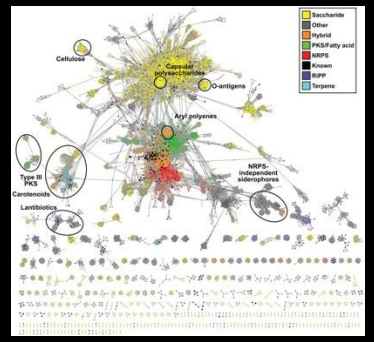
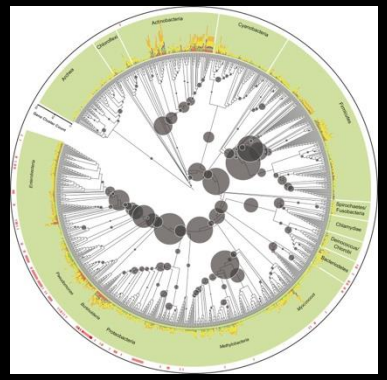
A Multi-Repository Scale Genomic and Chemical Search Engine to Enable the Discovery, Production, and Function of Natural Products

Chase Clark
Postdoctoral Associate
Jason Kwan's Group
University of Wisconsin-Madison

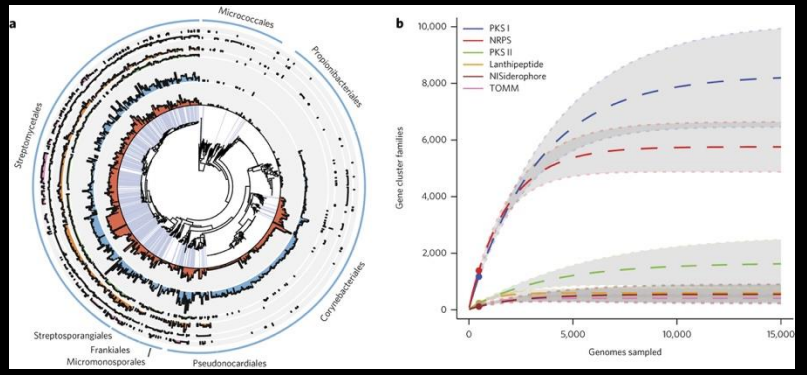
2014:
1,154 Genomes; 33,351 BGCs; 2,400 GCFs





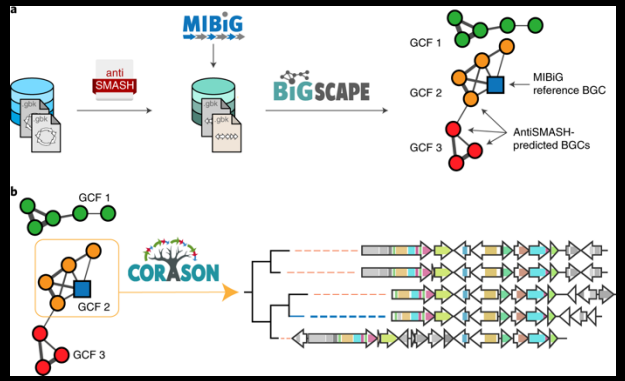
Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014

2014:
830 Genomes; 11,422 BGCs; 4,122 GCFs



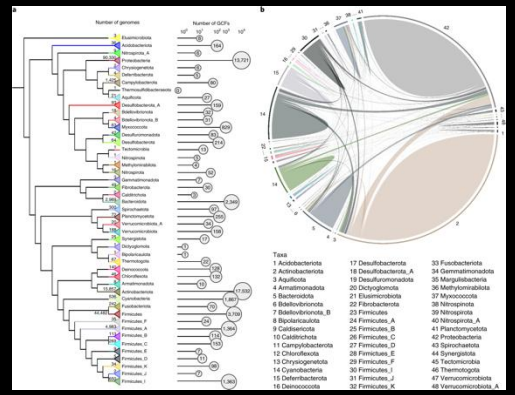
Doroghazi JR, Albright JC, Goering AW, Ju KS, Haines RR, Tchalukov KA, Labeda DP, Kelleher NL, Metcalf WW. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014

2020:
3,080 Genomes; 74,652 BGCs

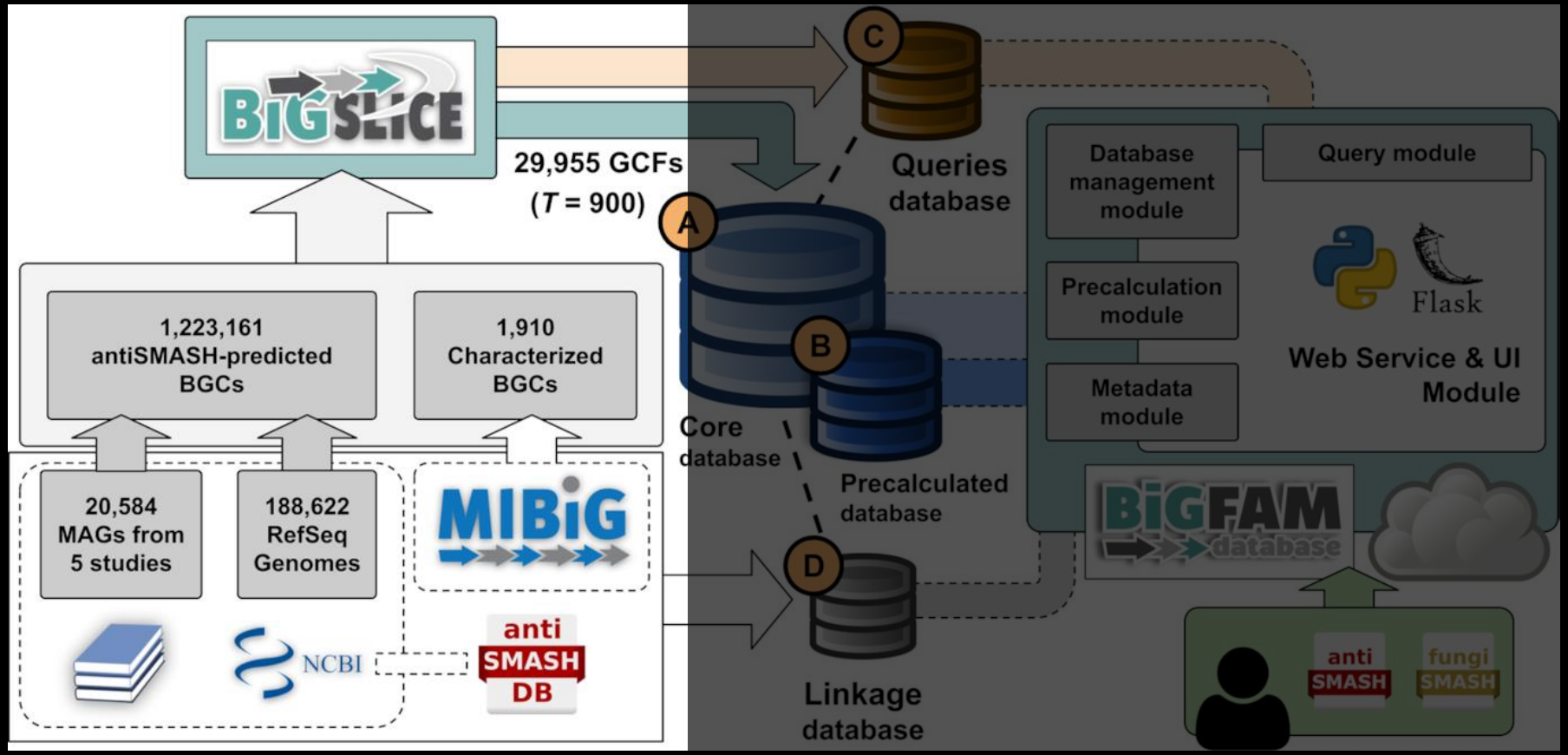


Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2020

2022:
170,000 Genomes, 1,185,995 BGCs; 62,449 GCFs



Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, Ziemert N. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nat Microbiol. 2022



