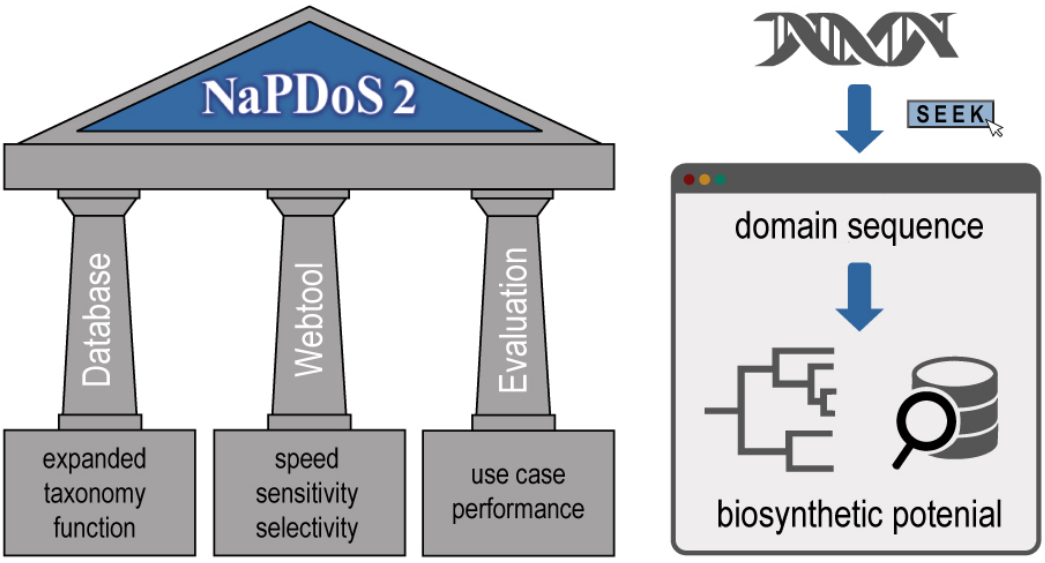
NaPDoS 2

The Natural Product Domain Seeker 2

NaPDoS2: Documentation

Documentation for:

Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, Moore BS, Ziemert N, Letzel AC, Jensen PR. **The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function**. 2022. *The Journal of Biological Chemistry.* https://doi.org/10.1016/j.jbc.2022.102480



Version 2.2

Updated: October 5, 2022

Documentation authored by Leesa J Klau, Kaitlin E Creamer, Sheila Podell, & Paul R Jensen

Table of Contents

Quick Start with NaPDoS2 4

Overview 9

Natural Products 9

Polyketides and Nonribosomal Peptides 10

Phylogeny of KS and C domains 10

Biosynthetic Gene Clusters in NaPDoS2 14

BGC Cards 16

Classification 19

KS Domains 19

Type I PKS 20

Class: Modular cis-AT (cisAT) 20

Subclass: Olefin synthases (cisOLS) 20

Subclass: Loading module (cisloading) 20

Subclass: Hybrid (cisHybridKS) 20

Subclass: Tandem ECH (cistandemECH) 20

Class: Iterative cis-AT (iPKS) 20

Subclass: Polyunsaturated fatty acids (iPKSPUFA) 21

Subclass: Enediynes (iPKSenediyne) 21

Subclass: Aromatic (iPKSaromatic) 21

Subclass: Polycyclic tetramate macrolactam-like (iPKSPTM) 21

Subclass: Non-reducing (iPKSNR) 21

Subclass: Partially reducing (iPKSPR) 21

Subclass: Highly reducing (iPKSHR) 21

Class: trans-AT 22

Subclass: beta-branching module (transBdomain) 22

Subclass: Hybrid (transHybridKS) 22

Subclass: Hybrid non-elongating KS (transHybridKS0) 22

Class: Metazoa (MetazoaPKS) 22

Class: Protist (ProtistPKS) 23

Type II PKS 23

Class: Aromatic (aromaticKSa) or (aromatic KSb) 23

Subclass: angucycline-derived I (angucyclineIKSa) or (angucyclineIKSb) 23

Subclass: angucycline-derived II (angucyclineIIKSa or angucyclineIIKSb) 24
Subclass: anthracycline-derived I (anthracyclineIKSa) or (anthracyclineIKSb) 24
Subclass: anthracycline-derived II (anthracyclineIIKSa or anthracyclineIIKSb) 24
Subclass: isochromanequinone-derived (isochromanequinoneKSa or isochromanequinoneKSb) 24
Subclass: pentangular polyphenol-derived (pentangularpolyphenolKSa or pentangularpolyphenolKSb) 24
Subclass: tetracenomycin-derived (tetracenomycinKSa or tetracenomycinKSb) 24
Subclass: tetracycline-derived (tetracyclineKSa or tetracyclineKSb) 25
Subclass: spore pigment (sporepigmentKSa or sporepigmentKSb) 25
Class: β-branching cassettes (betabranch) 25
Class: Polyenes (polyeneKSa or polyeneKSb) 25
Class: Aryl polyenes (arylpolyeneKSa or arylpolyeneKSb) 25
Class: Non-iterative (noniterative) 25
Class: Type II unclassified (typeII) 26
Fatty Acid Synthases 26
Class: Type I FAS 26
Subclass: Bacteria and fungi (bfFASI) 26
Subclass: Metazoa (MetazoaFASI) 26
Subclass: Protist (ProtistFASI) 26
Class: Type II FAS (FASII) 26
C Domains 28
Class: Starter (starter) 28
Class: LCL (LCL) 28
Class: DCL (DCL) 28
Class: Cyclization (cyclization) 28
Class: Epimerization (epimerization) 28
Class: Dual (dual) 28
Class: Modified amino acid (modifiedAA) 28
Class: Hybrid (hybridC) 29
Class: Condensation (condensation) 29

Tutorials: Using NaPDoS2 34
Analyzing a single genome with NaPDoS2: analyze a Salinispora bacterial genome 35
Genome tutorial A objective 35
Genome tutorial B objective 50
Analyzing multiple genomes with NaPDoS2: analyze multiple fungal genomes 64

Analyzing an animal genome with NaPDoS2: analyze the Elysia chlorotica genome 71
Analyzing a BGC from antiSMASH, PRISM, or MIBiG with NaPDoS2 80
Analyzing a metagenome with NaPDoS2 93
Analyzing individual genes/proteins & PCR clone products with NaPDoS2 108
Analyzing KS amplicon “tag” sequences with NaPDoS2 116
File Size Management Tools 126

The NaPDoS2 Analysis Workflow Details 131
Preliminary Candidate Screening 133
BLASTP search results output 133
Annotation of appended information to query hit results 134
Tree construction 137
Interpreting Results 138
BLASTP E-value cutoff 140
Sensitivity and accuracy versus alignment length 141
Further analyses beyond NaPDoS2 143

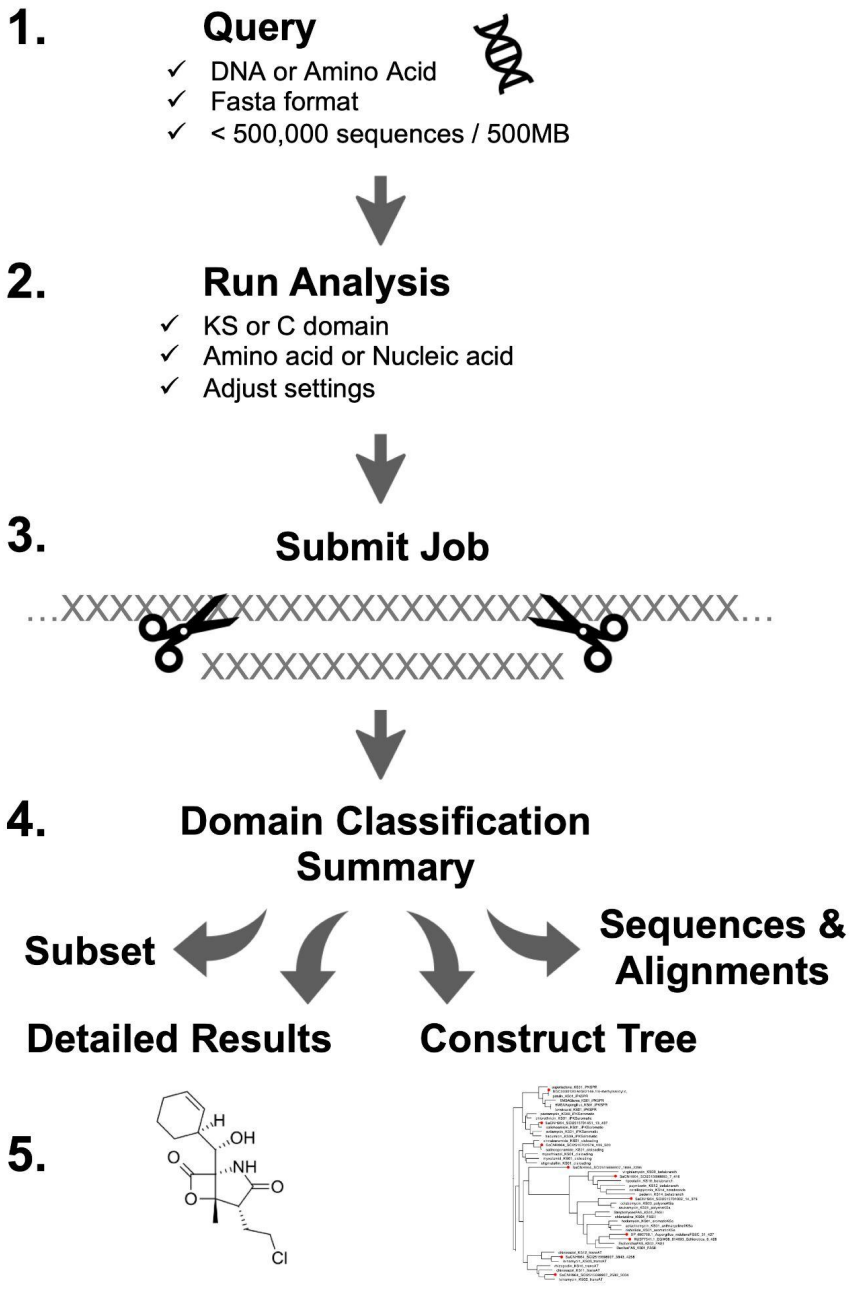
Frequently Asked Questions (FAQ) 146

Quick Start with NaPDoS2

Objective: A quick-start tutorial with base instructions for running a NaPDoS2 analysis.

These 5 quick start steps will help you get started, but please be sure to read the **interpretation** tips and detailed workflow **tutorial examples** for how to analyze genomes, metagenomes, and amplicon sequence datasets with NaPDoS2.

(These steps are also outlined on the "QuickStart" page on NaPDoS2)



1. **Gather Sequences**

    - Sequences can be either nucleic or amino acids, from genomic, metagenomic, or amplicon sources, but must be in FASTA format.

        - "A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is

distinguished from the sequence data by a greater-than (">") symbol at the beginning."

- FASTA format example:

```
>StCNB440_BA|640473291|salC
MSEAVERNYDPGDVAIIGASCRFPGARNKEQYWDNLLHGREGVTFYAKDEIEVD
ETLINSPAYVRATGALDGYDEFDPAVFGVSDRMAAAMTPEHRVFLEASWEVMED
ASYDPERVRGEVGVYASTNPQSAALYSSPPDWVSAGPEVMDRSNAWLPDTITSN
VLYYLGLTGEAVTVTAVCSGFHYAVHLACQSLLLGQTDMAIAGGVMVRLPQRRG
YLWEEGRILSRDGHCRPFDANGTGSVLASGVAVVLLKPLPQAVADRDHIYATVK
GTAINNNGISAMAYGLAQPERLSACIAGAMQAGDVAPETVSMYEANGFGMPITD
SLEVHAAHLAFGKQSGTCSIGAVKGNIGHAGVVAGGSGAVKAAFALYHRSLPPT
INLTELNEEIDFPRTPFVPQLEPAAWQPECGIRRAGITALGGGGYNAHLVLEEP
PRPVEREPENRRPRIVTLSALDDAAVSRQRAALSSWLAEHSDARLDDIAYSLNL
GRKALPSRWAAVISTRDELLEVLSGDGKSGRVSRFGQERRADLARFRRTEDGLA
LGSGDEMRDVQALTELAAAWVQGERVNFEVLHADERSHRISLPNYPFARRRFWR
TDW
```

- Please avoid using non-standard characters (e.g. commas, slashes, colons, parentheses, ampersands, etc) in file names, identifiers/description line, or sequences, as well as ambiguity codes "R" and "Y" for nucleic acid sequence files.

- Maximum file size limits are currently set at < 500,000 sequences or < 500 MB. For larger data sets (and faster analysis), please see our file size management page & the section below for how to to pre-filter and/or subdivide your data into smaller batches before submission.

2. **Select Parameters**

- Go to the Run Analysis page.

- Select KS or C domain (NaPDoS2 can only detect one domain type per analysis).

- Select query sequence type: either amino acid or nucleic acid. (Nucleic acids will be automatically translated into all 6 reading frames.)

- Choose a local sequence file to upload (sequences within must be in FASTA format), or paste FASTA-formatted sequences into the window.

❖ **Advanced Settings** can be used to adjust BLASTP stringency (default e-value = 1e-8) and minimum alignment (default length = 200aa/600nt).

- ❖ Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details.

3. **Run Analysis**

   - Click **SEEK**, wait for JOB ID assignment.
   - Review the search parameters and estimated completion time displayed.
   - Click **SUBMIT JOB** to continue the analysis.
   - See section "The NaPDoS2 Analysis Workflow" for the steps NaPDoS2 completes for analyses

4. **Select Display Options**

   - When the analysis is complete, a page will appear indicating how many domains were detected.
   - For nucleotide sequence queries: There is an option to **DOWNLOAD** a table listing the domain matches with their nucleotide coordinates and translation frames (see later section "Interpreting Results" for interpretation of this table). Click **CONTINUE ANALYSIS** to display the **Domain Classification Summary** page.
   - For amino acid sequence queries, the **Domain Classification Summary** will be displayed immediately after analysis is complete.
   - You can view and download the results summary table, select specific classes/subclasses and **VIEW A SUBSET** of the results, or click **VIEW ALL MATCHES** to see complete Database Search Results.
   - Detailed information about the classification scheme can be found on the **CLASSIFICATION** web page and in the detailed Classification section.

5. **Download Results**

   - From the **Database Search Results** page, sequences can be selected to **View nucleotide coordinates** for all trimmed domain candidates (nucleotide sequence queries only), **Output selected sequences** in fasta format, or **Output Alignment (**MUSCLE**) of your query sequences** to the top database hits.
   - All result tables can be downloaded using the **DOWNLOAD** button– click on **DOWNLOAD** to view the results in a new tab, or right-click on the **DOWNLOAD** link and select "Save Link As…".

- Click on **domain class** or **BGC match** links to learn more about related domain categories and biosynthetic gene clusters in the NaPDoS2 database.

- If results include < 100 KS or C domains, selected sequences can also be viewed in phylogenetic context relative to functionally characterized reference domains (**Construct Tree**), and downloaded as an SVG image (click to view, right-click to save) or in Newick tree format (click to view, right-click to save).

- The 100 domain tree-building limit has been set due to computational resource limitations. If more than 100 domains are detected initially, tree construction can be enabled by re-running the analysis using a smaller subset of the original sequences.

Want a more detailed tutorial? We have included multiple tutorials for using NaPDoS2:

1. **Analyzing a single genome with NaPDoS2**: analyze a *Salinispora* bacterial genome
2. **Analyzing multiple genomes with NaPDoS2**: analyze multiple fungal genomes
3. **Analyzing an animal genome with NaPDoS2**: analyze the *Elysia chlorotica* genome
4. **Analyzing a BGC from antiSMASH, PRISM, or MIBiG with NaPDoS2**
5. **Analyzing a metagenome with NaPDoS2**: analyze a marine sediment metagenome
6. Analyzing **individual genes/proteins & KS PCR clone products** with NaPDoS2
7. Analyzing **KS amplicon tag-sequencing datasets** with NaPDoS2
8. How to use the **File Size Management Tools**

These tutorials cover:

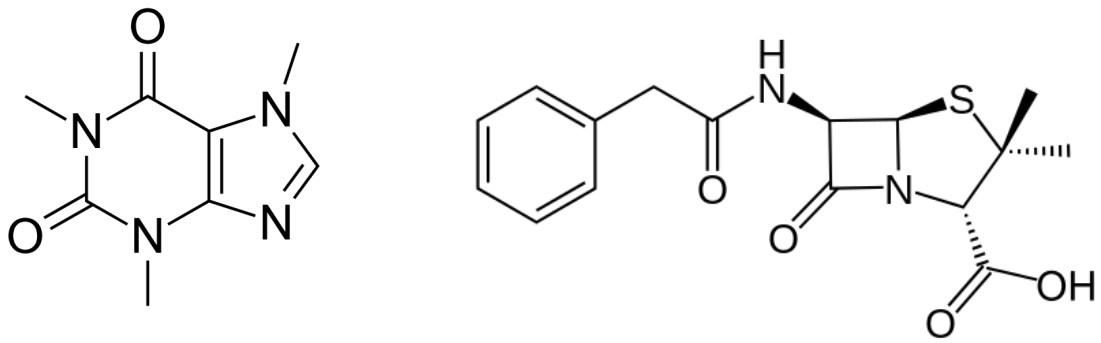
- How NaPDoS2 can be applied to objectives & goals of a research question and hypothesis building
- Interpretation of NaPDoS2 results
- Details about the NaPDoS2 webtool interface steps
- Suggestions for further analyses & our considerations for analyses

Overview

Natural Products

Natural products are organic compounds that are produced by living organisms, such as microorganisms, plants, and animals. While they are not considered essential for normal growth and survival, natural products serve ecological functions related to communication, defense, competition, and reproduction, among others.

Natural products have historically provided an important source of medicines including antibiotics and anticancer agents (Li and Vederas, 2009). Traditional natural product drug discovery efforts use bioassays to guide the isolation of active compounds, which were then checked for structural novelty. Decreasing returns using this approach led the pharmaceutical industry to move away from natural products in favor of other drug discovery platforms.



Caffeine (left) and Penicillin G (right) are natural products.

In the genomic era, access to DNA sequence data has revealed a wealth of unrealized biosynthetic potential and generated considerable excitement in the development of sequence-based approaches to natural product discovery (Zerikly and Challis, 2009).

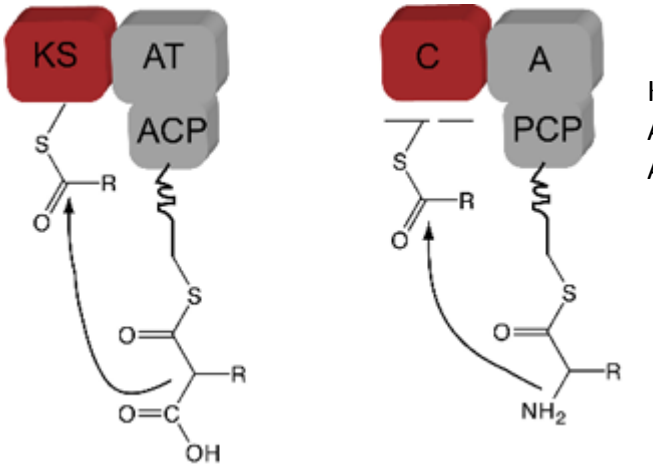
NaPDoS2 provides a simple, straightforward approach to assess polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) diversity based on the analysis of ketosynthase (KS) and condensation (C) domains derived from the considerably larger genes that encode these enzymes.

Polyketides and Nonribosomal Peptides

Two major classes of bacterial specialized/secondary metabolites are polyketides and nonribosomal peptides. These compounds are biosynthesized by large, multifunctional enzymes that sequentially construct products from small carboxylic acid and amino acid building blocks (Fischbach and Walsh, 2006).

Polyketides can be assembled by modular **polyketide synthases (PKSs)**. Contrary to peptides and their amino acid building blocks, polyketides are assembled from acyl units. Selection of the monomers is performed by acyltransferase (AT) domains. Domains responsible for the elongation step are termed **ketosynthase (KS)** domains (Hertweck, 2009).

Nonribosomal peptide synthetases (NRPSs) are large multi-enzyme complexes. They have a modular structure, with each module being responsible for the activation, thiolation, modification and condensation of one specific amino acid. Each module consists of a number of domains, two of which are adenylation (A) and **condensation (C)** domains. The A domain activates a specific amino acid (analogous to a t-RNA) and transfers it to the PCP (peptidyl carrier protein), which holds on to the growing peptide as a thioester. The C domain forms a peptide bond between the next amino acyl and the peptidyl unit. Modifying domains for epimerization, heterocyclization or oxidation can be additionally integrated (Schwarzer *et al.*, 2003).



Domains of PKSs and NRPSs

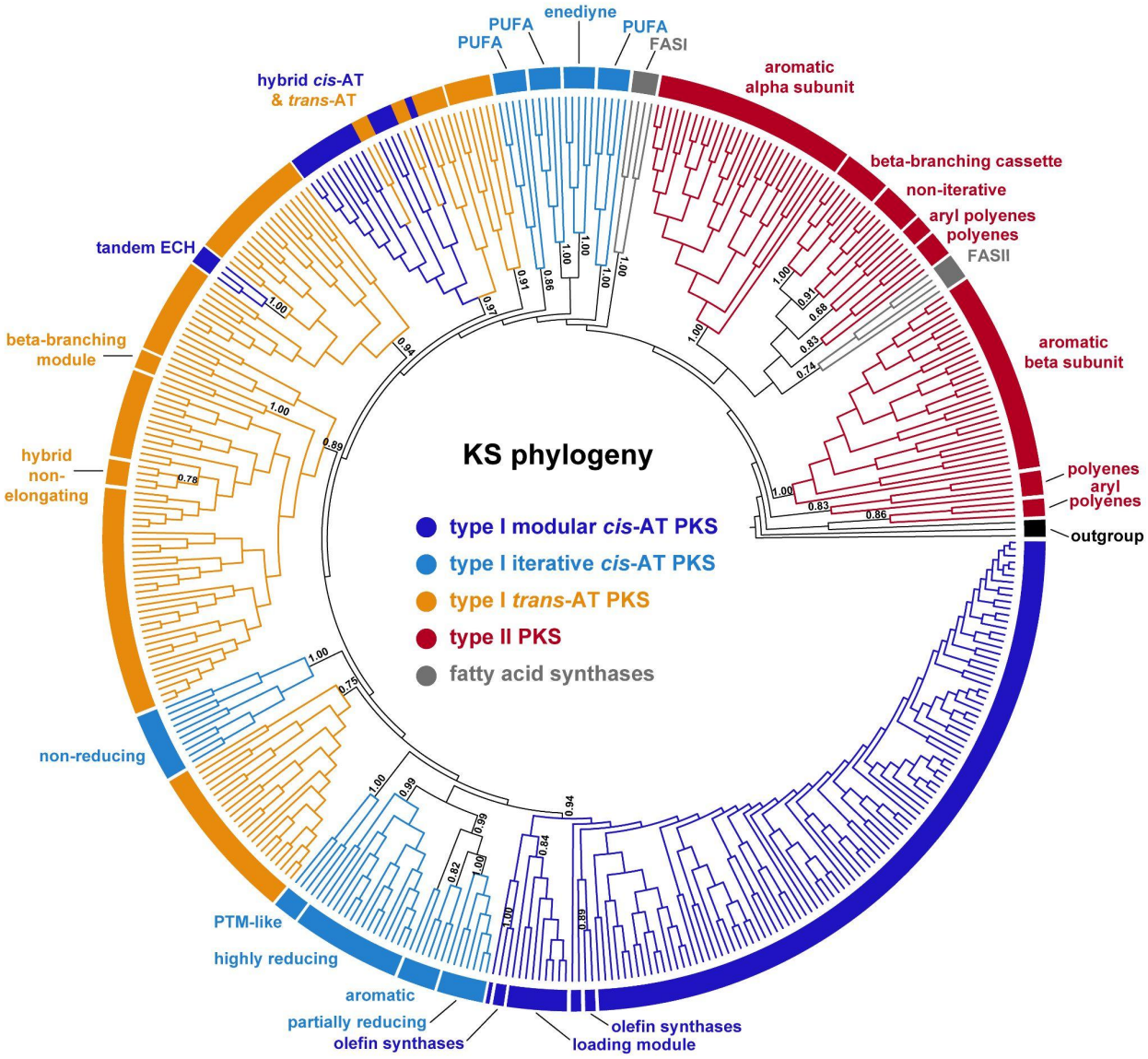
KS ketosynthase
AT acyltransferase
ACP acyl carrier protein

C condensation
A adenylation
PCP peptidyl carrier protein

Phylogeny of KS and C domains

Domain-specific phylogenetic analyses of the different modules have shown the elongation domains in PKS and NRPS assembly lines, called KS and C domains, respectively, to be the most informative in terms of predicting pathway associations. Protein sequences for the majority of these domains are grouped in pathway-specific clades according to their chemical activity (Rausch *et al.*, 2007; Jenke-Kodama and Dittmann, 2009).

The phylogeny of KS domains delineates the major classes of polyketide synthases, according to the architecture of their biosynthesis enzymes. Type I PKSs possess a multidomain architecture that consists of either sets of modules corresponding to the number of acyl units in the product, or a single set of catalytic domains that act iteratively. Type II PKS carry each catalytic site on a separate protein (Jenke-Kodama *et al.*, 2005).

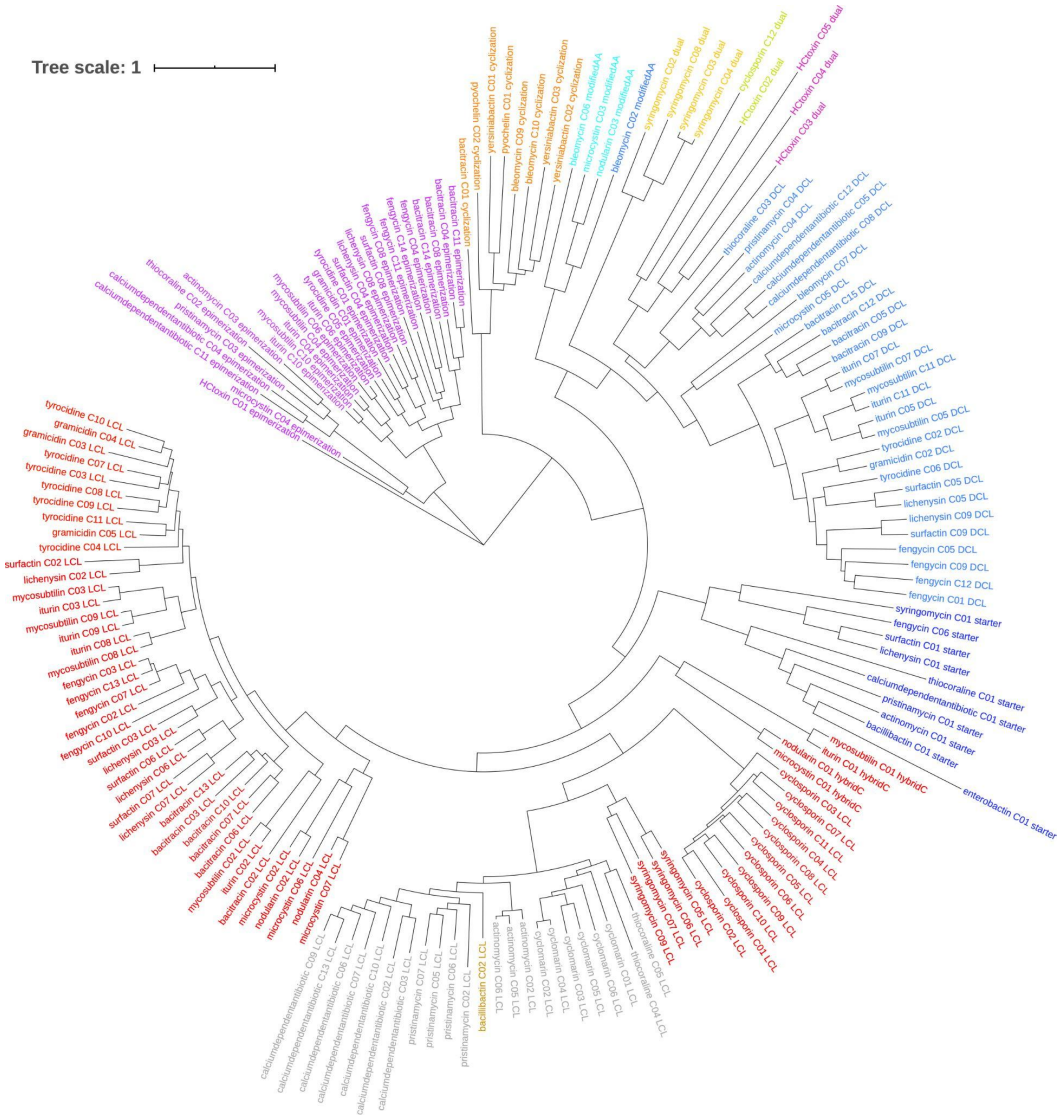


NaPDoS2 KS domain phylogeny-based classification.

Maximum likelihood phylogeny generated from 414 NaPDoS2 reference KS sequences. Clades are color-coded and labeled according to their NaPDoS2 classification. Transfer Bootstrap

Expectation (TBE) bootstrap support estimated using Booster. In the ***Supporting Information*** of the corresponding publication for the NaPDoS2 tool, the full name of each sequence can be viewed in ***Figure S7***, which can be used to link a query match to a specific location in the tree.

The phylogeny of NRPS C domains clearly reflects the six functional categories that are known. An LCL domain catalyzes a peptide bond between two L-amino acids, a DCL domain links an L-amino acid to a growing peptide ending with a D-amino acid, and heterocyclization domains catalyze both peptide bond formation and subsequent cyclization of cysteine, serine or threonine residues. Three other subtypes of C domains are starter domains, epimerization domains and dual epimerization/condensation domains (Rausch *et al.*, 2007).



NaPDoS2 C domain phylogeny-based classification.

Maximum likelihood phylogeny generated from 172 NaPDoS2 reference C sequences. Clades are color-coded and labeled according to their NaPDoS2 classification. Transfer Bootstrap Expectation (TBE) bootstrap support was estimated using Booster.

Domain Clades

Phylogenetic trees of trimmed amino acid sequences for PKS domains can be divided into three major clades:

1) Fatty acid synthases (FAS)
2) Type I PKS
3) Type II PKS

> In NaPDoS2, KS domains have been further subdivided into a much wider diversity, represented as classes and subclasses.

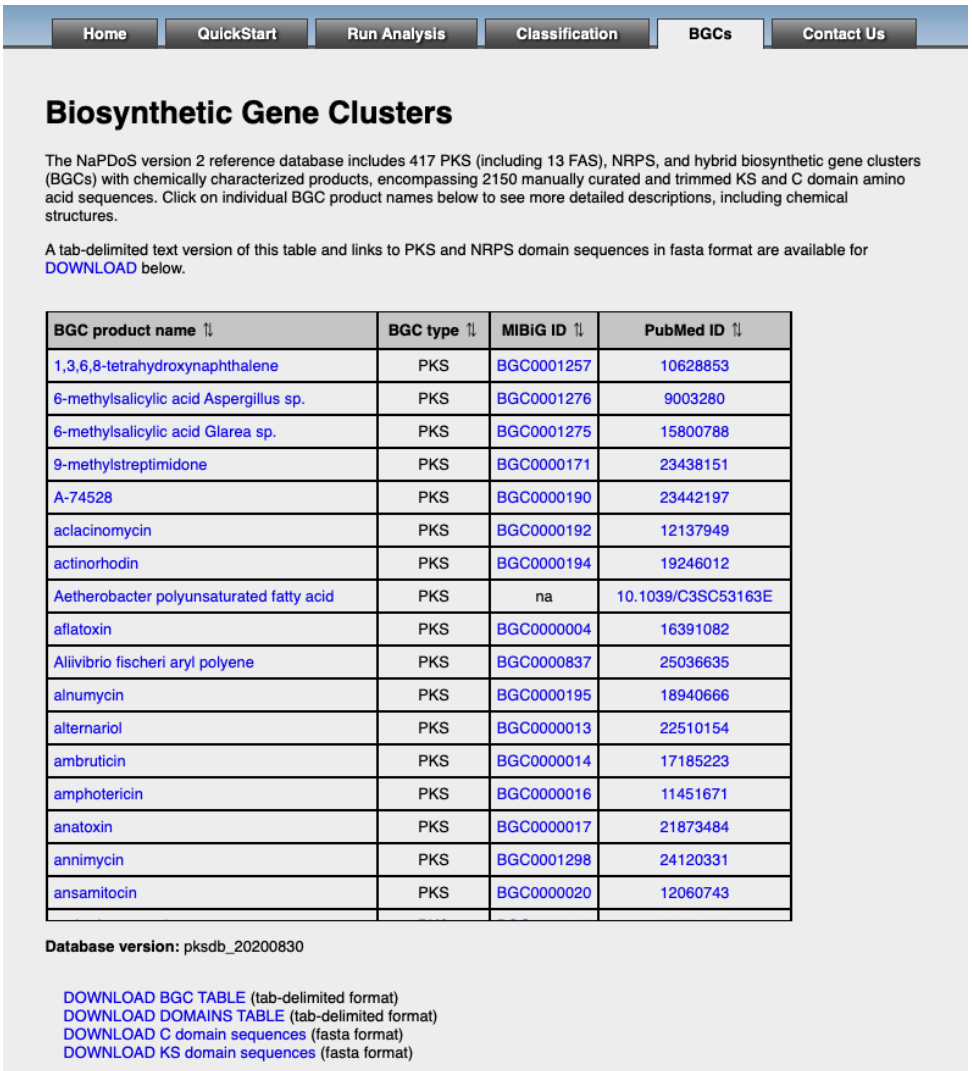
> The eight major C domain classes used in NaPDoS version 1 remain unchanged.

Please see the **CLASSIFICATION** tab on NaPDoS2 website for brief descriptions, and see **Classification section** below for detailed descriptions of **KS** and **C** domain classes and subclasses, including example chemical structures of example products of each type.

Biosynthetic Gene Clusters in NaPDoS2

The NaPDoS2 reference database includes a total of 2,150 KS and C domains from a total of 404 biosynthetic gene clusters (BGCs) with chemically characterized products. The BGCs comprise PKS, NRPS, and hybrid PKS-NRPS pathway from bacterial, fungal, and other eukaryotic sources.

Navigate to the **BGCs** tab on NaPDoS2 website (see screenshot below) for a list of all BGCs included in the database. The columns are sortable by: BGC product name, BGC type, MIBiG ID, and PubMed ID). Click on individual BGC product names to go to the 'BGC card' (see next section) to see the included domains, chemical structure, and biological source for each BGC. The "MIBiG ID" hyperlinks to the MIBiG repository of known biosynthetic gene clusters BGC page. The "PubMed ID" hyperlinks to the reference from which the sequence was sourced.



A tab-delimited text version of the BGC table and links to PKS and NRPS domain sequences in FASTA format are available for **DOWNLOAD** below the table, and linked here below.

Database version: pksdb_20200830

(to download: click the link to view; right-click on the link & select "Save Link As…" to save)

DOWNLOAD BGC TABLE (tab-delimited format)

- Lists all 404 BGCs in the NaPDoS2 database with their name, BGC type, MIBiG ID, and Pubmed reference ID.
- Columns: BGC name; BGC type; MIBiG ID; Pubmed reference ID.

DOWNLOAD DOMAINS TABLE (tab-delimited format)

- Lists all KS and C domains in the NaPDoS2 database and the BGC they are found in, the BGC type, name of the domain in NaPDoS2 database, the domain class and subclass, MIBiG BGC ID, Pubmed reference ID, species name, strain name, and gene product name.
- Columns: BGC_name; BGC_type; domain name; domain_type; domain_class; domain_subclass; gbk_prot_id (Genbank protein accession ID); mibig_id; pubmed_ref_id; species_name; strain_name; gene_prod_name

DOWNLOAD C domain sequences (FASTA format)

- Lists all of the 273 C domain sequences in the NaPDoS2 database.
- Each domain in the protein FASTA format is named with a three part notation separated with underscores.
- The three parts consist of: the BGC product name, the C domain number in the database, and a classification code relating to class
    - i.e., cyclomarin_C06_LCL

DOWNLOAD KS domain sequences (FASTA format)

- Lists all of the 1,877 KS-domain sequences in the NaPDoS2 database.
- Each domain in the protein FASTA format is named with a three part notation separated with underscores.
- The three parts consist of: the BGC product name, the KS domain number in the database, and a classification code relating to class/subclass
    - i.e., cyanosporaside_KS01_iPKSenediyne

BGC Cards

On the BGCs tab, you can click on each BGC product name to see the BGC card for that product.

On the BGC card page, we have listed the:

1. BGC type (PKS, PKS-NRPS, NRPS, FAS)
2. Example product name
3. Example product chemical structure
4. Source genus & species of the sequence for this BGC/product in the NaPDoS2 database
5. PubMed reference ID or DOI (directly links to the reference)
6. MIBiG ID for the BGC (directly links to MIBiG)
7. Domain classifications
    - This includes all of the KS and/or C domains found in the BGC listed in the current NaPDoS2 database. You can click on the class/subclass names for a more detailed description of that classification.
    - We have listed all KS and/or C domains found in each BGC because BGCs can have multiple KS and/or C domains classified in different classes/subclasses, so being able to distinguish them could be important for predicted domain function.
    - *Please Note*:
        - Each KS and C domain has a unique name in the NaPDoS2 database, so you can find the specific KS and/or C domain that your query is closest to, and then understand where in the reference BGC the KS and/or C domain exists and its putative function.
        - The NaPDoS2 database does not always include every KS and/or C domain found in an entire BGC. For example, the NaPDoS2 database does not yet include C-domains for some NRPS-PKS hybrid clusters; curation efforts are ongoing.
        - Numerical designations for some NaPDoS KS reference domains may differ from their functional biosynthetic order. Users are encouraged to consult the primary literature citations provided on the BGC descriptions page (URL link) to resolve any observed inconsistencies.

Examples of BGC cards:

anatoxin

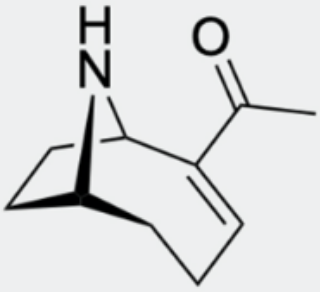
BGC type: pks

Example product: anatoxin A

Source species: *Anabaena sp. 37*

PubMed ref id or DOI: 21873484

MIBiG id: BGC0000017



Domain Classifications

KS and/or C domains for this BGC identified in the current NaPDoS database are listed below. Click on the class name link for a more detailed description.

Database ID	Domain type	Class	Subclass
anatoxin_KS01_cisAT	KS	type I modular cis-AT	-
anatoxin_KS02_cisAT	KS	type I modular cis-AT	-
anatoxin_KS03_cisAT	KS	type I modular cis-AT	-

1,3,6,8-tetrahydroxynaphthalene

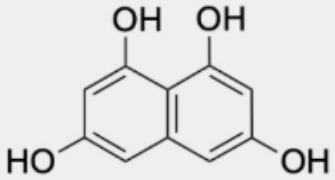
BGC type: pks

Example product: 1,3,6,8-tetrahydroxynaphthalene

Source species: *Nodulisporium sp.*

PubMed ref id or DOI: 10628853

MIBiG id: BGC0001257



Domain Classifications

KS and/or C domains for this BGC identified in the current NaPDoS database are listed below. Click on the class name link for a more detailed description.

Database ID	Domain type	Class	Subclass
tetrahydroxynaphthalene_KS01_iPKSNR	KS	type I iterative cis-AT	non-reducing

asukamycin

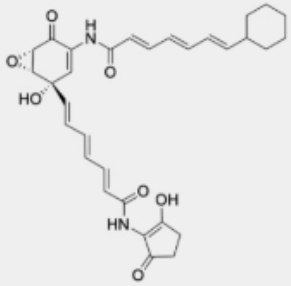
BGC type: pks

Example product: asukamycin A1

Source species: *Streptomyces nodosus subsp. Asukaensis*

PubMed ref id or DOI: 20522559

MIBiG id: BGC0000187

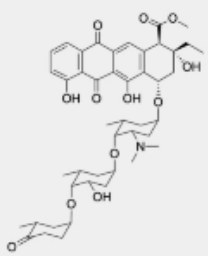


Domain Classifications

KS and/or C domains for this BGC identified in the current NaPDoS database are listed below. Click on the class name link for a more detailed description.

Database ID	Domain type	Class	Subclass
asukamycin_KS01_polyeneKSa	KS	type II polyene	KSa
asukamycin_KS02_typeII	KS	type II unclassified	-

aclacinomycin

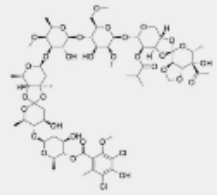


Domain Classifications

KS and/or C domains for this BGC identified in the current NaPDoS database are listed below. Click on the class name link for a more detailed description.

Database ID	Domain type	Class	Subclass
aclacinomycin_KS01_anthracycline-like KSa	KS	type II aromatic	anthracycline-like KSa
aclacinomycin_KS02_anthracycline-like KSa	KS	type II aromatic	anthracycline-like KSa

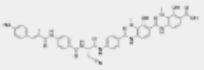
avilamycin



Domain Classifications

KS and/or C domains for this BGC identified in the current NaPDoS database are listed below. Click on the class name link for a more detailed description.

albicidin



Domain Classifications

KS and/or C domains for this BGC identified in the current NaPDoS database are listed below. Click on the class name link for a more detailed description.

Note: The NaPDoS database does not yet include C-domains for this nrps-pks hybrid cluster pending manual curation.

arthrofactin

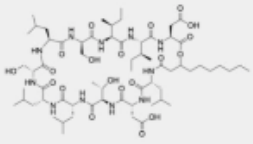
BGC type: nrps

Example product: arthrofactin

Source species: *Pseudomonas sp.*

PubMed ref id or DOI: 14522057

MIBiG id: BGC0000305



Domain Classifications

KS and/or C domains for this BGC identified in the current NaPDoS database are listed below. Click on the class name link for a more detailed description.

Database ID	Domain type	Class	Subclass
arthrofactin_C01_starter	C	starter	-
arthrofactin_C02_DCL	C	DCL	-
arthrofactin_C03_DCL	C	DCL	-
arthrofactin_C04_DCL	C	DCL	-
arthrofactin_C05_DCL	C	DCL	-
arthrofactin_C06_DCL	C	DCL	-
arthrofactin_C07_DCL	C	DCL	-
arthrofactin_C08_LCL	C	LCL	-
arthrofactin_C09_DCL	C	DCL	-
arthrofactin_C10_LCL	C	LCL	-
arthrofactin_C11_LCL	C	LCL	-

References

Fischbach, M.A. and Walsh, C.T., 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chemical Reviews*, 106(8), pp.3468-3496. https://doi.org/10.1021/cr0503097

Hertweck, C., 2009. The biosynthetic logic of polyketide diversity. *Angewandte Chemie International Edition*, 48(26), pp.4688-4716. https://doi.org/10.1002/anie.200806121

Jenke-Kodama, H., Sandmann, A., Müller, R. and Dittmann, E., 2005. Evolutionary implications of bacterial polyketide synthases. *Molecular Biology and Evolution*, 22(10), pp.2027-2039. https://doi.org/10.1093/molbev/msi193

Jenke-Kodama, H. and Dittmann, E., 2009. Evolution of metabolic diversity: insights from microbial polyketide synthases. *Phytochemistry*, 70(15-16), pp.1858-1866. https://doi.org/10.1016/j.phytochem.2009.05.021

Li, J.W.H. and Vederas, J.C., 2009. Drug discovery and natural products: end of an era or an endless frontier?. *Science*, 325(5937), pp.161-165. https://doi.org/10.1126/science.1168243

Rausch, C., Hoof, I., Weber, T., Wohlleben, W. and Huson, D.H., 2007. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evolutionary Biology*, 7(1), pp.1-15. https://doi.org/10.1186/1471-2148-7-78

Schwarzer, D., Finking, R. and Marahiel, M.A., 2003. Nonribosomal peptides: from genes to products. *Natural Product Reports*, 20(3), pp.275-287. https://doi.org/10.1039/B111145K

Zerikly, M. and Challis, G.L., 2009. Strategies for the discovery of new natural products by genome mining. *ChemBioChem*, 10(4), pp.625-633. https://doi.org/10.1002/cbic.200800389

Classification

Phylogenetic analyses of ketosynthase (KS) and condensation (C) domains derived from polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes, respectively, have proven highly informative of gene architecture and function (Rausch *et al.*, 2007, Jenke-Kodama and Dittmann, 2009). Following this logic, and as per the original NaPDoS version 1 release, the sequences in the NaPDoS2 reference database were used to generate KS and C domain maximum likelihood phylogenies from which cladding patterns and established gene functions and architectures were used to develop classification schemes. This resulted in an expansion of both the type I and type II PKS classification schemes. The type II aromatic class was further delineated into subclasses based on a separate phylogeny generated from concatenated sequences from the KS alpha (KSα) and beta (KSβ) subunits, which provided better classification than either subunit alone.

The C domain classification scheme has not changed from the original NaPDoS release, but the database and reference tree have been modified to reflect the updated naming system. Additionally, 84 C domains were added to the NaPDoS2 database.

After assigning a phylogeny-based classification to each sequence in the NaPDoS2 database, a subset of the sequences (414 KS and 172 C domains), representing all classes and subclasses identified, was used to construct reference trees within which query sequences can be placed within the webtool.

KS Domains

The primary bifurcation in the KS phylogeny delineates type I PKSs from type II PKSs and FASs. These lineages can be further resolved into classes and subclasses as described below.

- The abbreviations used in the database ID name, reference tree, and NaPDoS2 output are given in parentheses.

Type I PKS

Large multifunctional proteins with a domain and modular organization (Shen, 2003). Three broad classes can be recognized in the KS phylogeny based on function (i.e modular assembly-line or iterative) and position of the AT domain, either within (*cis*-AT) or outside of the module (*trans*-AT).

Class: Modular *cis*-AT (cisAT)

Canonical type I PKSs that function in an assembly-line fashion (Jenke-Kodama *et al.*, 2005). Some members of this class perform additional functions and can be further distinguished. These are represented as subclasses within the *cis*-AT class.

Subclass: Olefin synthases (cisOLS)

KSs in the two OLS clades are associated with the biosynthesis of a terminal olefin. These are best known from Cyanobacteria (Coates *et al.*, 2014).

Subclass: Loading module (cisloading)

KSs in the first module of *cis*-AT modular PKSs in which the catalytic cysteine has been replaced with glutamine (often denoted as: KS_Q). This monophyletic clade also includes KSs that are responsible for both loading and the first elongation reaction (e.g. the SalA KS in salinosporamide biosynthesis (Eustáquio *et al.*, 2009)).

Subclass: Hybrid (cisHybridKS)

Located immediately downstream of a peptidyl carrier protein (PCP) domain and facilitate condensation reactions on a PCP-tethered intermediate. Thus, they are found in genes that contain both PKS and NRPS modules. Hybrid KSs are present in both *cis*- and *trans*-AT PKSs and together form a monophyletic clade in the reference tree.

Subclass: Tandem ECH (cistandemECH)

Associated with gene cassettes involved in the introduction of a β-branch to a β-keto group (often referred to as β-branching cassettes). KSs within this subclass are located in modules immediately downstream of β-branching cassettes (gene cassettes involved in the introduction of a beta-branch) and are associated with enoyl CoA hydratase (ECH) and enoyl reductase (ER) domains (Nakamura *et al.*, 2012). Contain a *cis*-acting ECH domain that performs the final decarboxylation to produce an unsaturated β-branch.

Class: Iterative *cis*-AT (iPKS)

These *cis*-AT type I PKSs function iteratively. They have been documented in bacteria and fungi and can be further divided into seven subclasses (Chen and Du, 2016; Chooi and Tang, 2012).

