type="checkbox"/> SubClusterBlast
<input checked="" type="checkbox"/> smCoG analysis	<input checked="" type="checkbox"/> ActiveSiteFinder	<input type="checkbox"/> Whole-genome PFAM analysis
<input type="checkbox"/> Cluster-border prediction based on transcription factor binding sites (CASSIS)		

Extra features for the analysis of fungal sequences:

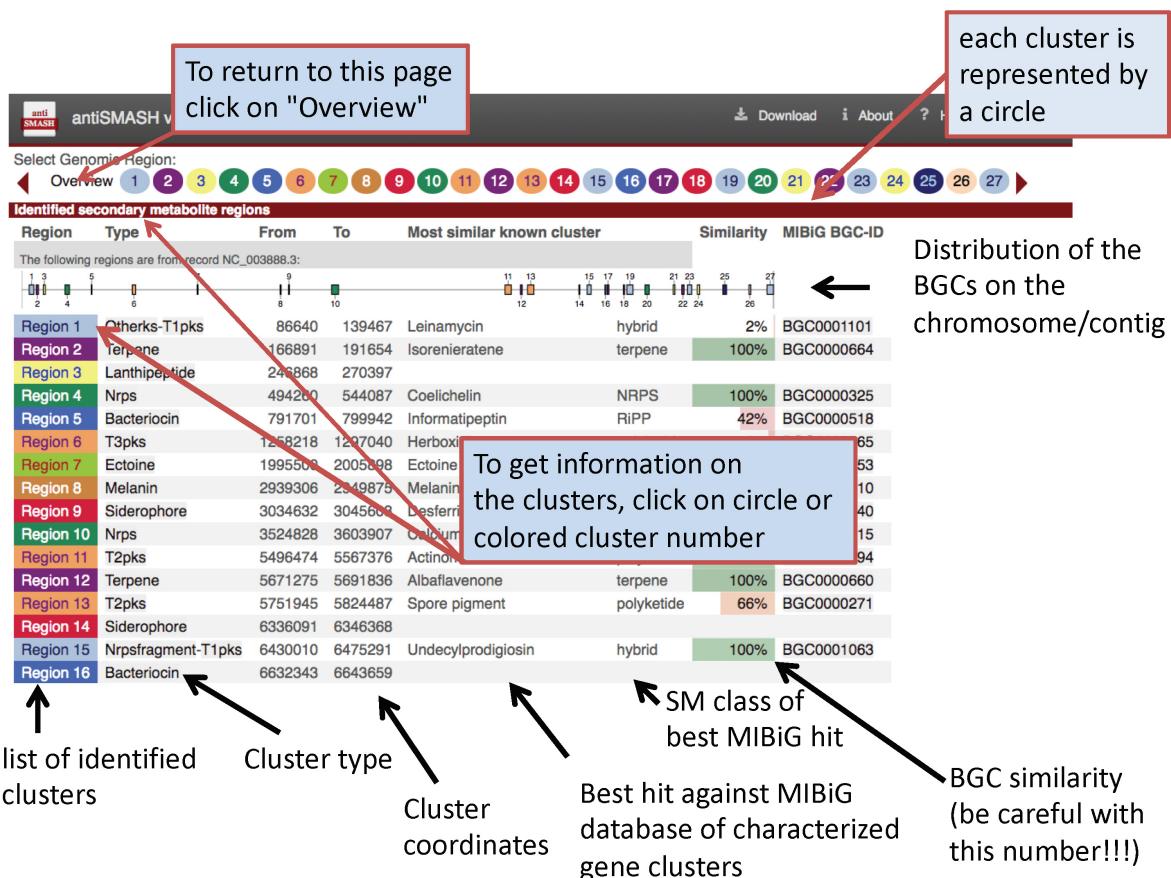
**Cluster-border prediction based on transcription factor binding sites (CASSIS):** In fungi, it is possible to predict clustered genes by identifying highly conserved regulator binding motifs in the promotor regions of the biosynthesis genes. The CASSIS algorithm (16) can detect such motifs and thus predict core-regions of fungal BGCs. In bacteria, these motifs unfortunately are less conserved, thus this approach only works on fungal sequences)

## 4.2 antiSMASH results

Once the antiSMASH job has been submitted on the webpage, the calculations for the identification and analysis of the clusters are carried out on our servers. When the job is finished – and an email address was provided – the user gets a notification email that the results are available.