Kautsar SA, Blin K, Shaw S, Weber T, Medema MH. BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res. 2021

Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. Gigascience. 2021

nature reviews drug discovery

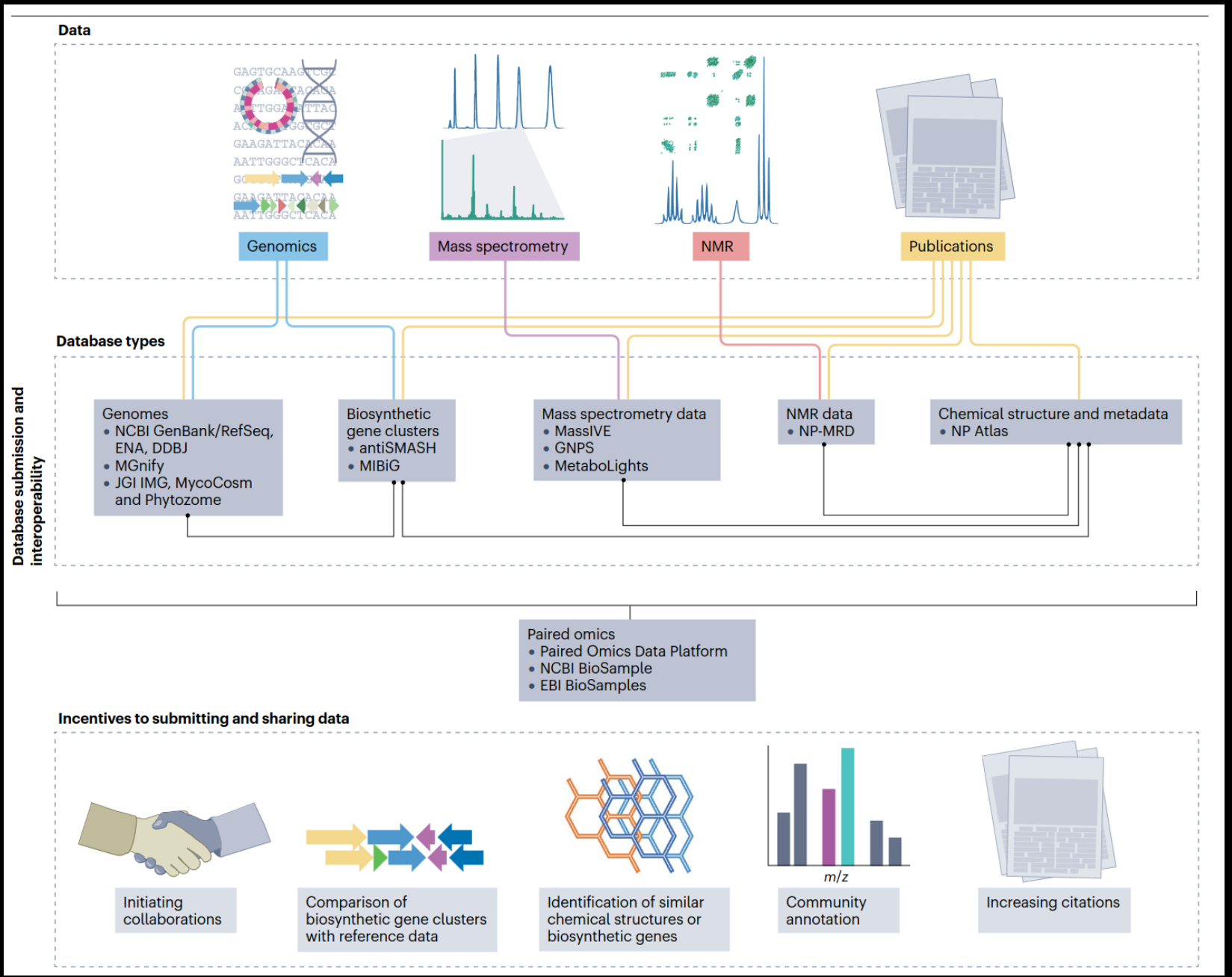
https://doi.org/10.1038/s41573-023-00774-7

Review article

Check for updates

Artificial intelligence for natural product drug discovery

Michael W. Mullowney[1,62], Katherine R. Duncan[2,62], Somayah S. Elsayed[3,62], Neha Garg[4,62], Justin J. J. van der Hooft[5,6,62], Nathaniel I. Martin[7,62], David Meijer[5,62], Barbara R. Terlouw[5,62], Friederike Biermann[5,8,9], Kai Blin[10], Janani Durairaj[11], Marina Gorostiola González[12,13], Eric J. N. Helfrich[8,9], Florian Huber[14], Stefan Leopold-Messer[15], Kohulan Rajan[16], Tristan de Rond[17], Jeffrey A. van Santen[18], Maria Sorokina[19,20], Marcy J. Balunas[21,22], Mehdi A. Beniddir[23], Doris A. van Bergeijk[3], Laura M. Carroll[24], Chase M. Clark[25], Djork-Arné Clevert[26], Chris A. Dejong[27], Chao Du[3], Scarlet Ferrinho[28], Francesca Grisoni[29,30], Albert Hofstetter[31], Willem Jespers[12], Olga V. Kalinina[32,33,34], Satria A. Kautsar[35], Hyunwoo Kim[36], Tiago F. Leao[37], Joleen Masschelein[38,39], Evan R. Rees[25], Raphael Reher[40,41], Daniel Reker[42,43], Philippe Schwaller[44], Marwin Segler[45], Michael A. Skinnider[27,46], Allison S. Walker[47,48], Egon L. Willighagen[49], Barbara Zdrazil[50], Nadine Ziemert[51], Rebecca J. M. Goss[28], Pierre Guyomard[52], Andrea Volkamer[34,53], William H. Gerwick[54], Hyun Uk Kim[55], Rolf Müller[32,56,57,58], Gilles P. van Wezel[3,59], Gerard J. P. van Westen[12] ✉, Anna K. H. Hirsch[32,56,57,58] ✉, Roger G. Linington[18] ✉, Serina L. Robinson[60] ✉ & Marnix H. Medema[5,61] ✉