Subclass: Polyunsaturated fatty acids (iPKSPUFA)

Produce polyunsaturated fatty acids that contain multiple *cis* double bonds. Most commonly observed in bacteria and are more closely related to PKSs than to FASs. PUFA PKSs contain three KS domains and repetitive (from five to nine) tandem ACP domains. The three PUFA KSs form three distinct clades in the reference tree.

Subclass: Enediynes (iPKSenediyne)

Produce molecules characterized by a nine- or ten-membered ring that contains a conjugated alkyne-alkene-alkyne moiety. Observed in bacteria, they are composed of a single module that contains a PPTase at the C-terminal (Chen and Du, 2016). Members of this subclass form a monophyletic clade in the reference tree.

Subclass: Aromatic (iPKSaromatic)

Produce simple aromatic compounds that usually consist of mono- or bicyclic rings. The domain organization of the single module that is observed in bacteria is homologous to that observed in partially reducing fungal PKSs (Chen and Du, 2016). In the reference tree members of this subclass form paraphyletic clades together with the non-reducing PKSs.

Subclass: Polycyclic tetramate macrolactam-like (iPKSPTM)

Produce compounds usually consisting of a tetramic acid moiety and 2-3 rings fused to a macrolactam. The BGCs that produce these molecules consist of a single PKS-NRPS gene despite the presence of two distinct polyketide moieties in the products (Chen and Du, 2016). These KS domains, observed in bacteria, form a monophyletic clade in the reference tree.

Subclass: Non-reducing (iPKSNR)

Produce mono- or polycyclic aromatic polyketides from non-reduced and reactive poly-β-keto chains. Characterized by the absence of all three β-keto processing domains; ketoreductase (KR), dehydratase (DH), and enoyl reductase (ER). Distinguished by the presence of starter unit ACP transacylase (SAT), product template (PT), and thioesterase (TE) domains (Chooi and Tang, 2012). Members of this subclass are observed in fungi and form a monophyletic clade in the reference tree.

Subclass: Partially reducing (iPKSPR)

Are reported from fungal sources and produce simple mono- or bicyclic aromatic compounds similar to the products of bacterial aromatic iPKSs. Partially reducing PKSs lack ER domains and have either KR, or both KR and DH domain(s) (Gallo *et al.*, 2013). Members of this subclass form a monophyletic clade in the reference tree.

Subclass: Highly reducing (iPKSHR)

Produce linear and cyclic non-aromatic compounds. Highly reducing PKS genes possess all three β-keto processing domains (KR, DH, and ER) and often contain a C-methylation (CMeT) domain responsible for α-carbon methylation (Chooi and Tang, 2012). Some

highly reducing iPKSs can be fused with a C-terminal NRPS module. Members of this subclass form a monophyletic clade in the reference tree and are observed in fungi.

Class: *trans*-AT

Distinguished from *cis*-AT PKSs by the absence of a cognate AT domain, this activity is instead provided by a discrete protein that functions in *trans* with the type I PKS gene(s) (Piel, 2010). Phylogenies of *trans*-AT KSs have been shown to reflect substrate-specificity (see Nguyen *et al.*, 2008, Helfrich *et al.*, 2019). Here, three additional subclasses that relate to the function of the KS domain are defined. They are observed in bacteria.

Subclass: beta-branching module (transBdomain)

Observed within modules that facilitate chain branching and comprise a KS domain, a cryptic "B" domain, and an ACP domain. The functionally divergent KS domain, together with the B domain, catalyzes the addition of a malonyl unit onto an α,β-unsaturated intermediate (Bretschneider *et al.*, 2013). These modules can also facilitate ring formation, as observed in rhizoxin biosynthesis (Bretschneider *et al.*, 2013). KS domains associated with B domains form a subclade in the reference tree.

Subclass: Hybrid (transHybridKS)

Are similar to *cis*-AT hybrid KS (see above) except the AT domain occurs in *trans*. Members of this subclass clade with the *cis*-AT KSs in the reference tree.

Subclass: Hybrid non-elongating KS (transHybridKS0)

Non-elongating KS domains (often denoted as: KS^0) that follow an NRPS module. Identified by the lack of histidine in the conserved HGTGT catalytic motif required for decarboxylative condensation. Proposed to chaperone the peptidyl-intermediate from the upstream PCP onto the downstream module, where the subsequent KS domain performs the condensation reaction (Nguyen *et al.*, 2008). Currently characterized members of this subclass follow NRPS modules that generate oxazole or thioazole moieties from the incorporation of either serine or cysteine, respectively.

Class: Metazoa (MetazoaPKS)

Detected in Metazoans including birds (*Melopsittacus undulatus*), fish (*Oryzias latipes*), sea urchins (*Strongylocentrotus purpuratus*), and nematodes (*Caenorhabditis elegans*) (Feng *et al.*, 2021). Representative PKS sequences are included in the NaPDoS2 database but not in the

reference tree due to a lack of experimental characterization and a newly emerging understanding of their evolutionary relationships (Sabatini *et al.*, 2018).

Class: Protist (ProtistPKS)

Reported in the amoeba *Dictyostelium discoideum* and the parasite *Cryptosporidium parvum*. Sequences are included in the NaPDoS2 database but not in the reference tree because few have been experimentally characterized.

Type II PKS

Discrete, monofunctional proteins (Hertweck *et al.*, 2007). The most commonly studied type II PKSs function iteratively and produce aromatic polyketides, while others function non-iteratively (Shen *et al.*, 2003). Five distinct classes can be recognized in the KS phylogeny with the aromatic type II PKSs further delineated into subclasses.

Class: Aromatic (aromaticKSa) or (aromatic KSb)

Produce polycyclic aromatic compounds through the iterative decarboxylative condensation of malonyl-CoA extender units. The nascent poly-β-keto intermediate is further cyclized, aromatized, and subjected to other post-processing modifications (Hertweck *et al.*, 2007). The KS domain encodes a heterodimer consisting of an alpha subunit (KSα) that catalyzes the condensation reaction and a beta subunit (KSβ) that lacks the catalytic cysteine residue and has been implicated in control of the chain length (Hertweck *et al.*, 2007). For classification a phylogeny was generated using a concatenated alignment of the alpha and beta subunits and was used to establish the subclass designations below. Cladding patterns were associated with the early biosynthetic steps in the formation of the polyketide intermediate (including the number of elongations, position of the first ring cyclization, and type of starting unit). Some sequences were unable to be assigned to a subclass and are annotated as "aromaticKSa" or "aromaticKSb". While concatenated sequences were used to establish the subclass classification, the alpha and beta sequences remain individual in the reference database and are phylogenetically distinct in the reference tree.

Subclass: angucycline-derived I (angucyclineIKSa) or (angucyclineIKSb)

Derived from or contain an angular tetracyclic structure comprising a benz[*a*]anthracene moiety. Biosynthesis is most often initiated with an acetyl-CoA starting unit followed by

nine extensions using malonyl-CoA to produce a decaketide intermediate (Kharel *et al.*, 2012). The C-C bond formed during first ring cyclization occurs between C-7 and C-12. Some compounds within this class do not retain the angular tetracyclic core structure despite passing through a common benz[*a*]anthracene-containing intermediate (e.g. kinamycin).

Subclass: angucycline-derived II (angucyclineIIKSa or angucyclineIIKSb)
Distinguished from angucycline-derived I by the incorporation of a non-acetyl stater unit. In the case of lomaiviticin the starter unit consists of methylmalonyl-CoA-derived propionyl-CoA (Waldman and Balskus, 2014). Otherwise, shares a similar benz[*a*]anthracene moiety produced by nine extensions (with malonyl-CoA) yielding a decaketide.

Subclass: anthracycline-derived I (anthracyclineIKSa) or (anthracyclineIKSb)
Possess a linear tetracyclic core structure derived from 7,8,9,10-tetrahydro-5,12-naphtacenoquinones (Metsä-Ketelä *et al.*, 2007). Initiated with an acetyl-CoA starting unit followed by nine extensions to yield a decaketide and a first ring cyclization between C-7 and C-12. Members of this group can follow different folding patterns during the subsequent three ring cyclization reactions and are polyphyletic.

Subclass: anthracycline-derived II (anthracyclineIIKSa or anthracyclineIIKSb)
Distinguished from anthracycline-derived I by the incorporation of a propionyl starter unit and additional enzymes for starter unit selection in the BGCs (Metsä-Ketelä *et al.*, 2007). Otherwise shares a tetracyclic core formed by nine extensions.

Subclass: isochromanequinone-derived (isochromanequinoneKSa or isochromanequinoneKSb)
Octaketides with a linear tricyclic core structure containing isochromane and quinone moieties (Ichinose *et al.*, 2001). Several members of this subclass form dimers (e.g. actinorhodin).

Subclass: pentangular polyphenol-derived (pentangularpolyphenolKSa or pentangularpolyphenolKSb)
Long-chain polyphenols that form angular polycyclic core structures (Lackner *et al.*, 2007). Use either an acetyl-CoA as starting unit or longer chain acyl units (e.g. hexanoyl- and butyryl-CoA). Distinguished from other aromatic type II subclasses by longer pentacyclic and hexacyclic aromatic core structures formed from dodecaketide and tridecaketide chains, respectively. The C-C bond of the first ring cyclization occurs between C-9 and C-14.

Subclass: tetracenomycin-derived (tetracenomycinKSa or tetracenomycinKSb)
Linear tetracyclic decaketide core structures resulting from nine elongations of a starting unit (i.e. acetyl-CoA as in tetracenomycin (Hutchinson, 1997) or glycine derive starting unit as in lactonamycin (Zhang, 2008)). While the core structure is

anthracycline-like, the first ring cyclization occurs between C-9 and C-14 rather than C-7 and C-12.

Subclass: tetracycline-derived (tetracyclineKSa or tetracyclineKSb)
Tetracyclic ring structure formed from a decaketide poly-β-keto intermediate as seen in the anthracyclines. However, members of this subclass are characterized by a carboxamido moiety resulting from a malonamyl-CoA starting unit (Pickens and Tang 2009).

Subclass: spore pigment (sporepigmentKSa or sporepigmentKSb)
Not well characterized but attributed to the biosynthesis of Streptomycete spore pigments (e.g. *whiE* in *S. coelicolor*).

Class: β-branching cassettes (betabranch)

Clustered sets of free-standing genes responsible for the introduction of β-keto branches (often collectively referred to as β-branching cassettes); introduce a beta-branch to a beta-keto group. Usually consist of KS, HMGS (3-hydroxy-3-methylglutaryl-CoA synthase), CR (crotonyl reductase), and/or ECH (enoyl-CoA dehydratases) genes. The free-standing KS is responsible for the decarboxylation of malonyl-ACP and lacks the catalytic residues (lack the active site cysteine) required to catalyze the condensation reaction and function to decarboxylate ACP-bound malonyl as an early step in beta branch formation (Piel, 2010). Members of this group form a monophyletic clade.

Class: Polyenes (polyeneKSa or polyeneKSb)

Produce reduced linear polyenes such as ishigamide (Du *et al*., 2016, Du *et al.*, 2020) rather than polycyclic aromatic compounds. Iteratively acting and include multiple KS domains that are similar to the alpha and beta ketosynthase subunits of type II aromatic PKSs. The two different KSs form separate clades in the reference phylogeny. No subclasses are identified.

Class: Aryl polyenes (arylpolyeneKSa or arylpolyeneKSb)

Produce a polyene chain with an aryl moiety that is often substituted (Cimermancic *et al.*, 2014). They act iteratively and include two KS domains that share similarity with alpha and beta ketosynthase subunits and form separate clades in the reference phylogeny. No subclasses are identified.

Class: Non-iterative (noniterative)

Function as an assembly line to produce compounds such as pamamycin (Rebets *et al.*, 2015) and nonactin (Walczak *et al.*, 2000, Walczak *et al.*, 2001, Kwon *et al.*, 2001). The associated BGCs may also contain type III KSs. Non-iterative KS domains form a monophyletic clade in the reference tree.

Class: Type II unclassified (typeII)

Type II KS domains that could not be classified into one of the five type II classes are annotated as "typeII" in the database. These are not included in the reference tree.

Fatty Acid Synthases

KS domains from type I and II FASs have been included in the NaPDoS2 database and can be distinguished in the reference tree.

Class: Type I FAS

Large multifunctional proteins with a modular organization similar to *cis*-AT type I PKSs (Schweizer and Hofmann, 2004).

Subclass: Bacteria and fungi (bfFASI)

Found in bacteria and fungi with the associated KSs forming a monophyletic clade in the reference tree.

Subclass: Metazoa (MetazoaFASI)

Observed to date in the phyla Chordata and Nematoda. Included in the reference database but not included in the reference tree.

Subclass: Protist (ProtistFASI)

Observed in the phylum Apicomplexa (Alveolata). Included in the reference database but not included in the reference tree.

Class: Type II FAS (FASII)

Discrete, monofunctional proteins (Campbell and Cronan, 2001). Commonly observed in bacteria and archaea.

C Domains

The phylogeny of NRPS condensation (C) domains reflects eight functional categories (Rausch *et al.,* 2007).

Class: Starter (starter)

Typically, the first module of a NRPS usually does not contain a C domain. But, when present, these starter C domains acylate the first amino acid with a fatty acid, polyketide, or other molecule.

Class: LCL (LCL)

Catalyze the formation of a peptide bond between two L-amino acids.

Class: DCL (DCL)

Catalyze the formation of a peptide bond between an L-amino acid and a growing peptide ending with a D-amino acid.

Class: Cyclization (cyclization)

Catalyze both peptide bond formation and the subsequent cyclization of cysteine, serine or threonine residues.

Class: Epimerization (epimerization)

Change the chirality of the last amino acid in the chain from L to D. These domains can occur either before or after a C domain that performs a condensation reaction.

Class: Dual (dual)

Catalyzes both condensation and epimerization reactions.

Class: Modified amino acid (modifiedAA)

Modifies the incorporated amino acid, for example the dehydration of serine to dehydroalanine.

Class: Hybrid (hybridC)

Located downstream of an aminotransferase domain. Involved in the condensation of an amino acid to a growing polyketide resulting in a hybrid PKS/NRPS metabolite.

Class: Condensation (condensation)

Condensation domains that could not be classified into one of the eight known functionalities are annotated as "condensation" in the database. These are not included in the reference tree.

References

Bretschneider, T., Heim, J.B., Heine, D., Winkler, R., Busch, B., Kusebauch, B., Stehle, T., Zocher, G. and Hertweck, C., 2013. Vinylogous chain branching catalysed by a dedicated polyketide synthase module. *Nature*, 502(7469), pp.124-128. https://doi.org/10.1038/nature12588

Campbell, J.W. and Cronan Jr, J.E., 2001. Bacterial fatty acid biosynthesis: targets for antibacterial drug discovery. *Annual Reviews in Microbiology*, 55(1), pp.305-332. https://doi.org/10.1146/annurev.micro.55.1.305

Chen, H. and Du, L., 2016. Iterative polyketide biosynthesis by modular polyketide synthases in bacteria. *Applied Microbiology and Biotechnology*, 100(2), pp.541-557. https://doi.org/10.1007/s00253-015-7093-0

Chooi, Y.H. and Tang, Y., 2012. Navigating the fungal polyketide chemical space: from genes to molecules. *The Journal of Organic Chemistry*, 77(22), pp.9933-9953. https://doi.org/10.1021/jo301592k

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Brown, L.C.W., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J. and Birren, B.W., 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, 158(2), pp.412-421. https://doi.org/10.1016/j.cell.2014.06.034

Coates, R.C., Podell, S., Korobeynikov, A., Lapidus, A., Pevzner, P., Sherman, D.H., Allen, E.E., Gerwick, L. and Gerwick, W.H., 2014. Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways. *PloS one,* 9(1), p.e85140. https://doi.org/10.1371/journal.pone.0085140

Du, D., Katsuyama, Y., Onaka, H., Fujie, M., Satoh, N., Shin-ya, K. and Ohnishi, Y., 2016. Production of a novel amide-containing polyene by activating a cryptic biosynthetic gene cluster in Streptomyces sp. MSC090213JE08. *ChemBioChem*, 17(15), pp.1464-1471. https://doi.org/10.1002/cbic.201600167

Du, D., Katsuyama, Y., Horiuchi, M., Fushinobu, S., Chen, A., Davis, T.D., Burkart, M.D. and Ohnishi, Y., 2020. Structural basis for selectivity in a highly reducing type II polyketide synthase. *Nature Chemical Biology*, 16(7), pp.776-782. https://doi.org/10.1038/s41589-020-0530-0

Eustáquio, A.S., McGlinchey, R.P., Liu, Y., Hazzard, C., Beer, L.L., Florova, G., Alhamadsheh, M.M., Lechner, A., Kale, A.J., Kobayashi, Y. and Reynolds, K.A., 2009. Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from S-adenosyl-L-methionine. *Proceedings of the National Academy of Sciences*, 106(30), pp.12295-12300. https://doi.org/10.1073/pnas.0901237106

Feng, L., Gordon, M.T., Liu, Y., Basso, K.B. and Butcher, R.A., 2021. Mapping the biosynthetic pathway of a hybrid polyketide-nonribosomal peptide in a metazoan. *Nature Communication,* 12, 4912. https://doi.org/10.1038/s41467-021-24682-9

Gallo, A., Ferrara, M. and Perrone, G., 2013. Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. *Toxins*, 5(4), pp.717-742. https://doi.org/10.3390/toxins5040717

Helfrich, E.J., Ueoka, R., Dolev, A., Rust, M., Meoded, R.A., Bhushan, A., Califano, G., Costa, R., Gugger, M., Steinbeck, C. and Moreno, P., 2019. Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nature Chemical Biology,* 15(8), pp.813-821. https://doi.org/10.1038/s41589-019-0313-7

Hertweck, C., Luzhetskyy, A., Rebets, Y. and Bechthold, A., 2007. Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. *Natural Product Reports*, 24(1), pp.162-190. https://doi.org/10.1039/B507395M

Hutchinson, C.R., 1997. Biosynthetic studies of daunorubicin and tetracenomycin C. *Chemical Reviews*, 97(7), pp.2525-2536. https://doi.org/10.1021/cr960022x

Ichinose, K., Taguchi, T., Bedford, D.J., Ebizuka, Y. and Hopwood, D.A., 2001. Functional complementation of pyran ring formation in actinorhodin biosynthesis in Streptomyces coelicolor A3 (2) by ketoreductase genes for granaticin biosynthesis. *Journal of Bacteriology*, 183(10), pp.3247-3250. https://doi.org/10.1128/JB.183.10.3247-3250.2001

Jenke-Kodama, H., Sandmann, A., Müller, R. and Dittmann, E., 2005. Evolutionary implications of bacterial polyketide synthases. *Molecular Biology and Evolution*, 22(10), pp.2027-2039. https://doi.org/10.1093/molbev/msi193

Jenke-Kodama, H. and Dittmann, E., 2009. Evolution of metabolic diversity: insights from microbial polyketide synthases. *Phytochemistry*, 70(15-16), pp.1858-1866. https://doi.org/10.1016/j.phytochem.2009.05.021

Kharel, M.K., Pahari, P., Shepherd, M.D., Tibrewal, N., Nybo, S.E., Shaaban, K.A. and Rohr, J., 2012. Angucyclines: Biosynthesis, mode-of-action, new natural products, and synthesis. *Natural Product Reports*, 29(2), pp.264-325. https://doi.org/10.1039/C1NP00068C

Kwon, H.J., Smith, W.C., Scharon, A.J., Hwang, S.H., Kurth, M.J. and Shen, B., 2002. CO bond formation by polyketide synthases. Science, 297(5585), pp.1327-1330. https://doi.org/10.1126/science.1073175

Lackner, G., Schenk, A., Xu, Z., Reinhardt, K., Yunt, Z.S., Piel, J. and Hertweck, C., 2007. Biosynthesis of pentangular polyphenols: deductions from the benastatin and griseorhodin pathways. *Journal of the American Chemical Society*, 129(30), pp.9306-9312. https://doi.org/10.1021/ja0718624

Metsä-Ketelä, M., Niemi, J., Mäntsälä, P. and Schneider, G., 2007. Anthracycline biosynthesis: genes, enzymes and mechanisms. *Anthracycline Chemistry and Biology I*, pp.101-140. https://doi.org/10.1007/128_2007_14

Nakamura, H., Hamer, H.A., Sirasani, G. and Balskus, E.P., 2012. Cylindrocyclophane biosynthesis involves functionalization of an unactivated carbon center. *Journal of the American Chemical Society*, 134(45), pp.18518-18521. https://doi.org/10.1021/ja308318p

Nguyen, T., Ishida, K., Jenke-Kodama, H., Dittmann, E., Gurgui, C., Hochmuth, T., Taudien, S., Platzer, M., Hertweck, C. and Piel, J., 2008. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nature Biotechnology*, 26(2), pp.225-233. https://doi.org/10.1038/nbt1379

Piel, J., 2010. Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural Product Reports*, 27(7), pp.996-1047. https://doi.org/10.1039/B816430B

Rausch, C., Hoof, I., Weber, T., Wohlleben, W. and Huson, D.H., 2007. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evolutionary Biology*, 7(1), pp.1-15. https://doi.org/10.1186/1471-2148-7-78

Rebets, Y., Brötz, E., Manderscheid, N., Tokovenko, B., Myronovskyi, M., Metz, P., Petzke, L. and Luzhetskyy, A., 2015. Insights into the pamamycin biosynthesis. *Angewandte Chemie International Edition*, 54(7), pp.2280-2284. https://doi.org/10.1002/anie.201408901

Sabatini, M., Comba, S., Altabe, S., Recio-Balsells, A.I., Labadie, G.R., Takano, E., Gramajo, H. and Arabolaza, A., 2018. Biochemical characterization of the minimal domains of an iterative eukaryotic polyketide synthase. *The FEBS Journal*, 285(23), pp.4494-4511. https://doi.org/10.1111/febs.14675

Schweizer, E. and Hofmann, J., 2004. Microbial type I fatty acid synthases (FAS): major players in a network of cellular FAS systems. *Microbiology and Molecular Biology Reviews*, 68(3), pp.501-517. https://doi.org/10.1128/MMBR.68.3.501-517.2004

Shen, B., 2003. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology*, 7(2), pp.285-295. https://doi.org/10.1016/S1367-5931(03)00020-6

Walczak, R.J., Woo, A.J., Strohl, W.R. and Priestley, N.D., 2000. Nonactin biosynthesis: the potential nonactin biosynthesis gene cluster contains type II polyketide synthase-like genes. *FEMS Microbiology Letters*, 183(1), pp.171-175. https://doi.org/10.1111/j.1574-6968.2000.tb08953.x

Walczak, R.J., Nelson, M.E. and Priestley, N.D., 2001. Nonactin Biosynthesis: Disruption of the Polyketide Synthase Genes, n onKJ, in Streptomyces griseus Leads to an Overall 96 Decrease in Macrotetrolide Production, Yet a Net Increase in Nonactin Analogues Which Incorporate Isobutyrate. *Journal of the American Chemical Society,* 123(42), pp.10415-10416. https://doi.org/10.1021/ja005749o

Waldman, A.J. and Balskus, E.P., 2014. Lomaiviticin biosynthesis employs a new strategy for starter unit generation. *Organic Letters*, 16(2), pp.640-643. https://doi.org/10.1021/ol403714g

Zhang, X., Alemany, L.B., Fiedler, H.P., Goodfellow, M. and Parry, R.J., 2008. Biosynthetic investigations of lactonamycin and lactonamycin z: cloning of the biosynthetic gene clusters and discovery of an unusual starter unit. *Antimicrobial Agents and Chemotherapy*, 52(2), pp.574-585. https://doi.org/10.1128/AAC.00717-07

Tutorials: Using NaPDoS2

We have included multiple detailed tutorials for using NaPDoS2:

1) **Analyzing a single genome with NaPDoS2**: analyze a *Salinispora* bacterial genome
2) **Analyzing multiple genomes with NaPDoS2**: analyze multiple fungal genomes
3) **Analyzing an animal genome with NaPDoS2**: analyze the *Elysia chlorotica* genome
4) **Analyzing a BGC from antiSMASH, PRISM, or MIBiG with NaPDoS2**
5) **Analyzing a metagenome with NaPDoS2**: analyze a marine sediment metagenome
6) Analyzing **individual genes/proteins & KS PCR clone products** with NaPDoS2
7) Analyzing **KS amplicon tag-sequencing datasets** with NaPDoS2
8) How to use the **File Size Management Tools**

These tutorials cover:

- How NaPDoS2 can be applied to objectives & goals of a research question and hypothesis building
- Interpretation of NaPDoS2 results
- Details about the NaPDoS2 webtool interface steps
- Suggestions for further analyses & our considerations for analyses

- While the tutorials include instructions to download the sequences for each tutorial, we have included all sequences and other analysis files for download on the **NaPDoS2 Tutorial OSF page**.

Still have questions? Please see the sections following the tutorials for:

- Additional details about the NaPDoS2 webtool interface steps
- Additional interpretation notes
- **FAQ (Frequently Asked Questions)**

Still stuck on something with the tutorials?
Contact us: email Kaitlin (creator of these tutorials; email: kcreamer at ucsd dot edu), or via our **Contact Us** page.

1. Analyzing a single genome with NaPDoS2: analyze a *Salinispora* bacterial genome

Objective: Analyze a single *Salinispora* bacterial genome to identify & classify KSs of interest, assess their novelty, and use the query domain hits to find what contigs the KSs are found on in the genomes.

Genome tutorial A objective

Analyze a single bacterial genome Salinispora tropica CNB-440 to uncover the unique salC KS domain, which clades separately from other KS domains and thus likely represents a novel biosynthetic class of KS (as reported in Bauman et al. Nat Chem Biol. 2022).

1. **Gather the *Salinispora tropica* CNB-440 genome sequence**

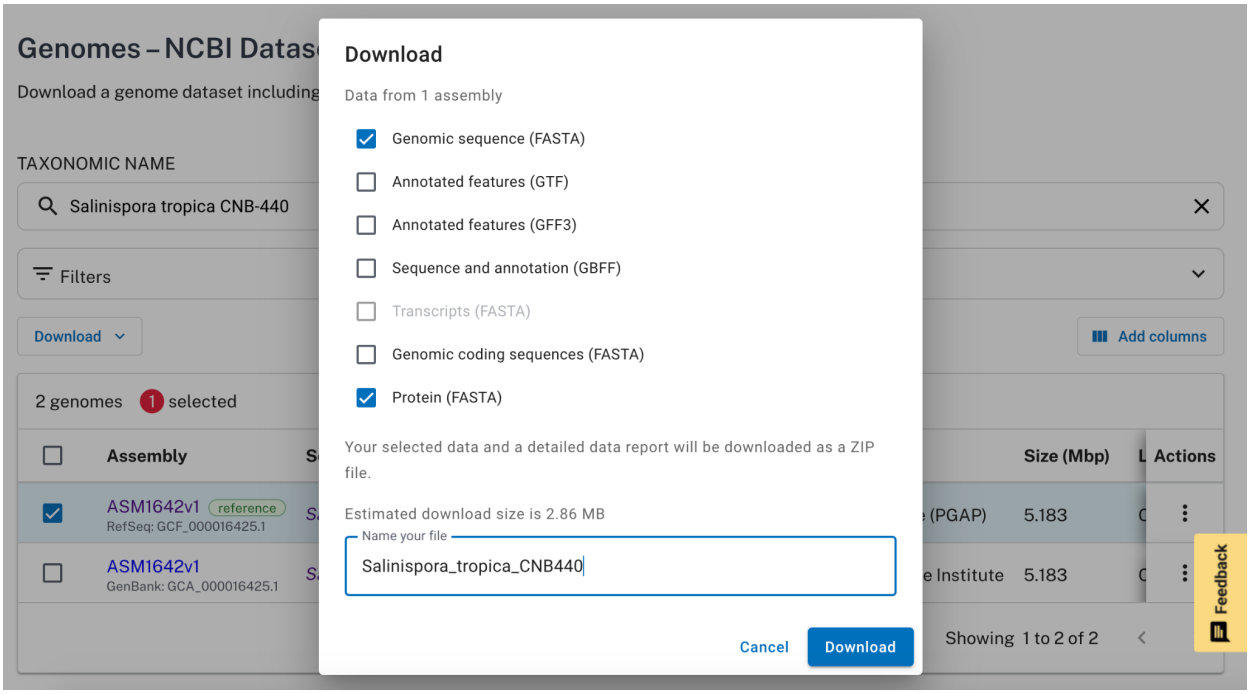
    - We'll download the FASTA-formatted genome sequence of *Salinispora tropica* CNB-440 using the NCBI Genomes Dataset tool.

        - We have also provided the genome sequences used in the tutorial on the NaPDoS2 Tutorial OSF page that you can download directly from there

        - On the NCBI Genomes Dataset tool, search "**CNB-440**" and select "**Salinispora tropica CNB-440**".

Bacterial genome	**Accession**	**Reference**
Salinispora tropica CNB-440	GCF_000016425.1	Millán-Aguiñaga et al. 2017. Sci Rep.

        - Check the box of the first result (assembly "ASM1642v1") and click "**Download dataset**". Select both "**Genomic sequence (FASTA)**" and "**Protein (FASTA)**"-- you can analyze either the genomic (nucleic acid) or protein (amino acid) FASTA genomes sequences with NaPDoS2.



- Save the downloaded zipped file in a place you can access it (for example, your Desktop or Downloads folder). Navigate to where you saved it and unzip the file
    - (Windows: double-click, or select it and click "Extract all"; Mac: double-click; Linux/Command line: navigate to the directory containing your zipped file and use the command "`unzip filename.zip`")

- The *Salinispora tropica* CNB-440 genome sequences are found in the folder "GCF_000016425.1" which can be found by navigating through the folder path: `~/ncbi_dataset/ncbi_dataset/data/GCF_000016425`
    - "GCF_000016425.1_ASM1642v1_genomic.fna": nucleic acid FASTA genome file
    - "protein.faa": protein FASTA genome file

2. **Submit a NaPDoS2 analysis of the single bacterial genome *Salinispora tropica* CNB-440**

- Navigate to the "Run Analysis" tab on NaPDoS2. Select "**Domain Type: KS domains**".

- If you are analyzing the *Salinispora tropica* CNB-440 "protein.faa" protein FASTA genome file, select "**Query type: Amino acid sequences**". If you are analyzing the *Salinispora tropica* CNB-440 "GCF_000016425.1_ASM1642v1_genomic.fna": nucleic acid FASTA genome file, select "**Query type: Nucleic acid sequences**".

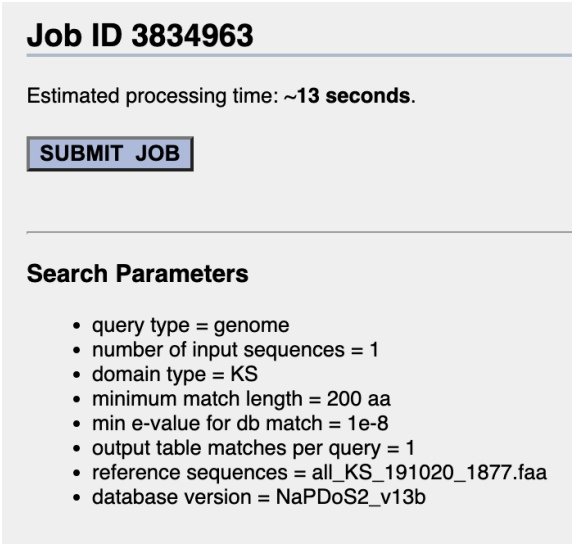
- Click "**Choose File**" to upload your genome file. Be sure to select the correct one depending if you selected "Query type" amino acid or nucleic acid sequences.

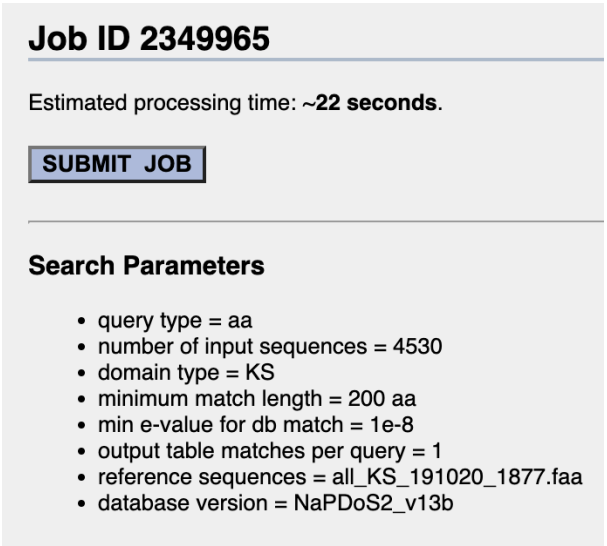
- Click on "**Advanced Settings**". Here, you can change the BLASTP e-value cutoff and the Minimum alignment length. For this tutorial, we will use the default settings of BLASTP e-value cutoff of 1e-8 and Minimum alignment length of 200aa (600nt).

    - Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details.

- Click **SEEK** to be assigned a job ID and review the job analysis settings.

Submitted job ID screens for *Salinispora tropica* CNB-440 genome nucleic acid (**left**) and amino acid protein (**right**) FASTA files:





- For "Query type: nucleic acid sequences", you will first proceed to the "**Nucleotide Sequence Search Results**" page which lists the initial number of KS domains that NaPDoS2 identified. On this page, you can **DOWNLOAD** (click link to view, or right-click & select "Save Link As" to save the table) a table of the nucleotide match locations in tab-delimited format.

    - This lists the domains numbered KS# ("**domain id**"), the sequence/contig name the domain was found on ("**parents seq**"), the base pair/nucleotide start and end location of the domain ("**start**" , "**end**"), the translation reading frame ("**read frame**"), and the final name of the domain ("**cand_id**").

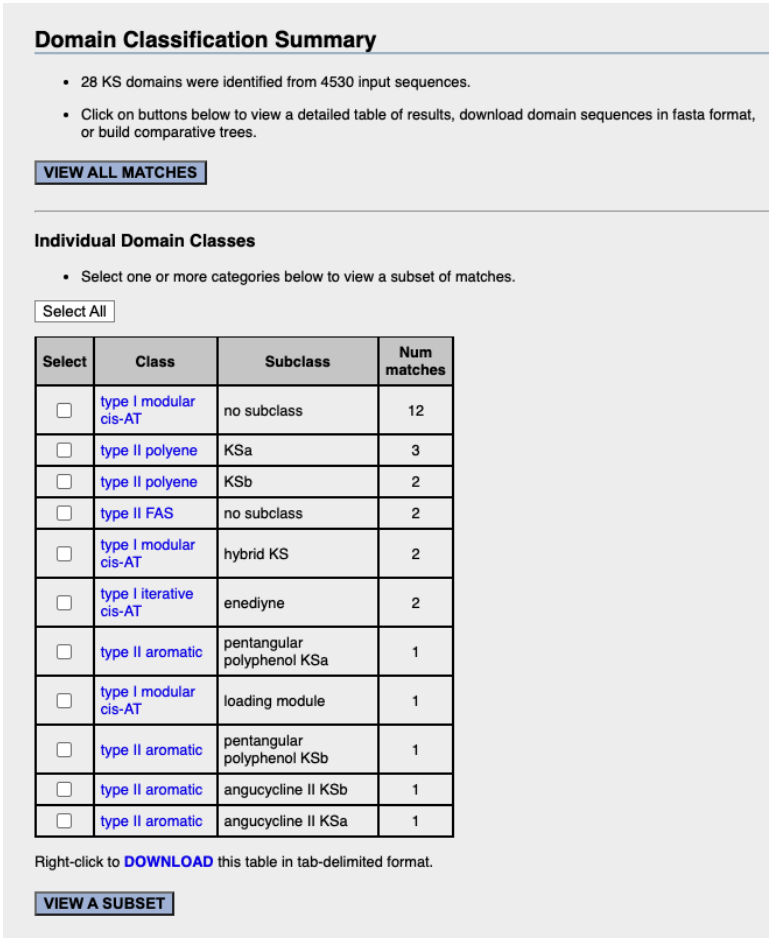
- This table can be used to find the location of the KS domains in your nucleic acid genome file via the starting and ending base pairs.

- To continue the analysis, click "**CONTINUE ANALYSIS**"

3. Select results display options for KSs that NaPDoS2 detected and classified

- The "**Domain Classification Summary**" page lists the total number of domains found from the number of sequences that were input. Clicking "**VIEW ALL MATCHES**" opens a new page with the hit table of all domains identified in the genome analysis.

    - In our example, 28 KS domains were found from the 1 nucleic acid *Salinispora tropica* CNB-440 genome and 28 KS domains were found from 4,530 genome amino acid proteins.



Domain Classification Summary

- 28 KS domains were identified from 4530 input sequences.
- Click on buttons below to view a detailed table of results, download domain sequences in fasta format, or build comparative trees.

VIEW ALL MATCHES

Individual Domain Classes

- Select one or more categories below to view a subset of matches.

Select All

Select	Class	Subclass	Num matches
☐	type I modular cis-AT	no subclass	12
☐	type II polyene	KSa	3
☐	type II polyene	KSb	2
☐	type II FAS	no subclass	2
☐	type I modular cis-AT	hybrid KS	2
☐	type I iterative cis-AT	enediyne	2
☐	type II aromatic	pentangular polyphenol KSa	1
☐	type I modular cis-AT	loading module	1
☐	type II aromatic	pentangular polyphenol KSb	1
☐	type II aromatic	angucycline II KSb	1
☐	type II aromatic	angucycline II KSa	1

Right-click to **DOWNLOAD** this table in tab-delimited format.

VIEW A SUBSET

- The "**Individual Domain Classes**" lists the number of specific classes of KS domains that were found in your query genome sequence, ordered by the class and subclass with the most number of domains.

    - This is helpful if you only are interested in a specific subset type of domain– you can click to select what class/subclasses you would like to view the results of and click "**VIEW A SUBSET**", which will open a new page with a table of only the classes/subclasses you selected.

    - You can also "**DOWNLOAD**" the "**Individual Domain Classes**" table (click link to view; or right-click link & select "Save Link As" to save the table).

4. Assess and download NaPDoS2 search results

- If you clicked "**VIEW ALL MATCHES**" or "**VIEW A SUBSET**", you will view the "**Database Search Results**" page, either for all hits in your query sequence or for a subset of hits of certain class/subclass that you selected.

- The table can be sorted by each of the column values by clicking on the column header; the table scrolls up and down and if your query hit names are long, the table scrolls left to right. You can **DOWNLOAD** the results table (click link to view; or right-click link & select "Save Link As" to save the table) in a tab-delimited format.

- The scrollable table lists the following information:

    - **cand_id**: name of the domain hit, which is the name of the sequence in the query FASTA file before spaces and appended with location information of where the domain hit was found (see below for details)

    - **database match**: the name of the domain in the NaPDoS2 database that the query hit is closest to
        - **names include**: BGCname_KS#_uniquetag
        - **BGC name**: name of the BGC
        - **KS#**: number KS in successive order from the BGC
        - **unique tag**: abbreviation that is unique for every class/subclassification

    - **percent identity:** percent identity that the query hit shares with the closets NaPDoS2 database match

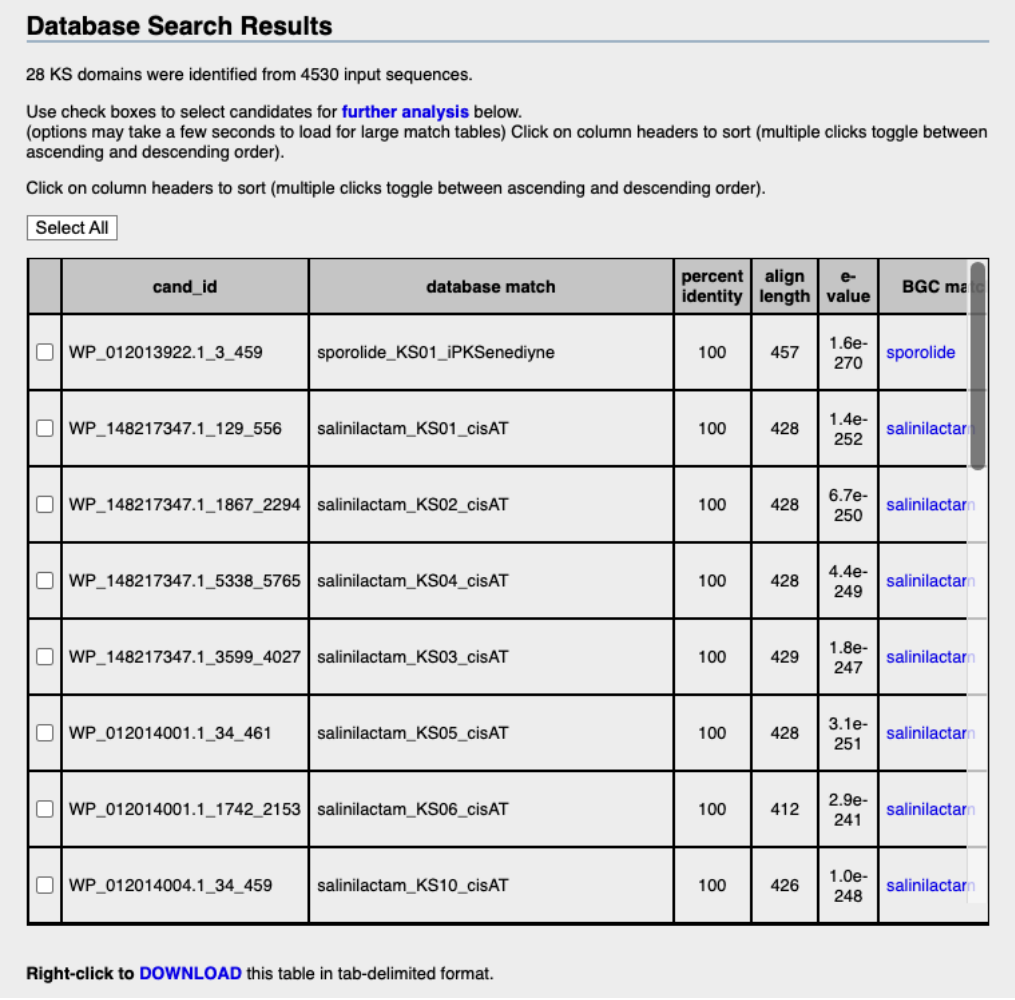
    - **align length:** alignment length that the query hit shares with the closest NaPDoS2 database match

- **e-value:** e-value of the query hit to the closest NaPDoS2 database match

- **BGC match**: name of the BGC of the closest NaPDoS2 database match to the query hit; click on the link to open a new page of the BGC card information

- **domain class:** classification of the closest NaPDoS2 database match to the query hit
    - This is the ***classification*** of your query hit

- **domain subclass:** specific subclass of the closest NaPDoS2 database match to the query hit
    - This is the ***complete specific subclass*** of your query hit (some classes do not have subclasses, ie “no subclass”)



Database Search Results

28 KS domains were identified from 4530 input sequences.

Use check boxes to select candidates for **further analysis** below.
(options may take a few seconds to load for large match tables) Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Select All

	cand_id	database match	percent identity	align length	e-value	BGC ma
☐	WP_012013922.1_3_459	sporolide_KS01_iPKSenediyne	100	457	1.6e-270	sporolide
☐	WP_148217347.1_129_556	salinilactam_KS01_cisAT	100	428	1.4e-252	salinilactar
☐	WP_148217347.1_1867_2294	salinilactam_KS02_cisAT	100	428	6.7e-250	salinilactar
☐	WP_148217347.1_5338_5765	salinilactam_KS04_cisAT	100	428	4.4e-249	salinilactar
☐	WP_148217347.1_3599_4027	salinilactam_KS03_cisAT	100	429	1.8e-247	salinilactar
☐	WP_012014001.1_34_461	salinilactam_KS05_cisAT	100	428	3.1e-251	salinilactar
☐	WP_012014001.1_1742_2153	salinilactam_KS06_cisAT	100	412	2.9e-241	salinilactar
☐	WP_012014004.1_34_459	salinilactam_KS10_cisAT	100	426	1.0e-248	salinilactar

Right-click to DOWNLOAD this table in tab-delimited format.

- Here, you can also select specific domain hits to download or further analyze (select all to download or analyze all results). You have the follow options for downloading the data:
    - **View nucleotide coordinates** for all trimmed domain candidates
        - **only an option if your query sequence was a nucleic acid file
        - This shows a table where each potential domain candidate has been given a unique candidate ID number, based on parent sequence id, reading frame number (1-6), gene number within the reading frame, and trimmed nucleotide start and stop coordinates within the reading frame.
        - The table columns are sortable by each column header. You can **DOWNLOAD** the table of nucleotide match locations in tab-delimited format (click link to view; or right-click link & select "Save Link As" to save)
    - **Output selected sequences** in FASTA format
        - Select which trimmed query domain hits or select all to output in FASTA format
            - **if your query was a nucleic acid, you have the option to output the query domain hits as either amino acid or nucleic acid
            - **if your query was an amino acid protein, you only have the option to output the query domain hits as amino acid sequences
        - Click **GET TRIMMED SEQS**
        - This is a very useful output for you to continue working with the KS or C domains that NaPDoS2 detected & classified. You can copy and paste the trimmed query KS or C domain hits into a new file, a sequencing editing program, etc.
        - For interpretation of the appended information to your query search (ie. >ProteinA_13_429 or >GeneB_5_1323_56_472), please see below.
    - **Output Alignment** with closest database matches

- Select alignment format (all alignments are created with MUSCLE but can be output in different formats suitable for many downstream applications):
    - FASTA (.fasta)
    - Msf (.msf)
    - Clustalw (.clw)

- This outputs an alignment of your selected sequences with the closest NaPDoS2 database matches. Note, the alignment output will include all closest NaPDoS2 database matches to all query hits, not just your selected sequences, but only your selected sequences will be present in the alignment.

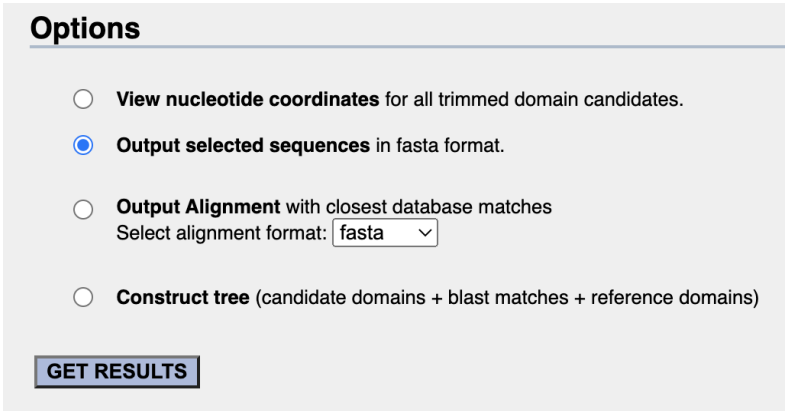
- You can **DOWNLOAD** (click link to view; or right-click link & select "Save Link As" to save) the alignment file for downstream use in sequence programs, etc.

■ **Construct tree** (candidate domains + blast matches + reference domains)

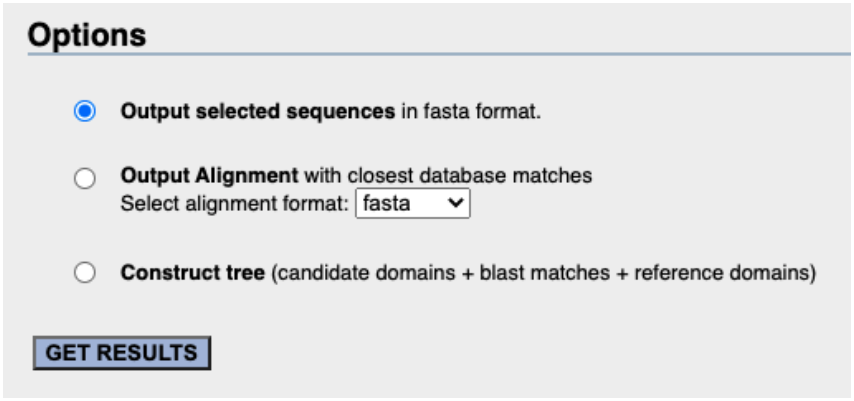
- Select the query domain hits that you would like to build a tree with. You can either build a tree with:
    - **Closest database matches** to your selected sequences, ie. # of selected queries + closest database references
    - **All classified KS domains** which would be the # of selected queries + all 414 NaPDoS2 database reference sequences.