### 4.2.1 Cluster overview table

Each identified gene cluster is represented as a colored circle at top of the screen. Below, a detailed list of clusters is displayed including coordinates of the identified gene cluster



#### 4.2.2 The antiSMASH 5 region concept

Currently, there is no good method available to accurately predict gene cluster borders (with exception with the CASSIS algorithm that can detect co-regulated genes in fungal genomes – however CASSIS doesn't work in prokaryotes)

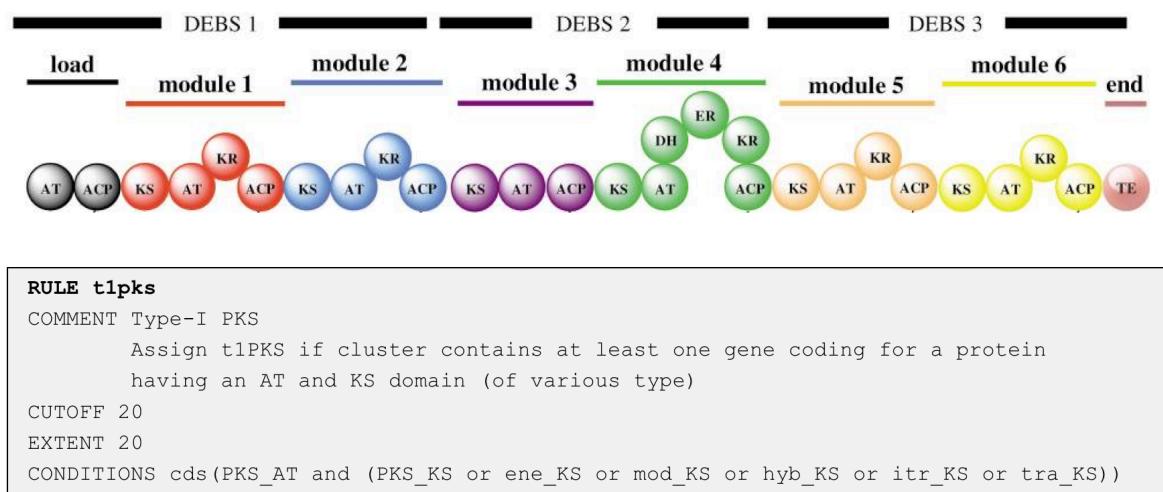
Therefore, antiSMASH5 changed the way gene clusters are displayed – to reflect the fact that the BGC borders are just offsets defined in the cluster detection rules we renamed the highest level that is displayed to “Region”. A “Region” in antiSMASH 5 corresponds to the “Gene Cluster” annotation in antiSMASH 4 and below.

#### How are antiSMASH 5 regions defined?

In the first step, all gene products of the analyzed sequence are searched against a database of highly conserved enzyme HMM profiles (core-enzymes), which are indicative of a specific BGC type.

In a second step, pre-defined cluster rules are employed to define individual “Clusters” encoded in the Region.

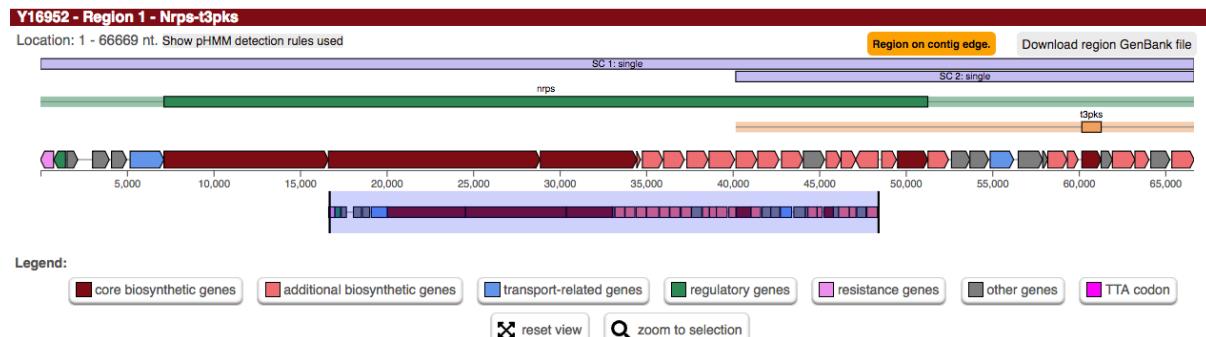
As an example an excerpt of the cluster rules to detect type I PKS:



Whenever antiSMASH finds a gene coding for a protein that has as PKS AT domain and a PKS KS domain of various sub-types, antiSMASH, a new “Type I PKS Cluster” feature is generated within the region; this feature comprises the core gene-product(s) that trigger the rule (i.e. the PKS encoding genes and extends to the left and right of the core genes by 20 kb to the left and right (as defined by the EXTENT parameter in the rule definition). The values for the different cluster types are empirically determined and generally tend to rather overpredict, i.e. included also adjacent genes.