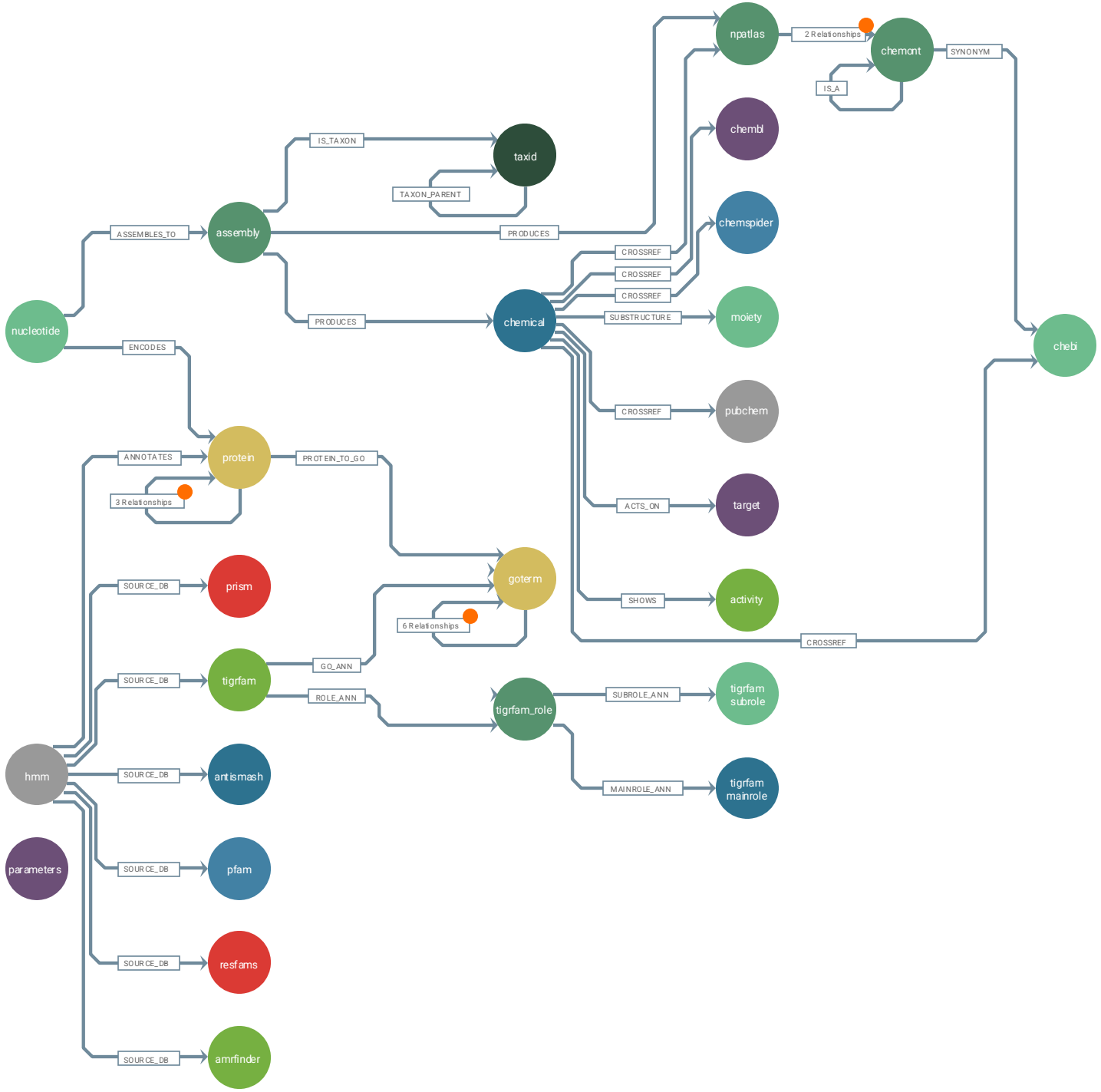
What if we just connected all the things?