- The estimated time to calculate the trees is given, usually the time to complete the trees is much faster than estimated.

- Details about tree building are continued below as the aim of this tutorial is put potentially novel KS sequences into phylogenetic context with known KS domains.



Data view options with a nucleic acid query sequence.



Data view options with a protein amino acid query sequence.

5. Constructing a tree to estimate biosynthetic novelty of domain hits

- Our goal of this tutorial was to find the salinosporamide biosynthetic KS domains in *Salinispora tropica* CNB-440 and assess how closely related it is to other biosynthetically characterized KSs in the NaPDoS2 database.

- Therefore, searching through the table, we see that there were 2 hits to the "salinosporamide_KS01_cisloading" and "salinosporamide_KS02_cisHybridKS" domains in our genome. We will select these 2 sequences to construct a tree.

    - By clicking on the "salinosporamide" BGC, we can see that the salinosporamide BGC has 3 domains in the NaPDoS2 database, and by clicking the MIBiG ID that links to the complete BGC, we can see that the BGC only has these 3 domains. We are interested in the two KS domains.

	cand_id	database match	percent identity	align length	e-value	BGC matcl
☑	WP_011904932.1_107_524	salinosporamide_KS01_cisloading	100	418	1.8e-248	salinosporam
☑	WP_011904930.1_13_429	salinosporamide_KS02_cisHybridKS	100	417	1.7e-247	salinosporam

- We want to put these 2 query KS domain hits into the context of the entire NaPDoS2 reference database. We'll select the second option of building a tree with our query hits and **all classified KS domains**.

Select Preferred Reference Alignment:

○ **Closest database matches**
2 queries + closest database references
Estimated processing time: **~18 seconds**.

◉ **all_classified_KS domains**
2 queries + all 414 database references
Estimated processing time: **~2 minutes**.

SUBMIT JOB

- The **Tree results** page has two options for viewing the calculated trees:
  - Tree image (SVG format file), click **DOWNLOAD** to view, right-click to save.
    - When viewing the tree, the query domain hits will have red circles of where they are placed in the tree.
  - Newick format (plain text file), click **DOWNLOAD** to view, right-click to save
    - This file format can be viewed in tree viewing programs like FigTree, iTol, TreeViewer, ETE Toolkit, etc.

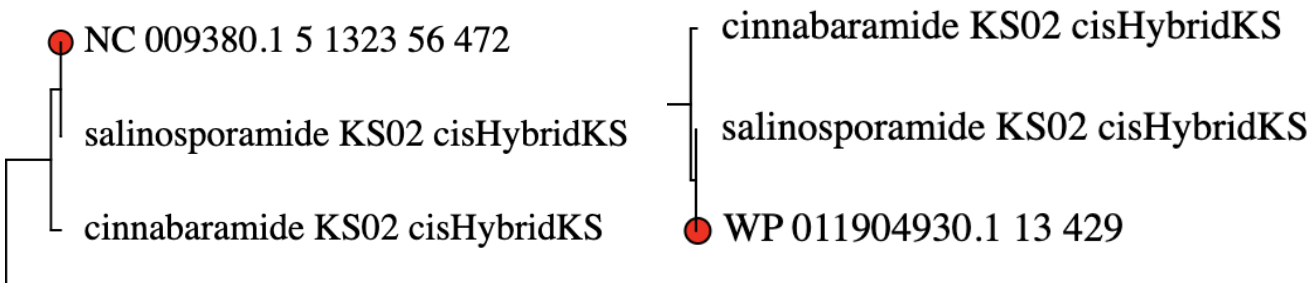
Tree results

Additional biosynthetic gene cluster information for reference database matches in these trees, is provided in the **BGCs** reference tab.

Tree image (svg format file) Right click to **DOWNLOAD**

Newick format (plain text file) Right click to **DOWNLOAD**

- We are interested in the *salC* KS domain that clades separately from other type 1 *cis*-AT hybrid KS pathways and was recently found to perform a unique function for a KS domain in *Bauman et al. Nat Chem Biol. 2022*. We see that WP_011904930.1 (protein *Salinispora tropica* CNB-440 genome) and "NC_009380.1_5_1323_56_472" (nucleic acid *Salinispora tropica* CNB-440 genome) are classified as type 1 *cis*-AT hybrid KSs and clade with the *salC* homolog in the cinnabaramide BGC.



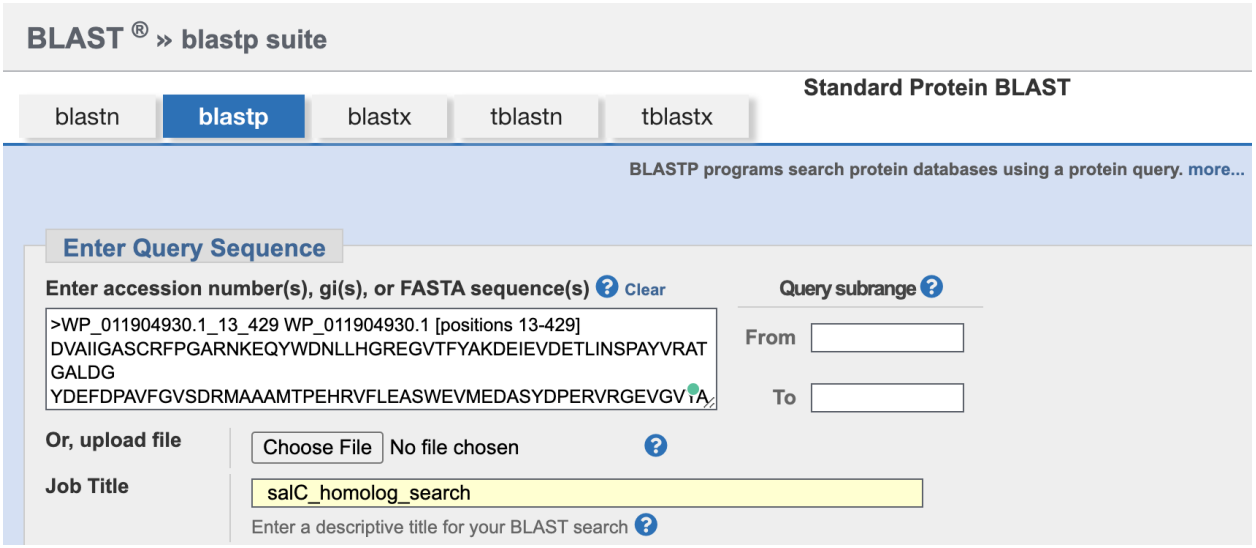
Nucleic acid (left) and protein (right) query hits (red circles)– these are the SalC sequences from the Salinispora tropica CNB-440 genomes in the context of the NaPDoS2 reference sequence database tree.

- To determine how closely related this *salC* "hybrid KS" domain is to other domains, we will re-run our NaPDoS2 analysis, but this time, use the query domain hit to:
  - 1) identify other potential homologs via NCBI BLASTP
  - 2) run all homologs and the original *salC* query hit through NaPDoS2 to detect and classify the domains and
  - 3) construct a tree to assess the relatedness of *salC*, close homologs, and known KS domains in the NaPDoS2 database.

- 1) To identify other potential *salC* homologs, we will select the WP_011904930.1 (protein *Salinispora tropica* CNB-440 genome) and "NC_009380.1_5_1323_56_472" (nucleic acid *Salinispora tropica* CNB-440 genome) which are the query hits with the closest database match of "***salinosporamide_KS02_cisHybridKS***". Output the selected sequences as an amino acid, and copy and paste into NCBI BLASTP search query box. Click **BLAST**.

```
>WP_011904930.1_13_429 WP_011904930.1 [positions 13-429]
DVAIIGASCRFPGARNKEQYWDNLLHGREGVTFYAKDEIEVDETLINSPAYVRATGALDG
YDEFDPAVFGVSDRMAAAMTPEHRVFLEASWEVMEDASYDPERVRGEVGVYASTNPQSAA
LYSSPPDWVSAGPEVMDRSNAWLPDTITSNVLYYLGLTGEAVTVTAVCSGFHYAVHLACQ
SLLLGQTDMAIAGGVMVRLPQRRGYLWEEGRILSRDGHCRPFDANGTGSVLASGVAVVLL
KPLPQAVADRDHIYATVKGTAINNNGISAMAYGLAQPERLSACIAGAMQAGDVAPETVSM
YEANGFGMPITDSLEVHAAHLAFGKQSGTCSIGAVKGNIGHAGVVAGGSGAVKAAFALYH
RSLPPTINLTELNEEIDFPRTPFVPQLEPAAWQPECGIRRAGITALGGGGYNAHLVL
```

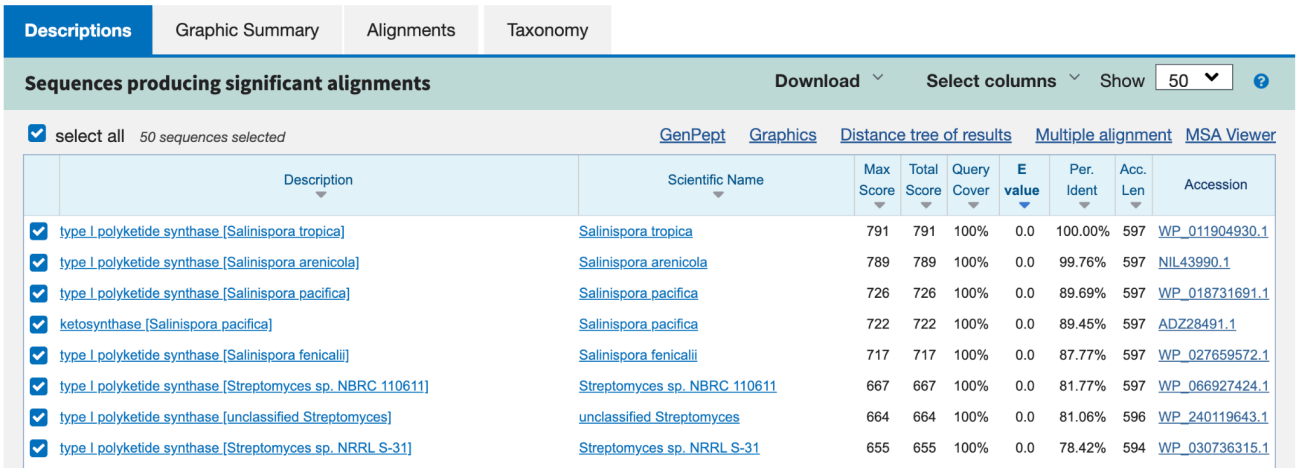
NaPDoS2 output of our salC sequence that we copy & paste into NCBI BLASTP.



NCBI BLASTP job for our salC sequence identified by NaPDoS2 from the Salinispora tropica CNB-440 genome.

- 2) We see that BLASTP has found the top 100 BLASTP hits, where many of these are potentially homologs to our search query. We will select the top 50 homologous sequences to analyze in the context of our NaPDoS2 database and original search query. Select the top 50 sequences (click dropdown "**Show**" and select **50**; click **select all** sequences check box). Click **download**, and select "**FASTA (complete sequence)**" and save the file. You can open this text file and it will include the top 50 homologs to our *salC* query domain hit search.

- We have also included this file of 50 BLASTP homologous sequences for download on the [NaPDoS2 Tutorial OSF page]– it also has the original *salC* BLASTP query sequence (n=51 sequences).



Descriptions | Graphic Summary | Alignments | Taxonomy

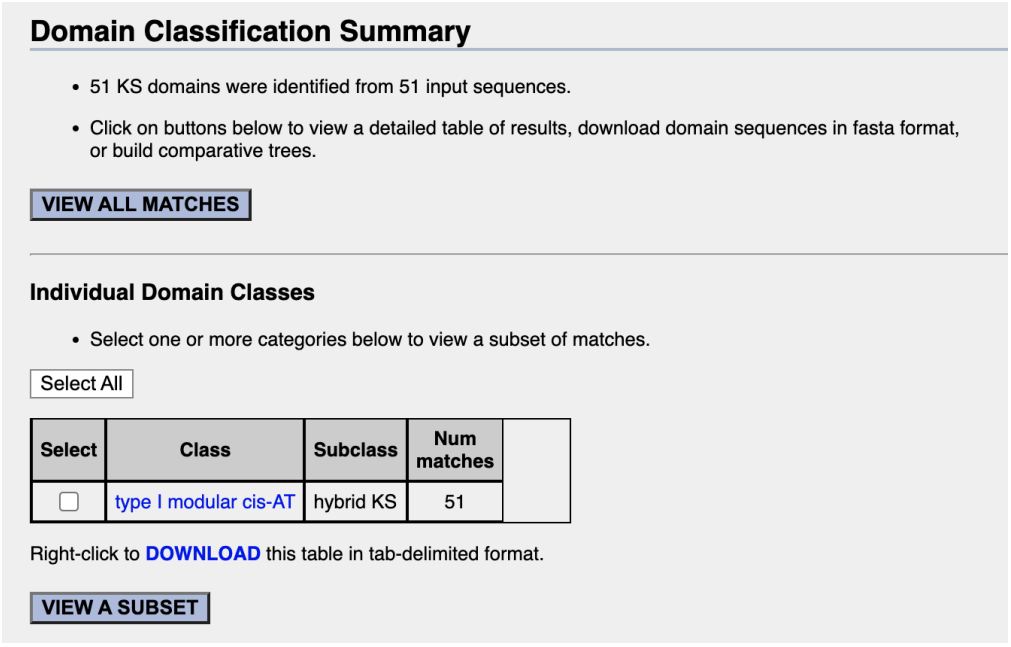
Sequences producing significant alignments — Download | Select columns | Show 50

select all *50 sequences selected* — GenPept | Graphics | Distance tree of results | Multiple alignment | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
type I polyketide synthase [Salinispora tropica]	Salinispora tropica	791	791	100%	0.0	100.00%	597	WP_011904930.1
type I polyketide synthase [Salinispora arenicola]	Salinispora arenicola	789	789	100%	0.0	99.76%	597	NIL43990.1
type I polyketide synthase [Salinispora pacifica]	Salinispora pacifica	726	726	100%	0.0	89.69%	597	WP_018731691.1
ketosynthase [Salinispora pacifica]	Salinispora pacifica	722	722	100%	0.0	89.45%	597	ADZ28491.1
type I polyketide synthase [Salinispora fenicalii]	Salinispora fenicalii	717	717	100%	0.0	87.77%	597	WP_027659572.1
type I polyketide synthase [Streptomyces sp. NBRC 110611]	Streptomyces sp. NBRC 110611	667	667	100%	0.0	81.77%	597	WP_066927424.1
type I polyketide synthase [unclassified Streptomyces]	unclassified Streptomyces	664	664	100%	0.0	81.06%	596	WP_240119643.1
type I polyketide synthase [Streptomyces sp. NRRL S-31]	Streptomyces sp. NRRL S-31	655	655	100%	0.0	78.42%	594	WP_030736315.1

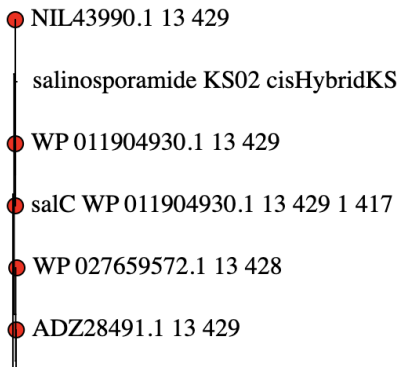
Selecting the top 50 salC homologs from our BLASTP search.

- 3) Next, we will analyze all of the BLASTP homologs and our original query sequence through NaPDoS2 using the same settings as previously (**repeat steps 1-4**). Copy and paste the 50 FASTA-formatted top homologous sequences, and the original *salC* query domain hit (must be in protein format, so either "**WP_011904930.1**" (protein *Salinispora tropica* CNB-440 genome) or "**NC_009380.1_5_1323_56_472**" (nucleic acid *Salinispora tropica* CNB-440 genome) output as an amino acid sequence). You should be analyzing a total of 51 sequences, and NaPDoS2 should detect and classify all 51 sequences as KS domains.



NaPDoS2 detected & classified all 51 SalC homologs.

- 4) Select all 51 query domain hits from the **Database Search Results** table and select "**Construct tree"**. Select "**all classified KS domains"** so we can see how these 51 KS domain sequences compare with all NaPDoS2 KS database reference sequences. We can see that our BLASTP homologs form a large clade with our original *salC* sequence!



Partial screenshot of the NaPDoS2 output tree of our salC homologs.

- To make this publication-ready, we can view this in a tree viewer like FigTree, color the BLAST matches blue, color our original KS *salC* query hit as red, and keep the NaPDoS2 sequences black as in this figure, published as ***Supplementary Figure 4*** in *Bauman et al. Nat Chem Biol. 2022*.



Screenshot of part of Supplementary Figure 4 from [Bauman et al. Nat Chem Biol. 2022](#).

- To keep the species information from the BLASTP file like seen in the figure, recall that NaPDoS2 will keep the ID of the sequence up until the first space in the sequence description of the FASTA file, so you can replace any spaces with underscores ("_") to keep the appended information. However, to build trees, all characters like ([ ] : ; ( ) etc) must be removed (we suggest using find and replace to take them out).

- We would ***interpret*** this tree as: the *salC* KS domain clades separately from other type 1 *cis*-AT hybrid KSs, and this follows what was discovered in [Bauman et al. Nat Chem Biol. 2022](#) that it performs an unique function unlike other hybrid KSs. The fact that it, with other closely related homologs, forms a clade that is separate from other hybrid KSs indicates that the *salC* class of KSs likely represent a novel functional class of KS. Robust sequence alignments and phylogenetic tree calculations can be used to further explore these hypotheses.

Genome tutorial B objective

Analyze a single bacterial genome Salinispora arenicola CNS-205 to uncover evidence of polycyclic tetramate macrolactam (PTM) KS domains and thus PTM biosynthetic potential, as reported in the NaPDoS2 publication (Supplementary Table S5).

1. **Gather the *Salinispora arenicola* CNS-205 genome sequence**

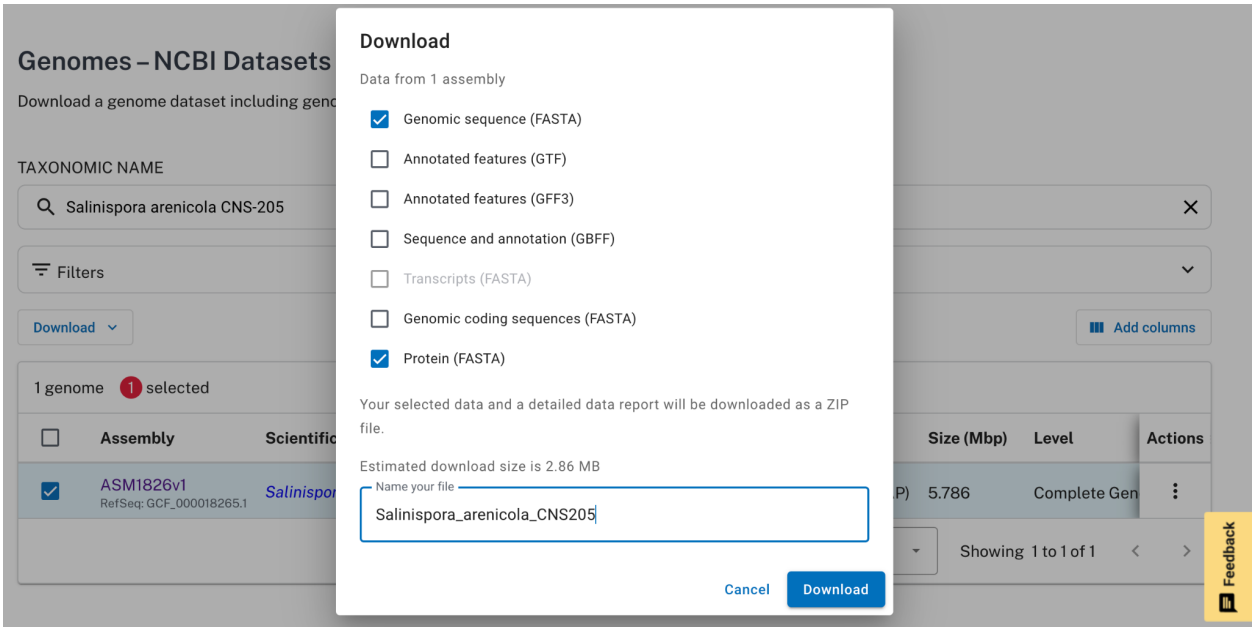
- We'll download the FASTA-formatted genome sequence of *Salinispora arenicola* CNS-205 using the NCBI Genomes Dataset tool.

    - We have also provided the genome sequences used in the tutorial on the NaPDoS2 Tutorial OSF page that you can download directly from there

    - On the NCBI Genomes Dataset tool, search "**CNS-205**" and select "***Salinispora arenicola* CNS-205**".

Bacterial genome	**Accession**	**Reference**
Salinispora arenicola CNS-205	GCF_000018265.1	Millán-Aguiñaga et al. 2017. Sci Rep.

    - Check the box of the first result (assembly "ASM1826v1") and click "**Download dataset**". Select both "**Genomic sequence (FASTA)**" and "**Protein (FASTA)**"-- you can analyze either the genomic (nucleic acid) or protein (amino acid) FASTA genomes sequences with NaPDoS2.



- Save the downloaded zipped file in a place you can access it (for example, your Desktop or Downloads folder). Navigate to where you saved it and unzip the file
  - (Windows: double-click, or select it and click "Extract all"; Mac: double-click; Linux/Command line: navigate to the directory containing your zipped file and use the command "unzip filename.zip")

- The *Salinispora arenicola* CNS-205 genome sequences are found in the folder "GCF_000018265.1" which can be found by navigating through the folder path: `~/ncbi_dataset/ncbi_dataset/data/GCF_000018265.1`
  - "GCF_000018265.1_ASM1826v1_genomic.fna": nucleic acid FASTA genome file
  - "protein.faa": protein FASTA genome file

2. Submit a NaPDoS2 analysis of the single bacterial genome *Salinispora arenicola* CNS-205

- Navigate to the "Run Analysis" tab on NaPDoS2. Select "Domain Type: KS domains".

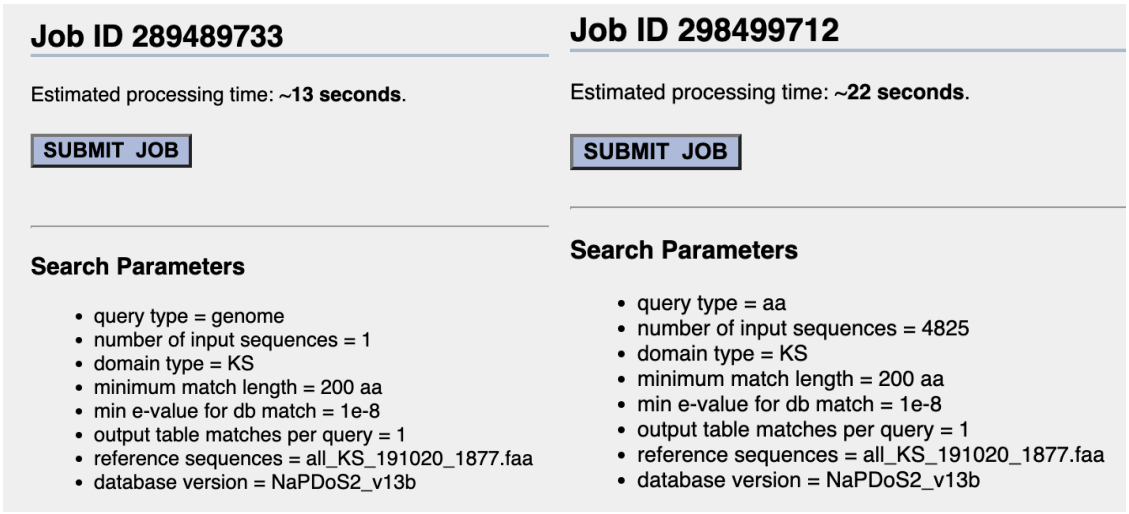
- If you are analyzing the *Salinispora arenicola* CNS-205 "protein.faa" protein FASTA genome file, select "**Query type: Amino acid sequences**". If you are analyzing the *Salinispora arenicola* CNS-205 "GCF_000018265.1_ASM1826v1_genomic.fna": nucleic acid FASTA genome file, select "**Query type: Nucleic acid sequences**".

- Click "**Choose File**" to upload your genome file. Be sure to select the correct one depending if you selected "Query type" amino acid or nucleic acid sequences.
    - ** For this tutorial, we recommend that you analyze the nucleic acid genome sequence as this will allow you to use the start and end coordinates of the PTM-type KS domains to find the location of the PTM-type BGC in the *Salinispora arenicola* CNS-205 genome. There are other ways to map back the location of KS domains in a protein genome (BLASTP, manual comparison to BGC-prediction software like antiSMASH, etc) but we will show the example with the nucleic acid genome query.

- Click on "**Advanced Settings**". Here, you can change the BLASTP e-value cutoff and the Minimum alignment length. For this tutorial, we will use the default settings of BLASTP e-value cutoff of 1e-8 and Minimum alignment length of 200aa (600nt).
        - Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details.

- Click **SEEK** to be assigned a job ID and review the job analysis settings.

Submitted job ID screens for *Salinispora arenicola* CNS-205 genome nucleic acid (**left**, recommended) and amino acid protein (**right**) FASTA files:



Job ID 289489733

Estimated processing time: ~**13 seconds**.

SUBMIT JOB

Search Parameters

- query type = genome
- number of input sequences = 1
- domain type = KS
- minimum match length = 200 aa
- min e-value for db match = 1e-8
- output table matches per query = 1
- reference sequences = all_KS_191020_1877.faa
- database version = NaPDoS2_v13b

Job ID 298499712

Estimated processing time: ~**22 seconds**.

SUBMIT JOB

Search Parameters

- query type = aa
- number of input sequences = 4825
- domain type = KS
- minimum match length = 200 aa
- min e-value for db match = 1e-8
- output table matches per query = 1
- reference sequences = all_KS_191020_1877.faa
- database version = NaPDoS2_v13b

- For "Query type: nucleic acid sequences", you will first proceed to the "**Nucleotide Sequence Search Results**" page which lists the initial number of KS domains that NaPDoS2 identified. On this page, you can **DOWNLOAD** (click link to view; or right-click link & select "Save Link As" to save) a table of the nucleotide match locations in tab-delimited format.
        - This lists the domains numbered KS# ("domain id"), the sequence/contig name the domain was found on ("parents seq"), the base pair/nucleotide

start and end location of the domain (“start” , “end”), the translation reading frame (“read frame”), and the final name of the domain (“cand_id”).

- This table can be used to find the location of the KS domains in your nucleic acid genome file via the starting and ending base pairs.

- To continue the analysis, click “**CONTINUE ANALYSIS**”

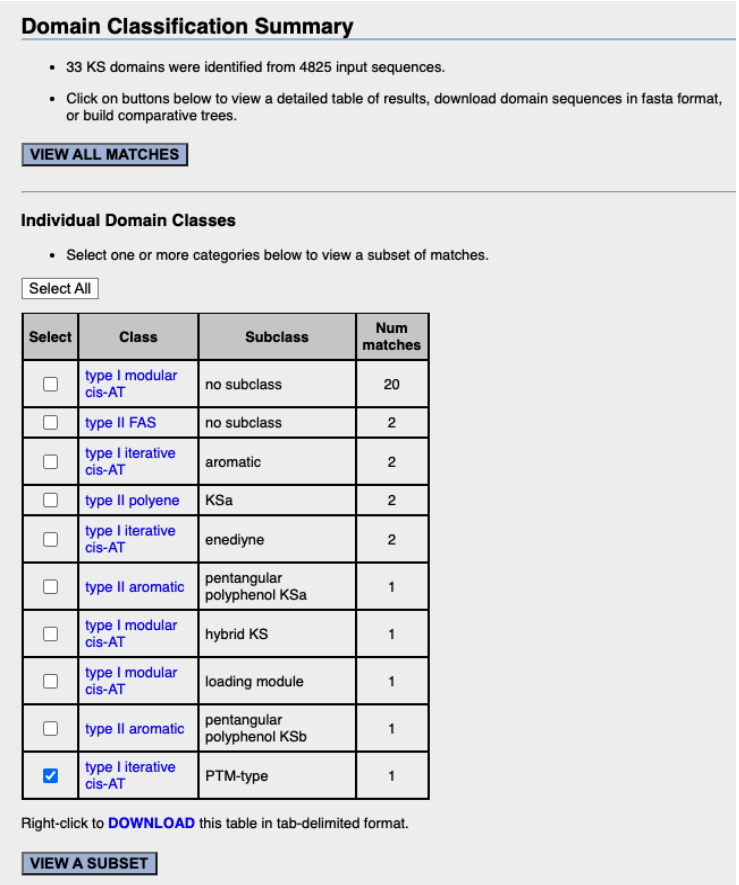
3. Select results display options for KSs that NaPDoS2 detected and classified

- The “**Domain Classification Summary**” page lists the total number of domains found from the number of sequences that were input. Clicking “**VIEW ALL MATCHES**” opens a new page with the hit table of all domains identified in the genome analysis.

    - In our example, 33 KS domains were found from the 1 nucleic acid *Salinispora arenicola* CNS-205 and 33 KS domains were found from 4,825 genome amino acid proteins.

- The “**Individual Domain Classes**” lists the number of specific classes of KS domains that were found in your query genome sequence, ordered by the class and subclass with the most number of domains.

    - For this tutorial, we are interested in the type 1 iterative *cis*-AT PTM-type KS domains! Select the specific “**PTM-type**” subclass and click “**VIEW A SUBSET**”, which will open a new page with our PTM-type KSs that NaPDoS2 identified in the *Salinispora arenicola* CNS-205 genome.



Domain Classification Summary

- 33 KS domains were identified from 4825 input sequences.
- Click on buttons below to view a detailed table of results, download domain sequences in fasta format, or build comparative trees.

VIEW ALL MATCHES

Individual Domain Classes

- Select one or more categories below to view a subset of matches.

Select All

Select	Class	Subclass	Num matches
☐	type I modular cis-AT	no subclass	20
☐	type II FAS	no subclass	2
☐	type I iterative cis-AT	aromatic	2
☐	type II polyene	KSa	2
☐	type I iterative cis-AT	enediyne	2
☐	type II aromatic	pentangular polyphenol KSa	1
☐	type I modular cis-AT	hybrid KS	1
☐	type I modular cis-AT	loading module	1
☐	type II aromatic	pentangular polyphenol KSb	1
☑	type I iterative cis-AT	PTM-type	1

Right-click to **DOWNLOAD** this table in tab-delimited format.

VIEW A SUBSET

4. View & download NaPDoS2 search results of a specific class of interest

- Since we clicked "**VIEW A SUBSET**", you will view the "**Database Search Results**" page, for a subset of hits of a certain class/subclass that you selected– the PTM-type KSs.

- The table can be sorted by each of the column values by clicking on the column header; the table scrolls up and down and if your query hit names are long, the table scrolls left to right. You can **DOWNLOAD** the results table (click link to view; or right-click link & select "Save Link As" to save the table) in a tab-delimited format.

- The scrollable table lists the following information:

    - **cand_id**: name of the domain hit, which is the name of the sequence in the query FASTA file before spaces and appended with location information of where the domain hit was found (see below for details)

    - **database match**: the name of the domain in the NaPDoS2 database that the query hit is closest to
        - **names include**: BGCname_KS#_uniquetag
        - **BGC name**: name of the BGC

- **KS#**: number KS in successive order from the BGC
- **unique tag**: abbreviation that is unique for every class/subclassification

- **percent identity:** percent identity that the query hit shares with the closets NaPDoS2 database match

- **align length:** alignment length that the query hit shares with the closest NaPDoS2 database match

- **e-value:** e-value of the query hit to the closest NaPDoS2 database match

- **BGC match**: name of the BGC of the closest NaPDoS2 database match to the query hit; click on the link to open a new page of the BGC card information

- **domain class:** classification of the closest NaPDoS2 database match to the query hit
  - This is the ***classification*** of your query hit

- **domain subclass:** specific subclass of the closest NaPDoS2 database match to the query hit
  - This is the ***complete specific subclass*** of your query hit (some classes do not have subclasses, ie “no subclass”)

Database Search Results

Results below are for 1 KS domain sequences, from 1 different categories.

Use check boxes to select candidates for **further analysis** below. (options may take a few seconds to load for large match tables).

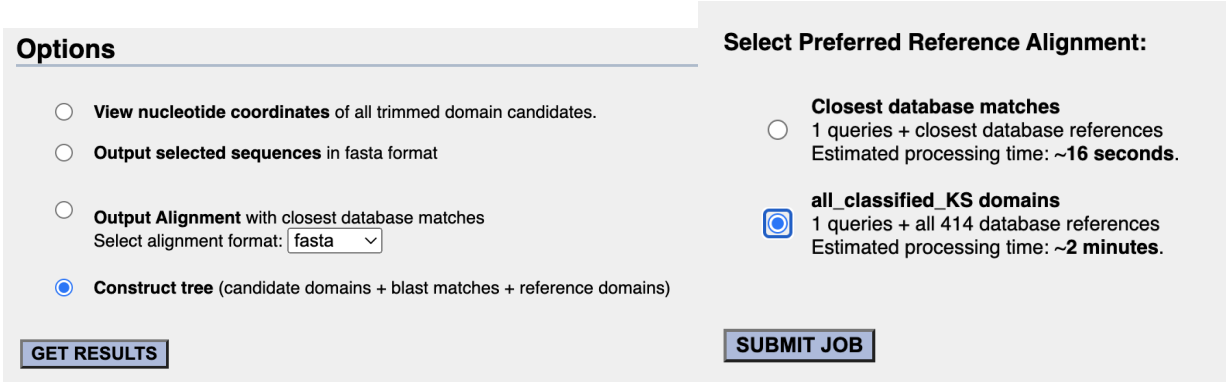
Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Select All

	cand_id	database match	percent identity	align length	e-value	BGC product match	domain class	domain subclass
☑	NC_009953.1_2_939_12_440	clifednamide_KS01_iPKSPTM	83	429	1.3e-218	clifednamide	type I iterative cis-AT	PTM-type

Right-click to DOWNLOAD this table in tab-delimited format.

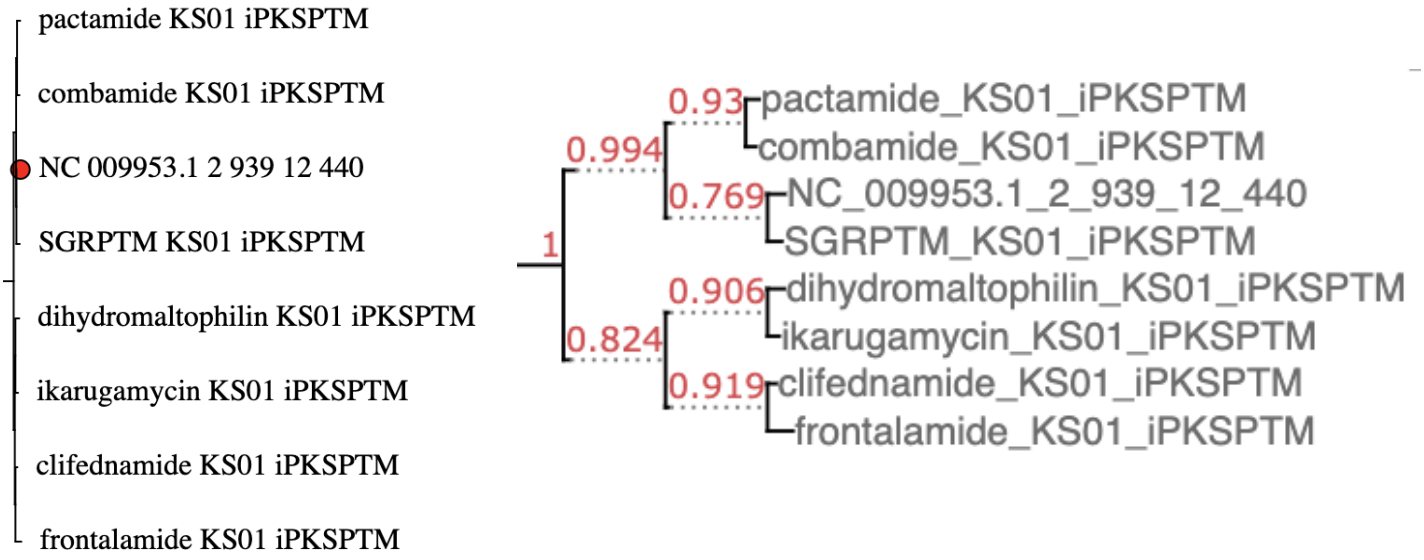
- Now that we have selected the specific type I iterative *cis*-AT PTM-type KS domain that NaPDoS2 found in the *Salinispora arenicola* CNS-205, we will analyze this specific domain hits further by 1) first building a tree of this hit in context with other characterized PTM-type KSs.
    - **Construct tree** (candidate domains + blast matches + reference domains)
        - Select the query domain hits that you would like to build a tree with. You can either build a tree with:
            - **Closest database matches** to your selected sequences, ie. # of selected queries + closest database references
            - **All classified KS domains** which would be the # of selected queries + all 414 NaPDoS2 database reference sequences.
        - The estimated time to calculate the trees is given, usually the time to complete the trees is much faster than estimated.
    - We will select "**all classified KS domains**" which will put our PTM-type KS in the context of all 414 NaPDoS2 database reference sequences.



- The **Tree results** page has two options for viewing the calculated trees:
    - Tree image (SVG format file), click **DOWNLOAD** to view, right-click to save.
        - When viewing the tree, the query domain hits will have red circles of where they are placed in the tree.
    - Newick format (plain text file), click **DOWNLOAD** to view, right-click to save
        - This file format can be viewed in tree viewing programs like FigTree, iTol, TreeViewer, ETE Toolkit, etc.
- We can see that our *Salinispora arenicola* CNS-205 PTM-like KS domain falls in a clade of other type 1 iterative *cis*-AT PMT-type KSs. While the top database hit from our initial

query search was “clifednamide KS01 iPKSPTM”, we can see that the taxa our PTM-type KS query hit clades most closely with a KS from the SGR810-815 BGC.

- We have also provided the *Salinispora arenicola* CNS-205 iPKSPTM KS query hit phylogenetic tree on the NaPDoS2 Tutorial OSF page that you can download directly from there



*Screenshot of the NaPDoS2 tree of our Salinispora arenicola CNS-205 PTM-like KS (left); expanded clade of the “iPKSPTM” clade viewed in more detail (right). Our Salinispora arenicola CNS-205 PTM-like KS query hit is “**NC_009953.1**”*

- We recommend downloading the sequences and alignments to complete robust phylogenetic analyses when reporting the closest cladding domains, ie. run robust alignments, compare amino acid active sites of the KSs, determine tree models, and calculate multiple phylogenetic trees using various models. Check out these references for an introduction to phylogenetics:
- Applied evolution: phylogeny-based approaches in natural products research (2019)
- Phylogenetics basics and best practices, for natural products; Chapter 8 (2012)

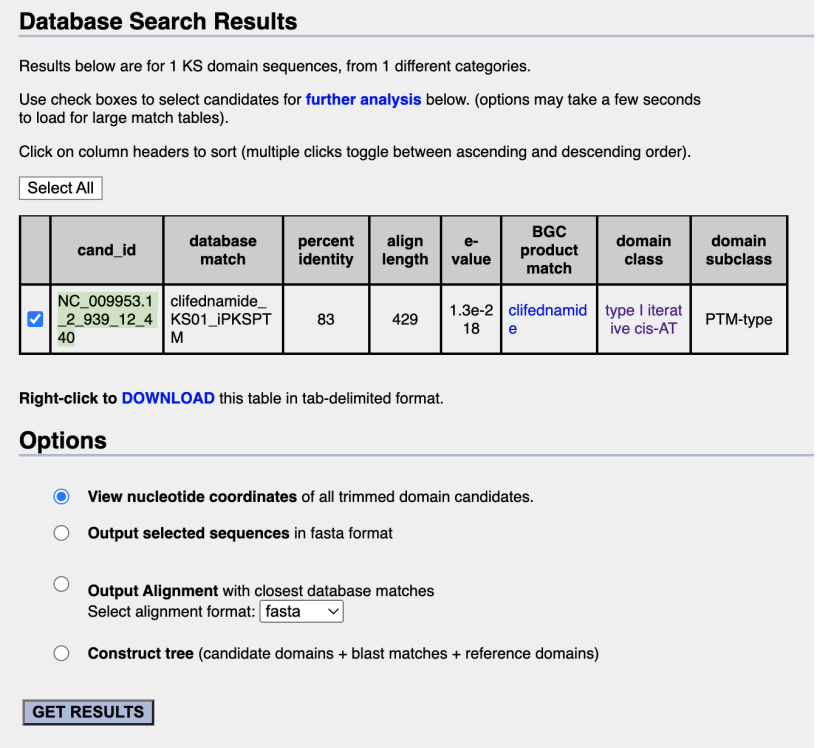
- We see that our query hit is similar yet sufficiently different from other PTM-type KSs that it would be useful for us to evaluate the broader BGC & gene neighborhood that this KS is part of to see if it is a new PTM BGC or potential PTM product.

5. Locating the KS-containing BGC by using our NaPDoS2 query hit as a search hook

- The goal of this tutorial is to analyze the *Salinispora arenicola CNS-205* bacterial genome to uncover evidence of polycyclic tetramate macrolactams (PTM) KS domains and thus PTM biosynthetic potential, which now we will further explore by finding the PTM-type KS containing BGC back in the genome.

- To go back to the genome, we will output the **nucleotide coordinates** of the trimmed domain query hits, which will give us a location range to compare BGCs in the same area. First, go back in your browser to the "**Database search results**" page and select the PTM-type KS and select "**View nucleotide coordinates** of all trimmed domain candidates" and click "**GET RESULTS**".

    - Note the "**cand_id**" specifically for the PTM-type KS query hit– we'll need this to select the correct nucleotide coordinate range. `NC_009953.1_2_939_12_440`



Database Search Results

Results below are for 1 KS domain sequences, from 1 different categories.

Use check boxes to select candidates for **further analysis** below. (options may take a few seconds to load for large match tables).

Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Select All

	cand_id	database match	percent identity	align length	e-value	BGC product match	domain class	domain subclass
☑	NC_009953.1_2_939_12_440	clifednamide_KS01_iPKSPTM	83	429	1.3e-218	clifednamide	type I iterative cis-AT	PTM-type

Right-click to DOWNLOAD this table in tab-delimited format.

Options

- ◉ **View nucleotide coordinates** of all trimmed domain candidates.
- ○ **Output selected sequences** in fasta format
- ○ **Output Alignment** with closest database matches
  Select alignment format: fasta
- ○ **Construct tree** (candidate domains + blast matches + reference domains)

GET RESULTS

- We see that NaPDoS2 found 33 KS domains in our *Salinispora arenicola* CNS-205 genome, however, our PTM-type KS is KS16. We can find that by searching the page for our **cand_id** "`NC_009953.1_2_939_12_440`".

    - The start coordinates and end coordinates of our PTM-type KS are: 2,762,759 to 2,764,045.
    - We will now find out what BGC or genes are predicted to contain the PTM-type KS.

- **View nucleotide coordinates** for all trimmed domain candidates
    - **only an option if your query sequence was a nucleic acid file
    - This shows a table where each potential domain candidate has been given a unique candidate ID number, based on parent sequence id, reading frame number (1-6), gene number within the reading frame, and trimmed nucleotide start and stop coordinates within the reading frame.
    - The table columns are sortable by each column header. You can **DOWNLOAD** the table of nucleotide match locations in tab-delimited format(click link to view; or right-click link & select "Save Link As" to save)
    - Or, for this tutorial, to find the coordinates of a specific "**cand_id**", control/command+F /search the webpage for the specific candidate ID.

Nucleotide Sequence Search Results

- 33 KS domains were detected from 1 input sequences.

Domain Locations

domain id	parent seq	start	end	read frame	cand_id
KS16	NC_009953.1	2762759	2764045	2	NC_009953.1_2_939_12_440
KS17	NC_009953.1	3050659	3049451	5	NC_009953.1_5_908_165_567
KS18	NC_009953.1	3053851	3052805	5	NC_009953.1_5_906_13_361
KS19	NC_009953.1	3058360	3057146	5	NC_009953.1_5_904_50_454
KS20	NC_009953.1	3059634	3058387	4	NC_009953.1_4_850_18_433
KS21	NC_009953.1	3595733	3594456	6	NC_009953.1_6_723_40_465
KS22	NC_009953.1	3600439	3599156	5	NC_009953.1_5_736_36_463
KS23	NC_009953.1	3606086	3604806	6	NC_009953.1_6_718_1854_2280
KS24	NC_009953.1	3611021	3609765	6	NC_009953.1_6_718_209_627
KS25	NC_009953.1	3617254	3616010	5	NC_009953.1_5_734_52_466
KS26	NC_009953.1	3623241	3621970	4	NC_009953.1_4_685_5811_6234
KS27	NC_009953.1	3627504	3626230	4	NC_009953.1_4_685_4390_4814
KS28	NC_009953.1	3632658	3631390	4	NC_009953.1_4_685_2672_3094
KS29	NC_009953.1	3637689	3636409	4	NC_009953.1_4_685_995_1421
KS30	NC_009953.1	3640617	3639379	4	NC_009953.1_4_685_19_431
KS31	NC_009953.1	3784500	3785780	3	NC_009953.1_3_1264_724_1150
KS32	NC_009953.1	4179625	4180824	1	NC_009953.1_1_1429_37_436

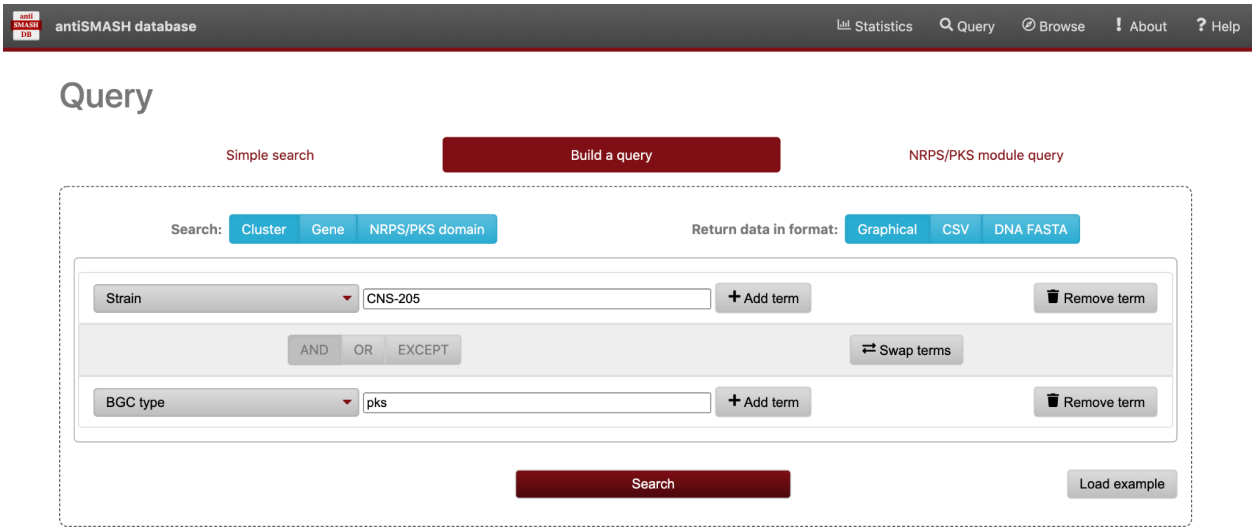
Note: Each potential domain candidate has been given a unique candidate ID number, based on parent sequence id, reading frame number (1-6), gene number within the reading frame, and trimmed nucleotide start and stop coordinates within the reading frame.