After the “Cluster” features are assigned (note: there can be multiple cluster features in a region!!), they are checked for overlaps (as defined by the CUTOFF parameter) and are grouped into several types of “Candidate Cluster” to reflect the observation that many BGCs

actually comprise several classes of biosynthetic machinery. For example Glycopeptides like vancomycin or balhimycin (shown) are synthesized via NRPS, but also contain a type III PKS as a precursor pathway.



Thus, the displayed region contains the “nrps-cluster” (green bar) and a “type III PKS cluster” yellow bar. As their extents overlap, they are assigned to Candidate Clusters.

### There are different types of Candidate Clusters:

#### chemical hybrid:

contains clusters which share Cluster-defining CDSFeatures, will also include clusters within that shared range that do not share a CDS provided that they are completely contained within the Candidate Cluster border, e.g.

```
--#A#####C---      <- Cluster 1 with definition CDSes A and C
--#A#---          <- Cluster 2 with definition CDS A
--#B#--          <- Cluster 3 with definition CDS B
--#C#--          <- Cluster 4 with definition CDS C
-#D#-           <- Cluster 5 with definition CDS D
```

Since clusters 1 and 2 share a CDS that defines those clusters, a chemical hybrid Candidate Cluster exists. Clusters 1 and 4 also share a defining CDS, so the hybrid Candidate Cluster now contains clusters 1, 2 and 4.

Cluster 3 does not share a defining CDS with either cluster 1, 2 or 4, but because it is interleaved into a chemical hybrid it is included under the assumption that it is relevant to the other clusters.

#### interleaved:

contains clusters which do not share Cluster-defining CDS features, but their core locations overlap, e.g.

```
--#A##A###A---      <- Cluster 1 with defining CDSes marked A
---B##B###B---      <- Cluster 2 with defining CDSes marked B
---C##C---          <- Cluster 3 with defining CDSes marked C
```

Since none of the clusters share any defining CDS with any other cluster, it is not a chemical hybrid. All three clusters would be part of an interleaved Candidate Cluster, since A overlaps with B and B overlaps with C.

### neighbouring:

contains clusters which transitively overlap in their neighbourhoods (the '-' sections in the examples above). In the chemical hybrid example, as all clusters overlap in some way, all 5 would be part of a neighbouring Candidate Cluster (with clusters 1-4 also being part of a hybrid Candidate Cluster). Every cluster in a 'neighbouring' cluster will also belong to one of the other kinds of Candidate Cluster.

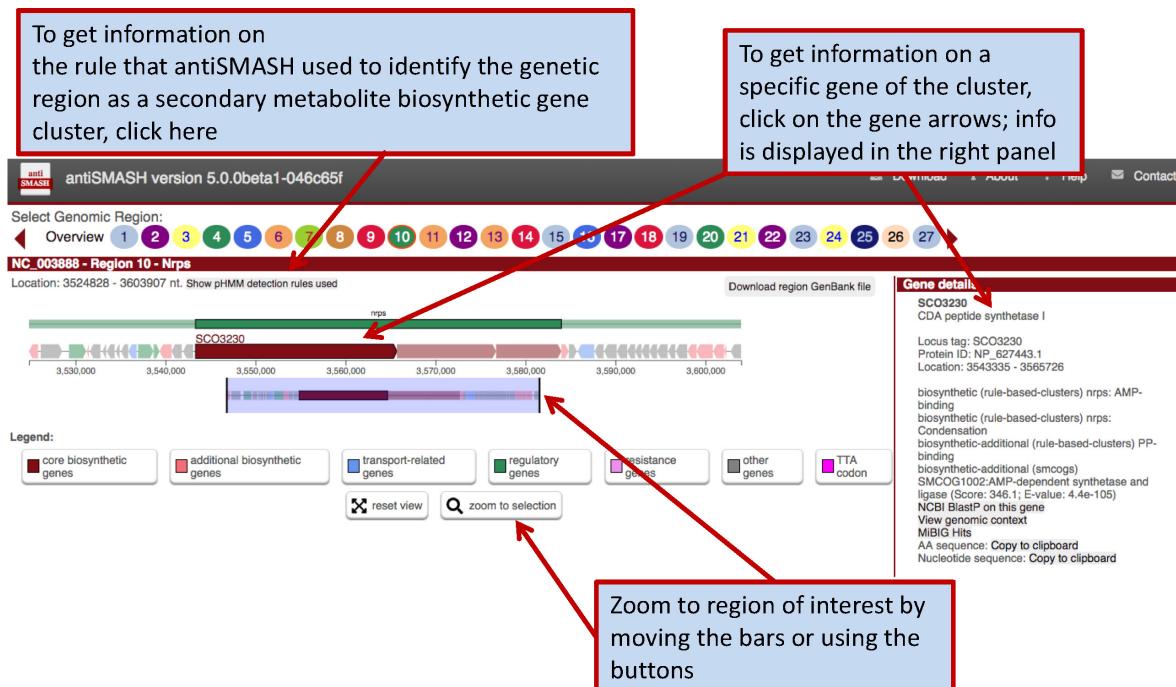
### single:

the kind for all Candidate Clusters where only one cluster is contained, only exists for consistency of access. A 'single' Candidate Cluster will not exist for a cluster which is contained in either a chemical hybrid or an interleaved Candidate Cluster. In the chemical hybrid example, only cluster 5 would be in a 'single' Candidate Cluster as well as in the 'neighbouring' Candidate Cluster.