SocialGene

nextflow

python™

neo4j

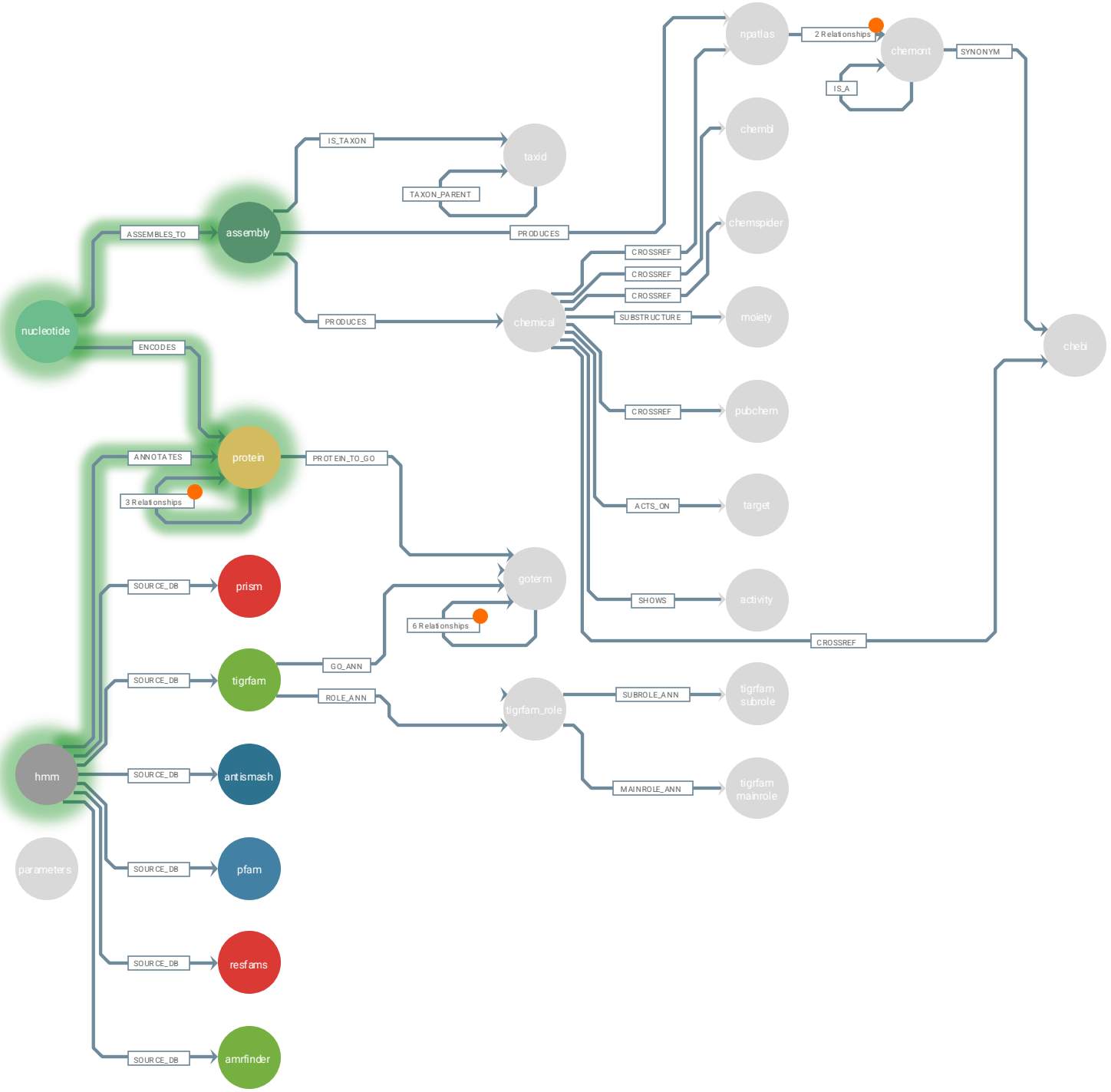


SocialGene

nextflow

python™

neo4j



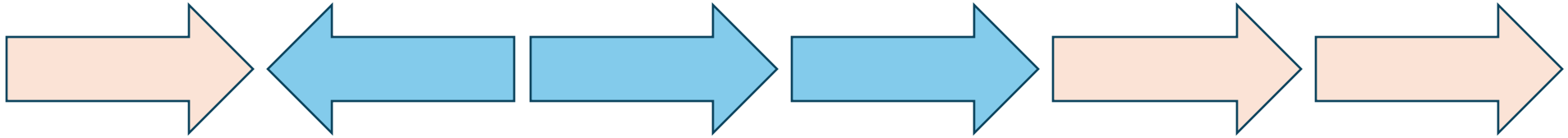
Molecular Biology in < 1 minute

DNA -> RNA, RNA -> protein

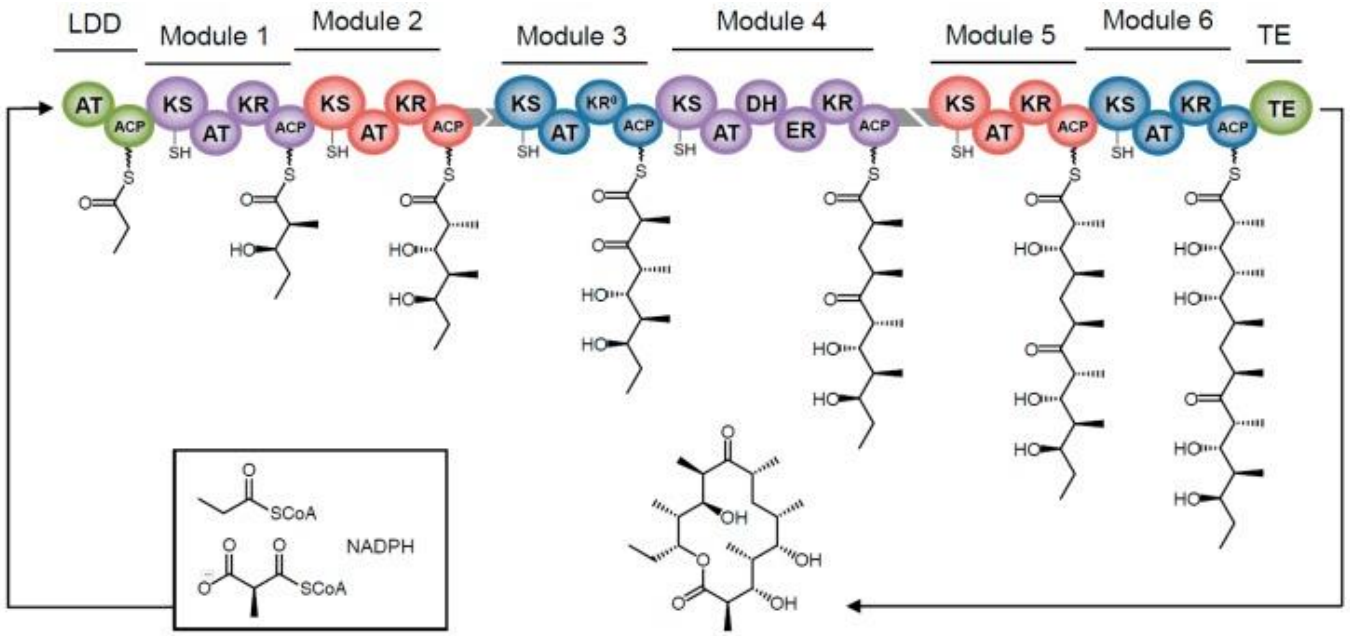
DNA (nucleotides)	ATGTCCAACGCC…	CGCGGCATCCTC…	GGCGCCGTGCTC…
RNA (nucleotides)	UTG|TCC|UUC|GCC|…	CGC|GGC|UTC|CTC|…	GGC|GCC|GTG|CTC|…
Proteins (amino acids)	MSNARATHLTRRGI…	KADVQAGDMDVSK…	AKSGPWTFKDDRGT…
Protein have domains	MSNARATHLTRRGI…	KADVQAGDMDVSK…	AKSGPWTFKDDRGT…

Genes encode proteins

Gene Cluster

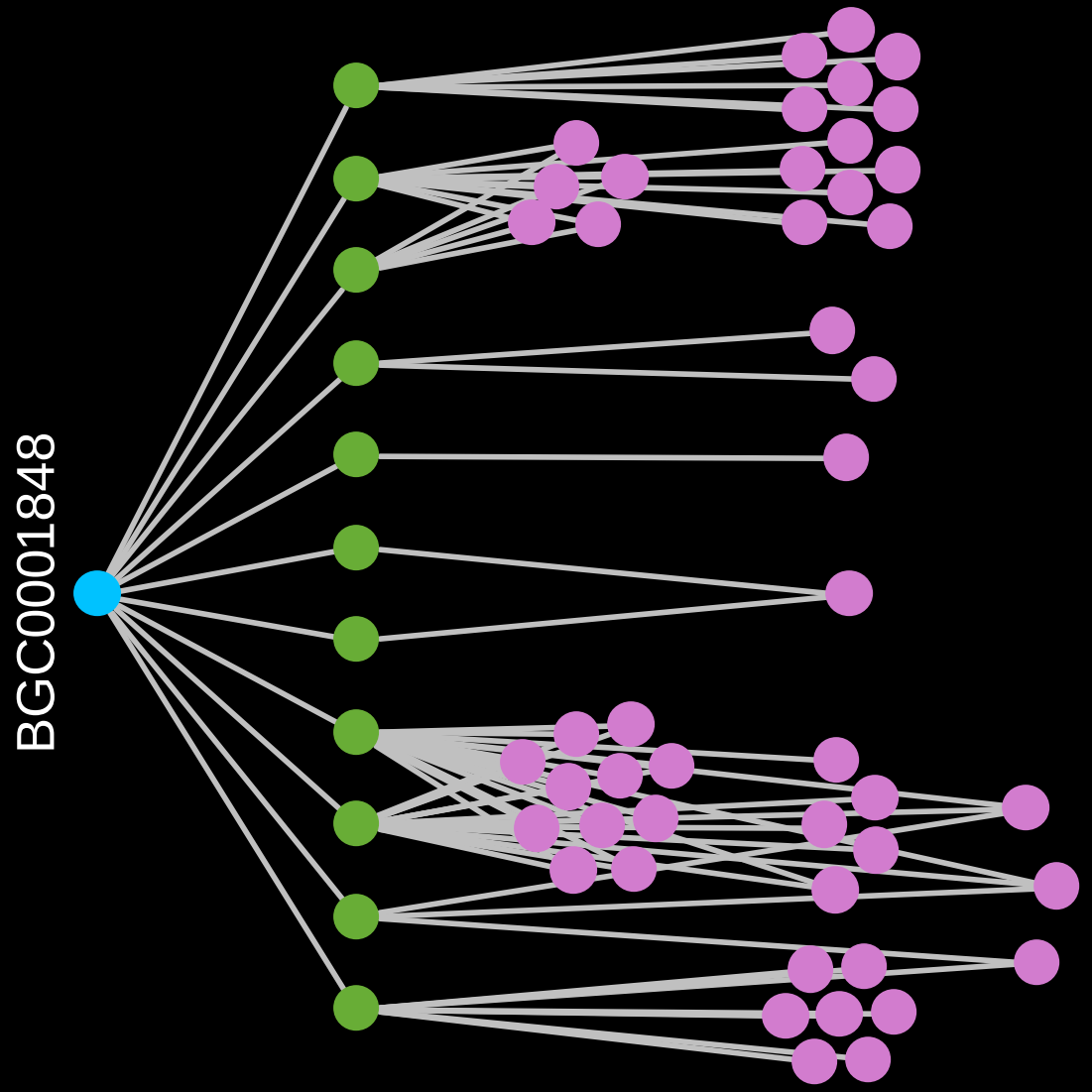


Polyketide synthases are large, multidomain proteins
“Beads on a String”