- Right-click to DOWNLOAD a table of nucleotide match locations in tab-delimited format.

- We will use the BGC prediction tool antiSMASH to analyze our genome to identify the BGCs. We could analyze our nucleic acid or protein *Salinispora arenicola* CNS-205 genome with antiSMASH (**upload file**, **select "all on"** for **"extra features"** options; **relaxed detection strictness** should be okay. Please see the excellent documentation for antiSMASH to learn more about these settings). However, for the focus of this tutorial, it turns out that *Salinispora arenicola* CNS-205 genome has already been run through antiSMASH and the results have been stored in antiSMASH-db.

    - antiSMASH-db "database" is a repository of "precomputed antiSMASH results and advanced queries". We can browse the database by phylogeny, or BGC type/specialized metabolite type, or query by building a BGC query using specific domains. See the reference for antiSMASH-db.

- Let's open the antiSMASH analysis of *Salinispora arenicola* CNS-205 by going to antiSMASH-db and building a query to find our 1) strain 2) BGC in the region of our PTM-type KS.
    - Go to **Query**, and then click "**Build a query**".
    - First, click "**Select a category**" and select **Strain**
    - Type "**CNS-205**" and select the CNS-205 option in the box.
    - Click "**Add term**".
    - Next, click "**Select a category**" on that new term line and select "**BGC type**".
    - Type "**pks**" in the box and select "**pks**" as the BGC type.
    - Click "**Search**".



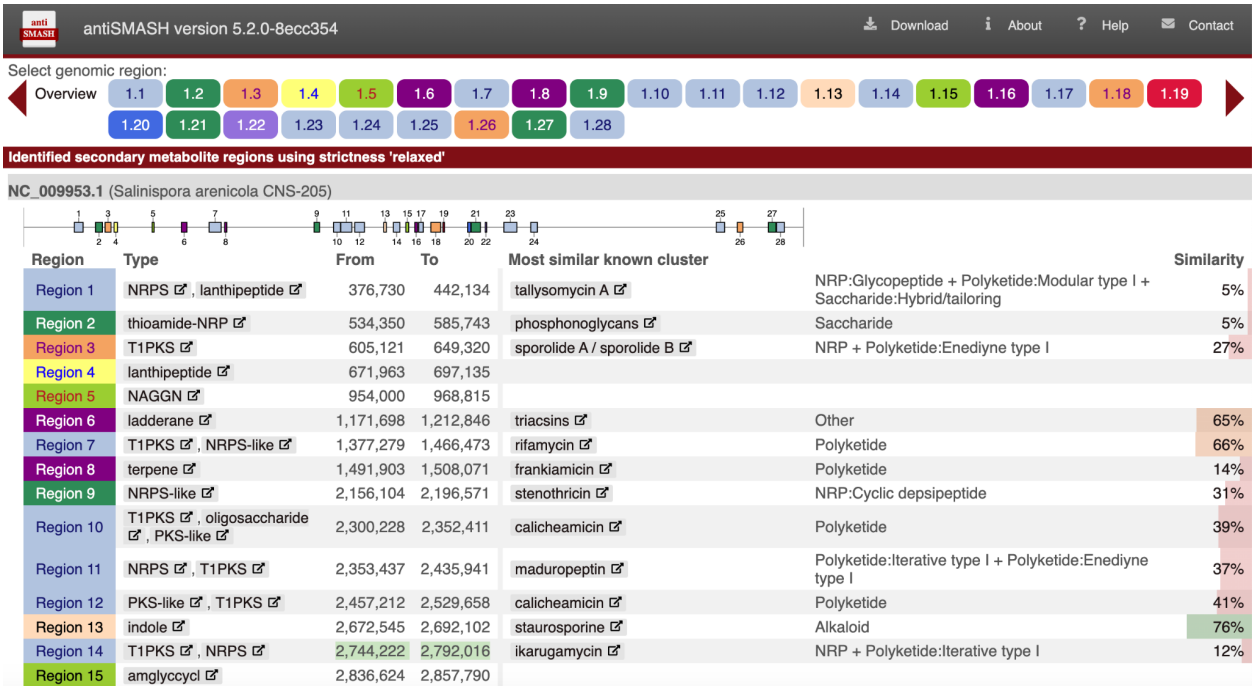
- We see that antiSMASH-db retrieved 11 PKS type BGCs from *Salinispora arenicola* CNS-205 genome. Click on the link "*Salinispora arenicola* CNS-205" – this will open up the antiSMASH results page of all BGCs found in the genome when you click "**Overview**" at the top left.

- Link to the *Salinispora arenicola* CNS-205 antiSMASH 5.2.0 overview analysis page
  - We have also provided the zipped antiSMASH-db results from the above link on the NaPDoS2 Tutorial OSF page that you can download directly from there

Your search gave **11** results, showing **1** to **11**.

Species	Region	Type	From	To	Edge	Most similar MIBiG cluser
Salinispora arenicola CNS-205	3	Type I polyketide	605120	649320	No	sporolide A / sporolide B

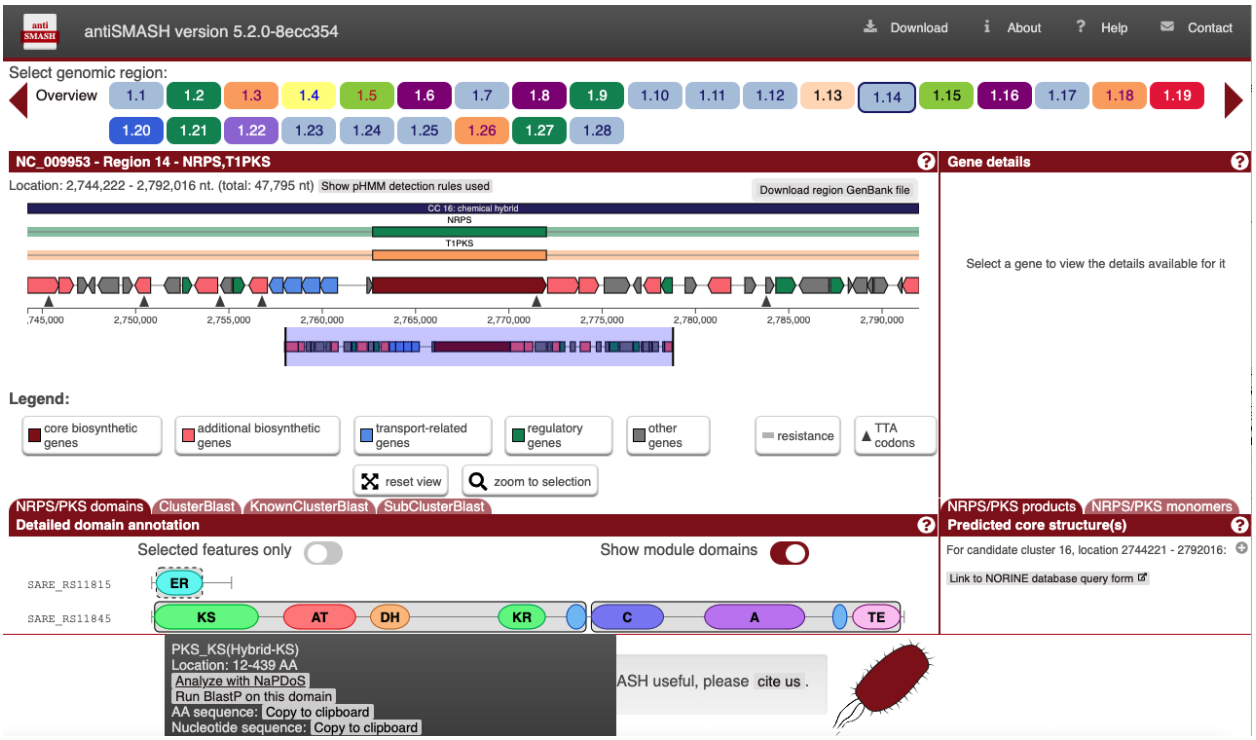
- The start coordinates and end coordinates of our PTM-type KS are: 2,762,759 to 2,764,045. We see that there is a BGC– called **Region 14** that overlaps with that region. It is a type 1 PKS/NRPS, which matches with the type 1 iterative *cis*-AT PTM-type KS from NaPDoS2! Click on **Region 14** to open a new page to look at the BGC in detail.
  - https://antismash-db.secondarymetabolites.org/output/GCF_000018265.1/index.html#r1c14



Region	Type	From	To	Most similar known cluster		Similarity
Region 1	NRPS, lanthipeptide	376,730	442,134	tallysomycin A	NRP:Glycopeptide + Polyketide:Modular type I + Saccharide:Hybrid/tailoring	5%
Region 2	thioamide-NRP	534,350	585,743	phosphonoglycans	Saccharide	5%
Region 3	T1PKS	605,121	649,320	sporolide A / sporolide B	NRP + Polyketide:Enediyne type I	27%
Region 4	lanthipeptide	671,963	697,135			
Region 5	NAGGN	954,000	968,815			
Region 6	ladderane	1,171,698	1,212,846	triacsins	Other	65%
Region 7	T1PKS, NRPS-like	1,377,279	1,466,473	rifamycin	Polyketide	66%
Region 8	terpene	1,491,903	1,508,071	frankiamicin	Polyketide	14%
Region 9	NRPS-like	2,156,104	2,196,571	stenothricin	NRP:Cyclic depsipeptide	31%
Region 10	T1PKS, oligosaccharide, PKS-like	2,300,228	2,352,411	calicheamicin	Polyketide	39%
Region 11	NRPS, T1PKS	2,353,437	2,435,941	maduropeptin	Polyketide:Iterative type I + Polyketide:Enediyne type I	37%
Region 12	PKS-like, T1PKS	2,457,212	2,529,658	calicheamicin	Polyketide	41%
Region 13	indole	2,672,545	2,692,102	staurosporine	Alkaloid	76%
Region 14	T1PKS, NRPS	2,744,222	2,792,016	ikarugamycin	NRP + Polyketide:Iterative type I	12%
Region 15	amglyccycl	2,836,624	2,857,790			

- This Type 1 PKS / NRPS BGC is our putative PTM-like BGC. We can toggle on "**show module**" domains and see that antiSMASH also predicted a KS domain in the key maroon biosynthetic gene. You can click on the "**KS**" domain and click "**Analyze with NaPDoS2**" to immediately run it back through the NaPDoS2 analysis to confirm that the antiSMASH annotation of "Hybrid-KS" actually is a more specific NaPDoS2 classification of type 1 ietrative *cis*-AT PTM-type KS.

- ***Note***: currently the antiSMASH website has not been updated yet to link to NaPDoS2, it instead says "**Analyze with NaPDoS**" and will run your domain through NaPDoS version 1, which will give different results than reported below. We hope to change the link soon! :)

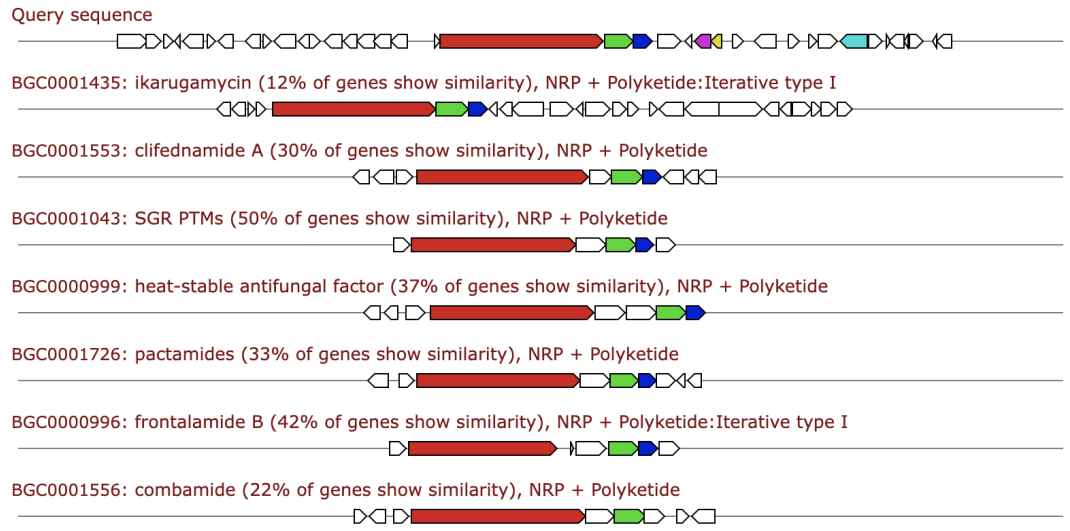


- Our ***interpretation*** of our results: This does look to be our target BGC as the key biosynthetic gene contains our PTM-type KS. We can click the "Known Cluster Blast" and we see that the closest top hits of this entire BGC are the other PTM BGCs, similar to what NaPDoS2 had predicted! However, this *Salinispora arenicola* CNS-205 BGC looks a little different from the known PTM BGCs, and thus could be new PTM-type biosynthetic chemical diversity, exciting!

NRPS/PKS domains | ClusterBlast | KnownClusterBlast | SubClusterBlast

Similar known gene clusters

All hits | Download graphic



2. Analyzing multiple genomes with NaPDoS2: analyze multiple fungal genomes

Objective: Analyze three eukaryotic fungal genomes to identify & classify KSs of interest, including KS types that are enriched in fungi, and compare the distribution of KS type in 3 different fungal species.

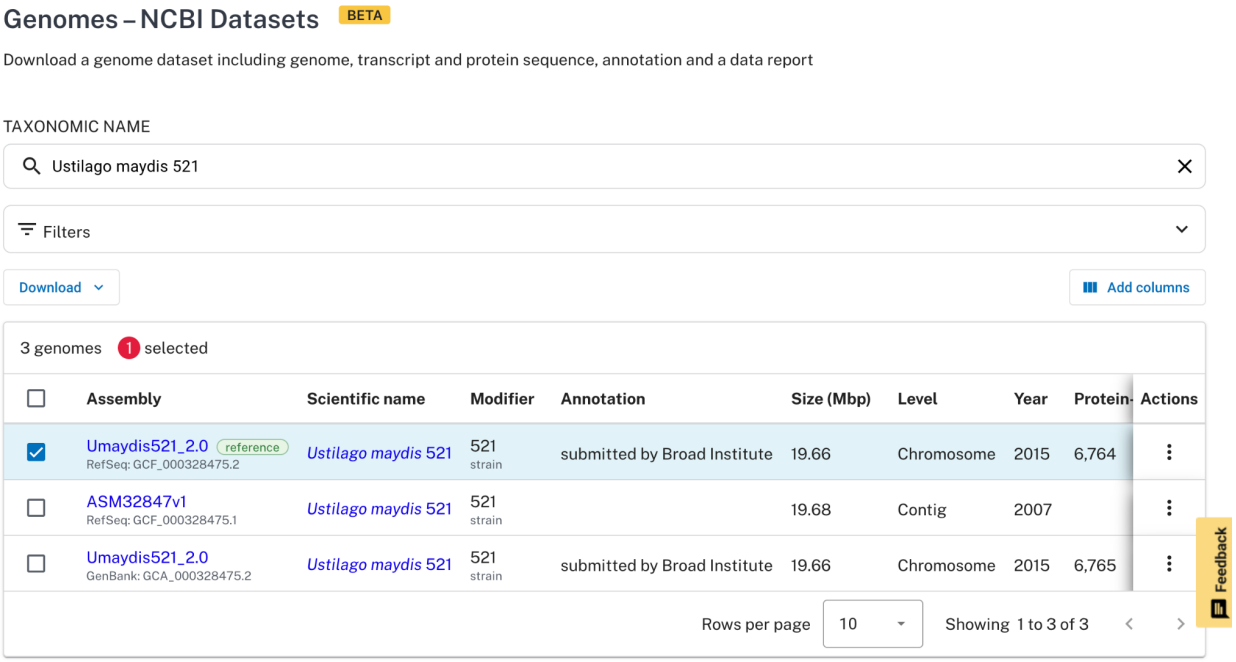
1. **Gather the three fungal genome sequence**

    - We'll download the FASTA-formatted genome sequence of three fungal genomes using the NCBI Genomes Dataset tool. These accession numbers were copied from ***Supplementary Table S3*** from the NaPDoS2 publication.

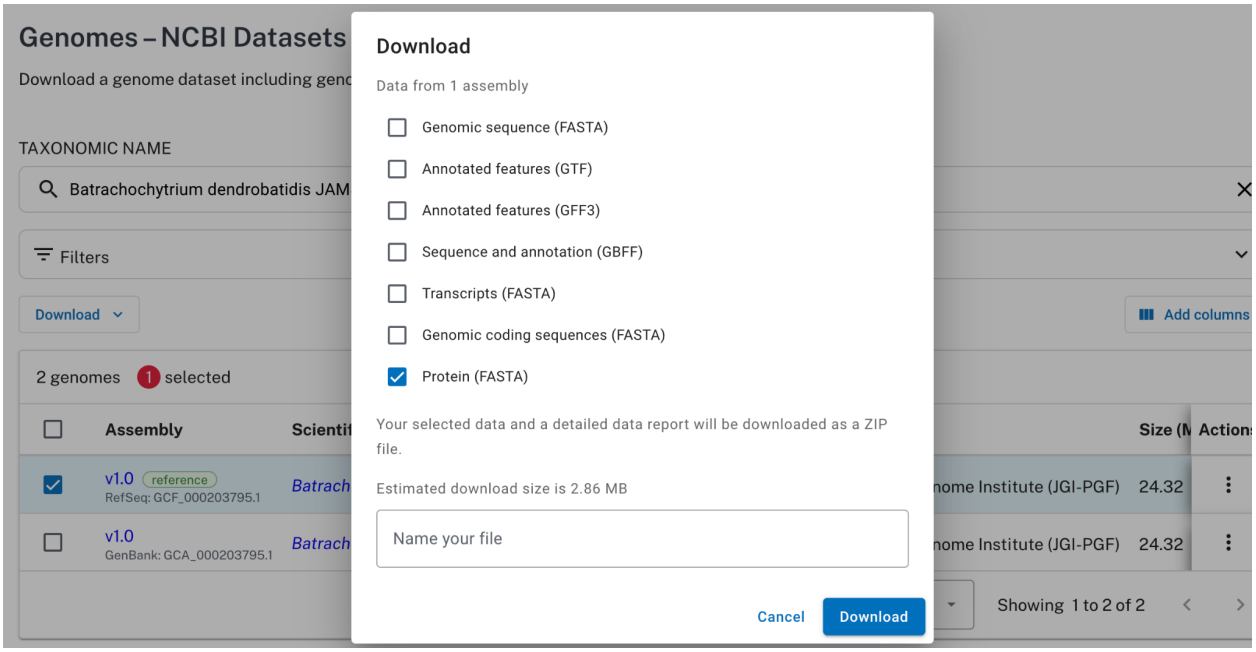
        - We have also provided the genome sequences used in the tutorial on the NaPDoS2 Tutorial OSF page that you can download directly from there.

Fungal genome	**Accession**	**Reference**
Aspergillus niger ATCC 1015 (JGI)	GCA_000230395.2	Almeida H. et al. 2019 IEEE BIBM.
Ustilago maydis 521 (smut fungi)	GCF_000328475.2	Almeida H. et al. 2019 IEEE BIBM.
Batrachochytrium dendrobatidis JAM81 (chytrids)	GCF_000203795.1	Almeida H. et al. 2019 IEEE BIBM.

        - On the NCBI Genomes Dataset tool, search each of the species listed below and select each with the accessions that match the table above
            - "*Aspergillus niger* ATCC 1015"
            - "*Ustilago maydis* 521"
            - "*Batrachochytrium dendrobatidis* JAM81"



- Check the box of each result and click "**Download dataset**". Select only "**Protein (FASTA)**"-- for Eukaryotes, NaPDoS2 analyses should only be performed with predicted proteins, translated coding sequences, or transcriptomes as NaPDoS2 cannot excise genomic introns associated with eukaryotic sequence data.



- Save the downloaded zipped file in a place you can access it (for example, your Desktop or Downloads folder). Navigate to where you saved it and unzip the file
    - (Windows: double-click, or select it and click "Extract all"; Mac: double-click; Linux/Command line: navigate to the directory containing your zipped file and use the command "`unzip filename.zip`")

- Each of the fungal genomes can be found in the folder within each folder labeled with the accession number, with the follow pat: `~/ncbi_dataset/ncbi_dataset/data/accession`
    - "protein.faa": protein FASTA genome file

- Rename each of the "protein.faa" files to something that is unique to each fungal genome, using no spaces:
    - *Aspergillus_niger*
    - U*stilago_maydis*
    - *Batrachochytrium_dendrobatidi*s

- Copy each of these protein genome files into one folder that is easily accessible using the command line (for example, a folder in your Desktop, Downloads; somewhere you can figure out the full path of the folder).

2. Combine multiple genome files into 1 file for a single NaPDoS2 analysis

- While we could analyze each of these 3 fungal genomes with NaPDoS2 one at a time, here is a suggestion we use when analyzing lots of genome sequences: concentrate your genome files into 1 file. Things to keep in mind if you do this:
    - All your genomes will be combined together, so check the labels of the FASTA sequences of your genomes (or rename the headers or add specific named headers to the beginning of your FASTA files (ie. find and replace ">" with ">strain_" to append a name, do not remove the ">" of the FASTAs!) before you combine them. For example:

        - *Aspergillus_niger:* 10,950 genes; each starts with ">EHA"
        - U*stilago_maydis*: 6,782 genes; each starts with ">XP_011"
        - *Batrachochytrium_dendrobatidi*s: 8,677 genes; each starts with ">XP_00

    - NaPDoS2 has a job limit of 500,000 sequences or 500MB file size. If your combined file is too large, we provide easy-to-use-scripts to help split your file into smaller files. See section below.

- Using the command line in Unix or Linux (including Mac) systems (Terminal), navigate to the folder with the three genomes.
    - List the files in the folder:
        - `ls -1`

```
Aspergillus_niger.faa*
Batrachochytrium_dendrobatidis.faa*
Ustilago_maydis.faa*
```

    - Count the number of sequences in each protein FASTA (*.faa) file:
        - `grep -c "^>" *.faa`

```
Aspergillus_niger.faa:10950
Batrachochytrium_dendrobatidis.faa:8677
Ustilago_maydis.faa:6782
```

    - Combine the files (either one below will work):
        - `awk 1 *.faa > 3_fungal_genomes.faa`
            - (ensures a new line for each sequence), or:
        - `cat *.faa > 3_fungal_genomes.faa`
    - Count the total number of sequences in the combined FASTA file:
        - `grep -c "^>" *.faa`

```
3_fungal_genomes.faa:26409
```

    - For Windows systems via command prompt, navigate to the folder (the folder you are in should only contain the required files you want to combine!) and use
        - `copy *.faa 3_fungal_genomes.faa`
            - "copy *.faa [space] new_file.faa"
    - Webtool option: Not recommended for large amounts of sequences, but there is a webtool that will combine FASTA sequences into one (linked here).

- Now that we have a single file with 26,409 sequences, we can analyze all 3 fungal genomes at the same time with NaPDoS2.
    - We have provided this combined file of the 3 fungal genome sequences with the files associated with this tutorial on the NaPDoS2 Tutorial OSF page.

3. Submit a NaPDoS2 analysis of the 3 combined fungal genomes

- Navigate to the "Run Analysis" tab on NaPDoS2. Select "**Domain Type: KS domains**".

- Since we are analyzing the protein FASTA genome files that have been combined into a single protein amino acid genome file, select "**Query type: Amino acid sequences**".

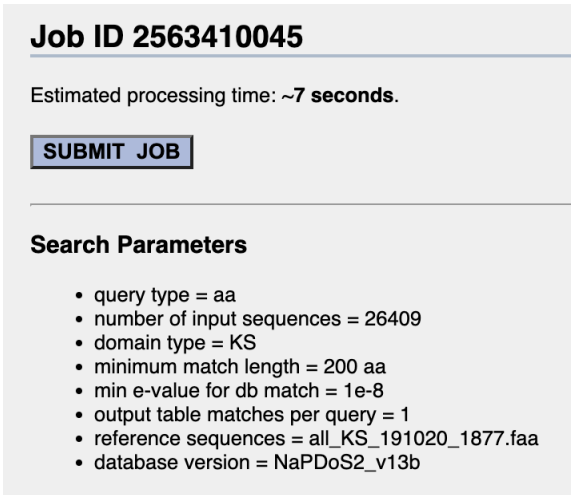
- Click "**Choose File**" to upload your genome file.

- Click on "**Advanced Settings**". Here, you can change the BLASTP e-value cutoff and the Minimum alignment length. For this tutorial, we will use the default settings of BLASTP e-value cutoff of 1e-8 and Minimum alignment length of 200aa (600nt).

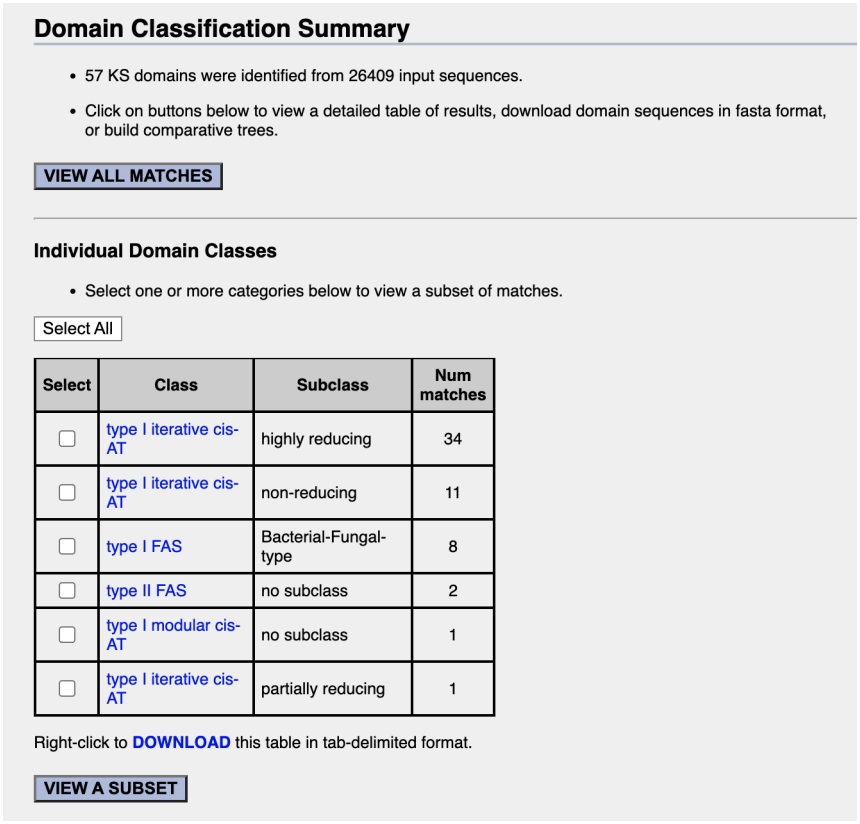
    - Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details.

- Click **SEEK** to be assigned a job ID and review the job analysis settings.

Submitted job ID screens for the three fungal amino acid protein genome FASTA files that were combined into 1:



- We see that from these 3 fungal genomes, NaPDoS2 detected and classified 57 KS domains. Consistent with the results reported in the NaPDoS2 publication, the type 1 iterative *cis*-AT highly reducing KSs were the most common type of KS domain in these fungal genomes.



Domain Classification Summary

- 57 KS domains were identified from 26409 input sequences.
- Click on buttons below to view a detailed table of results, download domain sequences in fasta format, or build comparative trees.

VIEW ALL MATCHES

Individual Domain Classes

- Select one or more categories below to view a subset of matches.

Select All

Select	Class	Subclass	Num matches
☐	type I iterative cis-AT	highly reducing	34
☐	type I iterative cis-AT	non-reducing	11
☐	type I FAS	Bacterial-Fungal-type	8
☐	type II FAS	no subclass	2
☐	type I modular cis-AT	no subclass	1
☐	type I iterative cis-AT	partially reducing	1

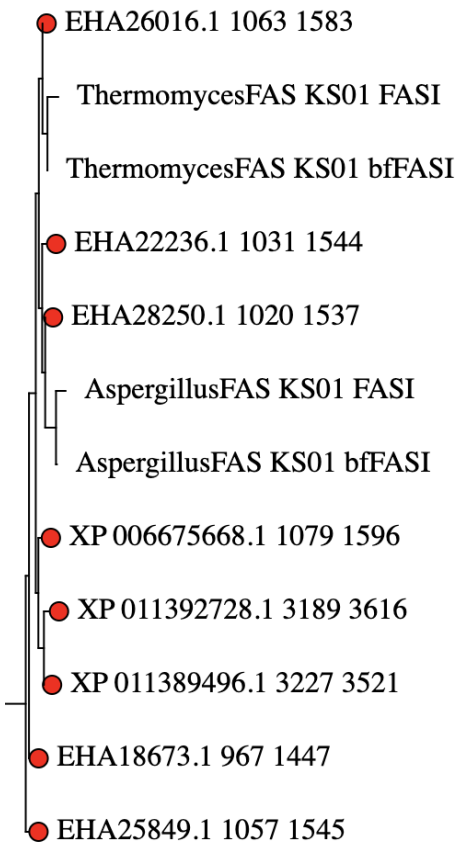
Right-click to **DOWNLOAD** this table in tab-delimited format.

VIEW A SUBSET

- The "**Domain Classification Summary**" page lists the total number of domains found from the number of sequences that were input.

4. Compare NaPDoS2 query hits between the 3 combined fungal genomes

- Clicking "**VIEW ALL MATCHES**" opens a new page with the hit table of all domains identified in the genome analysis.
  - In our example, we click the "cand_id" to sort the column. From our knowledge of what FASTA file header corresponds to each fungal genomes, we can see:
    - *Aspergillus_niger:* ">EHA": the most hits, 50 KSs!
    - U*stilago_maydis*: "XP_011": only 5 hits
    - *Batrachochytrium_dendrobatidis*: ">XP_00": only 2 hits
  - These results match with ***Supplementary* Table S6** from the NaPDoS2 publication where we analyzed 27 fungal genomes. We can further analyze these sequences, by constructing a tree to see if the various KSs between these fungal genomes are similar or different, but we can ***interpret*** these results that fungi each have different types of KS types.



- Our goal of this tutorial was to analyze 3 fungal genomes at the same time with NaPDoS2, and in doing so, we can immediately build a tree with KS domains from each of the 3 fungal genomes. We can see that the type 1 fatty acid synthases from the three fungi ("EHA" vs "XP_011" vs "XP_00") diverge a little bit on the phylogenetic tree.

- **Our recommendation: when working with combined genome files, **DOWNLOAD** the **Database Search results** table as a tab delimited file (*.tab), which then you can import into R, Microsoft Excel, or Google Sheets for further parsing & analysis (ie. splitting the results by sample and counting the number of hits/class of hits per sample).

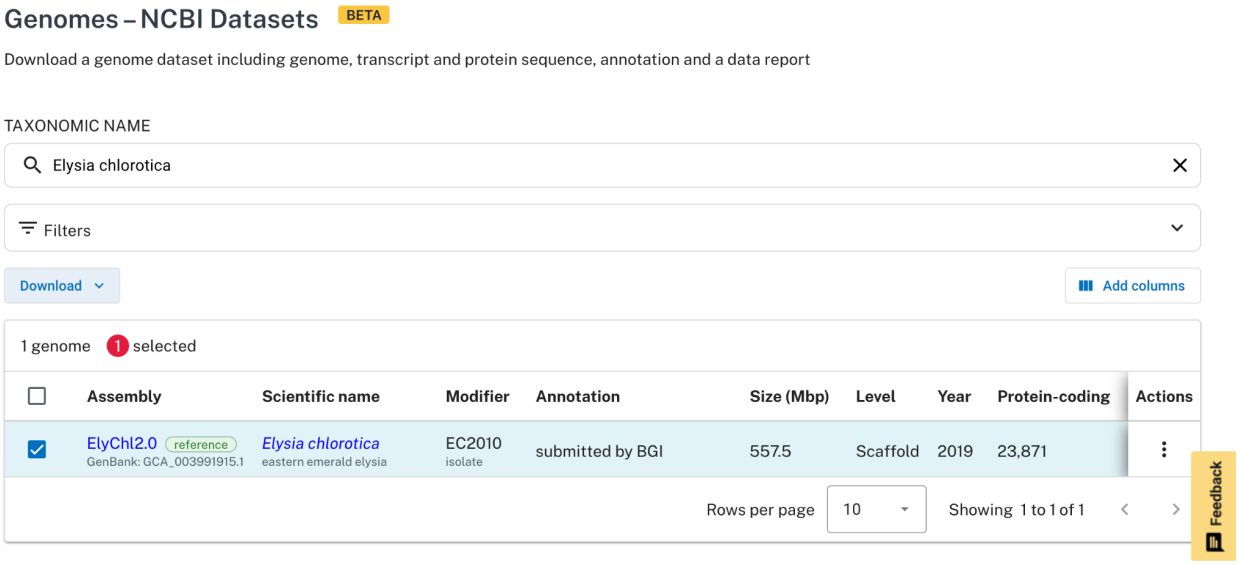
3. Analyzing an animal genome with NaPDoS2: analyze the *Elysia chlorotica* genome

Objective: Analyze an eukaryotic animal genome to identify & classify KSs of interest, including KS types that comprise a new biosynthetic group of FAS-like KSs from the sacoglossan "solar" emerald green sea slug *Elysia chlorotica* as reported in *Torres et al. Nature Comms. 2019*.

1. **Gather the animal genome sequence.**
    - We'll download the FASTA-formatted genome sequence of *Elysia chlorotica* using the NCBI Genomes Dataset tool.
        - We have also provided the genome sequences used in the tutorial on the NaPDoS2 Tutorial OSF page that you can download directly from there
        - On the NCBI Genomes Dataset tool, search & select "***Elysia chlorotica***".

Animal genome	**Accession**	**Reference**
Elysia chlorotica	GCA_003991915.1	Cai H. et al. 2019 Scientific Data.

        - Check the box of each result and click "**Download dataset**". Select only "**Protein (FASTA)**"-- for Eukaryotes, NaPDoS2 analyses should only be performed with predicted proteins, translated coding sequences, or transcriptomes as NaPDoS2 cannot excise genomic introns associated with eukaryotic sequence data.



- Save the downloaded zipped file in a place you can access it (for example, your Desktop or Downloads folder). Navigate to where you saved it and unzip the file
    - (Windows: double-click, or select it and click "Extract all"; Mac: double-click; Linux/Command line: navigate to the directory containing your zipped file and use the command "unzip filename.zip")

- The *Elysia chlorotica* genome sequences are found in the folder "GCA_003991915.1" which can be found by navigating through the folder path: `~/ncbi_dataset/ncbi_dataset/data/GCA_003991915.1`
    - "protein.faa": protein FASTA genome file

2. **Submit a NaPDoS2 analysis of the animal genome *Elysia chlorotica***

- Navigate to the "Run Analysis" tab on NaPDoS2. Select "**Domain Type: KS domains**".

- Because we are analyzing the *Elysia chlorotica* "protein.faa" protein FASTA genome file, select "**Query type: Amino acid sequences**".

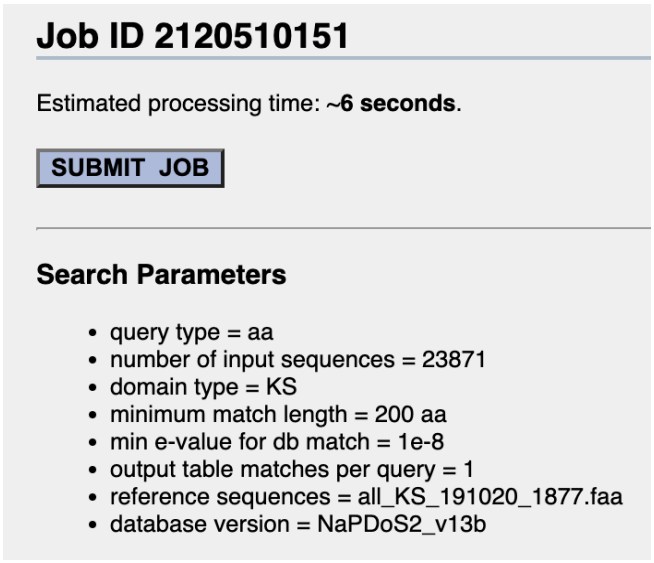
- Click "**Choose File**" to upload your genome file.

- Click on "**Advanced Settings**". Here, you can change the BLASTP e-value cutoff and the Minimum alignment length. For this tutorial, we will use the default settings of BLASTP e-value cutoff of 1e-8 and Minimum alignment length of 200aa (600nt).

    - Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details.

- Click **SEEK** to be assigned a job ID and review the job analysis settings.

Submitted job ID screens for *Elysia chlorotica* amino acid protein FASTA files:



- NaPDoS2 has a job limit of 500,000 sequences or 500MB file size. If your combined file is too large, we provide easy-to-use-scripts to help split your file into smaller files. See section below.

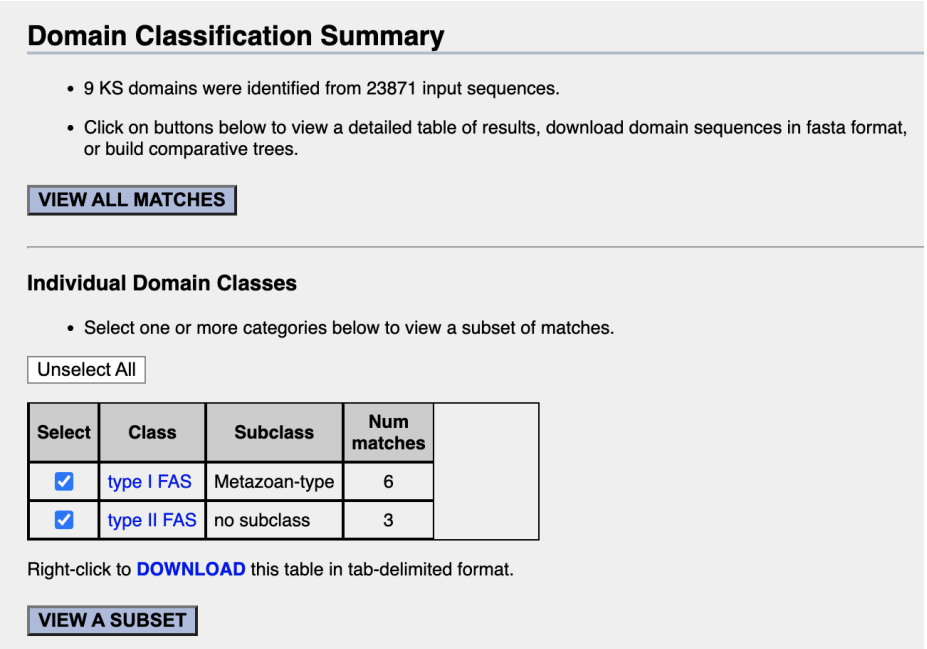
- We can see that the *Elysia chlorotica* amino acid genome has 23,871 sequences, so we do not need to split the file or sequences.

3. Assess and investigate KSs that NaPDoS2 detected and classified in the animal genome

- The goal of this tutorial is to analyze an eukaryotic animal genome to identify & classify KSs of interest, including KS types that comprise a new biosynthetic group of FAS-like KSs from the sacoglossan "solar" emerald green sea slug *Elysia chlorotica* as reported in *Torres et al. Nature Comms. 2019*.

- The "**Domain Classification Summary**" page lists the total number of domains found from the number of sequences that were input. Clicking "**VIEW ALL MATCHES**" opens a new page with the hit table of all domains identified in the genome analysis.

- In our example, 9 KS domains were found in the *Elysia chlorotica* genome out of the 23,871 proteins in the genome. They were all type 1 metazoan-type FASs and type II FASs (FAS = Fatty acid synthase)



Domain Classification Summary

- 9 KS domains were identified from 23871 input sequences.
- Click on buttons below to view a detailed table of results, download domain sequences in fasta format, or build comparative trees.

VIEW ALL MATCHES

Individual Domain Classes

- Select one or more categories below to view a subset of matches.

Unselect All

Select	Class	Subclass	Num matches
☑	type I FAS	Metazoan-type	6
☑	type II FAS	no subclass	3

Right-click to **DOWNLOAD** this table in tab-delimited format.

VIEW A SUBSET

- If you clicked "**VIEW ALL MATCHES**" or "**VIEW A SUBSET**", you will view the "**Database Search Results**" page, either for all hits in your query sequence or for a subset of hits of certain class/subclass that you selected.

- The table can be sorted by each of the column values by clicking on the column header; the table scrolls up and down and if your query hit names are long, the table scrolls left to right. You can **DOWNLOAD** the results table (click link to view; or right-click link & select "Save Link As" to save the table) in a tab-delimited format.

- The scrollable table lists the following information:

    - **cand_id**: name of the domain hit, which is the name of the sequence in the query FASTA file before spaces and appended with location information of where the domain hit was found (see below for details)

    - **database match**: the name of the domain in the NaPDoS2 database that the query hit is closest to
        - **names include**: BGCname_KS#_uniquetag
        - **BGC name**: name of the BGC

        - **KS#**: number KS in successive order from the BGC
        - **unique tag**: abbreviation that is unique for every class/subclassification

    - **percent identity:** percent identity that the query hit shares with the closets NaPDoS2 database match

    - **align length:** alignment length that the query hit shares with the closest NaPDoS2 database match

    - **e-value:** e-value of the query hit to the closest NaPDoS2 database match

    - **BGC match**: name of the BGC of the closest NaPDoS2 database match to the query hit; click on the link to open a new page of the BGC card information

    - **domain class:** classification of the closest NaPDoS2 database match to the query hit
        - This is the ***classification*** of your query hit

    - **domain subclass:** specific subclass of the closest NaPDoS2 database match to the query hit
        - This is the ***complete specific subclass*** of your query hit (some classes do not have subclasses, ie “no subclass”)

- Our next step to learn about these KS domains would be to place these KSs into a phylogenetic context to see if there are any unexpected patterns. Click “**VIEW ALL MATCHES**” so we can look closer at the KS domain query hits.

    - Here, you can also select specific domain hits to download or further analyze (select all to download or analyze all results). You have the follow options for downloading the data:

    - **Output selected sequences** in FASTA format

        - Select which trimmed query domain hits or select all to output in FASTA format

        - **if your query was an amino acid protein, you only have the option to output the query domain hits as amino acid sequences

        - Click **GET TRIMMED SEQS**

- This is a very useful output for you to continue working with the KS or C domains that NaPDoS2 detected & classified. You can copy and paste the trimmed query KS or C domain hits into a new file, a sequencing editing program, etc.

- For interpretation of the appended information to your query search (ie. >ProteinA_13_429 or >GeneB_5_1323_56_472), please see below.

- **Output Alignment** with closest database matches

  - Select alignment format (all alignments are created with MUSCLE but can be output in different formats suitable for many downstream applications):
    - FASTA (.fasta)
    - Msf (.msf)
    - Clustalw (.clw)

  - This outputs an alignment of your selected sequences with the closest NaPDoS2 database matches. Note, the alignment output will include all closest NaPDoS2 database matches to all query hits, not just your selected sequences, but only your selected sequences will be present in the alignment.

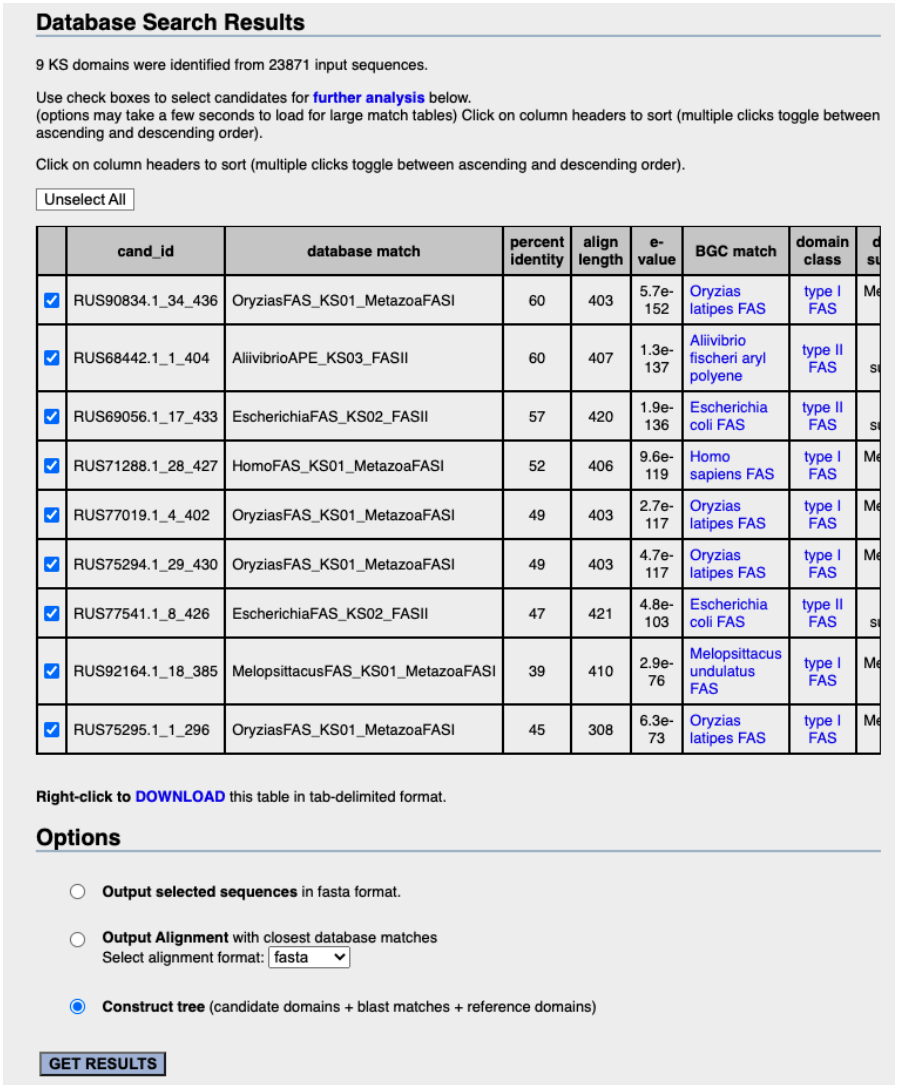
  - You can **DOWNLOAD** (click link to view; or right-click link & select "Save Link As" to save) the alignment file for downstream use in sequence programs, etc.

- **Construct tree** (candidate domains + blast matches + reference domains)

  - Select the query domain hits that you would like to build a tree with. You can either build a tree with:
    - **Closest database matches** to your selected sequences, ie. # of selected queries + closest database references
    - **All classified KS domains** which would be the # of selected queries + all 414 NaPDoS2 database reference sequences.

  - The estimated time to calculate the trees is given, usually the time to complete the trees is much faster than estimated.

  - Details about tree building are continued below for the objective aim of this tutorial to put novel KS sequences into phylogenetic context with known KS domains.