Finally, the largest Candidate Cluster defines the extent of the Region.

### 4.2.3 Detailed results view

A click at one of the clusters leads to an interactive webpage displaying a detailed view of the respective cluster



#### 4.2.4 Gene cluster details: Click on gene arrow:

##### Gene details

SCO3230  
CDA peptide synthetase I

RefSeq/GenBank annotation  
(not generated by antiSMASH)

Locus tag: SCO3230  
Protein ID: NP\_627443.1  
Location: 3543335 - 3565726

location

biosynthetic (rule-based-clusters) nrps: AMP-binding  
biosynthetic (rule-based-clusters) nrps: Condensation  
biosynthetic-additional (rule-based-clusters) PP-  
binding  
biosynthetic-additional (smcogs) SMCOG1002:AMP-  
dependent synthetase and ligase (Score: 346.1; E-  
value: 4.4e-105)  
NCBI BlastP on this gene  
View genomic context

Details of HMM hits

smCOG classification

Link to NCBI BLAST

Link to NCBI genome viewer  
(only works when genome  
was downloaded from NCBI)

MiBIG Hits

BLAST hits to MiBIG sequences

AA sequence: Copy to clipboard

copy DNA or amino acid sequence  
to clipboard for copy&pasting  
to other programs

Nucleotide sequence: Copy to clipboard

##### Details on domain architecture: click on domain symbol

SCO3227



Domain type

SCO3230



SCO3231



SCO3232



SCO3248



KS

AMP-binding  
Location: 1674-2097 AA  
Run BlastP on this domain  
Substrate predictions:  
-consensus: thr  
AA sequence: Copy to clipboard  
Nucleotide sequence: Copy to clipboard

Location of domain  
(with respect to full  
length protein)

Link to NCBI BlastP

A-domain  
specificity predictions

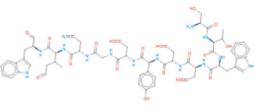
copy amino acid /DNA sequence of selected domain  
to clipboard for copy&pasting  
to other programs

## Core structure prediction and prediction details

At the right side of the window, a core structure deduced from the biosynthetic enzymes is displayed for PKS and NRPS gene clusters. CAUTION: This is just a rough prediction based on specificity prediction of the enzymatic domains and does not take into account any tailoring modifications or non-standard reactions

**Predicted core structure(s)**

For supercluster 10, location 3524827 - 3603907: ⓘ



Rough prediction of core scaffold based on assumed PKS/NRPS colinearity; tailoring reactions not taken into account.

**core structure prediction**

Polymer prediction:  
(ser - thr - trp - asp - hpg) + (asp - gly - asn) + (3-me-glu - trp)

Direct lookup in NORINE database: strict or relaxed

**prediction of the monomer sequence**

Link to NORINE database query form

**NRPS/PKS domain predictions**

SCO3230: ser - thr - trp - asp - asp - hpg ⓘ  
Search NORINE for peptide: strict or relaxed

AMP-binding (475..883): ser ⓘ  
NRPSPredictor2: ser ⓘ

SVM prediction details:

Predicted physicochemical class: hydrophobic-aliphatic  
Large clusters prediction: ser, thr, dhpg, hpg  
Small clusters prediction: ser  
Single AA prediction: ser  
Nearest Stachelhaus code: ser  
Stachelhaus code match: 100%

**Link to NORINE NRP database (for NRPs)**

AMP-binding (1674..2097): thr ⓘ  
AMP-binding (2759..3160): trp ⓘ  
AMP-binding (4283..4657): asp ⓘ  
AMP-binding (5323..5697): asp ⓘ  
AMP-binding (6360..6777): hpg ⓘ

**prediction details for the PKS/NRPS domains**

SCO3231: asp - gly - asn ⓘ  
Search NORINE for peptide: strict or relaxed

AMP-binding (472..845): asp ⓘ  
AMP-binding (1504..1884): gly ⓘ  
AMP-binding (2567..2973): asn ⓘ