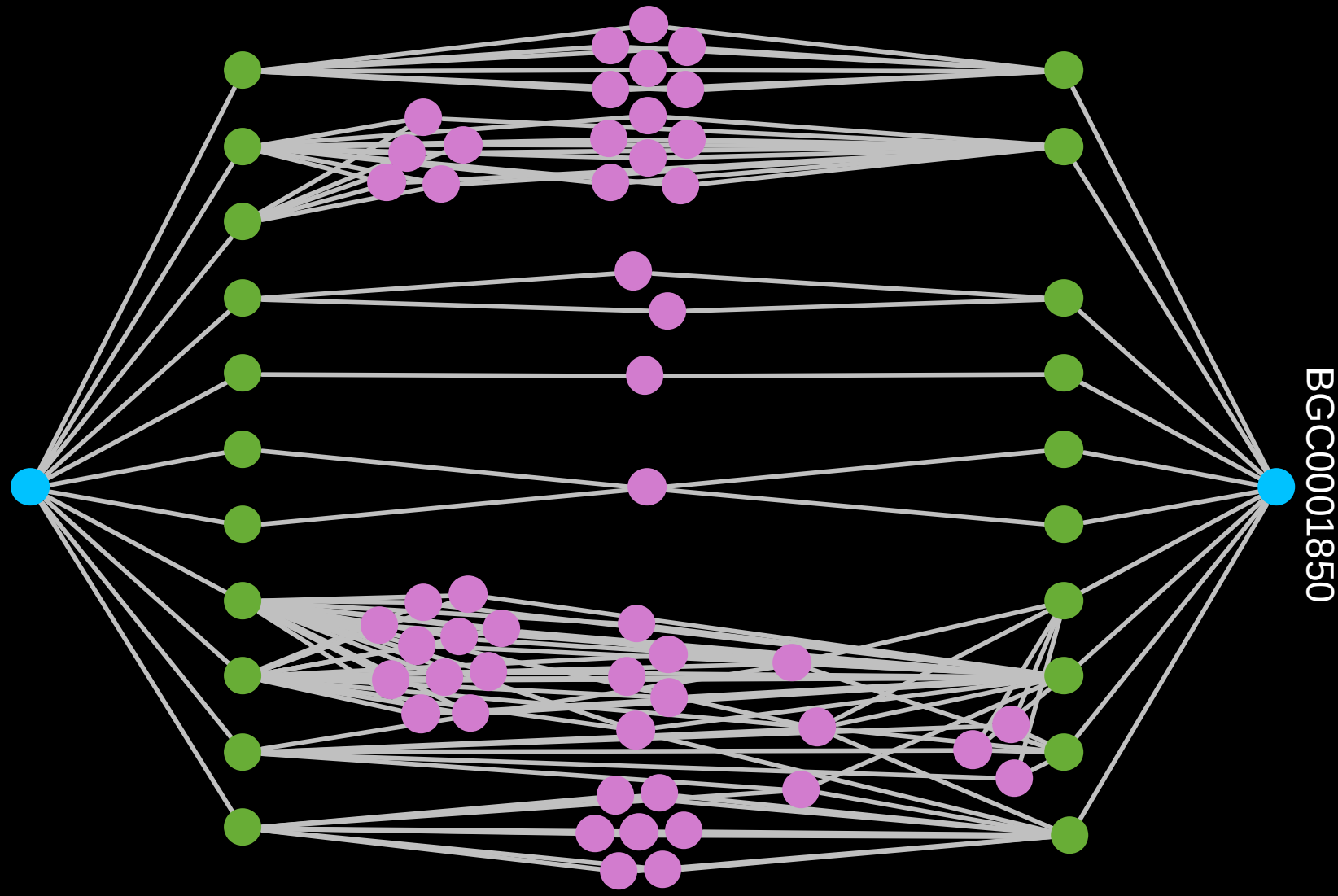
Bayly, Carmen L, and Vikramaditya G Yadav. “Towards Precision Engineering of Canonical Polyketide Synthase Domains: Recent Advances and Future Prospects.” *Molecules (Basel, Switzerland)*vol. 22,2 235. 5 Feb. 2017, doi:10.3390/molecules22020235