Database Search Results

9 KS domains were identified from 23871 input sequences.

Use check boxes to select candidates for **further analysis** below.
(options may take a few seconds to load for large match tables) Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Unselect All

	cand_id	database match	percent identity	align length	e-value	BGC match	domain class	d su
☑	RUS90834.1_34_436	OryziasFAS_KS01_MetazoaFASI	60	403	5.7e-152	Oryzias latipes FAS	type I FAS	Me
☑	RUS68442.1_1_404	AliivibrioAPE_KS03_FASII	60	407	1.3e-137	Aliivibrio fischeri aryl polyene	type II FAS	su
☑	RUS69056.1_17_433	EscherichiaFAS_KS02_FASII	57	420	1.9e-136	Escherichia coli FAS	type II FAS	su
☑	RUS71288.1_28_427	HomoFAS_KS01_MetazoaFASI	52	406	9.6e-119	Homo sapiens FAS	type I FAS	Me
☑	RUS77019.1_4_402	OryziasFAS_KS01_MetazoaFASI	49	403	2.7e-117	Oryzias latipes FAS	type I FAS	Me
☑	RUS75294.1_29_430	OryziasFAS_KS01_MetazoaFASI	49	403	4.7e-117	Oryzias latipes FAS	type I FAS	Me
☑	RUS77541.1_8_426	EscherichiaFAS_KS02_FASII	47	421	4.8e-103	Escherichia coli FAS	type II FAS	su
☑	RUS92164.1_18_385	MelopsittacusFAS_KS01_MetazoaFASI	39	410	2.9e-76	Melopsittacus undulatus FAS	type I FAS	Me
☑	RUS75295.1_1_296	OryziasFAS_KS01_MetazoaFASI	45	308	6.3e-73	Oryzias latipes FAS	type I FAS	Me

Right-click to DOWNLOAD this table in tab-delimited format.

Options

○ **Output selected sequences** in fasta format.

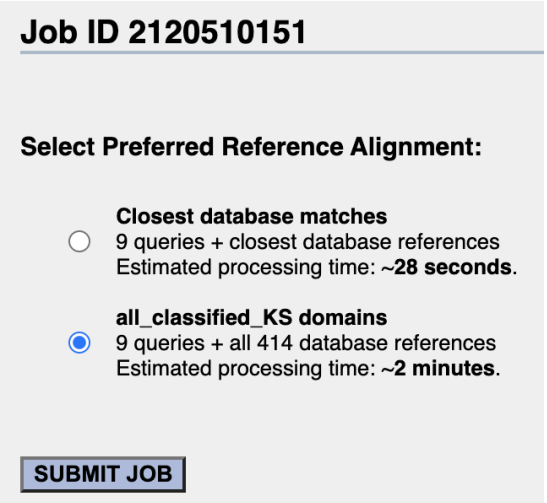
○ **Output Alignment** with closest database matches
Select alignment format: fasta

◉ **Construct tree** (candidate domains + blast matches + reference domains)

GET RESULTS

- Click “**Select All**” and then select “**Construct tree**”. We are interested in understanding if any of these KS domains clade differently than we are expecting, i.e. for example not with other metazoa type I and II fatty acid synthases (FASs). So, we will select the option of calculating a phylogenetic tree with all of our query domain hits “**all classified KS domains**”.
    - The **Tree results** page has two options for viewing the calculated trees:
        - Tree image (SVG format file), click **DOWNLOAD** to view, right-click to save.
            - When viewing the tree, the query domain hits will have red circles of where they are placed in the tree.

- Newick format (plain text file), click **DOWNLOAD** to view, right-click to save
    - This file format can be viewed in tree viewing programs like FigTree, iTol, TreeViewer, ETE Toolkit, etc.

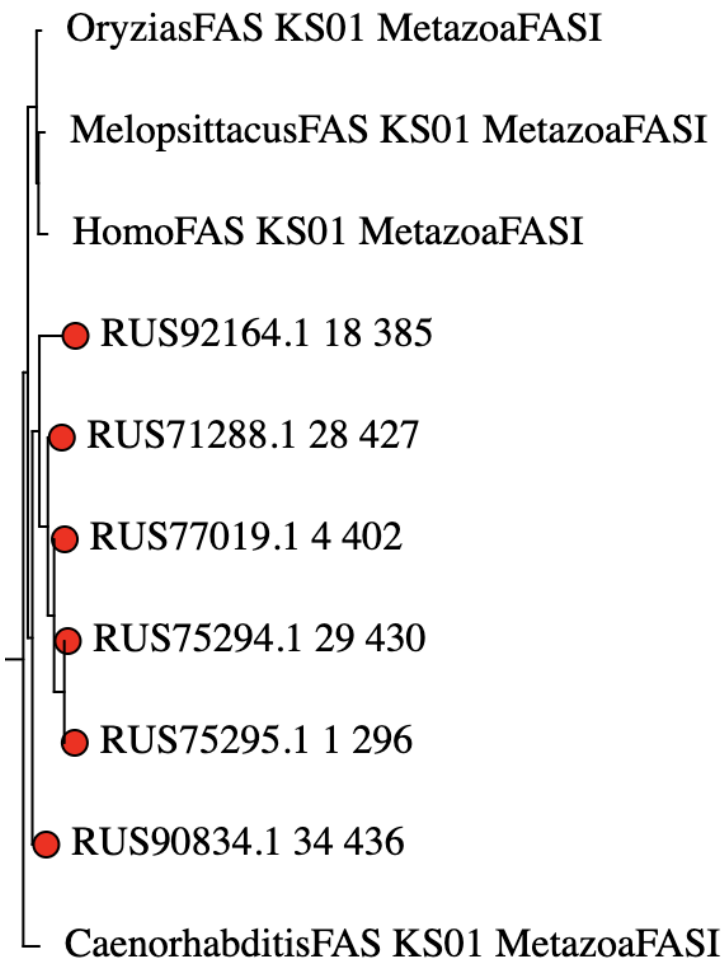


4. **Interpret the NaPDoS2 query hit and phylogenetic tree results**

    - In *Torres et al. Nature Comms. 2019*, they reported and experimentally characterized FAS and FAS-like KSs from the *Elysia chlorotica* sacoglossan genome. They found these were unique because they are animal FASs that perform chemical reactions similar to KSs. See Figure 2 of *Torres et al. Nature Comms. 2019* for the phylogenetic tree showing the divergence of the "Sacoglossan PKSs that group separate from the known FAS and PKS proteins from bacteria, fungi and animals".

        - *Specifically **RUS908341.1 ("EcFAS"), RUS71288.1 ("EcPKS1"), and RUS75294.1 ("EcPKS2")** were characterized by Torres et al. Nature Comms. 2019*

    - NaPDoS2 correctly classified these new clade of FAS-like KS sequences as metazoan-type type I FASs and from our NaPDoS2 generated phylogenetic tree, we can see that the sequenced reported by *Torres et al. Nature Comms. 2019* that confirmed their divergence from previously characterized animal FASs. In ***Supplementary Table S8*** in the NaPDoS2 publication, we highlight which of the sequences NaPDoS2 identified matched with the new EcFAS, EcPKS1, and EcPKS2 that were characterized by *Torres et al. Nature Comms. 2019*.



*Screenshot of the partial tree of protein query hits from the Elysia chlorotica genome– these are the new FAS-like KS sequences as they clade outside of other Metazoan-type type 1 FASs and specifically **RUS908341.1, RUS71288.1, and RUS75294.1** were characterized by Torres et al. Nature Comms. 2019, demonstrating that NaPDoS2 can find novel biosynthetic and enzymatic potential.*

- To ***interpret*** these results, although relatively few metazoan PKSs have been experimentally characterized, NaPDoS2 is able to recover not only the the recently described new clade of FAS-like KSs but can be used to analyze other animal genomes for discovery of more novel KS and FAS clades and new biosynthetic potential.

4. Analyzing a BGC from antiSMASH, PRISM, or MIBiG with NaPDoS2

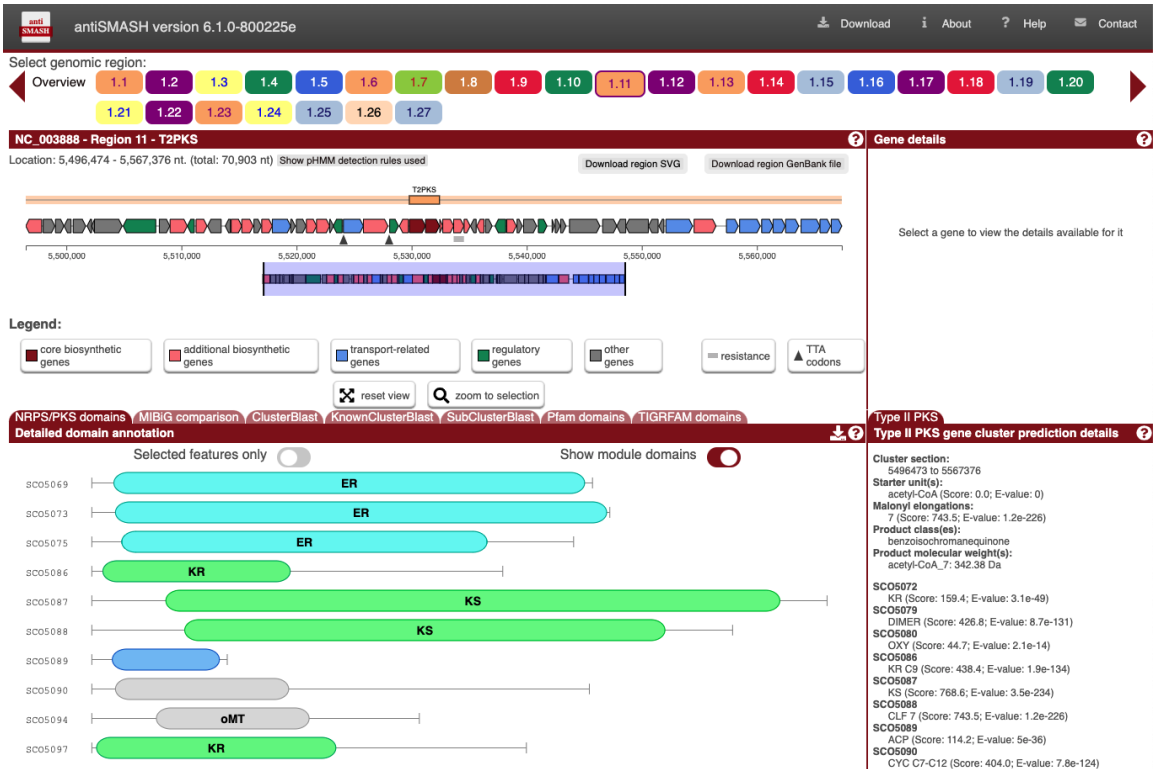
Objective: Analyze a BGC from antiSMASH, PRISM, and MIBiG to show how NaPDoS2 can give additional specific information about KS domains of interest in a BGC.

1. **Select your BGC to analyze.**

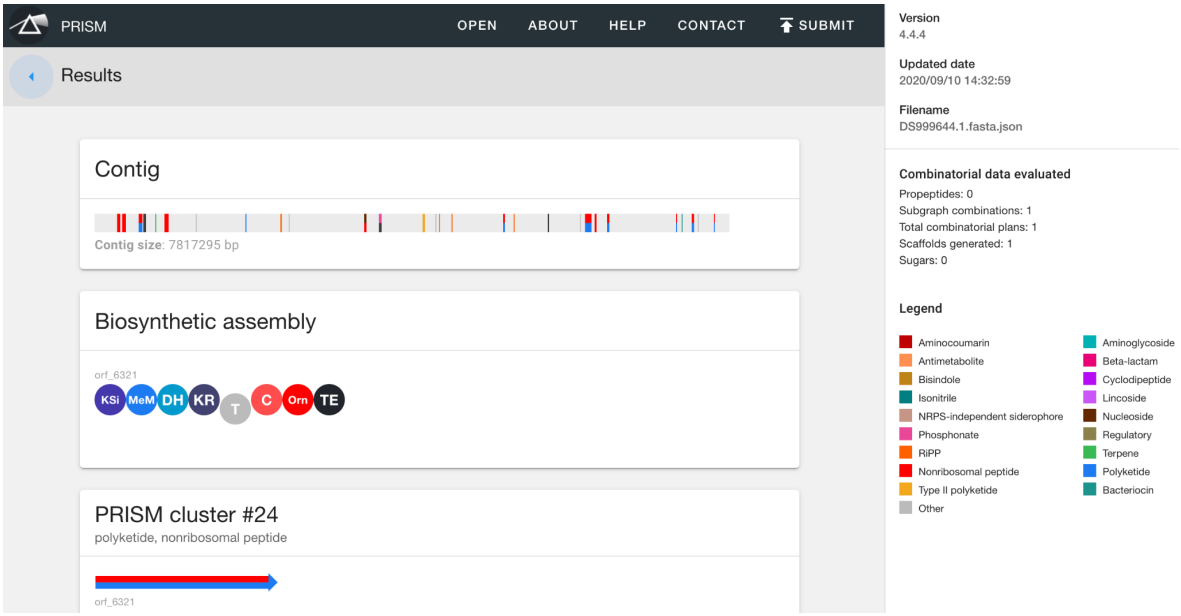
    - In a real-world example, you might want to analyze a BGC that you have found from your analysis, but for this tutorial we will pretend we have run an analysis with antiSMASH and PRISM and use example output to analyze a BGC with NaPDoS2.

BGC Source	**Accession**	**Tool Reference**
antiSMASH bacterial version *Streptomyces coelicolor A3(2)* (BGC: Region 11)	Example output link NC_003888.3	*Blin et al. 2021. Nucl Acids Res.*
PRISM *Streptomyces roseosporus* NRRL 15998 (BGC: Cluster #24)	Example annotated output DS999644.1	*Skinnider et al. 2020. Nat Commun.*
MIBiG 2.0 repository *Byssochlamys spectabilis* (BGC: Viriditoxin BGC0001866)	BGC0001866 NW_021167153.1	*Kautsar et al. 2020. Nucl Acids Res.*

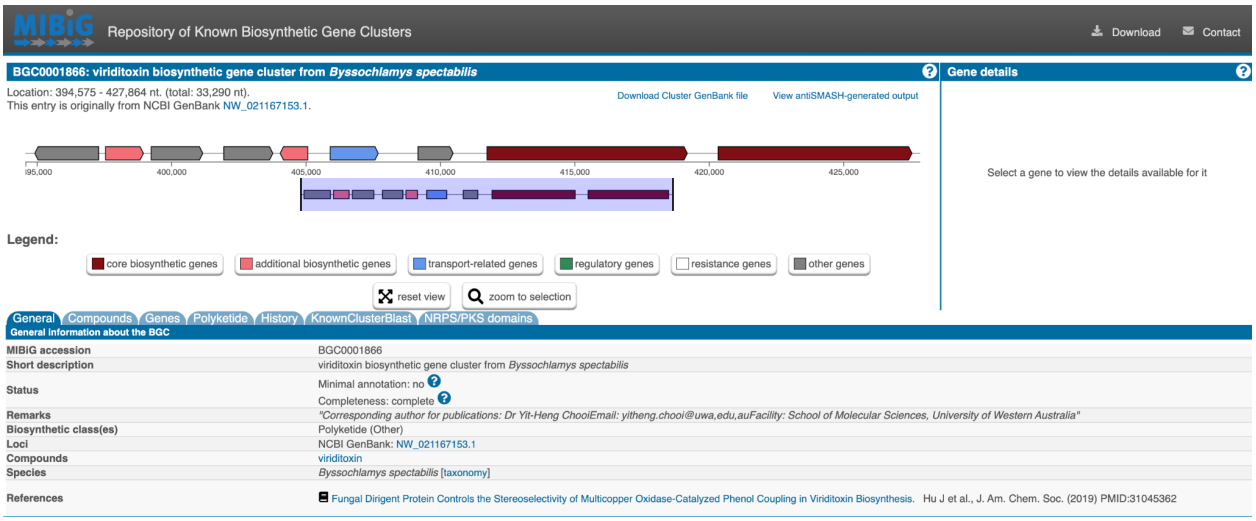
    - **antiSMASH**: Go to the antiSMASH bacterial version home page (you could also analyze a BGC from fungiSMASH or plantiSMASH). Click on the light gray "**Open example output**" button towards the top of the page to the right. This will open the antiSMASH 6.1 (at time of writing) version result for *Streptomyces coelicolor* A3(2).
        - The BGC we will analyze is BGC "**1.11**" which is a Type 2 PKS. Click on the orange "**Region 11**" box.



- **PRISM**: Go to the PRISM home page (proceed through the Terms of Service agreement). On the following page, click "**Open annotated output**" in blue. This will load an example annotated tutorial on interpreting PRISM output which we will use.
    - The BGC we will analyze in "**Cluster 24**" which is a single gene BGC called a "polyketide, nonribosomal peptide" with a putative iterative KS predicted.



- **MIBiG**: Go to the MIBiG home page. Click on "**Repository**" at the top of the page. In the filter box, type "**BGC0001866**" and click on the BGC result which will open a new tab.



- We have also provided the BGC sequences that we'll extract from these BGCs to analyze in the tutorial on the NaPDoS2 OSF page that you can download directly from there. The BGC accessions are linked above.

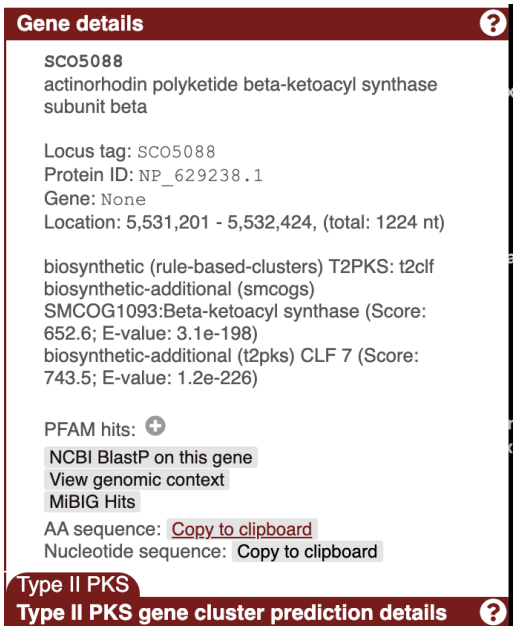
2. Make a FASTA sequence file for your BGC.

- In order to run a NaPDoS2 analysis, we need a FASTA formatted file of our sequences. For many of the BGC prediction programs, you can download your BGC results as a Genbank (*.gbk) file, which there are programs (Genbank to DNA FASTA file; Genbank to amino acid FASTA file) that can help extract sequences in FASTA format. However, in this tutorial, we'll show how copying the genes/proteins you are interested in (ie. play a biosynthetic role) and analyzing them with NaPDoS2 is a fast way to assess their KS content

- **antiSMASH**: In our **Region 11** BGC, there are 2 key biosynthetic genes predicted– in maroon. If the "**NRPS/PKS domains**" tab in red is selected, we can see where antiSMASH has predicted certain key biosynthetic domains. We see that there are 2 KS domains predicted, in genes **SCO5087, SCO5088**. These are the same genes that are predicted to be the core biosynthetic genes of this type 2 PKS.

    - We will copy the amino acid sequence from the 2 core biosynthetic genes to analyze with NaPDoS2. Click on each **SCO5087, SCO5088** maroon gene and in the "**Gene details**" box to the right, where it says "**AA sequence: Copy to clipboard**" click. Paste these sequences either into a plain text

document (TextEdit, Notepad, BBEdit, Sublime Text, etc) or copy them right into the NaPDoS2 query box.

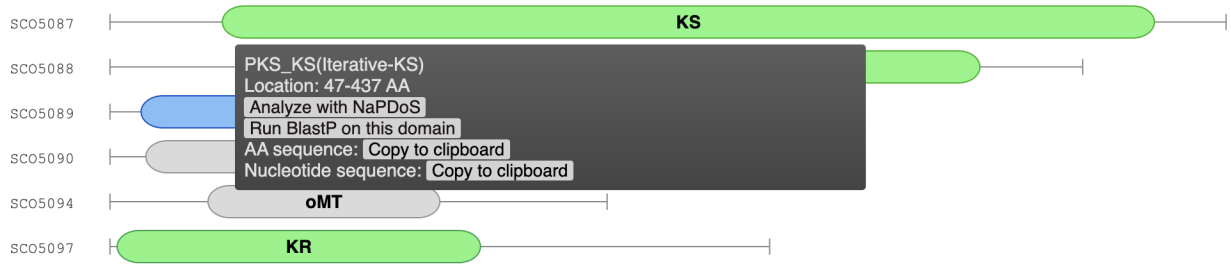


- Remember that you need to add a ">fasta_header" to each sequence line, so add the gene names:

```
>SCO5087_Region 11_antismash
MPLDAAPVDPAS…

>SCO5088_Region11_antismash
MSVLITGVGVV…
```

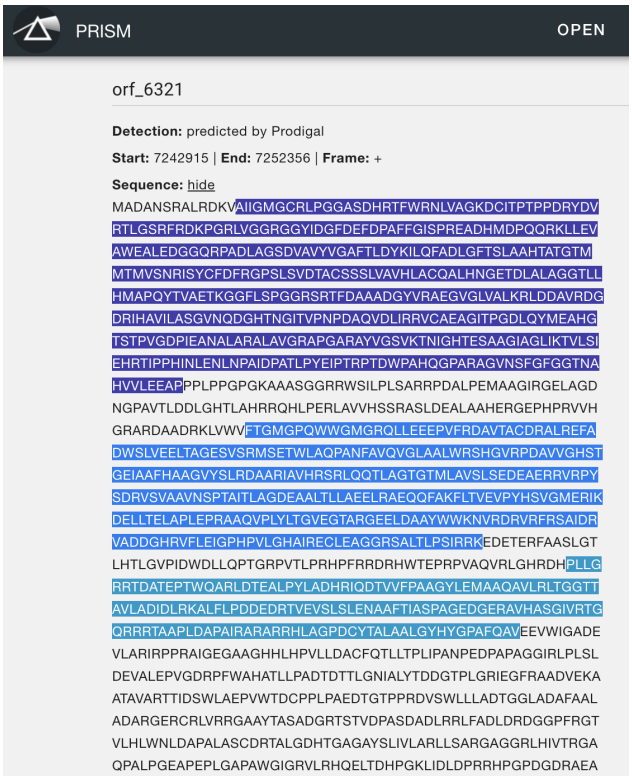
- **In antiSMASH, you can send an analysis directly to NaPDoS2! Click on the "**KS**" domain in the "**NRPS/PKS domains**" tab, and select "**Analyze with NaPDoS2**". This will automatically copy your protein amino acid sequence of the KS domain and bring you to the "**SUBMIT JOB**" page on NaPDoS2.
    - ***Note***: currently the antiSMASH website has not been updated yet to link to NaPDoS2, it instead says "**Analyze with NaPDoS**" and will run your domain through NaPDoS version 1, which will give different results than reported below. We hope to change the link soon! :)



- **PRISM**: In our **Cluster #24**, we see that the BGC is made of one biosynthetic gene "**orf_6321**" that is predicted to have both a KS and C domain. PRISM separates out the sequences of each "orf" in the "**Analysis**" section available to copy. So, we will copy the entire sequence of "**orf_6321**". Click "**Sequence: show**" and copy the sequence into a plain text document (TextEdit, Notepad, BBEdit, Sublime Text, etc) or copy them right into the NaPDoS2 query box.

    - Remember that you need to add a ">fasta_header" to each sequence line, so add the gene names:

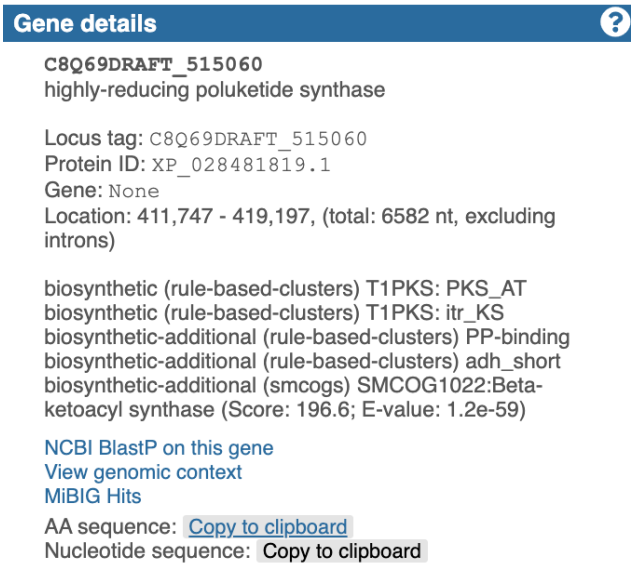
```
>orf_6321_Cluster24_Prism
MADANSRALRDKV…
```



- **MIBiG**: In our **Viriditoxin** BGC, there are 2 key biosynthetic genes predicted– in maroon. If the "**NRPS/PKS domains**" tab in blue is selected, we can see where antiSMASH has predicted certain key biosynthetic domains. We see that there

are 2 KS domains predicted, in genes **C8Q69DRAFT_515060, C8Q69DRAFT_480069**. These are the same genes that are predicted to be the core biosynthetic genes of this type 2 PKS.

- We will copy the amino acid sequence from the 2 core biosynthetic genes to analyze with NaPDoS2. Click on each **C8Q69DRAFT_515060, C8Q69DRAFT_480069** maroon gene and in the "**Gene details**" box to the right, where it says "**AA sequence: Copy to clipboard**" click. Paste these sequences either into a plain text document (TextEdit, Notepad, BBEdit, Sublime Text, etc) or copy them right into the NaPDoS2 query box.

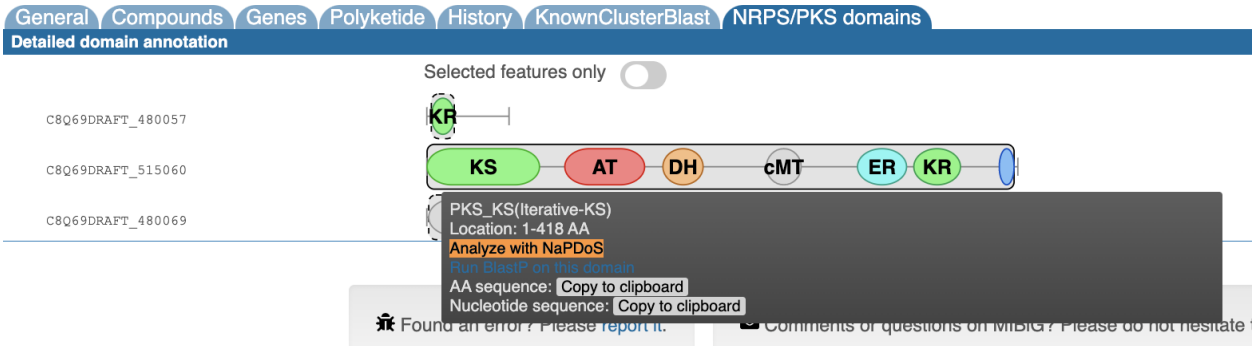


- Remember that you need to add a ">fasta_header" to each sequence line, so add the gene names:

```
>C8Q69DRAFT_515060_Viriditoxin_Mibig
MAICGIAVRLPG…

>C8Q69DRAFT_480069_Viriditoxin_Mibig
MAQKLRFYLFGD…
```

- **In MIBiG like antiSMASH, you can send an analysis directly to NaPDoS2! Click on the "**KS**" domain in the "**NRPS/PKS domains**" tab, and select "**Analyze with NaPDoS2**". This will automatically copy your protein amino acid sequence of the KS domain and bring you to the "**SUBMIT JOB**" page on NaPDoS2.
  - ***Note***: currently the antiSMASH website has not been updated yet to link to NaPDoS2, it instead says "**Analyze with NaPDoS**" and will run your domain through NaPDoS version 1, which will give different results than reported below. We hope to change the link soon! :)



3. Analyze your FASTA sequence file with NaPDoS2 and compare results .

- Now that we have FASTA sequences from the BGCs that we want to analyze, we can analyze the sequences with NaPDoS2. You can either copy & paste the sequences into the search box; or save your plain text file and upload it.

- Navigate to the "Run Analysis" tab on NaPDoS2. Select "**Domain Type: KS domains**".

- Since we are analyzing BGCs that we copied amino acid sequences from, select "**Query type: Amino acid sequences**".

- **Copy and paste** your sequences into the search box OR Click "**Choose File**" to upload your BGC FASTA file.

- Click on "**Advanced Settings**". Here, you can change the BLASTP e-value cutoff and the Minimum alignment length. For this tutorial, we will use the default settings of BLASTP e-value cutoff of 1e-8 and Minimum alignment length of 200aa (600nt).

    - Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details.

- Click **SEEK** to be assigned a job ID and review the job analysis settings.

- The "**Domain Classification Summary**" page lists the total number of domains found from the number of sequences that were input. Clicking "**VIEW ALL MATCHES**" opens a new page with the hit table of all domains identified in the genome analysis.

- The "**Individual Domain Classes**" lists the number of specific classes of KS domains that were found in your query genome sequence, ordered by the class and subclass with the most number of domains.

- **antiSMASH**: 2 KS domains were identified from 2 input sequences.

Select	Class	Subclass	Num matches
☑	type II aromatic	isochromanequinone KSb	1
☑	type II aromatic	isochromanequinone KSa	1

- **PRISM:** 1 KS domains were identified from 1 input sequence.

Select	Class	Subclass	Num matches
☑	type I iterative cis-AT	PTM-type	1

- **MIBiG**: 2 KS domains were identified from 2 input sequences.

Select	Class	Subclass	Num matches
☑	type I iterative cis-AT	highly reducing	1
☑	type I iterative cis-AT	non-reducing	1

6. **Compare your NaPDoS2 results to your BGC predictions & map the domains to the BGC**

- Clicking on "**VIEW ALL MATCHES**" will let us view the "**Database Search Results**" page, seeing all hits in our BGC query sequences.
- The table can be sorted by each of the column values by clicking on the column header; the table scrolls up and down and if your query hit names are long, the table scrolls left to right. You can **DOWNLOAD** the results table (click link to view; or right-click link & select "Save Link As" to save) in a tab-delimited format.

- The scrollable table lists the following information:

  - **cand_id**: name of the domain hit, which is the name of the sequence in the query FASTA file before spaces and appended with location information of where the domain hit was found (see below for details)

- **database match**: the name of the domain in the NaPDoS2 database that the query hit is closest to
    - **names include**: BGCname_KS#_uniquetag
    - **BGC name**: name of the BGC
    - **KS#**: number KS in successive order from the BGC
    - **unique tag**: abbreviation that is unique for every class/subclassification

- **percent identity:** percent identity that the query hit shares with the closets NaPDoS2 database match

- **align length:** alignment length that the query hit shares with the closest NaPDoS2 database match

- **e-value:** e-value of the query hit to the closest NaPDoS2 database match

- **BGC match**: name of the BGC of the closest NaPDoS2 database match to the query hit; click on the link to open a new page of the BGC card information

- **domain class:** classification of the closest NaPDoS2 database match to the query hit
    - This is the ***classification*** of your query hit

- **domain subclass:** specific subclass of the closest NaPDoS2 database match to the query hit
    - This is the ***complete specific subclass*** of your query hit (some classes do not have subclasses, ie “no subclass”)

- Here, you can also select specific domain hits to download or further analyze (select all to download or analyze all results). You have the follow options for downloading the data:

    - **Output selected sequences** in fasta format

        - Select which trimmed query domain hits or select all to output in FASTA format

            - **if your query was a nucleic acid, you have the option to output the query domain hits as either amino acid or nucleic acid

        - **if your query was an amino acid protein, you only have the option to output the query domain hits as amino acid sequences

    - Click **GET TRIMMED SEQS**

    - This is a very useful output for you to continue working with the KS or C domains that NaPDoS2 detected & classified. You can copy and paste the trimmed query KS or C domain hits into a new file, a sequencing editing program, etc.

    - For interpretation of the appended information to your query search (ie. >ProteinA_13_429 or >GeneB_5_1323_56_472), please see below.

- **Output Alignment** with closest database matches

    - Select alignment format (all alignments are created with MUSCLE but can be output in different formats suitable for many downstream applications):
        - FASTA (.fasta)
        - Msf (.msf)
        - Clustalw (.clw)

    - This outputs an alignment of your selected sequences with the closest NaPDoS2 database matches. Note, the alignment output will include all closest NaPDoS2 database matches to all query hits, not just your selected sequences, but only your selected sequences will be present in the alignment.

    - You can **DOWNLOAD** (click link to view; or right-click link & select "Save Link As" to save) the alignment file for downstream use in sequence programs, etc.

- **Construct tree** (candidate domains + blast matches + reference domains)

    - Select the query domain hits that you would like to build a tree with. You can either build a tree with:
        - **Closest database matches** to your selected sequences, ie. # of selected queries + closest database references
        - **All classified KS domains** which would be the # of selected queries + all 414 NaPDoS2 database reference sequences.

- The estimated time to calculate the trees is given, usually the time to complete the trees is much faster than estimated.

- Details about tree building are continued below for the objective aim of this tutorial to put novel KS sequences into phylogenetic context with known KS domains.

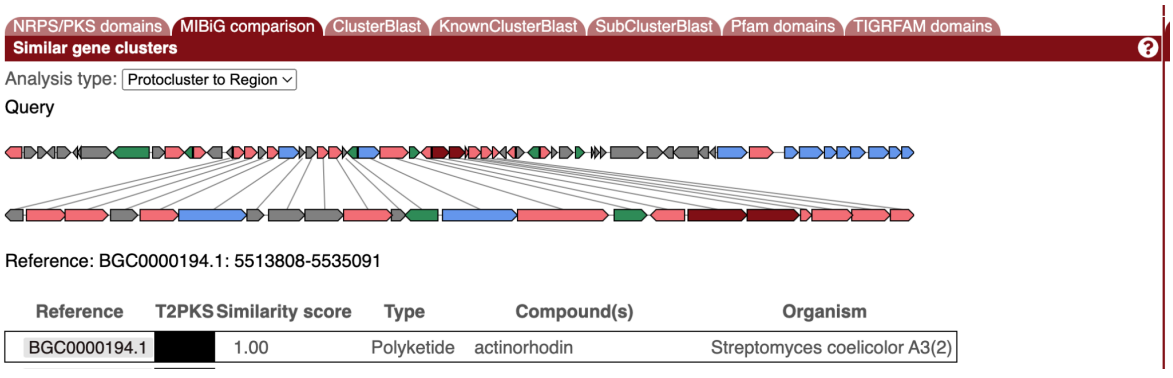
- ***Interpreting*** our results:

- **antiSMASH**: NaPDoS2 detected and classified 2 KS domains from 2 input sequences. The 2 KS domains were classified as
    - **SCO5087_Region_11_antismash**: type II aromatic, isochromanequinone KSa
    - **SCO5088_Region_11_antismash**: type II aromatic, isochromanequinone KSb

	cand_id	database match	percent identity	align length	e-value	BGC match	domain class	domain subclass
☑	SCO5087_Region_11_antismash_1_467	actinorhodin_KS01_isochromanequinoneKSa	100	467	6.3e-266	actinorhodin	type II aromatic	isochromanequinone KSa
☑	SCO5088_Region_11_antismash_1_407	actinorhodin_KS02_isochromanequinoneKSb	100	407	2.0e-239	actinorhodin	type II aromatic	isochromanequinone KSb

- These KS domains are predicted to create an isochromanequinone type of molecule, which matches the prediction for this BGC– this BGC has been characterized and its product is known to be the compound actinorhodin which belongs to that molecule class.

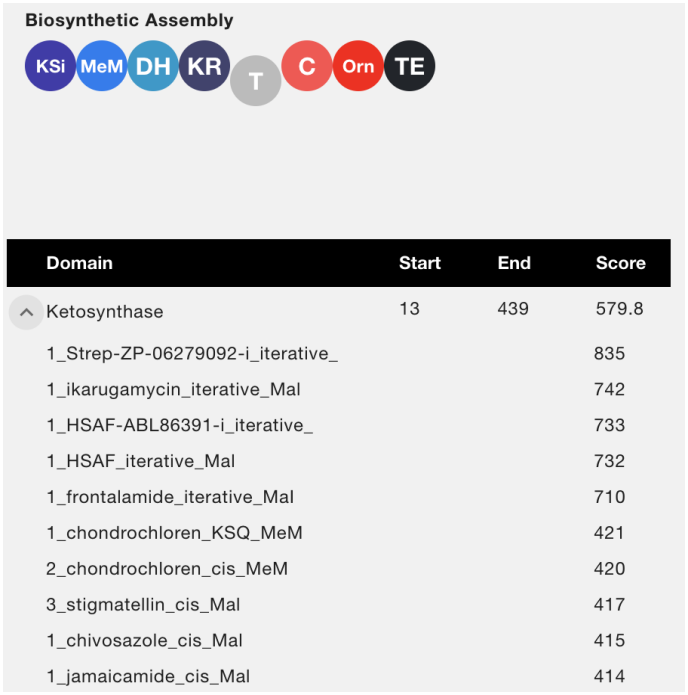
- We can see that NaPDoS2 helped to give a specific type of compound class for the key biosynthetic domains.



- **PRISM:** NaPDoS2 detected and classified 1 KS domains from 1 input sequence. This was the single BGC from PRISM which does not seem to have a predicted biosynthetic product. NaPDoS2 classified the KS domain as a type 1 iterative *cis*-AT PTM-type KS.
    - **orf_6321_Cluster24_Prism**: type 1 iterative *cis*-AT, PTM-type

	cand_id	database match	percent identity	align length	e-value	BGC match	domain class	domain subclass
☐	orf_6321_Cluster24_Prism_13_438	SGRPTM_KS01_iPKSPTM	96	426	5.0e-242	SGR810-815	type I iterative cis-AT	PTM-type

- While it might be that this BGC is incomplete, it is interesting that a PTM-type KS is nearby a NRPS C-domain and this cluster could be interesting to explore further. NaPDoS2 gave a classification and prediction that gave a little bit more information from the PRISM prediction of the closest homolog.



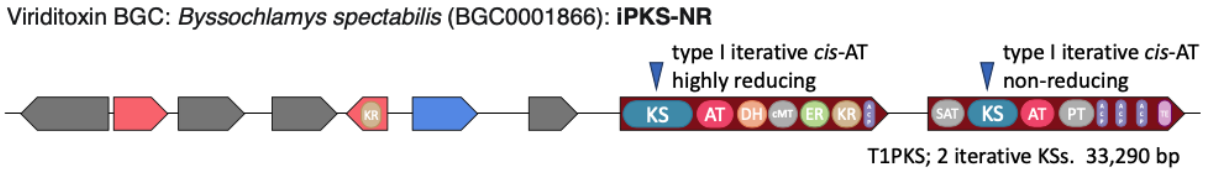
Domain	Start	End	Score
Ketosynthase	13	439	579.8
1_Strep-ZP-06279092-i_iterative_			835
1_ikarugamycin_iterative_Mal			742
1_HSAF-ABL86391-i_iterative_			733
1_HSAF_iterative_Mal			732
1_frontalamide_iterative_Mal			710
1_chondrochloren_KSQ_MeM			421
2_chondrochloren_cis_MeM			420
3_stigmatellin_cis_Mal			417
1_chivosazole_cis_Mal			415
1_jamaicamide_cis_Mal			414

- **MIBiG**: NaPDoS2 detected and classified 2 KS domains from 2 input sequences. Interestingly, the 2 KS domains that NaPDoS2 found were different iterative KSs, one predicted as highly and one non-reducing.

    - **C8Q69DRAFT_515060_Viriditoxin_Mibig:** type 1 iterative *cis*-AT, highly-reducing
    - **C8Q69DRAFT_480069_Viriditoxin_Mibig**: type 1 iterative *cis*-AT, non-reducing

	cand_id	database match	percent identity	align length	e-value	BGC match	domain class	domain subclass
☑	C8Q69DRAFT_480069_Viriditoxin_Mibig_375_806	fusarubin_KS01_iPKSNR	67	437	3.2e-182	fusarubin	type I iterative cis-AT	non-reducing
☑	C8Q69DRAFT_515060_Viriditoxin_Mibig_1_415	pneumocandin_KS01_iPKSHR	58	417	1.1e-137	pneumocandin	type I iterative cis-AT	highly reducing

- We can map these domains back to the described viriditoxin BGC, of which the biosynthesis is described, and these domain functions match what was reported in the paper (see screenshot below of ***Supplementary Figure S6*** from the NaPDoS2 paper for the annotated BGC with the NaPDoS2 hits). However, this information from NaPDoS2 was more detailed than the iterative prediction for both KS domains on the MIBiG entry.



5. Analyzing a metagenome with NaPDoS2

Objective: Analyze an assembled metagenome to identify & classify KSs of interest, getting a snapshot of the KS biosynthetic potential of a sample and using detected aryl polyene KS domains as a search hook to find the encoding aryl polyene BGC.

1. **Gather the metagenome sequence.**

    - There are many databases that contain metagenomes from NCBI Assemblies/SRA, JGI IMG/MER, MGnify, NMDC, MG-RAST, etc), but for this tutorial (and the analysis we performed for the NaPDoS2 publication), we will be downloading a metagenome submitted to the "**Paired Omics Data Platform (PoDP)**".

    - We'll download the FASTA-formatted marine sediment assembled metagenome from French Polynesia, Moorea sampled in 2018 ("**MO18_071**") from the ENA repository linked from the Paired Omics Data Platform (PoDP).

        - We have also provided the assembled metagenome FASTA file used in the tutorial on the NaPDoS2 Tutorial OSF page that you can download directly from there

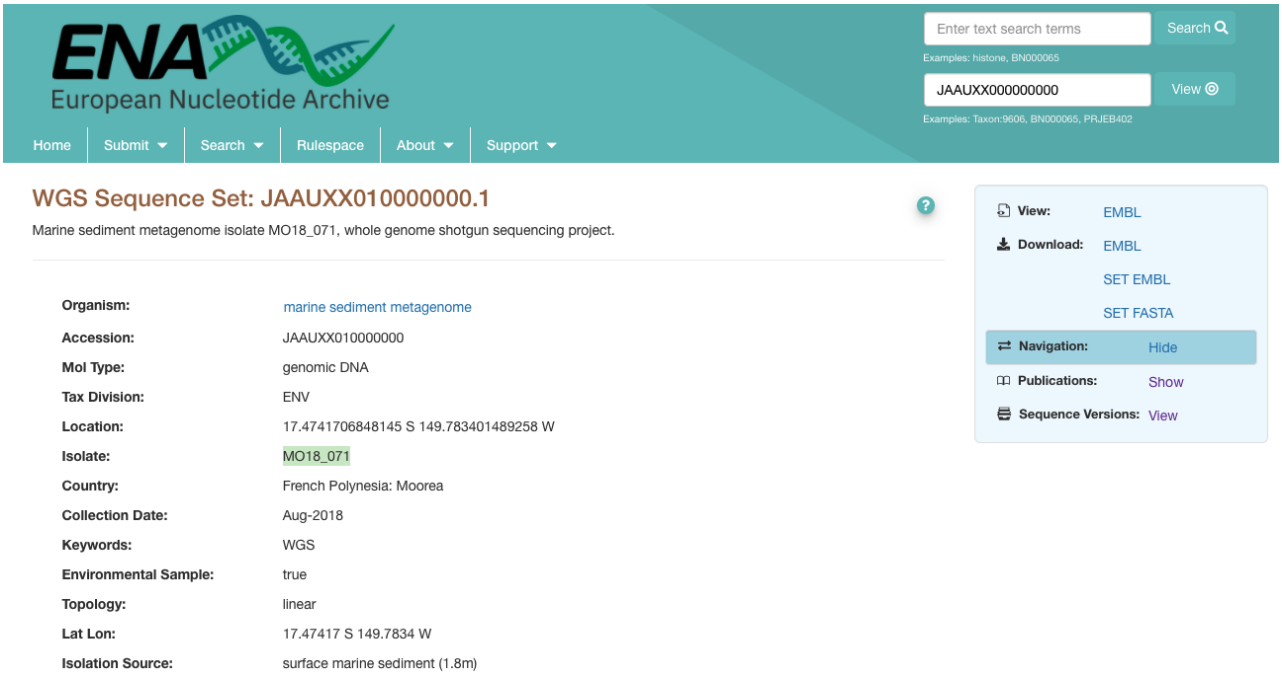
        - On the "Paired Omics Data Platform (PoDP)", click "**List**" and search "**MSV000083268**" in the search bar. Click on the top match to open the project page. Scroll down to the sample "**MO18_071**" and click on the **ENA/NCBI accession number** (JAAUXX000000000).

        -



Metagenome	Accession	Reference
Marine sediment sample MO18_071 from French Polynesia, Moorea	PoDP: MSV000083268 ENA/NCBI: JAAUXX000000000	*Schorn et al. 2021. Nat Chem Biol.*

- In the box to the right, click on "**Download: SET FASTA**". This will start a download of all the sequence records in this set in FASTA format. (If the download doesn't start automatically, right-click on the "**SET FASTA**" and open as a new tab; it is a direct download via a FTP) . The sequences will be in nucleic acid format.



- Save the downloaded zipped file ("**JAAUXX01.fasta.gz**") in a place you can access it (for example, your Desktop or Downloads folder). Navigate to where you saved it and **unzip the file**.
    - (Windows: double-click, or select it and click "Extract all"; Mac: double-click; Linux/Command line: navigate to the directory containing your zipped file and use the command "`unzip filename.zip`")

- The **MO18_071** Moorea marine sediment metagenome includes 37,150 nucleic acid contigs (the number of ">" in the FASTA file, see here (section "Combine multiple genome files") for how to count sequences in a FASTA file.

2. **Submit a NaPDoS2 analysis of the assembled metagenome.**

    - Navigate to the "Run Analysis" tab on NaPDoS2. Select "**Domain Type: KS domains**".

- Because the MO18_071 sample contains nucleic acid FASTA sequences, select "**Query type: Nucleic acid sequences**".

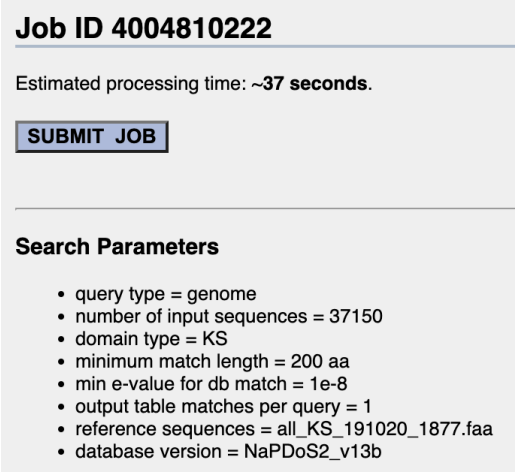
- Click "**Choose File**" to upload the metagenome file. NaPDoS2 cannot accept zipped files.

- Click on "**Advanced Settings**". Here, you can change the BLASTP e-value cutoff and the Minimum alignment length. For this tutorial, we will use the default settings of BLASTP e-value cutoff of 1e-8 and Minimum alignment length of 200aa (600nt).

    - Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details.

- Click **SEEK** to be assigned a job ID and review the job analysis settings.

    - NaPDoS2 has a job limit of 500,000 sequences or 500MB file size. If your metagenome file is too large, we provide easy-to-use-scripts to help split your file into smaller files. See section below.



- For our "**Query type: nucleic acid sequences**", we first proceed to the "**Nucleotide Sequence Search Results**" page which lists the initial number of KS domains that NaPDoS2 identified. On this page, you can **DOWNLOAD** (click link to view; or right-click link & select "Save Link As" to save) a table of the nucleotide match locations in tab-delimited format.

    - This lists the domains numbered KS# ("**domain id**"), the sequence/contig name the domain was found on ("**parents seq**"), the base pair/nucleotide start and end location of the domain ("**start**" , "**end**"), the translation reading frame ("**read frame**"), and the final name of the domain ("**cand_id**").

    - This table can be used to find the location of the KS domains in your nucleic acid genome file via the starting and ending base pairs.