SCO3232: 3-me-glu - trp ⓘ  
Search NORINE for peptide: strict or relaxed

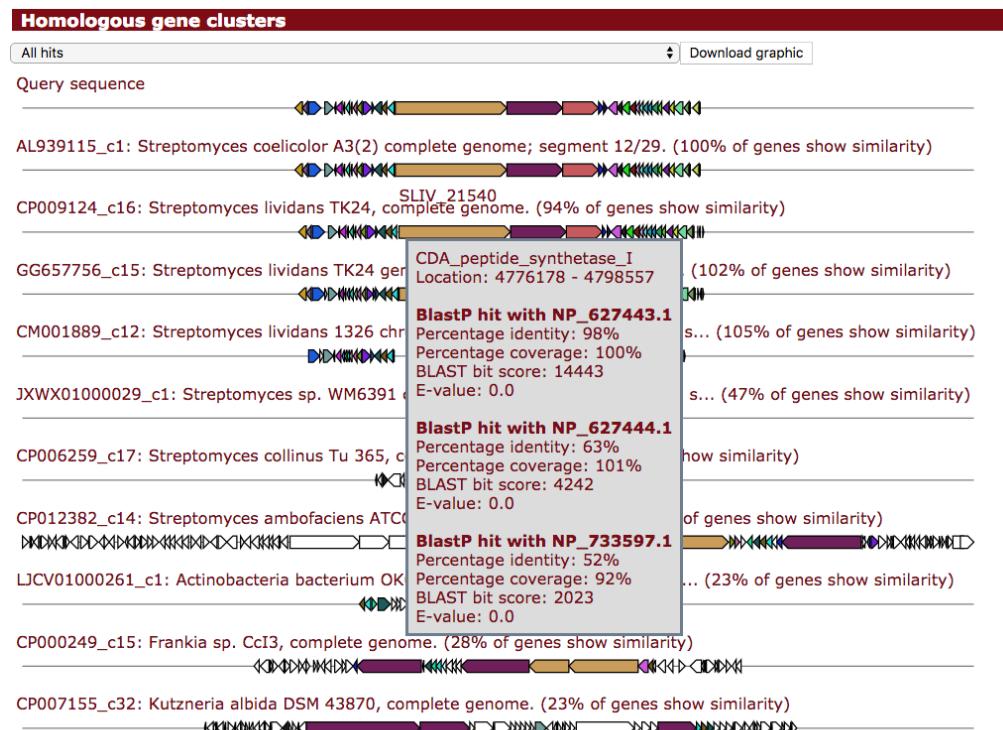
AMP-binding (467..874): 3-me-glu ⓘ  
AMP-binding (1541..1936): trp ⓘ

**Click “+” to unfold**

At the bottom of the panel, there is a link which directly links to the NORINE peptide database query form

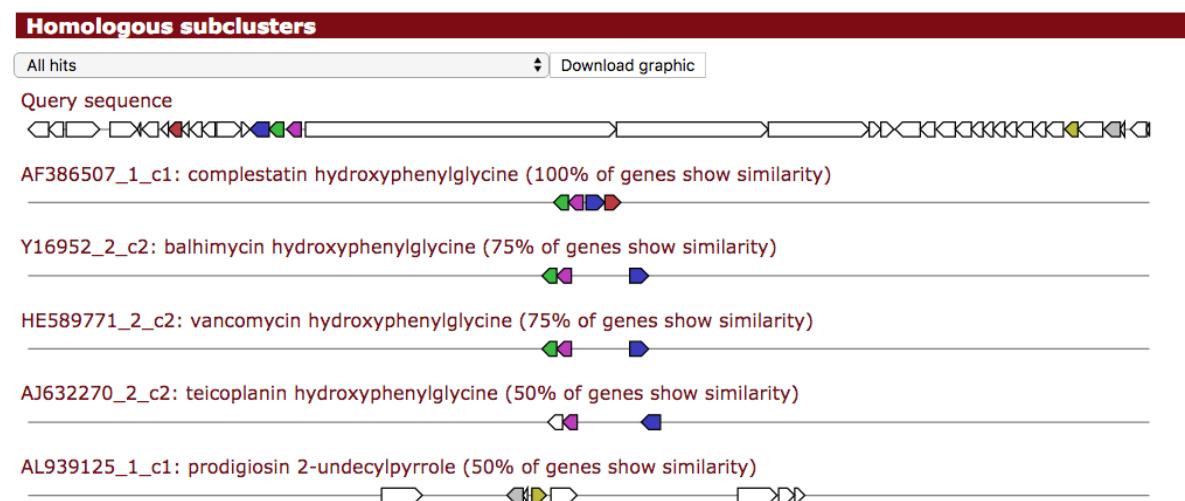
## Identification of similar gene clusters

Below the domain details panel, hits to secondary metabolite gene clusters are interactively displayed. The colors indicate BLAST matches of individual gene products.