Similar domains = similar function



DNA Sequence Protein HMM Model

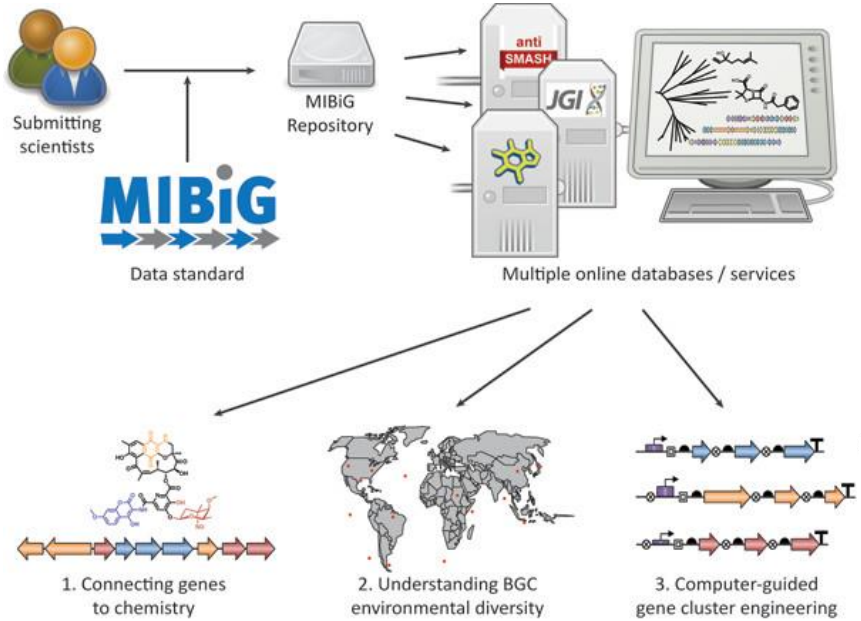
Similar domains = similar function



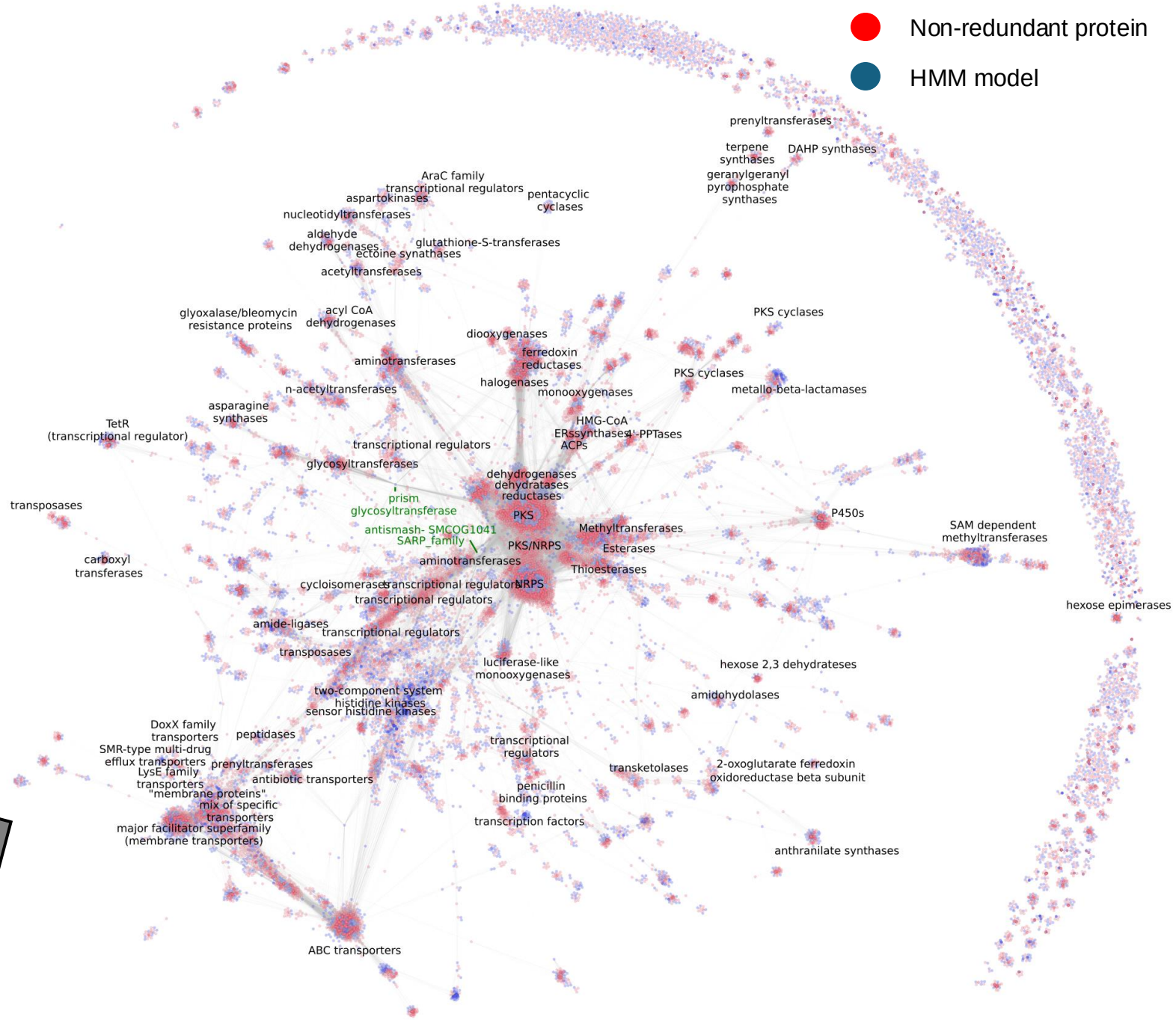
DNA Sequence Protein HMM Model

Application

Use domains to cluster proteins by function


Submitting scientists
MIBiG Repository
anti SMASH
JGI
MIBiG
Data standard
Multiple online databases / services
1. Connecting genes to chemistry
2. Understanding BGC environmental diversity
3. Computer-guided gene cluster engineering

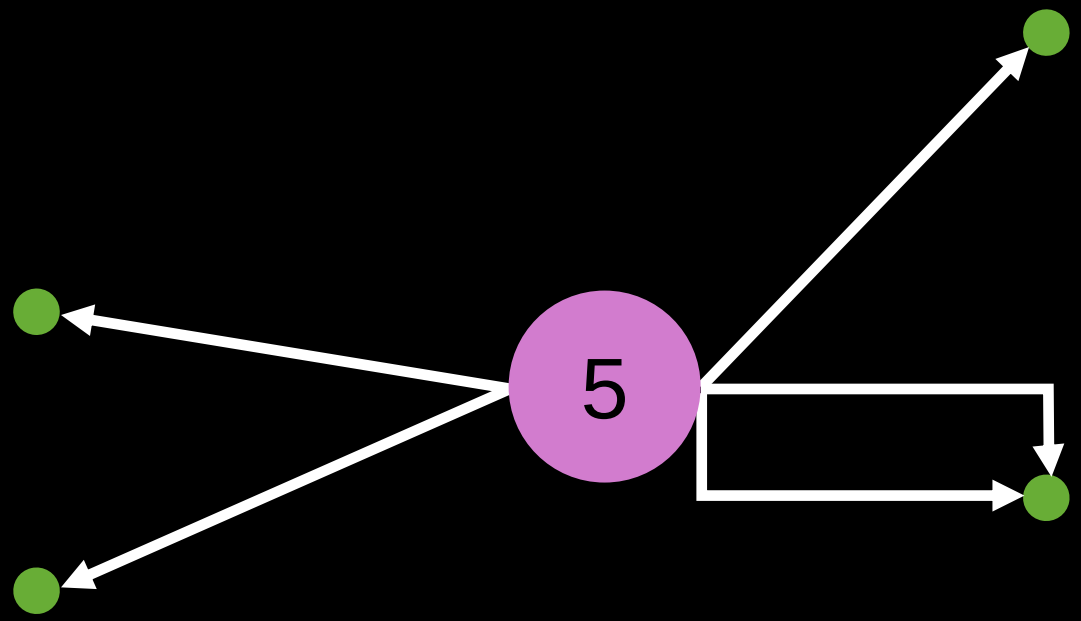
SocialGene


Non-redundant protein
HMM model
prenyltransferases
terpene synthases
DAHP synthases
geranylgeranyl pyrophosphate synthases
AraC family transcriptional regulators
aspartokinases
pentacyclic cyclases
nucleotidyltransferases
aldehyde dehydrogenases
glutathione-S-transferases
ectoine synathases
acetyltransferases
glyoxalase/bleomycin resistance proteins
acyl CoA dehydrogenases
PKS cyclases
dioxygenases
ferredoxin reductases
aminotransferases
halogenases
monooxygenases
PKS cyclases
metallo-beta-lactamases
n-acetyltransferases
asparagine synthases
TetR (transcriptional regulator)
HMG-CoA ERssynthases
4'-PPTases
ACPs
transcriptional regulators
glycosyltransferases
dehydrogenases
dehydratases
reductases
transposases
prism glycosyltransferase
PKS
P450s
antismash- SMCOG1041 SARP_family
Methyltransferases
SAM dependent methyltransferases
PKS/NRPS
Esterases
carboxyl transferases
aminotransferases
Thioesterases
cycloisomerases
transcriptional regulators
NRPS
transcriptional regulators
hexose epimerases
amide-ligases
transcriptional regulators
transposases
luciferase-like monooxygenases
hexose 2,3 dehydrateses
two-component system histidine kinases sensor histidine kinases
amidohydolases
DoxX family transporters
peptidases
SMR-type multi-drug efflux transporters
transcriptional regulators
LysE family transporters
prenyltransferases
antibiotic transporters
transketolases
2-oxoglutarate ferredoxin oxidoreductase beta subunit
"membrane proteins"
penicillin binding proteins
mix of specific transporters
transcription factors
major facilitator superfamily (membrane transporters)
anthranilate synthases
ABC transporters