- To continue the analysis, click "**CONTINUE ANALYSIS**"

3. Select results display options for KSs that NaPDoS2 detected and classified

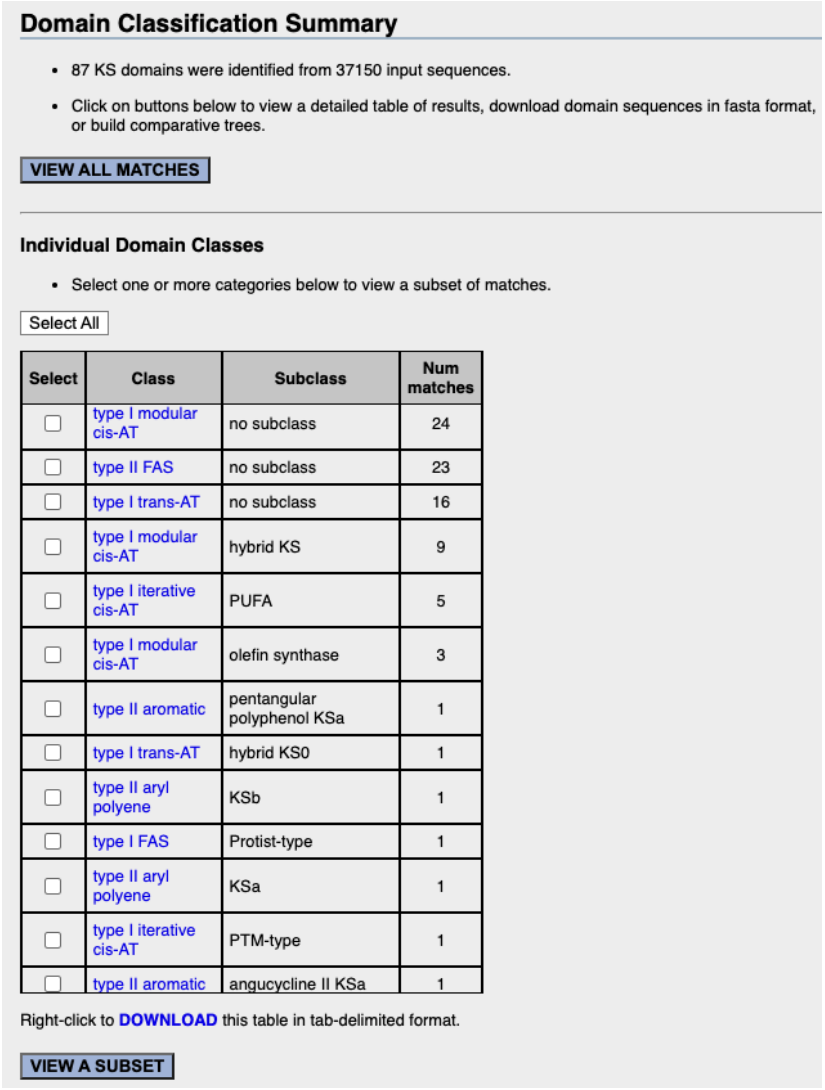
- The "**Domain Classification Summary**" page lists the total number of domains found from the number of sequences that were input. Clicking "**VIEW ALL MATCHES**" opens a new page with the hit table of all domains identified in the genome analysis.

    - In our metagenome example, 87 KS domains were found from 37,150 assembled metagenomic MO18_071 contigs.

- The "**Individual Domain Classes**" lists the number of specific classes of KS domains that were found in your query genome sequence, ordered by the class and subclass with the most number of domains.

    - This is helpful if you only are interested in a specific subset type of domain– you can click to select what class/subclasses you would like to view the results of and click "**VIEW A SUBSET**", which will open a new page with a table of only the classes/subclasses you selected.

    - You can also "**DOWNLOAD**" the "**Individual Domain Classes**" table for the counts of each KS type found (click link to view; or right-click link & select "Save Link As" to save the table).



Domain Classification Summary

- 87 KS domains were identified from 37150 input sequences.
- Click on buttons below to view a detailed table of results, download domain sequences in fasta format, or build comparative trees.

VIEW ALL MATCHES

Individual Domain Classes

- Select one or more categories below to view a subset of matches.

Select All

Select	Class	Subclass	Num matches
☐	type I modular cis-AT	no subclass	24
☐	type II FAS	no subclass	23
☐	type I trans-AT	no subclass	16
☐	type I modular cis-AT	hybrid KS	9
☐	type I iterative cis-AT	PUFA	5
☐	type I modular cis-AT	olefin synthase	3
☐	type II aromatic	pentangular polyphenol KSa	1
☐	type I trans-AT	hybrid KS0	1
☐	type II aryl polyene	KSb	1
☐	type I FAS	Protist-type	1
☐	type II aryl polyene	KSa	1
☐	type I iterative cis-AT	PTM-type	1
☐	type II aromatic	angucycline II KSa	1

Right-click to **DOWNLOAD** this table in tab-delimited format.

VIEW A SUBSET

- We have completed the first objective of this tutorial which is to identify and classify KSs of interest from a metagenome sample. In the NaPDoS2 publication, we used NaPDoS2 to capture the KS biosynthetic potential snapshot across 20 marine sediment metagenomes. We see in this sample some interesting type 1 iterative *cis*-AT PTM-type and PUFA KSs, among others.

- Let's continue on, focusing on the type 2 aryl polyene KS domains that NaPDoS2 detected in this MO18_071 sample– we will use it as a search hook to find its encoding BGC. Select the two "**type II aryl polyene KSa** and **KSb**" and click **VIEW A SUBSET** to look at just these 2 sequences.

	cand_id	database match	percent identity	align length	e-value	BGC product match	domain class	domain subclass
☑	ENAIJAAUXX010008132IJAAUXX010008132.1_5_12_14_406	xanthomonadin_KS01_arylpolyeneKSa	49	399	7.3e-101	xanthomonadin	type II aryl polyene	KSa
☑	ENAIJAAUXX010008132IJAAUXX010008132.1_6_7_120_368	arcuflavin_KS02_arylpolyene KSb	36	255	4.9e-30	arcuflavin	type II aryl polyene	KSb

- The table lists the following information:
    - **cand_id**: name of the domain hit, which is the name of the sequence in the query FASTA file before spaces and appended with location information of where the domain hit was found (see below for details)
    - **database match**: the name of the domain in the NaPDoS2 database that the query hit is closest to
        - **names include**: BGCname_KS#_uniquetag
        - **BGC name**: name of the BGC
        - **KS#**: number KS in successive order from the BGC
        - **unique tag**: abbreviation that is unique for every class/subclassification
    - **percent identity:** percent identity that the query hit shares with the closets NaPDoS2 database match
    - **align length:** alignment length that the query hit shares with the closest NaPDoS2 database match
    - **e-value:** e-value of the query hit to the closest NaPDoS2 database match
    - **BGC match**: name of the BGC of the closest NaPDoS2 database match to the query hit; click on the link to open a new page of the BGC card information
    - **domain class:** classification of the closest NaPDoS2 database match to the query hit
        - This is the ***classification*** of your query hit

        - **domain subclass:** specific subclass of the closest NaPDoS2 database match to the query hit
            - This is the ***complete specific subclass*** of your query hit (some classes do not have subclasses, ie "no subclass")

    - **Our recommendation: when working with metagenomes with many KS query domain hits, **DOWNLOAD** the **Database Search results** table as a tab delimited file (*.tab), which then you can import into R, Microsoft Excel, or Google Sheets for further parsing & analysis (ie. splitting the results by sample and counting the number of hits/class of hits per sample).

- We see that these 2 sequences have low percent identity to the top NaPDoS2 database matches (and high alignment length and good e-values), so these could be different KSs from the aryl polyene BGCs/KSs that have been described. Let's **select** the 2 sequences and click "**Construct tree**" to see how these sequences clade with other known KSs.

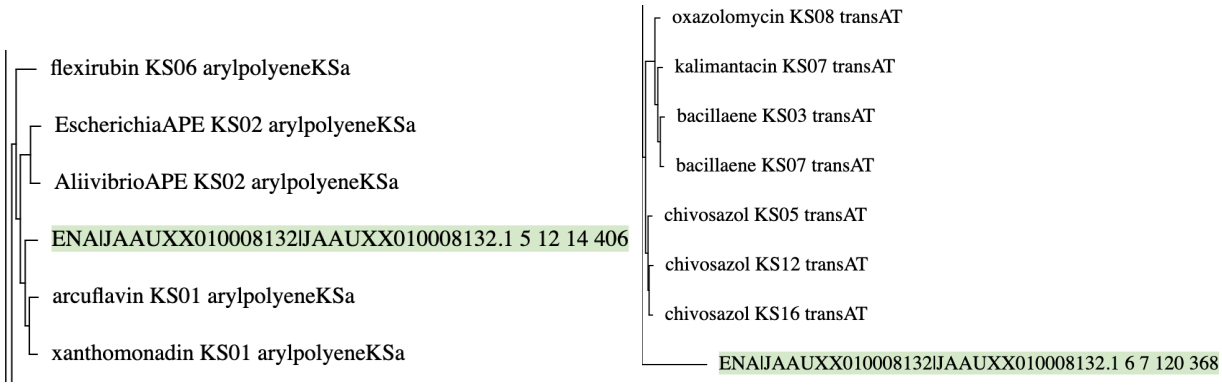
    - Select the "**all classified KS domains**" option which will put the two aryl polyene KSs in the context of the other KSs in the NaPDoS2 database.

    - The **Tree results** page has two options for viewing the calculated trees:

        - Tree image (SVG format file), click **DOWNLOAD** to view, right-click to save.
            - When viewing the tree, the query domain hits will have red circles of where they are placed in the tree.

        - Newick format (plain text file), click **DOWNLOAD** to view, right-click to save
            - This file format can be viewed in tree viewing programs like FigTree, iTol, TreeViewer, ETE Toolkit, etc.

    - We see that our "**ENA|JAAUXX010008132|JAAUXX010008132**" query domain hits clade with other type 2 aryl polyene KSa domains; and the domain query hit classified as a type 2 aryl polyene KSb domain clades very distantly from other KSb domains.



Our two aryl polyene KSa (left) and KSb (right) query domain hits from the metagenome in a phylogenetic context with the other NaPDoS2 database KS sequences. We can see that the KSa seems to clade with other aryl polyene KSa sequences whereas the aryl polyene KSb clades very distantly, which could mean these KSs comprise a unique BGC from what is in the NaPDoS2 database & known in the aryl polyene biosynthetic space.

4. **Using a metagenome query KS domain hit as a search hook to find the encoding BGC**

- Going back to the "**Database Search Results**" page, make sure the two aryl polyene KSa and KSb query hit sequences are selected. Click "**Output selected sequences** in FASTA format". Because our **MO18_071 metagenome JAAUXX01.fasta** file is nucleic acid, we will output our KS query hits we are going to use as a search hook as **Nucleic acid**.
  - **ENA|JAAUXX010008132|JAAUXX010008132.1_5_12_14_406**: type II aryl polyene KSa
  - **ENA|JAAUXX010008132|JAAUXX010008132.1_6_7_120_368**: type II aryl polyene KSb

- Open the MO18_071 metagenome **JAAUXX01.fasta** file in a plain-text editor (TextEdit, Notepad, BBEdit, Sublime Text, etc). From our output sequences, we can see that NaPDoS2 notes that these KS domains were found at certain positions, in the ***reverse complement***. That means we can't just search the KS sequences in our genome (if it wasn't the case, this would work). However, we are in luck we can use the specific FASTA header of "**ENA|JAAUXX010008132|JAAUXX010008132.1**" which is the same for both aryl polyene sequences, thus these sequences were found in the same FASTA sequence, or contig, in the metagenome.

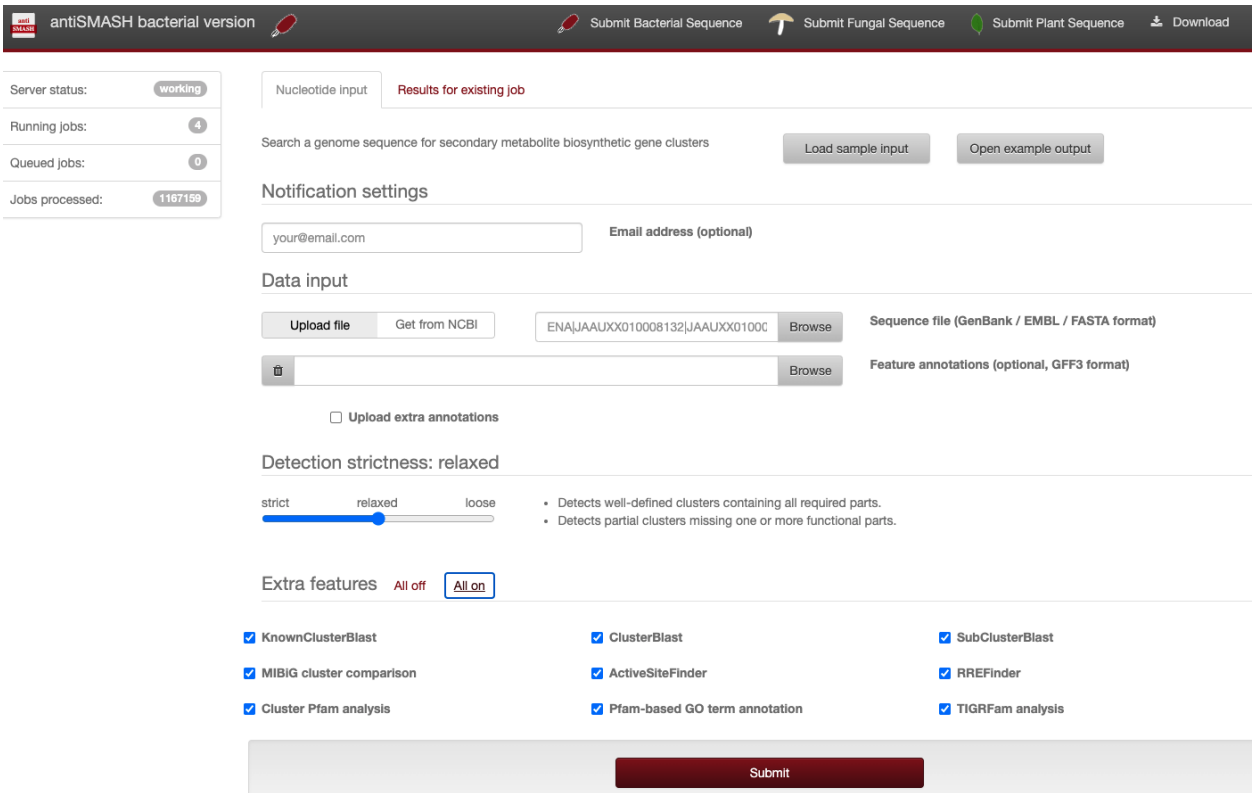
```
>ENA|JAAUXX010008132|JAAUXX010008132.1_6_7_120_368 ENA|JAAUXX010008132|JAAUXX010008132.1 [positions 4335-5081, reverse complement]
GCATGCCACGTGCTCGGCGCCGGGGCCTGGGGCCCGGGCTTCAGTAATTGGCCGCAACTC
AAAGCCCTGTTACAAAATTCCGACGGGCGGGATGAGCTGGTCGCGGTGCCGCCAAAACCG
GAAATTATTCCCGCCAACGAAAGGCGGCGCTCGCCGCTGCCAGTAAGAATCGCCGTCGAG
ACCTCCTGGCAGGCGGTGCAGCAAGCCGGTATCAACGCCGATCAGCTGGCTTGTGTGTTC
GGTTCGGGGCTGGGCGACACGCAAATCACTGATTACATGTGCCGCACGCTCACCACTGAG
CTGAAGCAACTGTCGCCGACCAAGTTTCACAACTCCGTGCATAACGCCGCCGCCGGCTAC
TGGACCATTAGTACCGGTTGCATGAAGTCGGCGACCTCCATCGCCGCCTATCACGAGACG
GCGGGGCTGGCCCTGCTGGAAGCCACCACCCAGGCCATTTACCAGGACGAGCCGGTGCTG
ATGACGCTGTTCGACACCAAGGCCCACAGCGTATACCACGATATCTTCGCCTGTGAGCTG
GATTTCGCCGGCGCCCTGGTGATCTCCCCCGCGGACGGCGGCGAGGCGGCGCTGGCAAGG
CTGGAAATCGCGCTGCAAACAGATAACAGCTACGGGCACTCGCCACTCGATCACCCACAG
TTGGCGCCGCTGTACCAAAGCAACCCGGCGGCCAAGGCCCTGGTCCTGTTTGCCCTGTTG
GCAGAAATTGACGCGGGAGCGCCGCGG
>ENA|JAAUXX010008132|JAAUXX010008132.1_5_12_14_406 ENA|JAAUXX010008132|JAAUXX010008132.1 [positions 5108-6286, reverse complement]
ATTGTCATTATGGACTATGAGGCCACTTGCGCCGCCGGCGCGGGGCTCGCGGCGCTGAGA
GAGGCTCTGTTGACACAAACGAGCGGCCTGCAGAAAAACGACCTTCCACATTCACACCTG
AATACCTGGATCGGCAAGGTGCGGGGATTGGATGACTACCACTGGCCTGAGCGGCTGCGA
CGCTGGCAGAGCCGCAACAACGCGCTGATTGCGCTGGGCCTTGCCCAGGGCCACTTCATG
ACCTCCGTGGCCCACCTCAAATCCCAATTCGACGCGCGACGCATCGGCGTGATTATCGGT
TCCAGCACCTCAAGTATCGATCGCGCCGAAGACGCCTACCGCGCGCTTGACGCCGACGGC
CATCTGCGCACCGAATTCCGCCAGGAACAGGTGCTCAACCCCCACGCGCCGGGACTGTTT
GTGGCCGAGCTGTGCCAGGTGCAAGGCCCCGCGGTGACCGTCAACACCGCCTGCTCGTCA
TCGGCCAAAGTCTTCGCCACCGCGGCGCGCTGGCTGCGTATGGGCCTGGTGGACGCGGTC
ATCGCCGGCGGCGCGGACACGCTGTGCCTGAGCGTGCTGCACGGCTTTCATTCCCTACAA
CTGGTGTCCGCACACCCCTGCCGCCCCTTTGACAAACAACGCGATGGCATCAACCTGGGT
GAGGCCGCCGGCTTTGCCCTGCTGGTGCGGGCGGAAGATATGGCGGAACACACGGGCATT
CAATTGCGCGGCTGTGGCGAAAGTTCGGACGCGCACCACATGTCGCACCCCCACCCCGAG
GGGCTGGGCGCGAAACTGGCGATAGAGCAGGCCCTGGGGGAAGCCGGCCTGGAAGCACAG
GACATTGACTACATCAACCTGCACGGCACCGCCAGTCGCGCAAACGACACCATCGAAGGC
CGGTTGGTGGCGGCCATGTTCCCACGCAGCACGCAGGCCAGCTCCACCAAAGGCTGGATG
GGCCACACCCTCGGCGCGGCGGGAATTATGGAGTCCATCGTCGCCCTGGACGCCCTAGTC
AGCGGGACCATTCCCGGCACACTCAACCTGGAAGAAGTGGACGACGATATCGCTCTGCCG
CTCACCAGCGCCAACCAGCAAGGGGATCTCAGCCACGTGATGTCGAATTCCTTTGGCTTC
GGCGGCAACAACTGCAGCGTGATCTTCAGCCGGGAGCGC
```

- Upon searching, we find a FASTA contig with the header:
  `>ENA|JAAUXX010008132|JAAUXX010008132.1 Marine sediment metagenome isolate MO18_071 MO18_071_k111_102111, whole genome shotgun sequence.`

- **Copy and paste** the entire sequence and the header into a plain-text document (TextEdit, Notepad, BBEdit, Sublime Text, etc). Save the file, using the file type ending as (***.fasta**). Our KSs should be in this contig– and to predict the BGC, we will run the entire contig through antiSMASH to predict the BGCs and assess the KS domains.

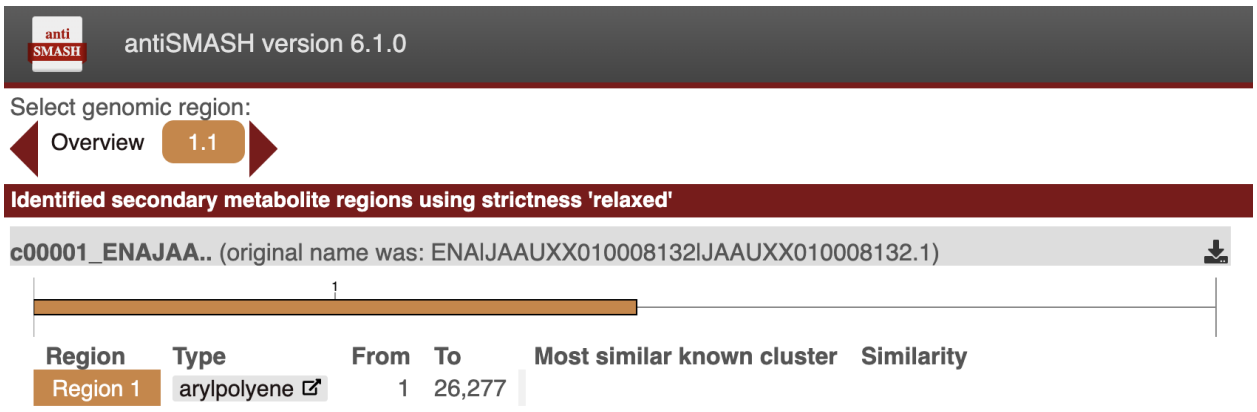
  - We have also provided this contig sequence for download on the NaPDoS2 tutorial OSF web page with other files for this tutorial.

- Go to the antiSMASH webtool website, bacterial version. Fill out your email, click **Upload file;** keep the **Detection strictness: relaxed**; click "**All on**" for **Extra features**. Click **Submit**.
  - Because our contig is pretty short, the antiSMASH job should complete relatively quickly, depending on how many other running jobs there are. (when testing this tutorial, it took 2 minutes to complete).



- Click on **"results**" from the page that shows when the job is finished, or the results link from the email.

- Fantastic– antiSMASH found an aryl polyene BGC on our contig! Click on the "**Region 1**" to view the cluster.

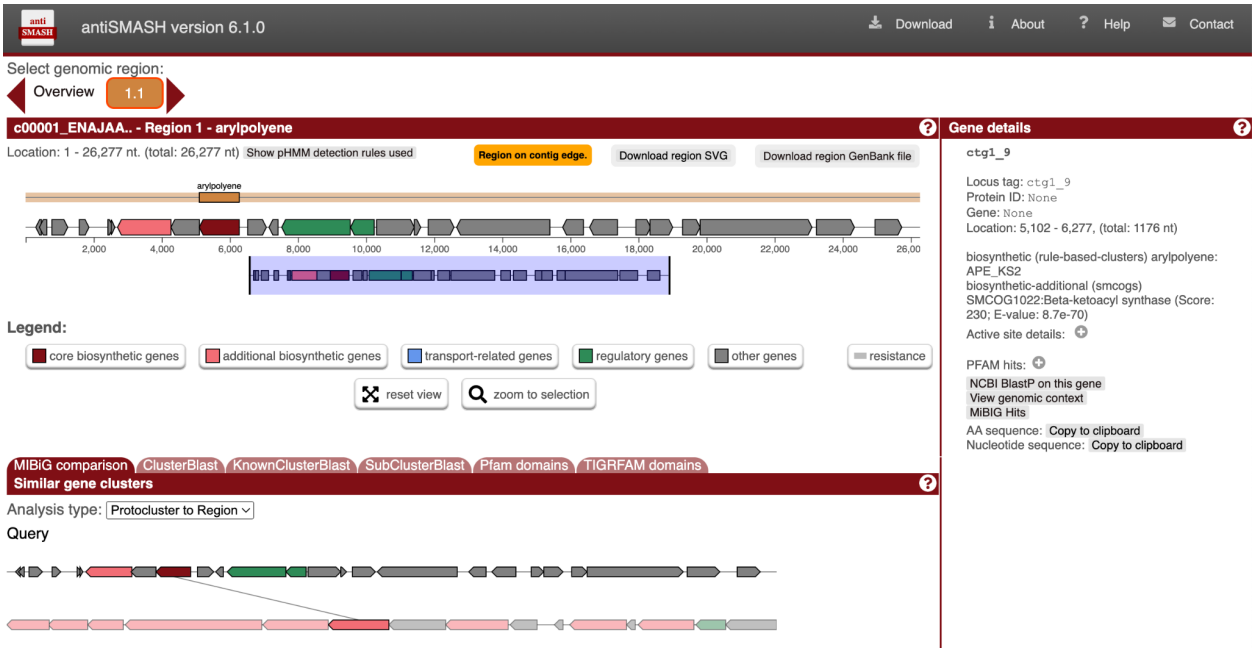


- We see that antiSMASH was only able to find 1 of the 2 key biosynthetic genes– we know from our NaPDoS2 analysis and output sequences that the aryl polyene KSa and KSb are next to one another and in reverse complement (ie. the arrow should be pointing to the left in the antiSMASH output). To link our results to our NaPDoS2 results, we will copy the amino acid sequences of the maroon gene and

the gene to the left of it, making a FASTA-formatted sequence file that we can copy into NaPDoS2 to confirm that these genes contain our aryl polyene KSa and KSb domains.

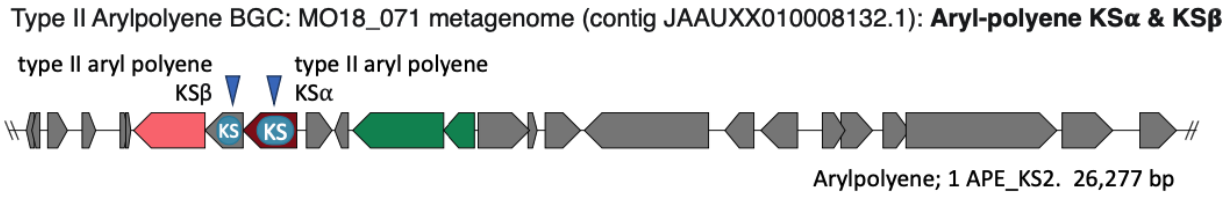
- >ctg1_9_maroon_ENA|JAAUXX010008132
  MDYEATCAAGAGLAALREALLTQTSGLQKNDLPHSHLNTWIGKVRGLDD
  YHWPERLRRWQSRNNALIALGLAQGHFMTSVAHLKSQFDARRIGVIIGS
  STSSIDRAEDAYRALDADGHLRTEFRQEQVLNPHAPGLFVAELCQVQGP
  AVTVNTACSSSAKVFATAARWLRMGLVDAVIAGGADTLCLSVLHGFHSL
  QLVSAHPCRPFDKQRDGINLGEAAGFALLVRAEDMAEHTGIQLRGCGES
  SDAHHMSHPHPEGLGAKLAIEQALGEAGLEAQDIDYINLHGTASRANDT
  IEGRLVAAMFPRSTQASSTKGWMGHTLGAAGIMESIVALDALVSGTIPG
  TLNLEEVDDDIALPLTSANQQGDLSHVMSNSFGFGGNNCSVIFSRERV

- >ctg1_8_grey_maroon_ENA|JAAUXX010008132
  MRRSHGKLACHVLGAGAWGPGFSNWPQLKALLQNSDGRDELVAVPPKPE
  IIPANERRRSPLPVRIAVETSWQAVQQAGINADQLACVFGSGLGDTQIT
  DYMCRTLTTELKQLSPTKFHNSVHNAAAGYWTISTGCMKSATSIAAYHE
  TAGLALLEATTQAIYQDEPVLMTLFDTKAHSVYHDIFACELDFAGALVI
  SPADGGEAALARLEIALQTDNSYGHSPLDHPQLAPLYQSNPAAKALVLF
  ALLAEIDAGAPRATTQMPLSDANHLSISAGGPR



- We have also provided the antiSMASH results file for the analysis of the metagenome, along with a FASTA file of the selected biosynthetic genes on the NaPDoS2 Tutorial OSF page that you can download.
- Analyze these with NaPDoS2 by following the above steps 2-3. Be sure to select "**Query type**: **Amino acid sequences**".

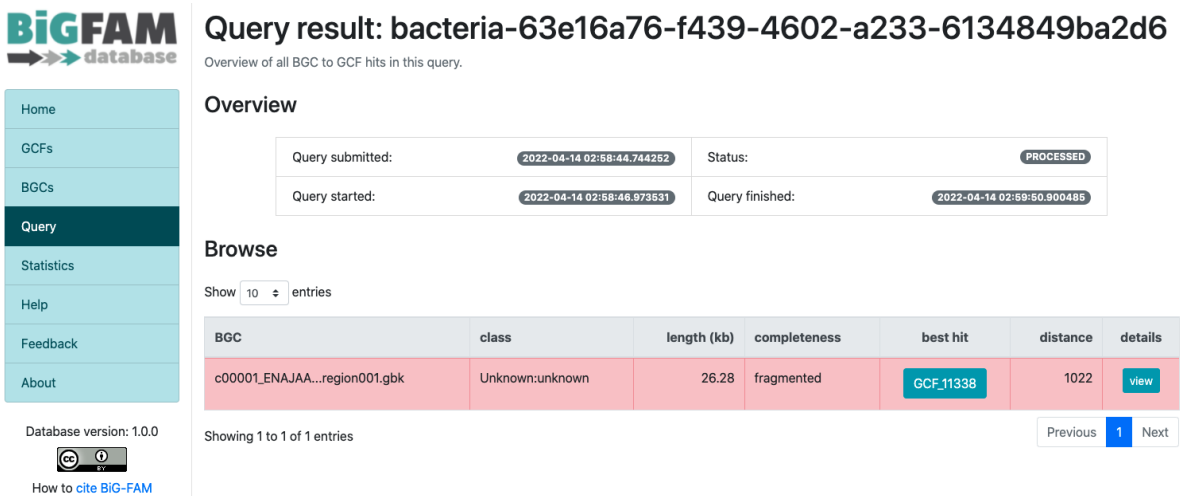
- Success! These two biosynthetic genes are indeed our type 2 aryl polyeke KSa and KSb subunits! We can use this information to draw what we know biosynthetically about the BGC, as we did in the NaPDoS2 publication ***Supplementary Figure S6***.



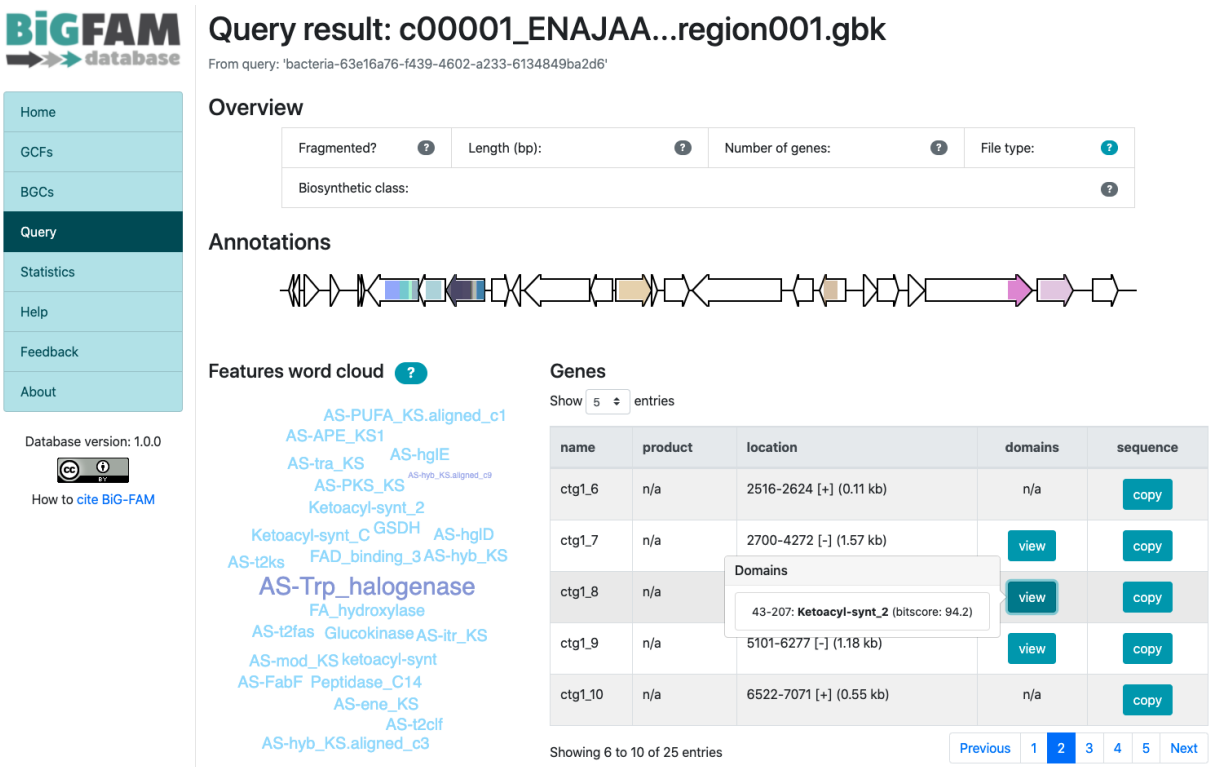
- For the last step, we've achieved our goal of using an interesting KS query domain hit from our metagenome sequence and linking it to a putative BGC. It is a BGC whose key KS domains were only half predicted by antiSMASH. Let's now see how unique this BGC is by submitting our antiSMASH output results to the BiG-FAM database to compare.
    - BiG-FAM is a database of Biosynthetic Gene Cluster Families (GCFs) for groups of biosynthetic gene clusters (BGCs) that exists as an "exporable atlas of microbial secondary metabolic diversity". The BGCs in the BiG-FAM database are from bacterial and archaeal genomes, fungal genomes, the MIBiG database, metagenome-assembled genomes from the human gut, chicken, the ocean, etc. At time of writing this tutorial, the BiG-FAM database has a total 122507 BGCs and 29955 GCFs.
    - BiG-FAM citation: *Kautsar et al 2020. Nucl Acids Res.*
- On the BiG-FAM homepage, click **Query a finished antiSMASH job** box. From our antiSMASH results page, **copy** the antiSMASH job ID which you can find in the web-address bar: it should start with "bacteria-" (example: bacteria-63e16a76-f439-4602-a233-6134849ba2d6). Paste this into the **Query** search field and click **Submit**.

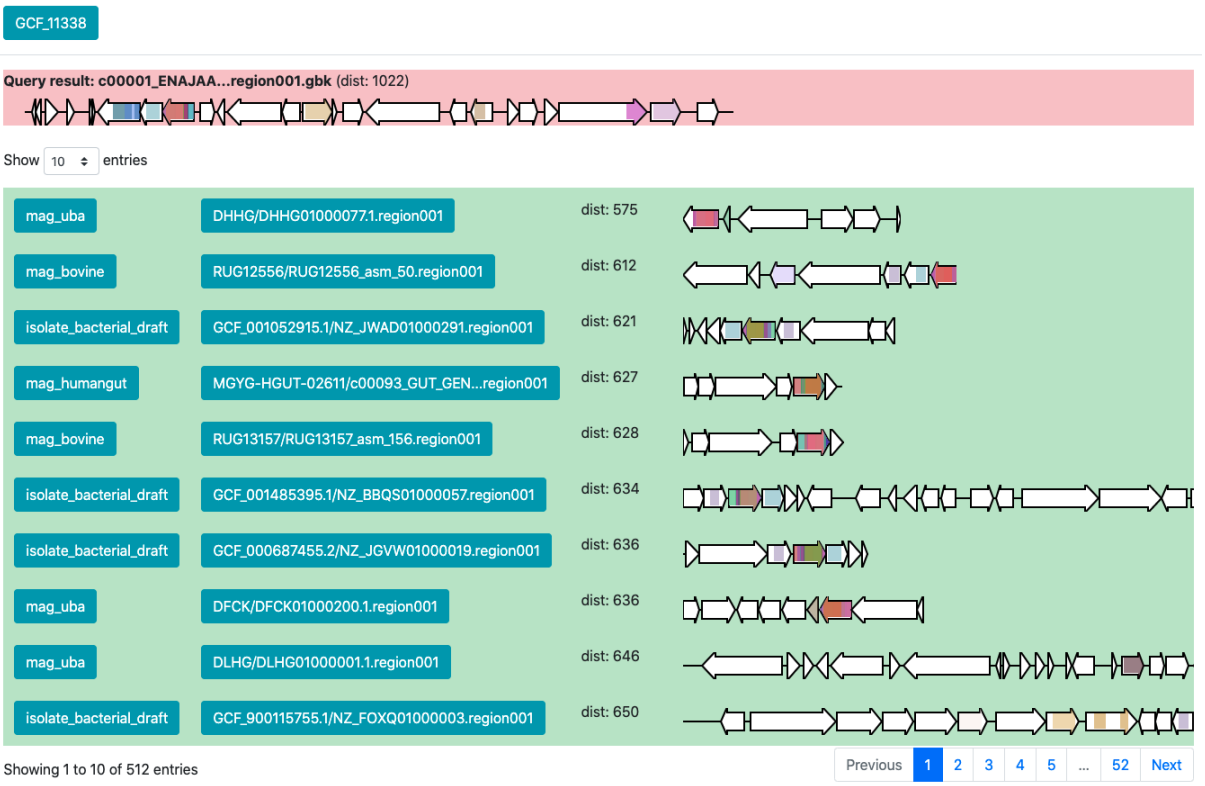


- BiG-FAM will return a results page with information for each BGC in the antiSMASH results. In our case, we only had the 1 BGC which had 1 distantly related hit. Click "**Details: View**" to see more about our aryl polyene BGC and the closest BGC hit.



- On this page, we can see our BGC and the highlighted biosynthetic domains. Our key genes, ctg1_8 and ctg1_9 are genes with domains we should watch for in the related BGCs. Scroll down to the GCF hit– we see by the red box, that this BGC is only ***putatively*** part of the GCF_11338 family of other aryl polyene BGCs (putatively due to the large distance of relatedness of 1022, red box). So, this BGC is unique but some other putative BGCs in this Gene Cluster Family might be similar to our BGC.

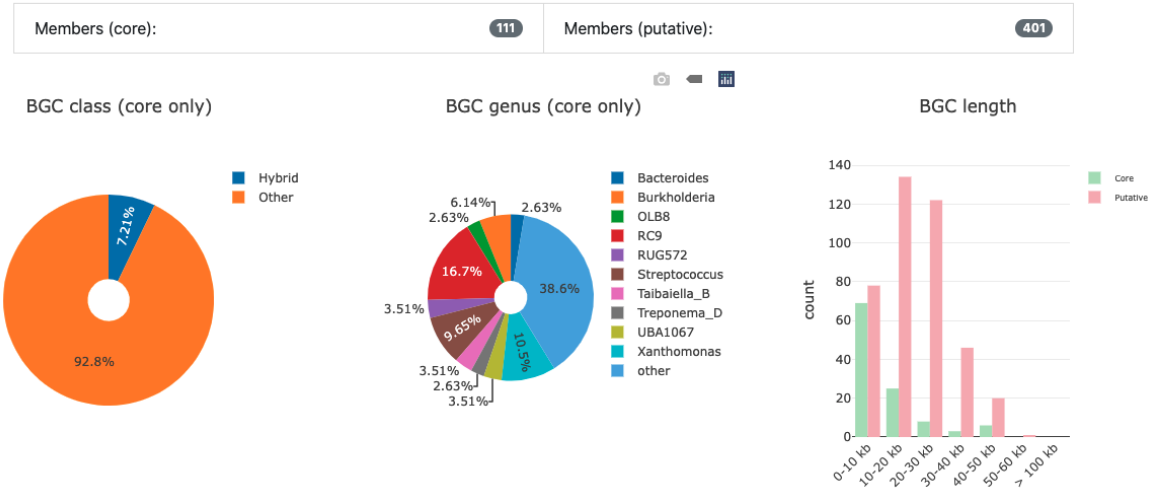




- Click on the **GCF_11338** to see our BGC in comparison with the BGCs in the Gene Cluster Family. To learn more about the GCF family, click the **GCF_11338 box** which will open the page about the distribution of this Gene Cluster Family, the BGC members, and the core (green) and putative (red) BGCs (measured by relatedness of BGC distance) that belong to this GCF.

GCF_11338

Statistics



- ***Interpreting*** these results, it seems as though our aryl polyene BGC is different from other aryl polyene BGCs and perhaps is only similar to some putative BGCs found in other genomic sources. We know, however, that the core aryl polyene KSa and KSb domains are present from our NaPDoS2 analysis and thus we could continue to analyze the GCF family to see which BGCs are most similar with the same domains and similar gene neighborhood organization.

6. Analyzing individual genes/proteins & PCR clone products with NaPDoS2

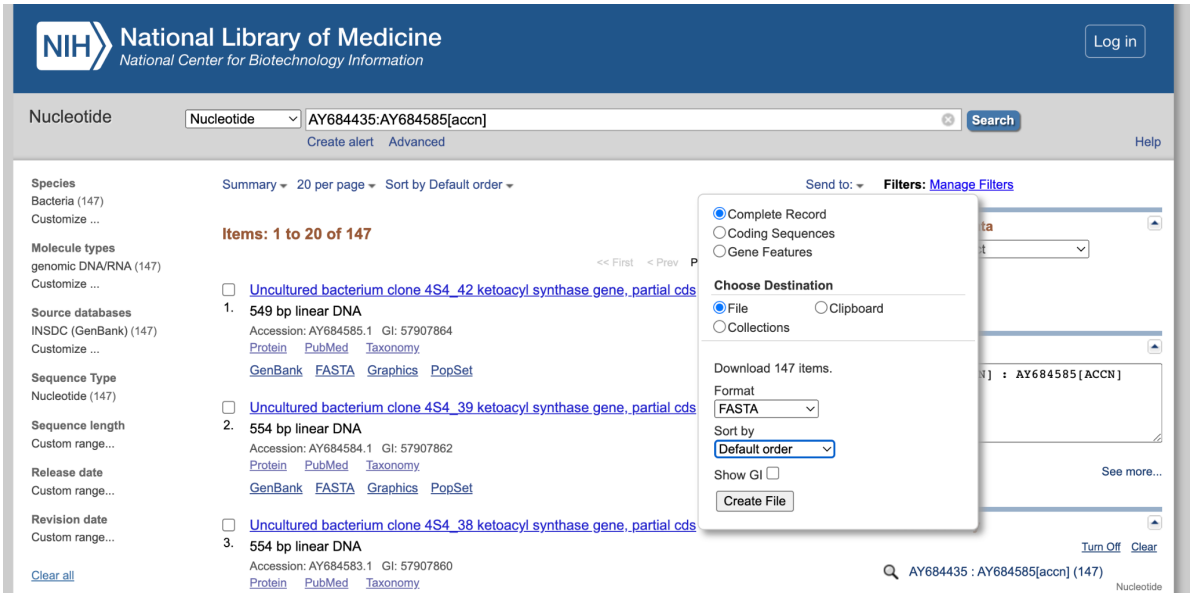
Objective: Analyze KS PCR clone products to identify & classify KSs that were amplified from environmental DNA (eDNA) samples to a specific subclass. The same steps would be followed for analysis of a single gene/protein with NaPDoS2.

1. **Gather the sequences**

- We'll download the KS PCR sequences from NCBI GenBank. These sequences were deposited from *Wawrik et al. 2005 AEM* study investigating type 2 KSs amplified from soil environmental DNA. We can download the sequences from NCBI GenBank.
    - We have also provided the genome sequences used in the tutorial on the NaPDoS2 Tutorial OSF page that you can download directly from there
    - On the NCBI GenBank, search "**AY684435:AY684585[accn]**"-- there are 147 KC PCR sequences from this study, and this query will list them all.

PCR sequences	**Accession**	**Reference**
Type II KS PCR clone sequences from soil eDNA	AY684435 - AY684585	*Wawrik et al. 2005 AEM.*

    - Click "**Send to**" and select "**Complete Record**", "**File**", "**Format: FASTA**", Sort by "**Default order**" and then click "**Create file**". This will create a FASTA file with 147 nucleotide sequences.



- Save the downloaded file in a place you can access it (for example, your Desktop or Downloads folder).

2. Submit a NaPDoS2 analysis of the KS PCR sequences

- Navigate to the "Run Analysis" tab on NaPDoS2. Select "**Domain Type: KS domains**".

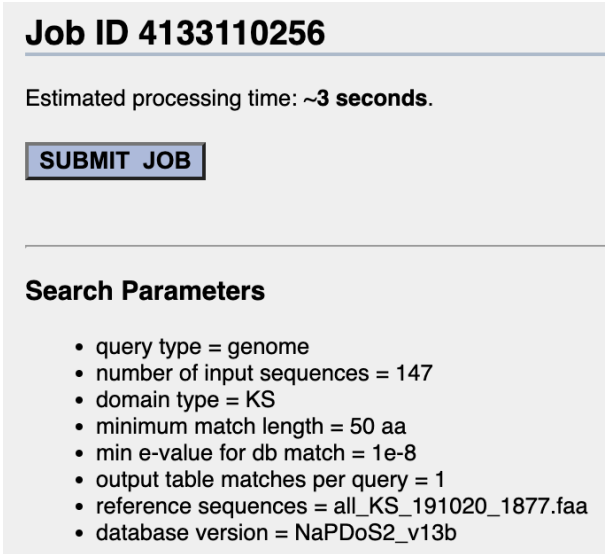
- The PCR sequences are nucleotides, so we will select "**Query type: Nucleic acid sequences**".

- Click "**Choose File**" to upload the FASTA file of sequences. Be sure to select the correct one depending if you selected "Query type" amino acid or nucleic acid sequences.

- Click on "**Advanced Settings**". Here, you can change the BLASTP e-value cutoff and the Minimum alignment length. For this tutorial, we will use the settings of BLASTP e-value cutoff of 1e-8 and Minimum alignment length of **50aa (150nt)**. **We are using 50aa (150nt) minimum alignment because most of the sequences are shorter PCR products <600nt long; you could increase the alignment length, however some KSs might not be found (ie. too short a query sequence).

    - Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details.

- Click **SEEK** to be assigned a job ID and review the job analysis settings.



- For the "Query type: nucleic acid sequences", we will first proceed to the "**Nucleotide Sequence Search Results**" page which lists the initial number of KS domains that NaPDoS2 identified. On this page, you can **DOWNLOAD** (click link to view; or right-click link & select "Save Link As" to save) a table of the nucleotide match locations in tab-delimited format.
    - This lists the domains numbered KS# ("**domain id**"), the sequence/contig name the domain was found on ("**parents seq**"), the base pair/nucleotide start and end location of the domain ("**start**" , "**end**"), the translation reading frame ("**read frame**"), and the final name of the domain ("**cand_id**").
    - This table can be used to find the location of the KS domains in your nucleic acid genome file via the starting and ending base pairs.
- To continue the analysis, click "**CONTINUE ANALYSIS**"

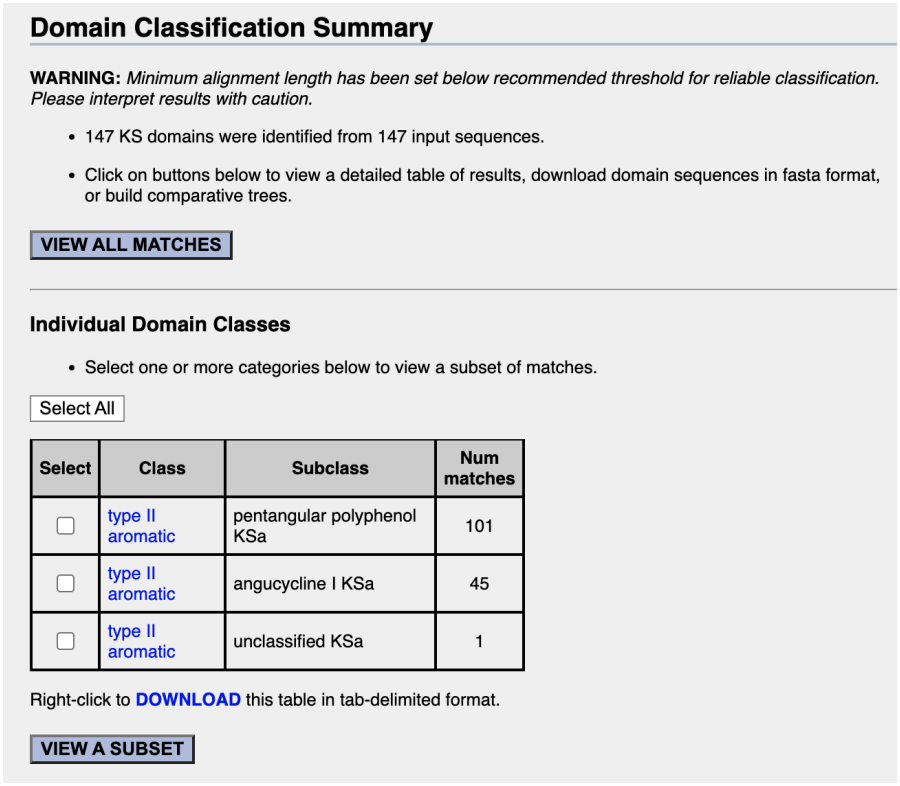
3. Select results display options for KSs that NaPDoS2 detected and classified

- The "**Domain Classification Summary**" page lists the total number of domains found from the number of sequences that were input. Clicking "**VIEW ALL MATCHES**" opens a new page with the hit table of all domains identified in the genome analysis.
    - In our example, 147 KS domains were found from all 147 nucleic acid PCR sequences. They are all type 2 KS types, which is what we expected from the PCR design of the study.