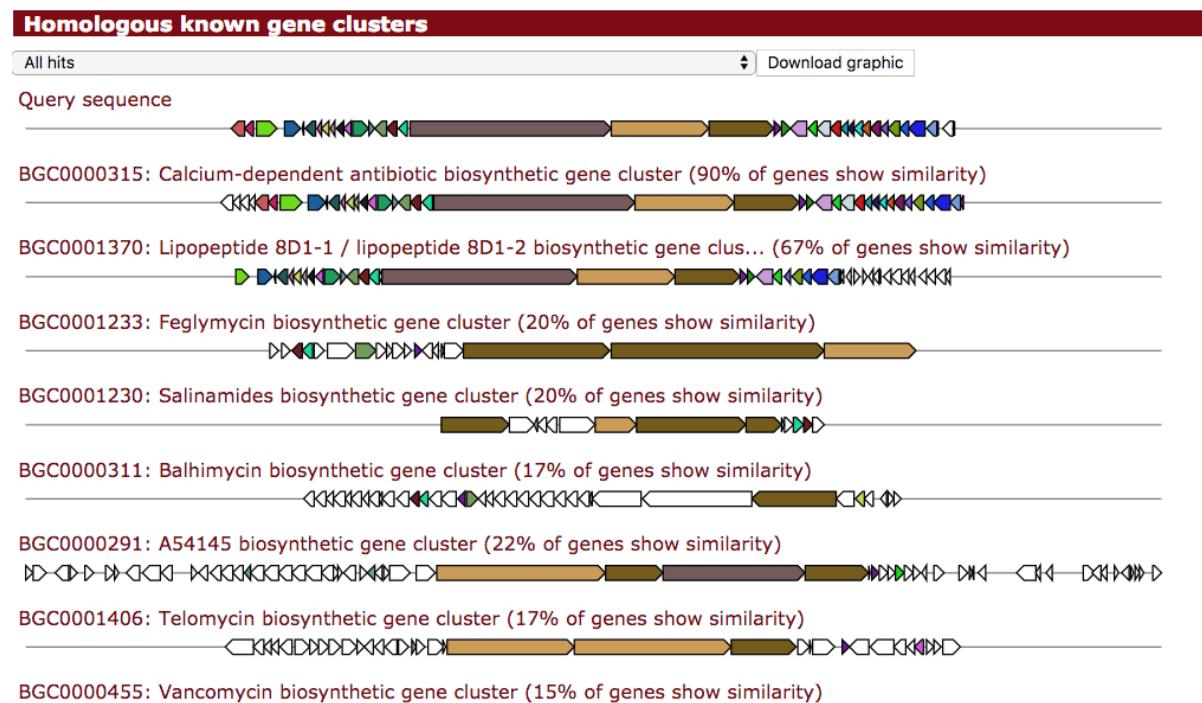
## Identification of conserved subclusters (e.g. biosynthetic operons)

The same strategy is applied to identify similar operons, which for example encode the biosynthesis of specific building blocks like non-proteinogenic amino acids.



### Identification of known clusters (from MIBiG dataset)

At the end of the results page, the results of the KnownClusterBlast, i.e. the comparison against the known (and studied) biosynthetic gene clusters is displayed:



#### 4.4 Downloading results

The results of the antiSMASH analysis are stored on our server for 1 month and are deleted afterwards. You can download all or subsets of the results using the download button at the top of the antiSMASH results page.



**Download all results:** Download a ZIP file containing all results.

Take the following steps to open the downloaded results:

1. Extract complete ZIP file to a folder on your disc
2. Open the “index.html” file with your web browser  
(we recommend Firefox for local use)

**Download XLS overview file:** Download MS Excel file with cluster identification summary

**Download EMBL summary file / GenBank summary file:** Download EMBL/Genbank file containing all the annotation for use in 3<sup>rd</sup> party sequence analysis Software.

We recommend using Artemis (17), which can be freely downloaded at

<http://www.sanger.ac.uk/resources/software/artemis/>

**Download log file:** Downloads the log file of the antiSMASH run (for Debug purposes)

#### **4.5 Obtaining antiSMASH**

antiSMASH is available via Bioconda / the conda package managing system for LINUX and MacOS.

Installers for standalone antiSMASH for Debian LINUX can be downloaded from the antiSMASH homepage using the download button at the antiSMASH start page. In addition, a Docker container is provided with a full antiSMASH installation.

## 6 The antiSMASH database

The antiSMASH software is dedicated to run a genome mining analysis on a single genome and by itself does not provide any cross-genome comparisons beyond the (x)ClusterBlast information. Each antiSMASH run requires several hours computing time – if different scientists are interested in antiSMASH results of the same published sequence, e.g. *S. coelicolor*, they have each to run the genome through antiSMASH using a lot of computing resource.