Application

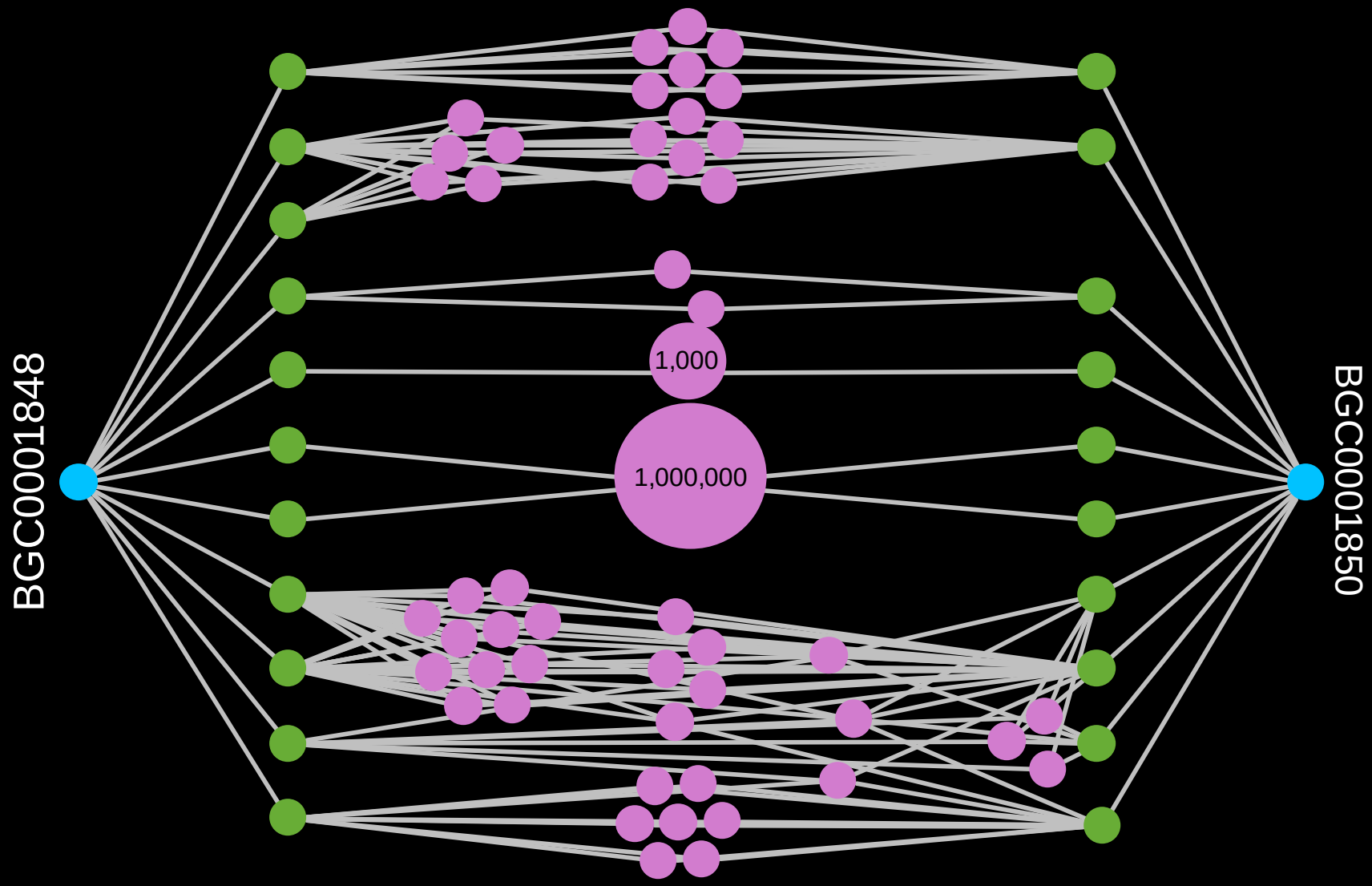
Fast Search for BGCs
with
Remote Homology
at
Really-Big-Scale™

Outdegree = number of annotations that have to be looked at



Protein HMM Model

Outdegree = number of annotations that have to be looked at



Search genomes for remotely-homologous BGCs

- Annotate input BGC domains

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains
- Search database for domains

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains
- Search database for domains
- Find proteins with those domains

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains
- Search database for domains
- Find proteins with those domains
- Find protein locations on genomes

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains
- Search database for domains
- Find proteins with those domains
- Find protein locations on genomes
- Create putative BGCs

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains
- Search database for domains
- Find proteins with those domains
- Find protein locations on genomes
- Create putative BGCs
- Reciprocal Best Hit BLASTp

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains
- Search database for domains
- Find proteins with those domains
- Find protein locations on genomes
- Create putative BGCs
- Reciprocal Best Hit BLASTp
- Evaluate and rank
  - # shared genes
  - Synteny
  - BLASTp scores

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains
- Search database for domains
- Find proteins with those domains
- Find protein locations on genomes
- Create putative BGCs
- Reciprocal Best Hit BLASTp
- Evaluate and rank
  - # shared genes
  - Synteny
  - BLASTp scores
- Create Clinker plot

Search genomes for remotely-homologous BGCs

- Annotate input BGC domains
- Prioritize domains
- Search database for domains
- Find proteins with those domains
- Find protein locations on genomes
- Create putative BGCs
- Reciprocal Best Hit BLASTp
- Evaluate and rank
  - # shared genes
  - Synteny
  - BLASTp scores
- Create Clinker plot

```
2024-02-26 12:04:33 INFO    Running search with args: {'input': '/home/chase/Documents/data/mibig/3_1/mibig_gbk_3.1/BGC0001040.gbk',                sea.py:32
                            'hmm_dir': '/home/chase/Documents/socialgene_data/v0_4_1/refseq_antismash_bgcs/socialgene_per_run/hmm_cache',
                            'use_neo4j_precalc': True, 'assemblies_must_have_x_matches': 0.4, 'nucleotide_sequences_must_have_x_matches':
                            0.4, 'gene_clusters_must_have_x_matches': 0.4, 'break_bgc_on_gap_of': 20000, 'target_bgc_padding': 10000,
                            'max_domains_per_protein': 3, 'max_outdegree': 1000000, 'max_query_proteins': 5, 'scatter': False,
                            'locus_tag_bypass_list': None, 'protein_id_bypass_list': ['CAA60459.1'], 'only_culture_collection': False,
                            'frac': 0.75, 'run_async': True, 'analyze_with': 'blastp', 'outpath_clinker':
                            '/home/chase/Downloads/clinker/plot/data.json'}
                INFO    Parsing: /home/chase/Documents/data/mibig/3_1/mibig_gbk_3.1/BGC0001040.gbk                                          genbank.py:322
                INFO    '/home/chase/Documents/data/mibig/3_1/mibig_gbk_3.1/BGC0001040.gbk' features {'CDS_motif': 246,                     genbank.py:351
                            'aSDomain': 85, 'gene': 28, 'CDS': 28, 'aSModule': 16, 'source': 1, 'misc_feature': 1, 'subregion': 1,
                            'region': 1}
                INFO    Input BGC has 28 proteins and/or pseudogenes                                                                        base.py:179
                INFO    Annotating input proteins domains                                                                                  hmmer.py:273
                INFO    Connected to Neo4j database at bolt://localhost:7687                                                                neo4j.py:62
2024-02-26 12:04:35 INFO    Finding the outdegree of all domains from all input proteins                                                   hmmer.py:319
```