- The "**Individual Domain Classes**" lists the number of specific classes of KS domains that were found in your query genome sequence, ordered by the class and subclass with the most number of domains.

    - This is helpful if you only are interested in a specific subset type of domain– you can click to select what class/subclasses you would like to view the results of and click "**VIEW A SUBSET**", which will open a new page with a table of only the classes/subclasses you selected.

    - You can also "**DOWNLOAD**" the "**Individual Domain Classes**" table (click link to view; or right-click link & select "Save Link As" to save the table).



Domain Classification Summary

WARNING: *Minimum alignment length has been set below recommended threshold for reliable classification. Please interpret results with caution.*

- 147 KS domains were identified from 147 input sequences.
- Click on buttons below to view a detailed table of results, download domain sequences in fasta format, or build comparative trees.

VIEW ALL MATCHES

Individual Domain Classes

- Select one or more categories below to view a subset of matches.

Select All

Select	Class	Subclass	Num matches
☐	type II aromatic	pentangular polyphenol KSa	101
☐	type II aromatic	angucycline I KSa	45
☐	type II aromatic	unclassified KSa	1

Right-click to DOWNLOAD this table in tab-delimited format.

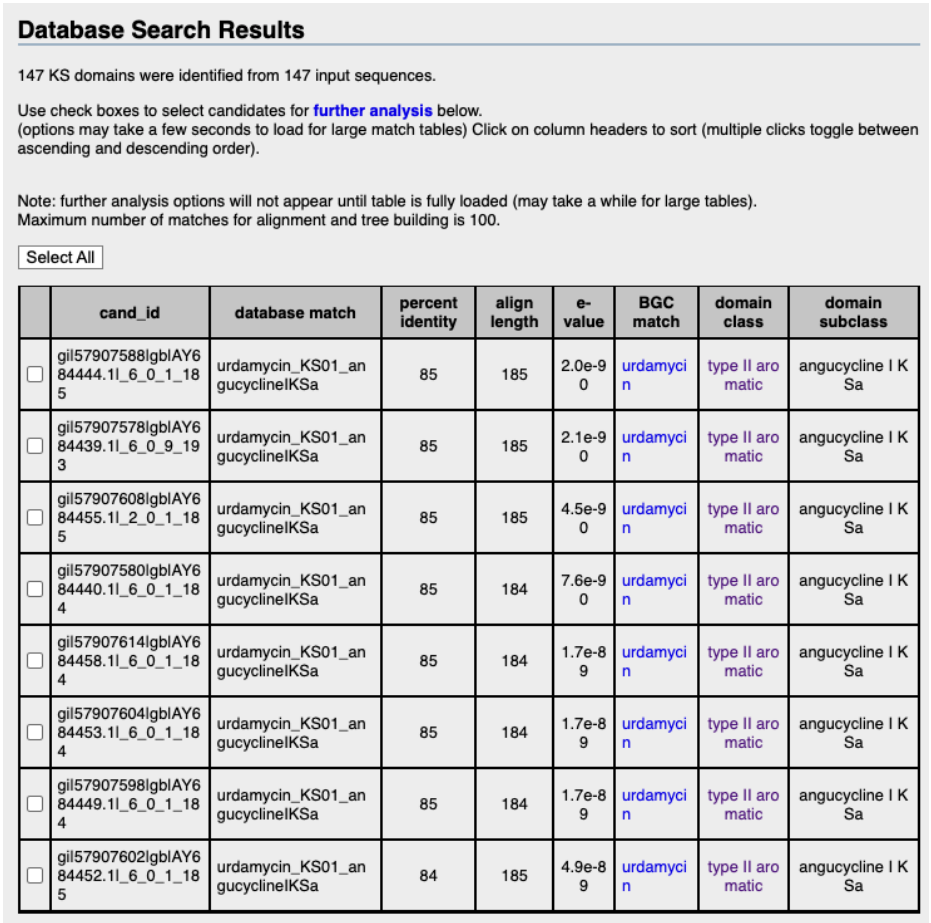
VIEW A SUBSET

4. Assess and download NaPDoS2 search results

- If you clicked "**VIEW ALL MATCHES**" or "**VIEW A SUBSET**", you will view the "**Database Search Results**" page, either for all hits in your query sequence or for a subset of hits of certain class/subclass that you selected.

- The table can be sorted by each of the column values by clicking on the column header; the table scrolls up and down and if your query hit names are long, the table scrolls left to right. You can **DOWNLOAD** the results table (click link to view; or right-click link & select "Save Link As" to save) in a tab-delimited format.

- The scrollable table lists the following information:
    - **cand_id**: name of the domain hit, which is the name of the sequence in the query FASTA file before spaces and appended with location information of where the domain hit was found (see below for details)
    - **database match**: the name of the domain in the NaPDoS2 database that the query hit is closest to
        - **names include**: BGCname_KS#_uniquetag
        - **BGC name**: name of the BGC
        - **KS#**: number KS in successive order from the BGC
        - **unique tag**: abbreviation that is unique for every class/subclassification
    - **percent identity:** percent identity that the query hit shares with the closets NaPDoS2 database match
    - **align length:** alignment length that the query hit shares with the closest NaPDoS2 database match
    - **e-value:** e-value of the query hit to the closest NaPDoS2 database match
    - **BGC match**: name of the BGC of the closest NaPDoS2 database match to the query hit; click on the link to open a new page of the BGC card information
    - **domain class:** classification of the closest NaPDoS2 database match to the query hit
        - This is the ***classification*** of your query hit
    - **domain subclass:** specific subclass of the closest NaPDoS2 database match to the query hit
        - This is the ***complete specific subclass*** of your query hit (some classes do not have subclasses, ie “no subclass”)



Database Search Results

147 KS domains were identified from 147 input sequences.

Use check boxes to select candidates for **further analysis** below.
(options may take a few seconds to load for large match tables) Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Note: further analysis options will not appear until table is fully loaded (may take a while for large tables).
Maximum number of matches for alignment and tree building is 100.

Select All

	cand_id	database match	percent identity	align length	e-value	BGC match	domain class	domain subclass
☐	gil57907588lgblAY684444.1l_6_0_1_185	urdamycin_KS01_angucyclineIKSa	85	185	2.0e-90	urdamycin	type II aromatic	angucycline I KSa
☐	gil57907578lgblAY684439.1l_6_0_9_193	urdamycin_KS01_angucyclineIKSa	85	185	2.1e-90	urdamycin	type II aromatic	angucycline I KSa
☐	gil57907608lgblAY684455.1l_2_0_1_185	urdamycin_KS01_angucyclineIKSa	85	185	4.5e-90	urdamycin	type II aromatic	angucycline I KSa
☐	gil57907580lgblAY684440.1l_6_0_1_184	urdamycin_KS01_angucyclineIKSa	85	184	7.6e-90	urdamycin	type II aromatic	angucycline I KSa
☐	gil57907614lgblAY684458.1l_6_0_1_184	urdamycin_KS01_angucyclineIKSa	85	184	1.7e-89	urdamycin	type II aromatic	angucycline I KSa
☐	gil57907604lgblAY684453.1l_6_0_1_184	urdamycin_KS01_angucyclineIKSa	85	184	1.7e-89	urdamycin	type II aromatic	angucycline I KSa
☐	gil57907598lgblAY684449.1l_6_0_1_184	urdamycin_KS01_angucyclineIKSa	85	184	1.7e-89	urdamycin	type II aromatic	angucycline I KSa
☐	gil57907602lgblAY684452.1l_6_0_1_185	urdamycin_KS01_angucyclineIKSa	84	185	4.9e-89	urdamycin	type II aromatic	angucycline I KSa

- Here, you can also select specific domain hits to download or further analyze (select all to download or analyze all results). You have the follow options for downloading the data:
    - **View nucleotide coordinates** for all trimmed domain candidates
        - **only an option if your query sequence was a nucleic acid file
        - This shows a table where each potential domain candidate has been given a unique candidate ID number, based on parent sequence id, reading frame number (1-6), gene number within the reading frame, and trimmed nucleotide start and stop coordinates within the reading frame.
        - The table columns are sortable by each column header. You can **DOWNLOAD** the table of nucleotide match locations in tab-delimited format (click link to view; or right-click link & select “Save Link As” to save)

- **Output selected sequences** in fasta format
    - Select which trimmed query domain hits or select all to output in FASTA format
        - **if your query was a nucleic acid, you have the option to output the query domain hits as either amino acid or nucleic acid
        - **if your query was an amino acid protein, you only have the option to output the query domain hits as amino acid sequences
    - Click **GET TRIMMED SEQS**
    - This is a very useful output for you to continue working with the KS or C domains that NaPDoS2 detected & classified. You can copy and paste the trimmed query KS or C domain hits into a new file, a sequencing editing program, etc.
    - For interpretation of the appended information to your query search (ie. >ProteinA_13_429 or >GeneB_5_1323_56_472), please see below.
- **Output Alignment** with closest database matches
    - Select alignment format (all alignments are created with MUSCLE but can be output in different formats suitable for many downstream applications):
        - FASTA (.fasta)
        - Msf (.msf)
        - Clustalw (.clw)
    - This outputs an alignment of your selected sequences with the closest NaPDoS2 database matches. Note, the alignment output will include all closest NaPDoS2 database matches to all query hits, not just your selected sequences, but only your selected sequences will be present in the alignment.
    - You can **DOWNLOAD** (click link to view; or right-click link & select “Save Link As” to save) the alignment file for downstream use in sequence programs, etc.
- **Construct tree** (candidate domains + blast matches + reference domains)

- Select the query domain hits that you would like to build a tree with. You can either build a tree with:
    - **Closest database matches** to your selected sequences, ie. # of selected queries + closest database references
    - **All classified KS domains** which would be the # of selected queries + all 414 NaPDoS2 database reference sequences.

- The estimated time to calculate the trees is given, usually the time to complete the trees is much faster than estimated.

- If you wanted to build a tree with these sequences the **Tree results** page has two options for viewing the calculated trees:

    - Tree image (SVG format file), click **DOWNLOAD** to view, right-click to save.
        - When viewing the tree, the query domain hits will have red circles of where they are placed in the tree.

    - Newick format (plain text file), click **DOWNLOAD** to view, right-click to save
        - This file format can be viewed in tree viewing programs like FigTree, iTol, TreeViewer, ETE Toolkit, etc.

- Our goal of this tutorial was to analyze KS PCR sequences from soil eDNA *Wawrik et al. 2005 AEM*. ***Interpreting the results:*** When the original study was performed, the KSs could only be identified as type 2 KS similar to a couple of known type 2 aromatic BGC products. Now with NaPDoS2, each sequence could now be classified to a specific subclass and larger comparisons can be performed (top database matches, phylogenetic tree calculation) with other related KS sequences.

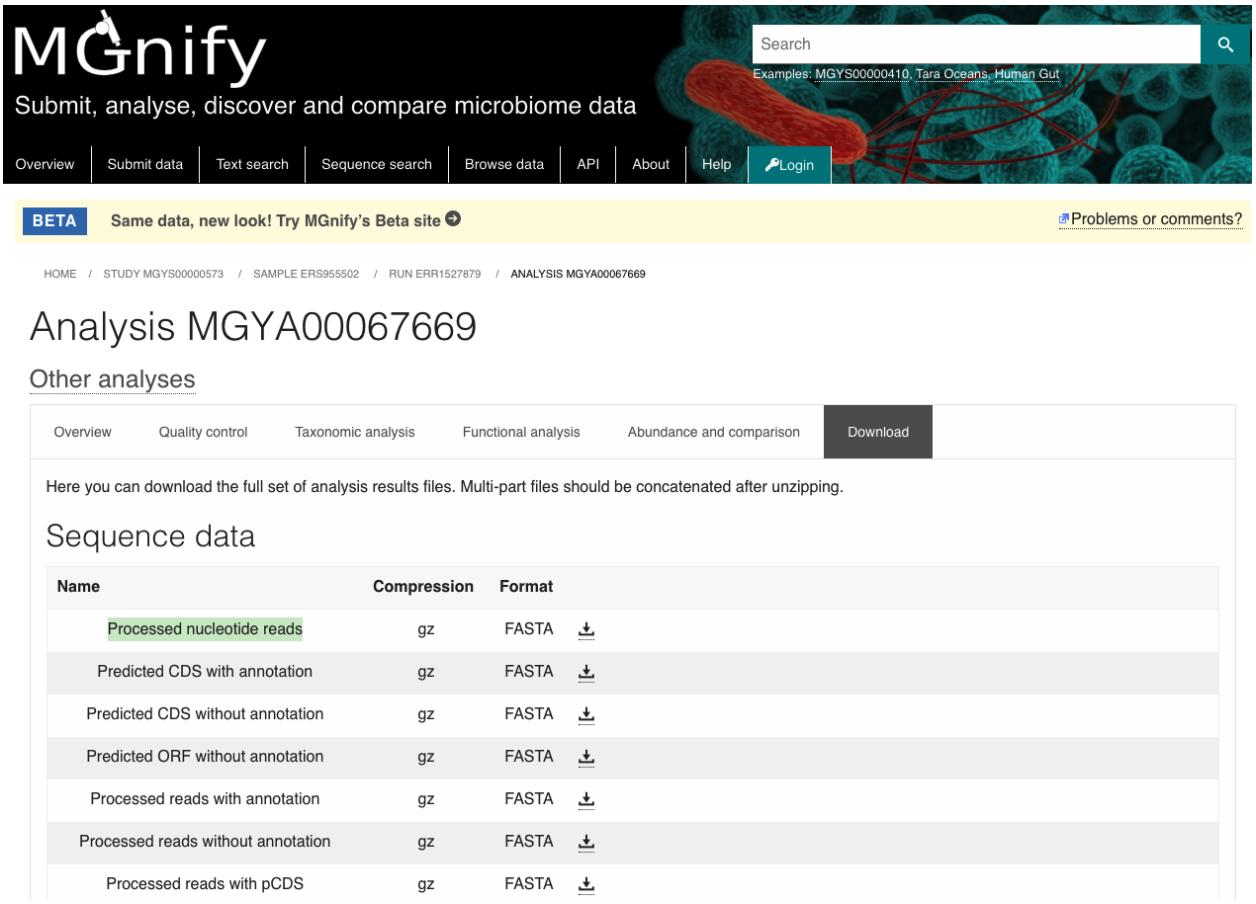
7. Analyzing KS amplicon "tag" sequences with NaPDoS2

Objective: Analyze targeted KS amplicon "tag" sequencing data to identify & classify the distribution of KSs in a sample.

1. **Gather the sequences**

- We'll download a KS PCR amplicon dataset from MGnify. These sequences were deposited from *Borsetto et al. 2019 Microbiome* study exploring the microbial and KS polyketide diversity of 13 soils. For this tutorial, we will analyze the "**S31 Antarctic Mars Oasis"** KS PCR soil sample– sample accession **ERR1527879**.
    - We have also provided the KS amplicon dataset used in the tutorial on the NaPDoS2 Tutorial OSF page that you can download directly from there
    - On MGnify, search "**ERR1527879**" in the search bar. Click on "**Analyses (1)**". Click on the **Analysis MGYA00067669**. Click on the "**Download**" tab– we will download the "**Processed nucleotide reads**" which will be in ".gz" zipped FASTA format. Click the download icon to start the download (size: zipped file 1.1MB, unzipped 9.6MB).

Amplicon KS PCR dataset	**Accession**	**Reference**
S31 Antarctica Mars Oasis soil type II KS PCR sequences	S31 Antarctica: ERR1527879 NCBI BioProject PRJEB11689	*Borsetto et al. 2019 Microbiome*



- Save the downloaded zipped file ("**ERR1527879_MERGED_FASTQ.fasta.gz**") in a place you can access it (for example, your Desktop or Downloads folder). Navigate to where you saved it and **unzip the file**.
    - (Windows: double-click, or select it and click "Extract all"; Mac: double-click; Linux/Command line: navigate to the directory containing your zipped file and use the command "`gunzip -k filename.gz`")
    - The **ERR1527879_MERGED_FASTQ.fasta** file contains 34,600 nucleic acid KS PCR sequences (the number of ">" in the FASTA file, see here (section "Combine multiple genome files") for how to count sequences in a FASTA file.

2. Submit a NaPDoS2 analysis of the KS PCR sequence dataset.

- Navigate to the "Run Analysis" tab on NaPDoS2. Select "**Domain Type: KS domains**".

- Because the **ERR1527879** sample contains nucleic acid FASTA sequences, select "**Query type: Nucleic acid sequences**".

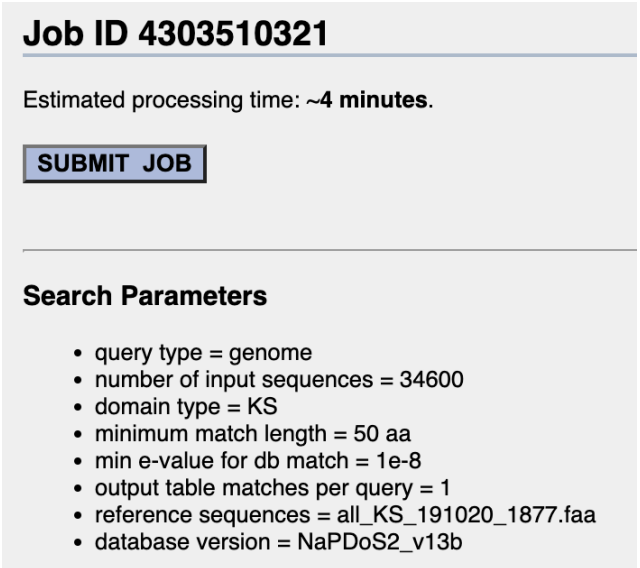
- Click "**Choose File**" to upload the KS PCR amplicon dataset file. NaPDoS2 cannot accept zipped files.

- Click on "**Advanced Settings**". Here, you can change the BLASTP e-value cutoff and the Minimum alignment length. For this tutorial, we will use the settings of BLASTP e-value cutoff of 1e-8 and Minimum alignment length of **50aa (150nt)**. ****We are using 50aa (150nt) minimum alignment because most of the sequences are shorter PCR products <600nt long; you could increase the alignment length, however some KSs might not be found (ie. too short a query sequence)

    - Caution: selecting less rigorous match criteria may decrease reliability of results. See below and the NaPDoS2 publication for more details. However, for shorter KS PCR sequences, evaluating what minimum alignment length captures the most amount of correct sequences should be considered.

- Click **SEEK** to be assigned a job ID and review the job analysis settings.

- NaPDoS2 has a job limit of 500,000 sequences or 500MB file size. If your KS PCR amplicon sequence file is too large, we provide easy-to-use-scripts to help split your file into smaller files. See section below.



- For our "**Query type: nucleic acid sequences**", we first proceed to the "**Nucleotide Sequence Search Results**" page which lists the initial number of KS domains that NaPDoS2 identified.

- To continue the analysis, click "**CONTINUE ANALYSIS**"

3. Select results display options for KSs that NaPDoS2 detected and classified

- The "**Domain Classification Summary**" page lists the total number of domains found from the number of sequences that were input. Clicking "**VIEW ALL MATCHES**" opens a new page with the hit table of all domains identified in the genome analysis.

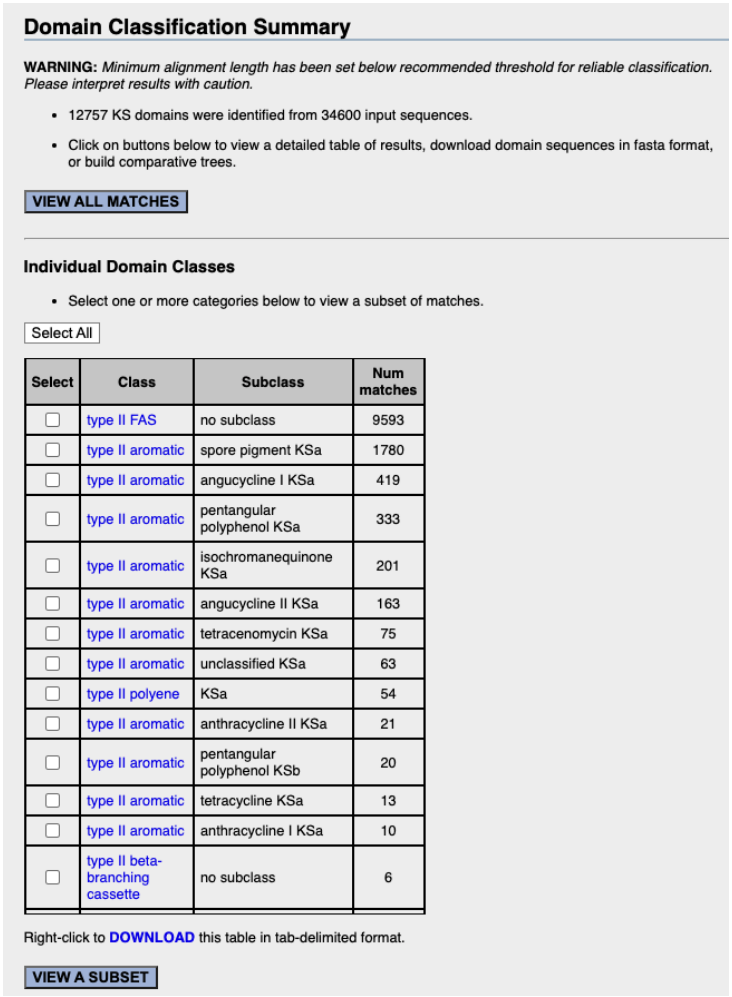
    - In our example, 12,757 KS domains were found from all 34,600 KS PCR sequences (~37% detection & identification of KS in the total sample at 50aa minimum alignment length). They are all type 2 KS types, which is what we expected from the PCR design of the study.

- The "**Individual Domain Classes**" lists the number of specific classes of KS domains that were found in your query genome sequence, ordered by the class and subclass with the most number of domains.

    - This is helpful if you only are interested in a specific subset type of domain– you can click to select what class/subclasses you would like to view the results of and click "**VIEW A SUBSET**", which will open a new page with a table of only the classes/subclasses you selected.

        - In our example, the largest proportion of KS PCR sequences in this sample were type II FAS KS domains (n= 9,593), followed by type II aromatic spore pigment KSa domains (n= 1,780).

    - You can also "**DOWNLOAD**" the "**Individual Domain Classes**" table (click link to view; or right-click link & select "Save Link As" to save the table).



Domain Classification Summary

WARNING: *Minimum alignment length has been set below recommended threshold for reliable classification. Please interpret results with caution.*

- 12757 KS domains were identified from 34600 input sequences.
- Click on buttons below to view a detailed table of results, download domain sequences in fasta format, or build comparative trees.

VIEW ALL MATCHES

Individual Domain Classes

- Select one or more categories below to view a subset of matches.

Select All

Select	Class	Subclass	Num matches
☐	type II FAS	no subclass	9593
☐	type II aromatic	spore pigment KSa	1780
☐	type II aromatic	angucycline I KSa	419
☐	type II aromatic	pentangular polyphenol KSa	333
☐	type II aromatic	isochromanequinone KSa	201
☐	type II aromatic	angucycline II KSa	163
☐	type II aromatic	tetracenomycin KSa	75
☐	type II aromatic	unclassified KSa	63
☐	type II polyene	KSa	54
☐	type II aromatic	anthracycline II KSa	21
☐	type II aromatic	pentangular polyphenol KSb	20
☐	type II aromatic	tetracycline KSa	13
☐	type II aromatic	anthracycline I KSa	10
☐	type II beta-branching cassette	no subclass	6

Right-click to **DOWNLOAD** this table in tab-delimited format.

VIEW A SUBSET

4. Assess and download NaPDoS2 search results

- If you clicked "**VIEW ALL MATCHES**" or "**VIEW A SUBSET**", you will view the "**Database Search Results**" page, either for all hits in your query sequence or for a subset of hits of certain class/subclass that you selected.
    - ** In this example, it might take a little bit longer for all 12,757 results to load when you click to view all matches. We've experienced that with lots of results, loading the full results table may fail.
    - ** We ***highly*** recommend when working with amplicon sequences with lots of KS query domain hits, **DOWNLOAD** the **Database Search results** table as a tab delimited file (*.tab), which then you can import into R, Microsoft Excel, or Google Sheets for further parsing & analysis (ie. splitting the results by sample and counting the number of hits/class of hits per sample). Or, only view a couple of subclasses ("**View a Subset**") at a time.

- The table can be sorted by each of the column values by clicking on the column header; the table scrolls up and down and if your query hit names are long, the table scrolls left to right. You can **DOWNLOAD** the results table (click link to view; or right-click link & select “Save Link As” to save) in a tab-delimited format.

- The scrollable table lists the following information:

    - **cand_id**: name of the domain hit, which is the name of the sequence in the query FASTA file before spaces and appended with location information of where the domain hit was found (see below for details)

    - **database match**: the name of the domain in the NaPDoS2 database that the query hit is closest to
        - **names include**: BGCname_KS#_uniquetag
        - **BGC name**: name of the BGC
        - **KS#**: number KS in successive order from the BGC
        - **unique tag**: abbreviation that is unique for every class/subclassification

    - **percent identity:** percent identity that the query hit shares with the closets NaPDoS2 database match

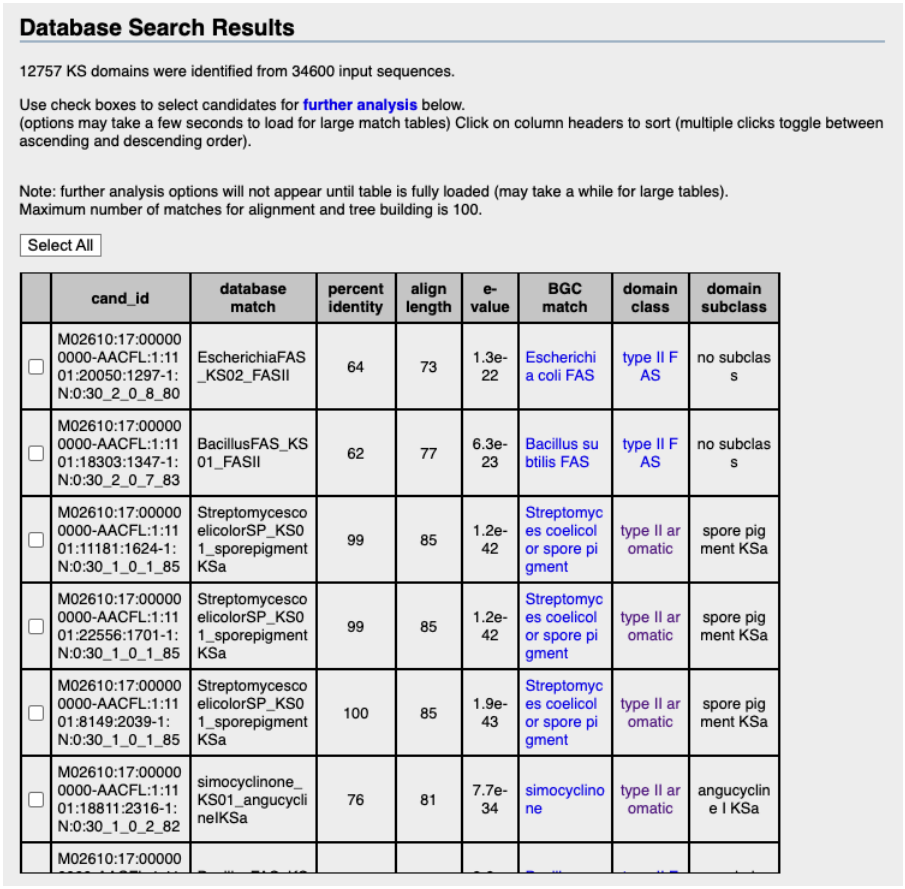
    - **align length:** alignment length that the query hit shares with the closest NaPDoS2 database match

    - **e-value:** e-value of the query hit to the closest NaPDoS2 database match

    - **BGC match**: name of the BGC of the closest NaPDoS2 database match to the query hit; click on the link to open a new page of the BGC card information

    - **domain class:** classification of the closest NaPDoS2 database match to the query hit
        - This is the ***classification*** of your query hit

    - **domain subclass:** specific subclass of the closest NaPDoS2 database match to the query hit
        - This is the ***complete specific subclass*** of your query hit (some classes do not have subclasses, ie “no subclass”)



Database Search Results

12757 KS domains were identified from 34600 input sequences.

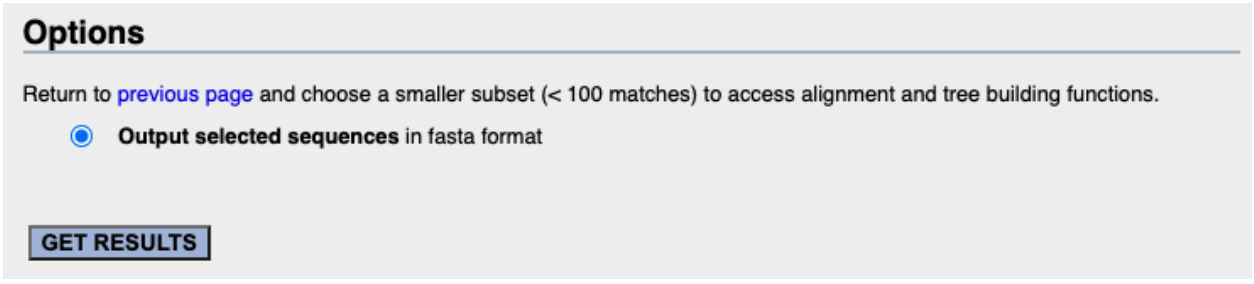
Use check boxes to select candidates for **further analysis** below.
(options may take a few seconds to load for large match tables) Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Note: further analysis options will not appear until table is fully loaded (may take a while for large tables).
Maximum number of matches for alignment and tree building is 100.

Select All

	cand_id	database match	percent identity	align length	e-value	BGC match	domain class	domain subclass
☐	M02610:17:000000000-AACFL:1:1101:20050:1297-1:N:0:30_2_0_8_80	EscherichiaFAS_KS02_FASII	64	73	1.3e-22	Escherichia coli FAS	type II FAS	no subclass
☐	M02610:17:000000000-AACFL:1:1101:18303:1347-1:N:0:30_2_0_7_83	BacillusFAS_KS01_FASII	62	77	6.3e-23	Bacillus subtilis FAS	type II FAS	no subclass
☐	M02610:17:000000000-AACFL:1:1101:11181:1624-1:N:0:30_1_0_1_85	StreptomycescoelicolorSP_KS01_sporepigmentKSa	99	85	1.2e-42	Streptomyces coelicolor spore pigment	type II aromatic	spore pigment KSa
☐	M02610:17:000000000-AACFL:1:1101:22556:1701-1:N:0:30_1_0_1_85	StreptomycescoelicolorSP_KS01_sporepigmentKSa	99	85	1.2e-42	Streptomyces coelicolor spore pigment	type II aromatic	spore pigment KSa
☐	M02610:17:000000000-AACFL:1:1101:8149:2039-1:N:0:30_1_0_1_85	StreptomycescoelicolorSP_KS01_sporepigmentKSa	100	85	1.9e-43	Streptomyces coelicolor spore pigment	type II aromatic	spore pigment KSa
☐	M02610:17:000000000-AACFL:1:1101:18811:2316-1:N:0:30_1_0_2_82	simocyclinone_KS01_angucyclineIKSa	76	81	7.7e-34	simocyclinone	type II aromatic	angucycline I KSa
☐	M02610:17:00000							

- Here, you can also select specific domain hits to download or further analyze (select all to download or analyze all results). You have the follow options for downloading the data:
    - ** Because there are so many results, the analysis options are limited to one option (> 100 hits) or three options (< 100 hits). We recommend re-running the subset of samples through NaPDoS2 if you would like to utilize the other output options (viewing sequence alignments, constructing a tree, etc)



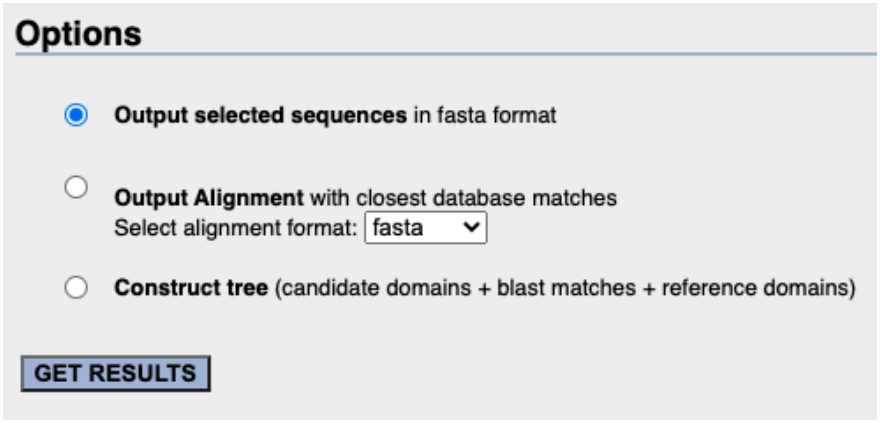
Options

Return to previous page and choose a smaller subset (< 100 matches) to access alignment and tree building functions.

◉ **Output selected sequences** in fasta format

GET RESULTS

One option for additional analyses when you have >100 results.



Three options for additional analyses when you have <100 results.

- **Output selected sequences** in fasta format
    - Select which trimmed query domain hits or select all to output in FASTA format
        - **if your query was a nucleic acid, you have the option to output the query domain hits as either amino acid or nucleic acid
    - Click **GET TRIMMED SEQS**
    - This is a very useful output for you to continue working with the KS or C domains that NaPDoS2 detected & classified. You can copy and paste the trimmed query KS or C domain hits into a new file, a sequencing editing program, etc.
    - For interpretation of the appended information to your query search (ie. >ProteinA_13_429 or >GeneB_5_1323_56_472), please see below.
- **Output Alignment** with closest database matches
    - ** Only an option with < 100 KS query hits; in this example, this is not an option.
    - Select alignment format (all alignments are created with MUSCLE but can be output in different formats suitable for many downstream applications):
        - FASTA (.fasta)
        - Msf (.msf)
        - Clustalw (.clw)

- This outputs an alignment of your selected sequences with the closest NaPDoS2 database matches. Note, the alignment output will include all closest NaPDoS2 database matches to all query hits, not just your selected sequences, but only your selected sequences will be present in the alignment.

- You can **DOWNLOAD** (click link to view; or right-click link & select "Save Link As" to save) the alignment file for downstream use in sequence programs, etc.

- **Construct tree** (candidate domains + blast matches + reference domains)

  - ** Only an option with < 100 KS query hits; in this example, this is not an option.

  - Select the query domain hits that you would like to build a tree with. You can either build a tree with:
    - **Closest database matches** to your selected sequences, ie. # of selected queries + closest database references
    - **All classified KS domains** which would be the # of selected queries + all 414 NaPDoS2 database reference sequences.

  - The estimated time to calculate the trees is given, usually the time to complete the trees is much faster than estimated.

  - If you wanted to build a tree with these sequences the **Tree results** page has two options for viewing the calculated trees:

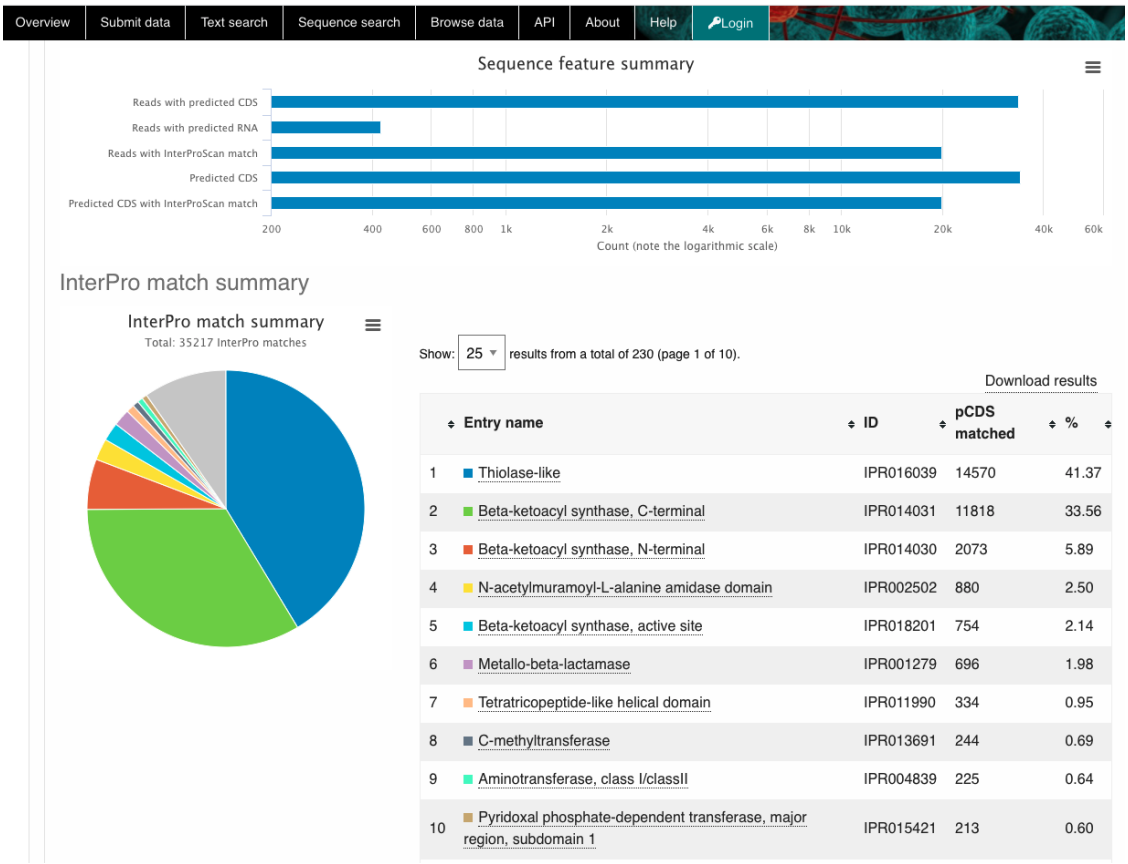
    - Tree image (SVG format file), click **DOWNLOAD** to view, right-click to save.
      - When viewing the tree, the query domain hits will have red circles of where they are placed in the tree.

    - Newick format (plain text file), click **DOWNLOAD** to view, right-click to save.
      - This file format can be viewed in tree viewing programs like FigTree, iTol, TreeViewer, ETE Toolkit, etc.

Our goal of this tutorial was to identify & classify the distribution of KSs in an amplicon tag-sequencing KS PCR sample. ***Interpreting the results:*** When the original study was performed,

the estimated biosynthetic potential was based on if the KS PCR sequences matched any known BGC in the MIBiG repository. However, we see that NaPDoS2 can add to this analysis by helping to assess primer specificity and identify potential non-target KS sequences prior to downstream analyses while also classifying the query hit KS sequences to a specific subclass.

Additionally, a cool feature of the MGnify portal is that they performed a functional analysis on the **S31 ERR1527879** sample. We see that 41% of the sequences were identified as "**Thiolase-like**" (blue) domains by MGnify/InterPro– this thiolase group likely comprises the KS-type of enzymes that are present in type 3 PKSs, not type 1 or 2 PKSs!, and thus, perhaps it makes sense why we see such a low recovery rate of KSs from this sample using NaPDoS2. Still, the KSs we do recover are a wide diversity within the type 2 PKS group.



	Entry name	ID	pCDS matched	%
1	Thiolase-like	IPR016039	14570	41.37
2	Beta-ketoacyl synthase, C-terminal	IPR014031	11818	33.56
3	Beta-ketoacyl synthase, N-terminal	IPR014030	2073	5.89
4	N-acetylmuramoyl-L-alanine amidase domain	IPR002502	880	2.50
5	Beta-ketoacyl synthase, active site	IPR018201	754	2.14
6	Metallo-beta-lactamase	IPR001279	696	1.98
7	Tetratricopeptide-like helical domain	IPR011990	334	0.95
8	C-methyltransferase	IPR013691	244	0.69
9	Aminotransferase, class I/classII	IPR004839	225	0.64
10	Pyridoxal phosphate-dependent transferase, major region, subdomain 1	IPR015421	213	0.60

>> However, we understand that PCR amplicon tag-sequencing studies are sequencing samples deeper and deeper, so if you have a sample that is more than >500,000 sequences / 500MB in size, you will need to split your file before analysis with NaPDoS2. (and in the cases where you have a lot of hits in your sample, you might need to split your sample as large tables of results might not load in a reasonable amount of time)-- see the **next section for File Size Management Tools** for how to do this.

8. File Size Management Tools

Objective: Use the provided file size management tools. This is needed if: 1) your query files contain sequences that aren't the correct length for analysis 2) if your query files are too large for a single NaPDoS2 analysis job 3) you want to take a random sampling approach of a larger dataset to get an initial NaPDoS2 snapshot of a large sequence dataset.

We have included our recommended best practices for file size management on the NaPDoS2 website. Below are some additional details of using the provided scripts, but we do not cover all specific recommended protocols/tools on the website.

Scripts can be downloaded from: 1) File size management page or 2) NaPDoS2 GitHub

If your query files are too large for NaPDoS website submission, several options are available for pre-filtering and/or subdividing the data into smaller batches on your own computer.

1. **Assembly**

- Short sequencing reads such as those obtained using Illumina technology (typically 150 nucleotides or shorter) cannot be used for NaPDoS classification unless they are first assembled into longer contigs, with a program such as Spades or IDBA-UD. This is because short query sequences do not provide enough information to determine whether or not they match PKS and NRPS reference domains, which are typically about 1250 nucleotides (425 amino acids) long. Query sequences should ideally cover at least half of that length.

2. **Size Filtering**

- Assembled contigs often include sequences that are too short to allow detection of PKS and NRPS domains. The size_limit_seqs.pl perl script can be downloaded (by clicking the link; GitHub link) and used on the unix command line on your local computer to pre-filter a FASTA file of contig sequences according to size, using the unix command line.

  `size_limit_seqs.pl`
  Usage: `./size_limit_seqs.pl fasta_filename minimum_seq_length`
  -Selects FASTA sequences of a specified minimum length and puts them in a new FASTA formatted file.
  Input: a file containing multiple FASTA format sequences;
  minimum sequence size to select
  Output: selected records to a new file
  -Notes: suitable for really large (gigabyte) FASTA source files and large numbers of retrieval sequences

Example commands:

```
gunzip size_limit_seqs.pl.gz

chmod 755 size_limit_seqs.pl

./size_limit_seqs.pl fasta_filename minimum_seq_length > out_filename
```

3. **File Splitting**

- Files can be split up into multiple parts by downloading the serialize_seqs.pl perl script (by clicking the link; GitHub link) and running it from the unix command line on your local computer.

  serialize_seqs.pl
  Usage: `./serialize_seqs.pl fasta_filename num_seqs_per_subfile`
  -Breaks a file containing multiple FASTA format sequences into sequential subfile chunks, each containing a specified number of the original sequences.
  -Output: Writes each chunk to output file(s) in FASTA format.
  -Notes: suitable for really large (multi-gigabyte) FASTA source files

  Example commands:

```
gunzip serialize_seqs.pl.gz

chmod 755 serialize_seqs.pl

./serialize_seqs.pl fasta_filename num_seqs_per_subfile
```

  Example commands using the KS PCR amplicon dataset from above:

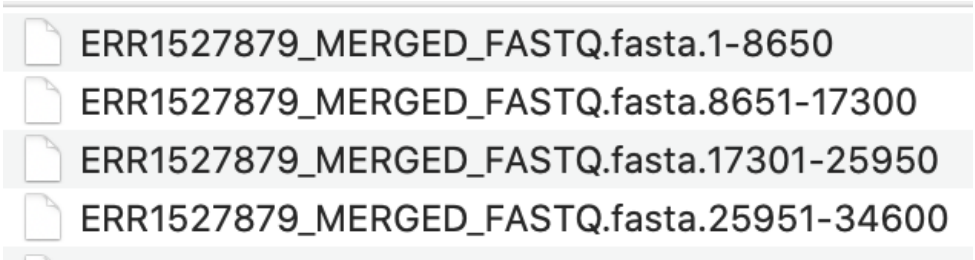
```
gunzip serialize_seqs.pl.gz

chmod 755 serialize_seqs.pl

./serialize_seqs.pl ERR1527879_MERGED_FASTQ.fasta 8650
```

```
Number of sequences/subfile = 8650
Total number of subfiles = 4
input file size = 9.6 MB
Finished processing 34600 sequences
```

>this resulted in 4 files:



- We have also provided these split KS amplicon sequence files that were the product of this script on the NaPDoS2 Tutorial OSF page. The complete KS amplicon sequence file can be found on the same page in the previous amplicon dataset tutorial folder.

4. De-replication

- Identical or nearly-identical sequences can be consolidated using the CD-hit program. This approach can be particularly helpful for PCR or transcriptome data sets with high levels of sequence replication.

5. Amino acid translation

- Nucleic acid sequence files can be pre-translated into amino acid sequence files, which will often be smaller in total size, using programs such as Prokka.

6. Preliminary BLASTX filtering

- Assembled contigs can be pre-selected based on a unix command-line Diamond BLASTX search against the downloadable domain reference sequences at the bottom of the NaPDoS2 BGCs page. Candidate sequences identified by the pre-filtering blast step can be selected on the unix command line with the perl script getseq_multiple.pl (download by clicking on link; GitHub link).

Example commands:

```
diamond makedb --in all_KS_191020_1877.faa -d
all_KS_191020_1877.dmnd -t temp_directory

diamond blastx -d all_KS_191020_1877.dmnd -q
orig_query_filename -e 1e-5 -p num_processors --max-target-
seqs 1 -t temp_directory -o results_filename

cut -f1 results_filename > candidate_query_ids
```

```
gunzip getseq_multiple.pl.gz

chmod 755 getseq_multiple.pl
./getseq_multiple.pl candidate_query_ids
orig_query_filename > selected_candidate_sequences.fasta
```

- This could be useful for NaPDoS2 targeted searches in large genomes, metagenomes, and long-read amplicon sequences.

7. Random sampling

- If you want to take a random sampling approach of a larger dataset to get an initial NaPDoS2 snapshot of a large sequence dataset, we provide scripts to do that. This type of analysis was used for a large KS PCR amplicon dataset analysis in the NaPDoS2 publication, reported in ***Table S12*** and ***Figure S13***.
- Scripts needed:
  - *To download from GitHub, click on the file below, click on "**Raw**", then "**File > Save Page As**".*
  - *Don't forget to:*
    ```
    gunzip script.pl
    chmod 755 script.pl
    ```

get_seq_info.pl
Usage: `./get_seq_info.pl fasta_filename`
-Input: Takes in a sequence file in FASTA format
-Output: outputs to STDOUT a tab-delimited table:
sequence_id seq_length annotation

randomize_lines.pl
Usage: `./randomize_lines.pl filename`
-Input: takes in a list of lines
-Function: randomizes order for output (to STDOUT)
-Notes: suitable for really long lists.

serialize_large_list.pl
Usage: `./serialize_large_list.pl input_filename num_lines_per_subfile`
-Inpu: Breaks a file containing sequential subfile chunks each containing a specified number of lines.
-Output: writes each chunk to output file
-Notes: suitable for really large (gigabyte) source files

getseq_multiple.pl
Usage: `./getseq_multiple.pl id_list fasta_filename`

-Input: a file containing multiple FASTA format sequences;
a list of accession numbers to retrieve (newline separated).
-Output: selected records to STDOUT
-Notes: suitable for really large (gigabyte) FASTA source files and large numbers of retrieval sequences

- Command steps to create a random subset of X number of sequences:
  - *We'll use the KS PCR amplicon dataset from above ("**ERR1527879_MERGED_FASTQ.fasta**") and split it into 5,0000 random sequences. If you wanted all 34,600 sequences randomized, you would repeat the "./getseq_multiple.pl" last command with each ERR1527879_random.txt1-#### files.*

```
./get_seq_info.pl ERR1527879_MERGED_FASTQ.fasta | cut -f1 >
ERR1527879_seqinfo.txt

./randomize_lines.pl ERR1527879_seqinfo.txt >
ERR1527879_random.txt

./serialize_large_list.pl ERR1527879_random.txt 5000

./getseq_multiple.pl ERR1527879_random.txt.1-5000
ERR1527879_MERGED_FASTQ.fasta > ERR1527879_random.1-
5000.fasta
```

- **Other provided scripts**: referenced in the NaPDoS2 publication for creation of sliding window subsequences for ***Figure 7, Figure S11***.