In order to address these use-cases, the antiSMASH database was developed, that currently contains pre-computed antiSMASH v4 results from over 6000 completed bacterial genomes and ~18.000 draft genomes totaling in more than 100,000 BGCs.

The URL of the antiSMASH database is: <https://antismash-db.secondarymetabolites.org/>

### 6.1 Browsing the antiSMASH database

1. Select „Browse“ on the top menu bar
2. Either select “Secondary metabolite type”, i.e. browse for the different BGC types that antiSMASH detects or “Taxonomy” to browse by taxonomy of the producer

#### Browse

Secondary Metabolite Type		Taxonomy				
Cluster	Type	From	To	Most similar MIBIG cluster	Similarity	MIBIG BGC-ID
1	Nonribosomal peptide	117649	166244	Incedine	9 %	BGC0000078_c1
2	Nonribosomal peptide	185138	273613	Lipopeptide 8D1-1 / lipopeptide 8D1-2 biosynthetic gene clus...	11 %	BGC0001370_c1
3	Hybrid cluster: Nonribosomal peptide-Type I polyketide-Butyrolactone	284627	428545	Salinomycin	100 %	BGC0000144_c1
4	Aminoglycoside/aminocyclitol	746864	768097	Pyralomicin	22 %	BGC0001038_c1
5	Hybrid cluster: hglE-type polyketide-Type I polyketide	843886	1009285	PM100117 / PM100118	60 %	BGC0001359_c2
6	Hybrid cluster: Terpene-Bacteriocin or other unspecified RPP	1051607	1089082			
7	Type I polyketide	1174260	1276844	Oligomycin	55 %	BGC0000117_c1
8	Thiopeptide	1281646	1316387			
9	Hybrid cluster: Terpene-Type I polyketide	1433883	1519438	Piericidin A1	50 %	BGC0000124_c1
10	Hybrid cluster: Nonribosomal peptide-Terpene-Lanthipeptide	1512525	1617897	Isorenieratene	85 %	BGC0000664_c1
11	Lanthipeptide	1622375	1647828			
12	Siderophore	1912796	1924731			
13	Other	2456611	2498069	Arginomycin	20 %	BGC0000863_c1

3. Click on the BGC of interest to be directly transferred to the respective antiSMASH results page

### 6.2 Querying the antiSMASH database

The antiSMASH database query functionality is accessed by clicking “Query” on the top menu bar.

### 6.2.1 Simple search

For simple queries, such as “lantipeptide streptomyces” or searching for a specific strain “*Streptomyces collinus*” you can use the “Simple search” Google-style search functionality.

### 6.2.2 Complex searches

The antiSMASH database provides an sophisticated query builder that allows querying on all antiSMASH annotation. To enable this function, click on “Build a query”.

1. Select the type of results you’re interested in
  - Cluster: The query results links to the BGCs that match your query
  - Gene: The query results in tables/files with the genes / gene products (DNA / amino acids) your query matches
  - NRPS/PKS Domain: The query results in tables/files with PKS or NRPS domain sequences your query matches
2. Select the type of “return data”, i.e. results you’re interested in. For searches on Cluster level, the database provides graphical output with links to the respective antiSMASH cluster. For the other search types the results are returned as tables (CSV) or FASTA sequence files. As example, we here search for all “RiPPs in Burkholderia”
3. In the “Select a category” select the antiSMASH annotation that you want to query (for example: Genus = Burkholderia; the search box has a type-ahead feature on the database content for this category)
4. If you want to combine several queries, press “Add term” button
5. Enter next query element (for example: BGC type = ripp); search terms can be grouped, swapped, added or removed by pressing the respective buttons