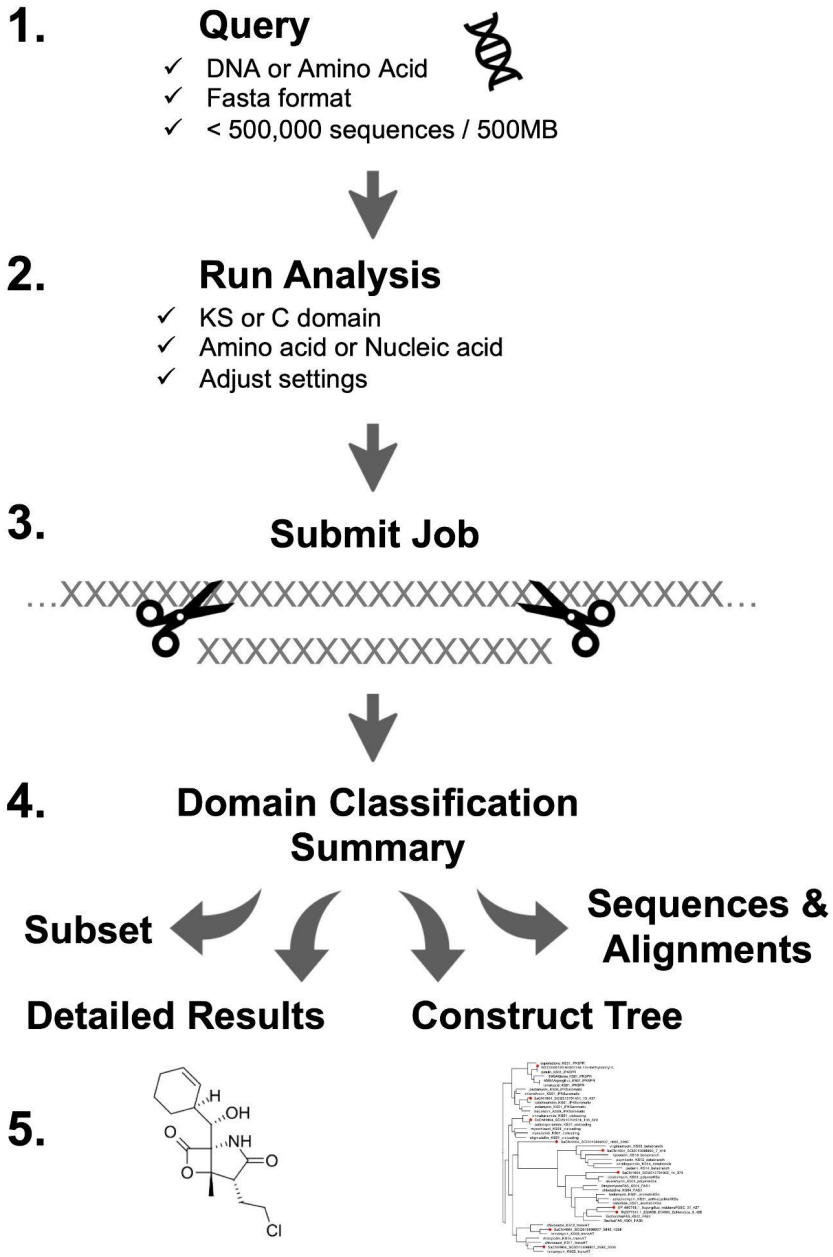
sequence_subdivider.pl
Usage: `./sequence_subdivider.pl <options> fasta_filenames`
Options:
-p prefix (default = "CNTRL")
-n subsequence length (default = 500 residues)
-d coverage depth (default = 1X)
-o overlap length (default = 50 nt)
-Generation of sliding window subsequences

>to download from GitHub, navigate to file, click on "Raw", then "File > Save Page As". Don't forget to:

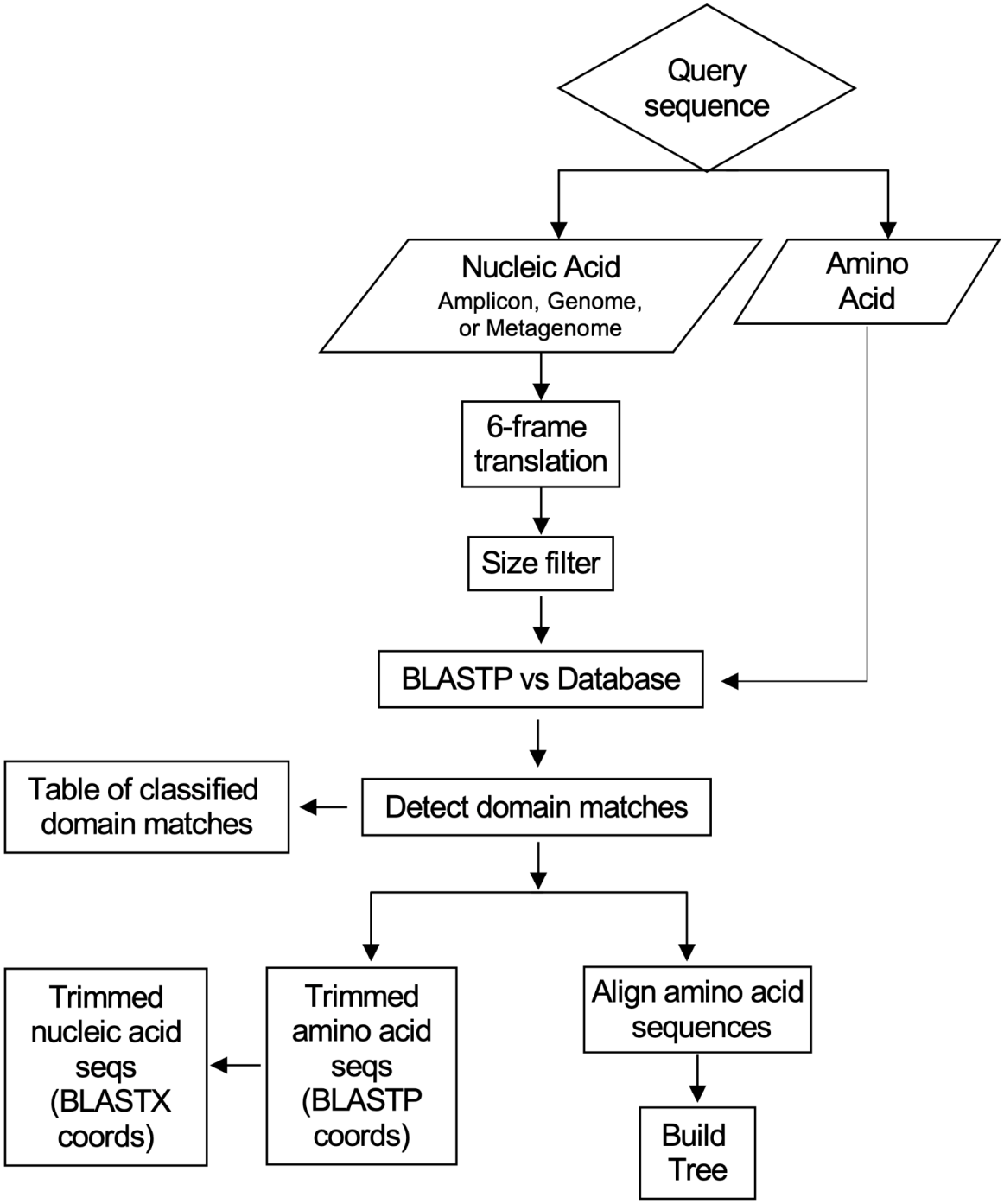
```
gunzip sequence_subdivider.pl
chmod 755 sequence_subdivider.pl
```

The NaPDoS2 Analysis Workflow Details

The NaPDoS2 webtool workflow is shown in the following diagram. This outlines the steps of a NaPDoS2 analysis that users interact with directly.



Behind the scenes, the NaPDoS2 bioinformatic pipeline is shown in the following diagram. The web interface to this pipeline is divided into four consecutive steps. The following sections describe details of each of these steps.



NaPDoS2 Web Interface Steps: Outline

1. **Preliminary Candidate Screening**
2. **BLASTP search results output**
    1. Annotation of appended information to query hit results
3. **Tree construction**
4. **Interpretation of Results**
    1. BLASTP E-value cutoff
    2. Sensitivity and accuracy versus alignment length
5. **Further analyses beyond NaPDoS2**

1. Preliminary Candidate Screening

- We have explained in detail the initial analysis steps in the tutorial sections above, please see the Quick Start and other tutorials above.

- A NaPDoS2 analysis can include either 1) select a domain type (KS or C) and 2) query type (amino acid or nucleic acid).
    - Nucleic acid query sequences are translated into predicted amino acids using all 6 possible reading frames.

- Default parameters can be adjusted by clicking on the **Advanced Settings** link, but caution should be used in interpreting results obtained with modified parameters.
    - Less stringent BLAST selection criteria may increase false positives.
    - Shorter minimum alignment lengths can decrease sensitivity and classification accuracy, especially for novel domains without exact matches in the reference database.
    - Quantitative validation results supporting these caveats are presented below.

2. BLASTP search results output

- A BLAST search is performed against curated reference database examples to identify matches to known PKS/NRPS/PKS-NRPS BGCs. Some suggested guidelines for interpreting BLAST scores are presented below.

- **Annotation of appended information to query hit results**

    - All query domain hits sequences are given a "cand_id" candidate ID. This includes appended information to the name of the sequence, ie. added on to the end of the FASTA file header description.

    - **Amino acid**: >ProteinA_13_429
        - "ProteinA": name of the FASTA sequence a domain was found in; the name is from the FASTA description line (>description_up_until_first_space) up until the first space. Avoid the use of special characters that are not allowed when constructing a tree.
        - "13": amino acid residue START of the query domain hit
        - "429": amino acid resident END of the query domain hit

        - The appended numbers are the range of amino acid residue positions from your original query sequence (ie. if there were multiple KS domains found within a larger query then the positions would be the KS domains within that sequence).

        - When you select this sequence and "**Output selected sequences in FASTA format**", it will output a FASTA file with this header:
          >ProteinA_13_429 ProteinA [positions 13-429]
          RIAIVGMANRYPDATSPGELWENAVSGRRAFRRLPDVR…

    - **Nucleic Acid**: >GeneB_5_1323_56_472
        - "GeneB": name of the FASTA sequence a domain was found in; the name is from the FASTA description line (>description_up_until_first_space) up until the first space. Avoid the use of special characters that are not allowed when constructing a tree.
        - "5": reading frame number (1-6) from query translation
        - "1323": gene number within the translated reading frame
        - "56": translated amino acid residue START of the query domain hit; if you want the nucleic acid residue START, output the trimmed sequence as nucleic acid FASTA file
        - "472": translated amino acid resident END of the query domain hit; if you want the nucleic acid residue END, output the trimmed sequence as nucleic acid FASTA file

        - Each potential domain candidate has been given a unique candidate ID number, based on parent sequence id, reading frame number (1-6), gene number within the reading frame, and trimmed nucleotide start and stop coordinates within the reading frame.

        - When you select this sequence and "**Output selected sequences in FASTA format**", it will output a FASTA file with this header:

            - *Amino acid trimmed output:*
              >GeneB_5_1323_56_472 GeneB_5_1323 [positions 56-472]
              DVAIIGASCRFPGARNKEQYWDNLLHGREGVTFYAKDEIEVDE…

            - *Nucleic acid trimmed output:*
              >GeneB_5_1323_56_472 GeneB [positions 1151792-1153042, reverse complement]
              GACGTCGCCATCATCGGCGCCTCGTGCAGATTCCCTGGCGCAC…

- To proceed with further analysis, one or more candidate sequences must be selected using check boxes. Three, or four (nucleotide queries), different output options are available as described in the tutorials above.

    - **View nucleotide coordinates** for all trimmed domain candidates

        - **only an option if your query sequence was a nucleic acid file

        - This shows a table where each potential domain candidate has been given a unique candidate ID number, based on parent sequence id, reading frame number (1-6), gene number within the reading frame, and trimmed nucleotide start and stop coordinates within the reading frame.

        - **Limits**: Due to computational limitations, if the initial number of domain matches detected exceeds 2000, the nucleotide sequence output will be untrimmed, regardless of the number selected for further analyses, thus this table/option will not be shown.

        - The table columns are sortable by each column header. You can **DOWNLOAD** the table of nucleotide match locations in tab-delimited format (click link to view; or right-click link & select "Save Link As" to save)

    - **Output selected sequences** in FASTA format

        - Select which trimmed query domain hits or select all to output in FASTA format

            - **if your query was a nucleic acid, you have the option to output the query domain hits as either amino acid or nucleic acid

        - **if your query was an amino acid protein, you only have the option to output the query domain hits as amino acid sequences

    - **Limits**: Due to computational limitations, if the initial number of domain matches detected exceeds 2000, nucleotide sequence output will be untrimmed, regardless of the number selected for further analyses.

    - Click **GET TRIMMED SEQS**

    - This is a very useful output for you to continue working with the KS or C domains that NaPDoS2 detected & classified. You can copy and paste the trimmed query KS or C domain hits into a new file, a sequencing editing program, etc.

    - For interpretation of the appended information to your query search (ie. >ProteinA_13_429 or >GeneB_5_1323_56_472), please see above.

- **Output Alignment** with closest database matches

    - Select alignment format (all alignments are created with MUSCLE but can be output in different formats suitable for many downstream applications):
        - FASTA (.fasta)
        - Msf (.msf)
        - Clustalw (.clw)

    - This outputs an alignment of your selected sequences with the closest NaPDoS2 database matches. Note, the alignment output will include all closest NaPDoS2 database matches to all query hits, not just your selected sequences, but only your selected sequences will be present in the alignment.

    - **Limits**: Due to computational limitations, the maximum number of matches for alignment is 100.

    - You can **DOWNLOAD** (click link to view; or right-click link & select "Save Link As" to save) the alignment file for downstream use in sequence programs, etc.

- **Construct tree** (candidate domains + blast matches + reference domains)

    - Select the query domain hits that you would like to build a tree with. You can either build a tree with:

        - **Closest database matches** to your selected sequences, ie. # of selected queries + closest database references
        - **All classified KS domains** which would be the # of selected queries + all 414 NaPDoS2 database reference sequences.

    - **Limits**: Due to computational limitations, the maximum number of matches for alignment is 100.

    - The estimated time to calculate the trees is given, usually the time to complete the trees is much faster than estimated.

3. Tree construction

- Tools are provided for construction of phylogenetic trees from predicted KS and C domain sequences, because these trees consistently out-perform simple BLAST searches in predicting the type of compound produced and/or its relative novelty.

- Selected candidate sequences plus their BLAST matches are trimmed and inserted into a manually curated amino acid reference alignment, keeping the original reference alignment intact. This alignment is used to build a tree, which is often more useful than BLAST results alone in predicting whether pathway products for candidate domains are likely to be similar or different from previously known examples, as demonstrated in the tutorials above.

- Please avoid using non-standard characters (e.g. commas, slashes, colons, parentheses, ampersands, etc) in file names, identifiers/description line, or sequences, as well as ambiguity codes "R" and "Y" for nucleic acid sequence files. Additionally, please avoid overly long sequence names as this will also prevent a phylogenetic tree from being calculated.

Tree output options:

- Phylogenetic domain trees are built using FastTree to estimate maximum likelihood.
- Tree building does not actually begin until the **"Submit job**" button is clicked.
- After the tree is built, users choose either Newick (plain text) format or an SVG graphics image as an output format.
- User sequences are highlighted in the SVG graphics image format with red dots, as shown in the tutorial screenshots.
- FastTree output does not include bootstrap values. However, the program does provide confidence values, which are included in the Newick format output. These values can be

visualized by opening the Newick file with most stand-alone GUI interface tree viewing programs, like FigTree, iTol, TreeViewer, ETE Toolkit, etc.

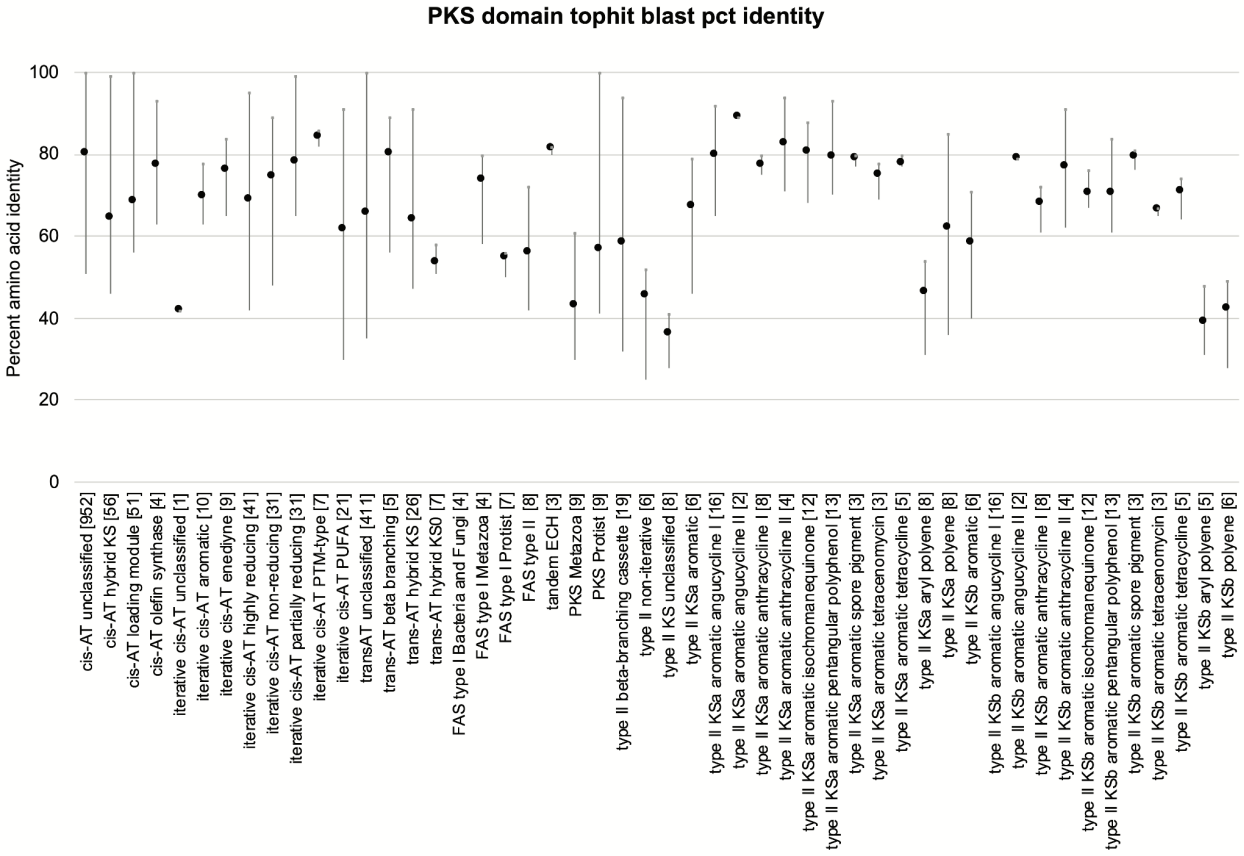
Newick format output:
(((salinilactam_KS07_cisAT:0.21578,meilingmycin_KS08_cisAT:0.11659)0.942:0.03688,((salinilactam_KS06_cisAT:0.00439,salinilactam_KS11_cisAT:0.01069)1.000:0.13503,(salinilactam_KS08_cisAT:0.13310,salinilactam_KS09_cisAT:0.15077)0.985:0.05611)0.982:0.04721,(salinilactam_KS10_cisAT:0.16240,(sceliphrolactam_KS03_cisAT:0.14473,(salinilactam_KS03_cisAT:0.02117,(salinilactam_KS02_cisAT:0.06450,salinilactam_KS04_cisAT:0.00205)0.641:0.03708)1.000:0.12769)0.905:0.03483)0.376:0.02993)0.867:0.02525)0.919:0.02878,(vicenistatin_KS09_cisAT:0.16297, …

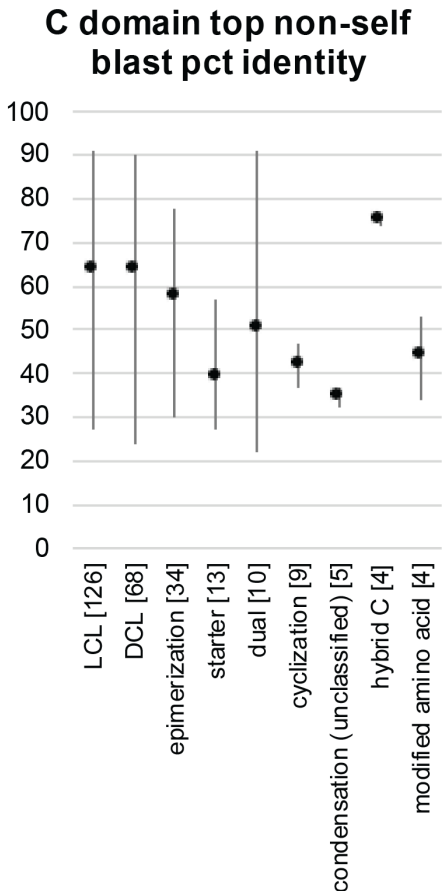
4. Interpreting Results

- BLAST matches to reference domain sequences provide a good starting point for predicting domain functionality, but are less reliable than multiple sequence alignments and the phylogenetic trees built from them. This is because BLAST scores represent average, overall similarities distributed over the entire domain, considering only two sequences at a time, whereas multiple sequence alignments and tree topologies combine information from many sequences together to reveal relationships based on more specific, potentially localized features they may have in common.

- Typical amino acid percent identities vary widely within different functional domain categories, due at least partly to variability in the number and diversity of reference examples available.

- The following chart of leave-one-out (jack-knife) cross validation scores, shows averages and expected ranges for members of different functional classes. These values were obtained by finding the closest, non-self database match for each NaPDoS2 reference domain, and comparing amino acid match statistics to other members of the same functional classification category. Brackets after each category on the x-axis indicate the number of database reference sequences analyzed for that category. These results provide the current sequence identity ranges within each class and subclass.

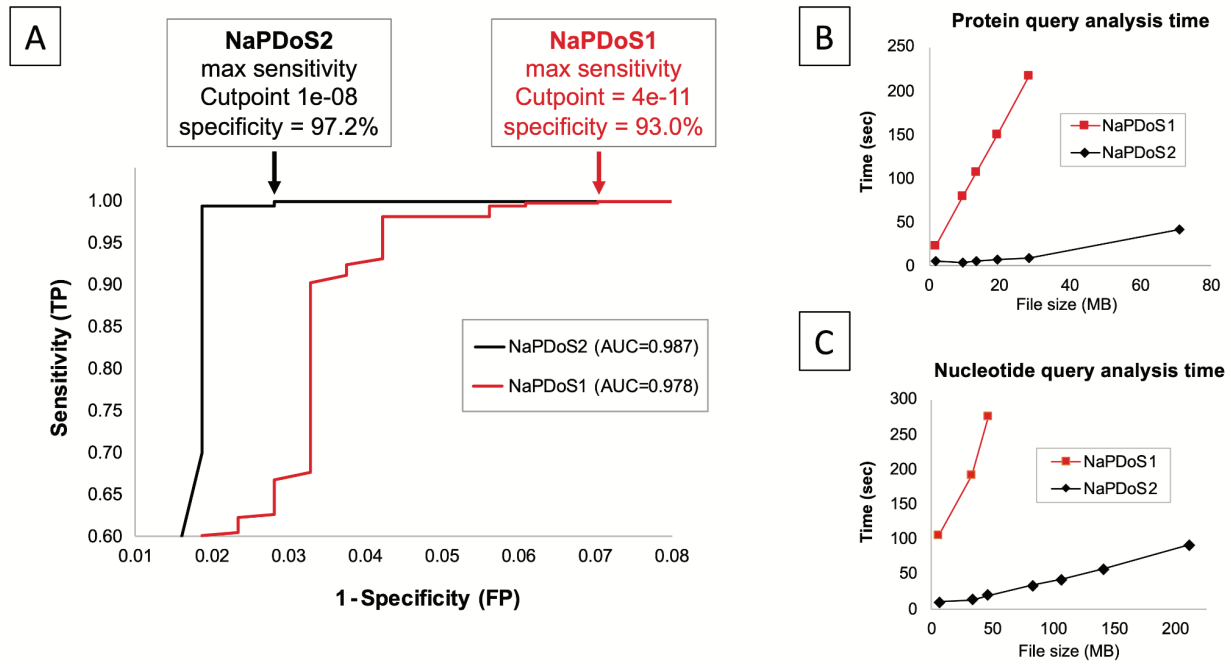

PKS domain tophit blast pct identity
Percent amino acid identity
100
80
60
40
20
0
cis-AT unclassified [952]
cis-AT hybrid KS [56]
cis-AT loading module [51]
cis-AT olefin synthase [4]
iterative cis-AT unclassified [1]
iterative cis-AT aromatic [10]
iterative cis-AT enediyne [9]
iterative cis-AT highly reducing [41]
iterative cis-AT non-reducing [31]
iterative cis-AT partially reducing [31]
iterative cis-AT PTM-type [7]
iterative cis-AT PUFA [21]
transAT unclassified [411]
trans-AT beta branching [5]
trans-AT hybrid KS [26]
trans-AT hybrid KS0 [7]
FAS type I Bacteria and Fungi [4]
FAS type I Metazoa [4]
FAS type I Protist [7]
FAS type II [8]
tandem ECH [3]
PKS Metazoa [9]
PKS Protist [9]
type II beta-branching cassette [19]
type II non-iterative [6]
type II KS unclassified [8]
type II KSa aromatic [6]
type II KSa aromatic angucycline I [16]
type II KSa aromatic angucycline II [2]
type II KSa aromatic anthracycline I [8]
type II KSa aromatic anthracycline II [4]
type II KSa aromatic isochromanequinone [12]
type II KSa aromatic pentangular polyphenol [13]
type II KSa aromatic spore pigment [3]
type II KSa aromatic tetracenomycin [3]
type II KSa aromatic tetracycline [5]
type II KSa aryl polyene [8]
type II KSb polyene [8]
type II KSb aromatic [6]
type II KSb aromatic angucycline I [16]
type II KSb aromatic angucycline II [2]
type II KSb aromatic anthracycline I [8]
type II KSb aromatic anthracycline II [4]
type II KSb aromatic isochromanequinone [12]
type II KSb aromatic pentangular polyphenol [13]
type II KSb aromatic spore pigment [3]
type II KSb aromatic tetracenomycin [3]
type II KSb aromatic tetracycline [5]
type II KSb aryl polyene [5]
type II KSb polyene [6]



KS domain sequence diversity. Percent sequence identities were determined for each KS class in the NaPDoS2 database using an all-against-all BLASTP comparison. Mean values (points) and ranges (vertical lines) are shown for top non-self matches within each NAPDoS2 class (x-axis). Bracketed values indicate the total number of database sequences in that class. (**Supplementary Figure S12 from the NaPDoS2 publication**).

BLASTP E-value cutoff

- Recommended e-value settings for routine use are based on Receiver Operating Characteristic (ROC) curves, calculated using leave-one-out (jack-knife) cross validation statistics for 213 non-redundant known positive and 308 negative control examples (see the NaPDoS2 publication for details).
- The default cutoff for NaPDoS version 2 has been set at 1e-8, an e-value that maximizes sensitivity with the minimum possible number of false positives.
- Users willing to accept higher false positive rates may select less stringent cutoff values using "**Advanced settings**".



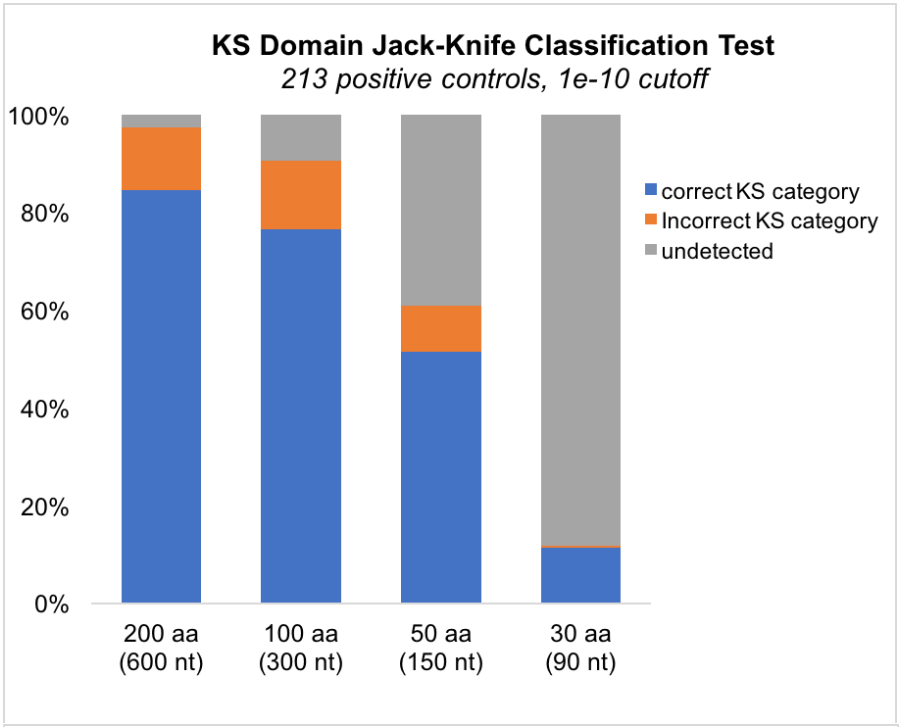
Comparative Performance Evaluations of NaPDoS and NaPDoS2. A) Receiver Operating Characteristic (ROC) curves were used to evaluate maximum obtainable sensitivity and specificity, and select optimal BLASTP cutoff scores. Non-redundant sets of 213 positive and 308 negative control KS domains were compared based on the BLASTP e-values of their closest non-self match in each NaPDoS database; FP, false positives; TP, true positives. B) Comparative analysis times for protein queries, C) Comparative analysis times for nucleic acid genomic and metagenomic queries (complete analysis time data provided in NaPDoS2 Supplementary Table S1; sec, seconds; MB, megabytes). ***(Figure 6 from the NaPDoS2 publication)***

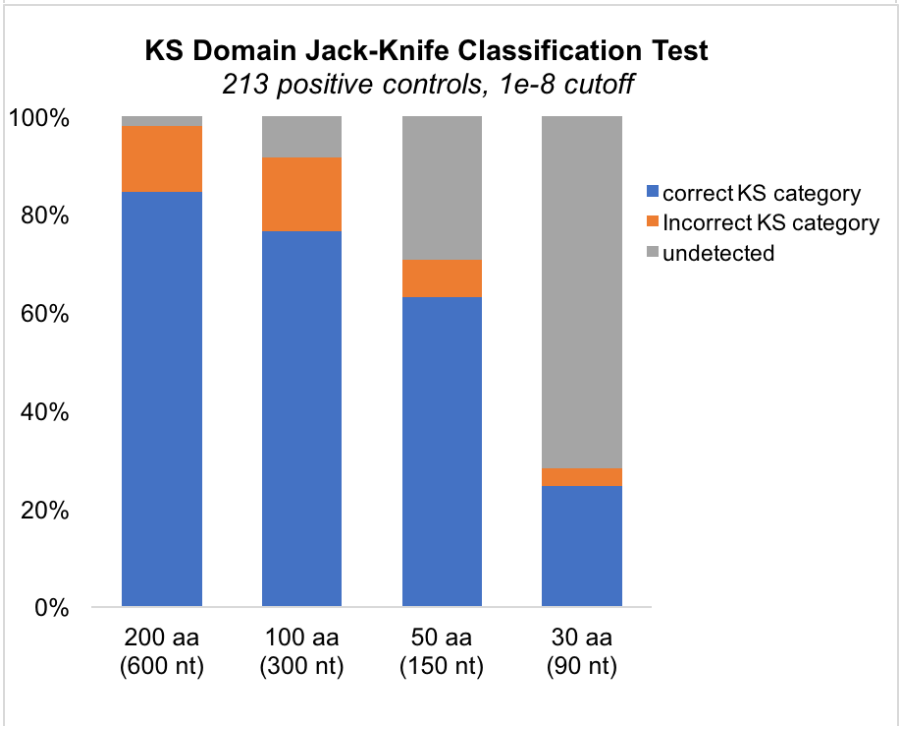
Sensitivity and accuracy versus alignment length

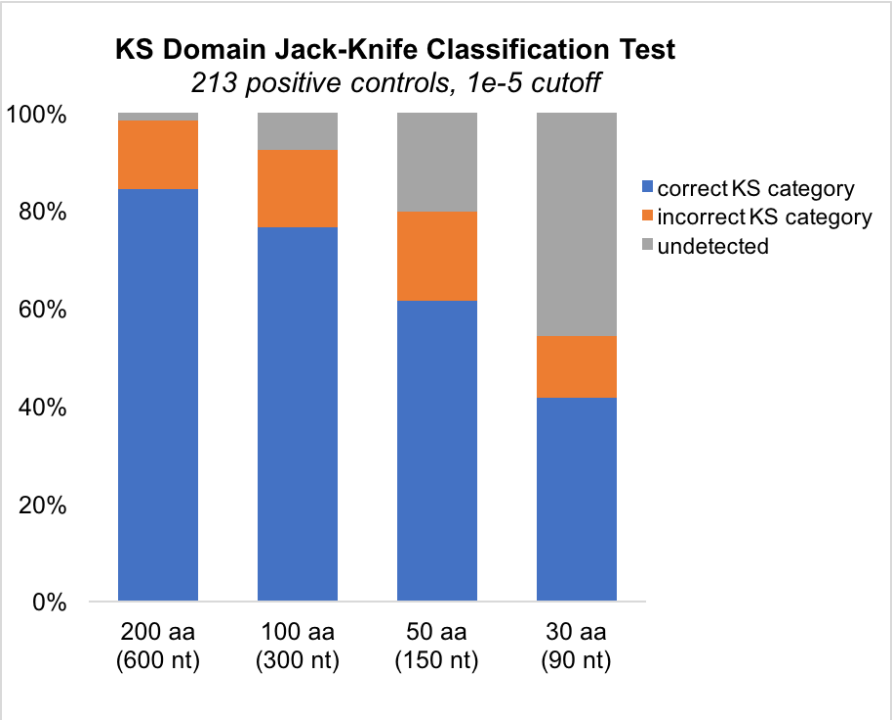
- Default minimum alignment lengths for NaPDoS2 have been set at 200 amino acids, but users should be aware that genomic or metagenomic sequences covering randomly placed domain fragments cannot provide the same sensitivity and accuracy as full length KS and C domain sequences, which are typically around 425 amino acids (1275 nucleotides) long. Unassembled, raw Illumina sequencing reads from genomes and metagenomes (typically < 150 nucleotides, or 30 translated amino acids) are poorly suited for obtaining useful PKS and NRPS domain predictions. Targeted PCR amplicons provide better sensitivity than similarly sized random fragments, but may still be too short for detailed, accurate functional classification.

- Large declines in sensitivity and classification accuracy were observed when full-length, non-redundant positive control domains were split into smaller, overlapping

subsequences of less than 100 amino acids (300 nucleotides), as shown below. Relatively novel domain classes, with fewer available reference examples, were the ones most likely to be mis-classified in leave-one-out cross validation tests. Sensitivity declines can be partially offset by decreasing BLAST stringency, but this adjustment may also increase false positives and mis-classifications, depending on the nature of the data being analyzed.



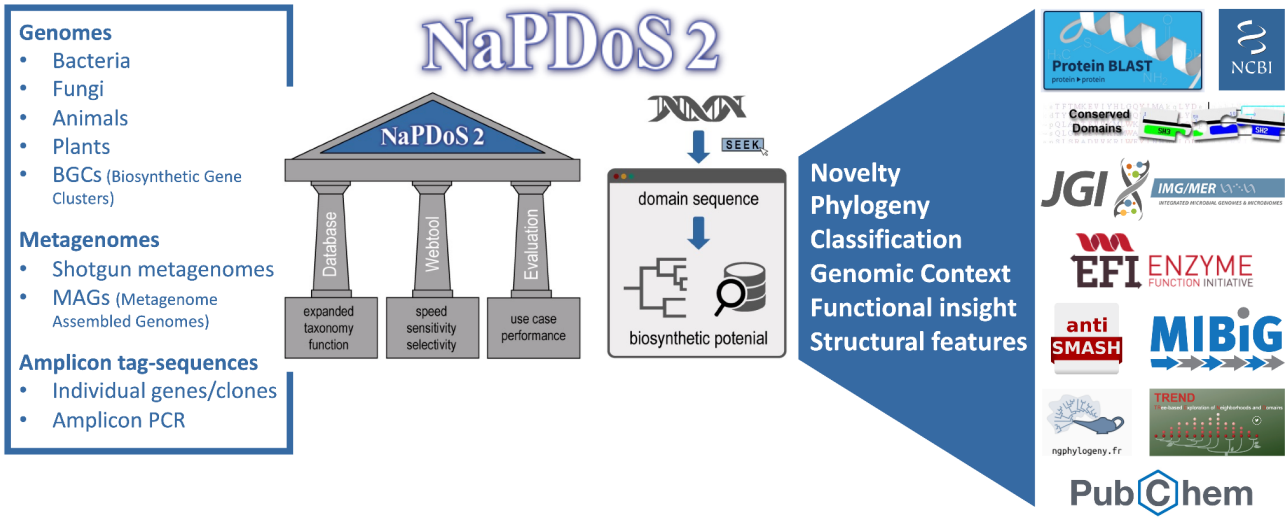




Effect of query size on KS detection and classification accuracy. Classifications were based on varying BLASTP e-value cutoff scores (1e-10, 1e-8, 1e-5) for the closest non-self database match. Test sequences of varying lengths were obtained as overlapping sliding window subsequences covering the full length of 213 non-redundant, positive control KS domains. *(**Figure 7 & Supplementary Figure S11 from the NaPDoS2 publication**)*.

5. Further analyses beyond NaPDoS2

- We have provided examples of further analyses that can be performed following a NaPDoS2 analysis in the tutorial sections above. For additional examples, this is how we imagine how NaPDoS2 fits into a roadmap of predicting biosynthetic potential from various samples & further analysis and hypothesis-building that could be performed with additional excellent tools.



- **Some tools that we brainstormed for use after a NaPDoS2 analysis:**
- *For finding similar sequences (homologs)*-- a way to assess novelty & functional insight
    - NCBI BLASTP
    - JGI IMG/MER
    - ThYme database
    - Conserved Domain Database (CD-Search)
- *Phylogeny & assessing relatedness, novelty, functional insight of homologous KS/C sequences*
    - NGPhylogeny.fr
    - EFI (Sequence similarity networks)
    - TREND (phylogeny and conserved domain prediction)
- *Genomic context*– linking the KS/C domains to their encoding BGCs and surrounding gene neighborhood
    - antiSMASH, fungiSMASH
    - BiG-FAM
    - MIBiG (database of characterized BGCs)
    - antiSMASH-db (database)
    - PRISM
    - STRING-db (protein interactions)
- *Structural features based on class/subclass characteristic features*
    - PubChem
    - NPAtlas (find related structural families of natural product molecules)
- General list of other excellent tools:
    - Secondary Metabolite Bioinformatics Portal

Frequently Asked Questions (FAQ)

1. No matches were found in my search. What can I do to increase sensitivity?

Default cutoff values for E-value (1e-8) and match length (200 amino acids) can be made less stringent using the "**Advanced Settings**" search parameters, but caution should be used in interpreting these results because reducing default parameter stringency has been shown to increase false positives and decrease classification accuracy. In this Documentation, see section "Sensitive and accuracy versus sequence length" for more details.

2. Can I download a stand-alone version of NaPDoS2?

No, there is no existing stand-alone version of the program.

3. What limitations are there on input data?

Sequences must be in FASTA format. Please avoid using sequence IDs (ie. description line of the FASTA sequences) containing non-standard characters (e.g. commas, slashes, colons, parentheses, ampersands, tab characters, etc), as well as ambiguity codes "R" and "Y" for nucleic acid sequence files. The current size limit for uploaded files is 500,000 sequences or 500 MB. For larger files, please consider our suggestions for file size management, described in this section of the documentation. Due to computational limits, if there are over 2,000 matches found in an analysis, other outputs/further analyses options will be limited; this is described in this section.

4. Where can I find all of the reference sequences for doing my own BLAST searches and making my own trees?

FASTA format sequences for our curated KS and C domain reference sets can be downloaded from the link at the bottom of the BGC description page: https://npdomainseeker.sdsc.edu/napdos2/pathways_v2.html. We describe these various files here.

Additionally, you can download the alignment of your query hits and the top NaPDoS2 database hits by selecting the query hit matches on the "**Database Search Results**" page and selecting "**Output Alignment**" or "**Output selected sequences in fasta format**". The MUSCLE alignments can be downloaded in FASTA, MSF, or ClustalW format.

We give an example in this tutorial of how you can go from NaPDoS2 results to find additional sequences via BLASTP, and then build a phylogenetic tree.

5. When were the KS and C reference domain databases last updated? When will it be updated next?

For NaPDoS2:
KS domains were last updated in October 2020, and include references/recent KS domains up through studies published in 2019.
C domains were last updated in April 2011.

NaPDoS2 will be updated in the future pending continued funding of this tool. We always appreciate community support– if you use the tool, be sure to cite it, but also– send us KS and C domains you would like to be included in future updates!

6. Why is an analysis sometimes not completed within the estimated time?

Estimated times for NaPDoS website analyses are only rough approximations, and can take considerably longer if the server is busy, or for non-amplicon (e.g. protein or metagenomic) data sets containing large numbers (> 100) of positive matches. Additionally, large amplicon datasets of longer lengths can take longer than expected for results to load and tabulate. In rare cases, if the number of sequences and file size are below the upload/analysis limits but your submitted job runs, spins for a long time, and then nothing loads, try re-running the dataset. If the same thing happens, we recommend splitting the submitted sequences into files with a smaller amount of sequences.

7. What criteria should be used to determine potential novelty for predicted domains?

This is an interesting question that we are also interested in exploring, and we don't yet have a definite answer– but these are some things we consider when assessing biosynthetic novelty:

1) For each class and subclass of predicted domains, there is variation in the percent identity, or similarity, between the domains. For an example of this, see the description of this in the above section. This means that a KS/C domain "novelty" cutoff is likely specific to each specific class/subclass. And, for classes/subclasses without many characterized representatives from limited or even broad taxonomic breadth, defining the "novelty" cutoff can be even more tricky.
2) Another consideration should be the taxonomy of the top database match and the taxonomic diversity in each class/subclass– this also affects the range of KS/C domain "novelty" cutoff for each class/subclass as perhaps the majority of characterized domains from a certain class/subclass are from a certain taxa, etc.
3) When comparing top NaPDoS2 database matches, percent identity can be a way to assess how similar your query domain hit is to sequences in the database. This should be

expanded upon by finding other homologs, performing alignments, and construction of phylogenetic trees, as we describe in one of the genome analysis tutorials.

8. Can I analyze unassembled metagenomic reads?

First, does your raw unassembled metagenomic data fit the requirements to run? The next thing that should be considered is what the length of your unassembled metagenomic reads are and what could be the false positive/false negative discovery rate. Be aware of the caveats: (copied here from this section) the default minimum alignment lengths for NaPDoS2 have been set at 200 amino acids, but users should be aware that genomic or metagenomic sequences covering randomly placed domain fragments cannot provide the same sensitivity and accuracy as full length KS and C domain sequences, which are typically around 425 amino acids (1275 nucleotides) long. Unassembled, raw Illumina sequencing reads from genomes and metagenomes (typically < 150 nucleotides, or 30 translated amino acids) are poorly suited for obtaining useful PKS and NRPS domain predictions. Targeted PCR amplicons provide better sensitivity than similarly sized random fragments, but may still be too short for detailed, accurate functional classification.

Large declines in sensitivity and classification accuracy were observed when full-length, non-redundant positive control domains were split into smaller, overlapping subsequences of less than 100 amino acids (300 nucleotides), as shown below. Relatively novel domain classes, with fewer available reference examples, were the ones most likely to be mis-classified in leave-one-out cross validation tests. Sensitivity declines can be partially offset by decreasing BLAST stringency, but this adjustment may also increase false positives and mis-classifications, depending on the nature of the data being analyzed.

9. How can I analyze large datasets?

We've asked this question too!. Like any large dataset analysis project, large analyses with NaPDoS2 takes patience, a decent internet connection (to help with those dataset upload speeds!), and a well-thought-out data management plan.

We recommend:

1) Use the provided file management scripts to break up large files >500MB or >500,000 sequences into multiple parts. You can then run each file, with waiting a minute between each job.
2) Each user can run 10GB of data per day. The webtool cache will clear every 24 hours.
3) We unfortunately do not have the capacity to run large analyses for you. But contact us if you're looking for advice or specific questions on how to tackle large NaPDoS2 analyses.
4) We hope that future versions of this tool will have a stand-alone analysis version, but that is not yet available..

10. What if the top database match and the closest result in the phylogenetic tree does not match?

Yes, it is possible that the top database match and the closest related KS/C domain sequence in the phylogenetic tree are different. This is because BLAST matches to reference domain sequences provide a good starting point for predicting domain functionality, but are less reliable than multiple sequence alignments and the phylogenetic trees built from them. BLAST scores represent average, overall similarities distributed over the entire domain, considering only two sequences at a time, whereas multiple sequence alignments and tree topologies combine information from many sequences together to reveal relationships based on more specific, potentially localized features they may have in common.

Alignments that are used to calculate phylogenetic trees are often more useful than BLAST results alone in predicting whether pathway products for candidate domains are likely to be similar or different from previously known examples. Please read more in the "Interpreting Results" and details about the NaPDoS2 workflow section.

11. My phylogenetic tree construction failed, what happened and how can I fix it?

There are a couple if things that could have gone wrong:

1) See the limitations for tree building here– trees can only be calculated on the NaPDoS2 webtool if you select <100 query domain hits.
2) However, if your initial number of query domain hits is >2,000, due to computational limits, a phylogenetic alignment will not be calculated, even if you select a subset of <100 hits. Output those 100 hits as FASTA sequences and re-run them through NaPDoS2 to build the tree!
3) Please avoid using non-standard characters (e.g. commas, slashes, colons, parentheses, ampersands, etc) in sequence names, identifiers/description line, or sequences, as well as ambiguity codes "R" and "Y" for nucleic acid sequence files. These characters are not supported by FastTree. Additionally, please avoid overly long sequence names as this will also prevent a phylogenetic tree from being calculated in FastTree.

12. What are the numbers & information added to the end of my query domain hit, ie. "cand_id"?

The information that is appended to each "cand_id" differs if your query was a protein amino acid or nucleic acid– please see this section for a detailed breakdown of the appended information.