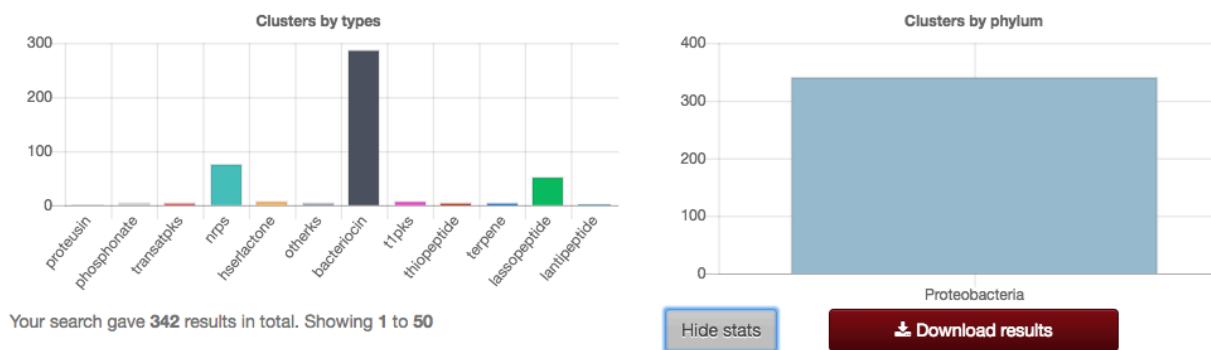
#### Query

The screenshot shows the 'Build a query' interface. At the top, there are tabs for 'Simple search' (selected) and 'Build a query'. Below the tabs, there are two search boxes: 'Search:' with dropdown options 'Cluster', 'Gene', and 'NRPS/PKS Domain', and 'Return data in format:' with options 'Graphical', 'CSV', and 'DNA FASTA'. The main search area contains two rows of search terms. The first row has a 'Genus' dropdown set to 'Burkholderia' and an 'Add term' button. The second row has a 'BGC type' dropdown set to 'ripp' and an 'Add term' button. Between these rows are buttons for 'AND', 'OR', and 'EXCEPT'. There are also 'Remove term' and 'Swap terms' buttons. At the bottom left is a 'Q. Search' button, and at the bottom right is a 'Load example search' link.

## 6. Press “Search” button

Species	Accession	Cluster	Type	From	To	Most similar MIBiG cluster	Similarity	MiBIG BGC-ID
Burkholderia cepacia GG4	NC_018514	6	Bacteriocin or other unspecified RIPP	2554336	2565227			
Burkholderia cepacia JBK9	NZ_CP013731	9	Bacteriocin or other unspecified RIPP	1487816	1500870			
Burkholderia cepacia JBK9	NZ_CP013732	17	Bacteriocin or other unspecified RIPP	1019365	1030181			
Burkholderia cepacia DDS 7H-2	NZ_CP007786	8	Bacteriocin or other unspecified RIPP	1743928	1756986			
Burkholderia cepacia DDS 7H-2	NZ_CP007785	18	Bacteriocin or other unspecified RIPP	856370	867186			
Burkholderia cepacia DDS 7H-2	NZ_CP011301	1	Bacteriocin or other unspecified RIPP	741952	752768			

## 7. By clicking “Show stats”, statistics about your query are displayed



Further examples for complex queries can be found in the exercises and at

<https://antismash-db.secondarymetabolites.org/#!/help>

## 7 References

1. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes-a review. *Nat. Prod. Rep.*, **33**.
2. Blin,K., Medema,M.H., Kazempour,D., Fischbach,M.A., Breitling,R., Takano,E. and Weber,T. (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.*, **41**.
3. Medema,M.H., Blin,K., Cimermancic,P., De Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**.
4. Weber,T., Blin,K., Duddela,S., Krug,D., Kim,H.U., Brucolieri,R., Lee,S.Y., Fischbach,M.A., Müller,R., Wohlleben,W., *et al.* (2015) AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.*, **43**.
5. Blin,K., Wolf,T., Chevrette,M.G., Lu,X., Schwalen,C.J., Kautsar,S.A., Suarez Duran,H.G., De Los Santos,E.L.C., Kim,H.U., Nave,M., *et al.* (2017) AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**.
6. Blin,K., Kazempour,D., Wohlleben,W. and Weber,T. (2014) Improved Lanthipeptide Detection and Prediction for antiSMASH. *9*, 1–7.
7. Blin,K., Kim,H.U., Medema,M.H. and Weber,T. (2017) Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief. Bioinform.*, 10.1093/bib/bbx146.
8. Kautsar,S.A., Suarez Duran,H.G., Blin,K., Osbourn,A. and Medema,M.H. (2017) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.*, **45**, W55–W63.
9. Grüning,B., Dale,R., Sjödin,A., Chapman,B.A., Rowe,J., Tomkins-Tinch,C.H., Valieris,R. and Köster,J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475–476.
10. 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 Bioinformatics*, **11**, 119.
11. Majoros,W.H., Pertea,M. and Salzberg,S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.
12. Cimermancic,P., Medema,M.H., Claesen,J., Kurita,K., Wieland Brown,L.C., Mavrommatis,K., Pati,A., Godfrey,P. a., Koehrsen,M., Clardy,J., *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, **158**, 412–421.
13. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., De Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C., *et al.* (2015) Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.*, **11**.

14. Medema,M.H., Takano,E. and Breitling,R. (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol. Biol. Evol.*, **30**, 1218–1223.
15. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A., *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
16. Wolf,T., Shelest,V., Nath,N. and Shelest,E. (2016) CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics*, **32**, 1138–1143.
17. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.