Outdegree of input protein domains

Protein	Locus/Descripton	Unique HMM models	Mean	Min	25%	50%	75%	Max	Sum
3aUcLZlofJ196v15TfvlMaxGpnFc6rMs	CAA60468.1 | no-locus-tag | pteridine-dependent dioxygenase	1	221	221	221	221	221	221	221
mZKlUBeQ7BRLGL8XlsYiYKslIxOgngr2	CAA60470.1 | no-locus-tag | methyltransferase	2	6735	6670	6702	6735	6768	6801	13471
sRl0ScCQMVcCyERmnvBlQ6H4s1MztBmm	CAA60475.1 | no-locus-tag | None	1	30410	30410	30410	30410	30410	30410	30410
MMibcQQViQhy0DxdmZs1ShVMe1ppHNT4	CAA60467.1 | no-locus-tag | lysine cyclodeaminase	3	13983	10392	11031	11670	15779	19889	41951
ZXl6t-PXlcYKdwKFGK7vzacYkhB53OCN	CAA60464.1 | no-locus-tag | ferredoxin	4	23639	13134	22087	26542	28094	28337	94556
7uTHFx2UJjgexIdgvaSIZD3B70xzUeJt	CAA60454.1 | no-locus-tag | membrane transport protein	1	132073	132073	132073	132073	132073	132073	132073
VpIU-bRqlVQfE_7wUDLHlZGDDJrJwaqh	CAA60453.1 | no-locus-tag | None	1	132073	132073	132073	132073	132073	132073	132073
ftWdDCKsDWr8ctijvGQdmPK9_GPdhobM	CAA60476.1 | no-locus-tag | cystathione synthase	3	77232	60272	70668	81065	85712	90360	231697
ipCX16XSA2Xjsfb0l-7zX9KkpxqTIR4D	CAA60469.1 | no-locus-tag | cytochrome P450	4	156969	134825	155495	163854	165328	165343	627877
h9ZYftht6AohbGZlGLV8rCgm3RdkU7wq	CAA60465.1 | no-locus-tag | cytochrome P450	4	156969	134825	155495	163854	165328	165343	627877
AMhYlgV_OzjQmc_pm2Jxag6BxG9gfHGy	CAA60472.1 | no-locus-tag | helix-turn-helix DNA binding protein	5	143578	92094	149242	157078	159721	159758	717893
RF_b3nP09jIdDFc87mHLsn54Zh42QvJA	CAA60456.1 | no-locus-tag | sensory protein kinase	5	193755	83075	167049	179315	237946	301392	968777
vqLK2oqmHMsc6IsnKJdYE-M0ykHZlrEE	CAA60471.1 | no-locus-tag | regulator of cholesterol oxidase	4	314414	80322	265902	330347	378860	516642	1257659
GRzZ-X8ZQJPyOIVhR695deKSl9TdIPF9	CAA60455.1 | no-locus-tag | response regulator	4	326690	185874	207945	302123	420867	516642	1306762
Ig3RycUetwS7t3zMBwjV-xYvoOE3qu8p	CAA60450.1 | no-locus-tag | None	5	279823	3890	279261	340896	373765	401305	1399117
EUQ8eDj23gr3dkBZ4vBZJBUFlxadj6Ye	CAA60473.1 | no-locus-tag | membrane transport protein	3	716873	608600	672791	736982	771010	805039	2150621
T76p9UrLfBWZm5GsfXNx4O3DoKhjALV8	CAA60449.1 | no-locus-tag | monosaccharide transporter	4	644651	323135	615869	725215	753996	805039	2578604
FUXZ6ANJ4Q-kioq2KA8eHP7GwHnHZ5dA	CAA60466.1 | no-locus-tag | methyltransferase	13	203643	84395	120319	208972	255533	360429	2647359
kw7A__VwycaKL0KAjNWhZkQ1OpNQBRUh	CAA60463.1 | no-locus-tag | methyltransferase	15	178774	16799	102678	144653	249502	360429	2681611
IFr-Iy2LA73cIouJR4l-WwCZ6DMFN1fO	CAA60451.1 | no-locus-tag | regulator of antibiotic transport complexes	6	524097	42715	508847	618408	663888	713677	3144584
s9HRwBkT10uC6uRfVw6cjuCmnc4BXWpC	CAA60452.1 | no-locus-tag | ABC-transporter	4	955107	17624	925889	1256626	1285843	1289552	3820428
-RQfBfx-0B83QwCIDPPavLGIQVirCHdI	CAA60458.1 | no-locus-tag | ketoreductase/dehydrogenase	14	546655	332394	520336	560294	602080	685782	7653170
k382V1a_Uw8VQDXGLm_rX3ntk7EiO6HR	CAA60461.1 | no-locus-tag | pipecolate incorporating enzyme	38	246096	2027	127413	208274	296634	944217	9351666
a0IUrl7zPvLuJxqCD6vtU4pPpIW6um_R	CAA60459.1 | no-locus-tag | polyketide synthase	59	206147	20625	131559	172651	239874	685782	12162718
kkfyLl_YDzZGAVMvir6dTmRDU5M4YQkr	CAA60462.1 | no-locus-tag | polyketide synthase	62	201669	16567	122501	172346	233673	685782	12503517
KFQSeRRgS631GJRB_Qcou2BEmtH6bz0w	CAA60460.1 | no-locus-tag | polyketide synthase	69	234345	960	130116	176911	294729	944217	16169809

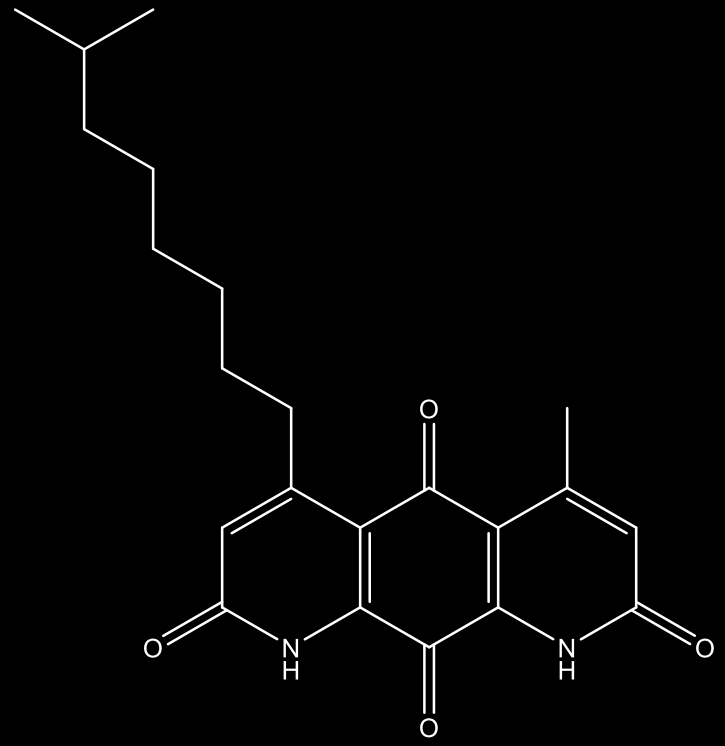
```
                INFO    Prioritizing input proteins by outdegree                                                                           hmmer.py:375
                INFO    'max_outdegree' is set to 1,000,000, will remove any domains with a higher outdegree                               hmmer.py:412
                INFO    'max_outdegree' reduced the total outdegree from 82,446,501 to 78,643,697                                          hmmer.py:418
                INFO    'max_domains_per_protein' is set to 3, will remove domains from proteins from highest to lowest outdegree          hmmer.py:428
                INFO    'max_domains_per_protein' reduced the total outdegree from 78,643,697 to 11,544,444                                hmmer.py:438
                INFO    'max_query_proteins' is set to 5, will limit search to 5 of 28 input proteins                                      hmmer.py:462
                INFO    'max_query_proteins' reduced the total outdegree from 11,544,444 to 158,979                                        hmmer.py:480
                INFO    The total outdegree was 82,446,501; now it's 12,243,753                                                            hmmer.py:488
```

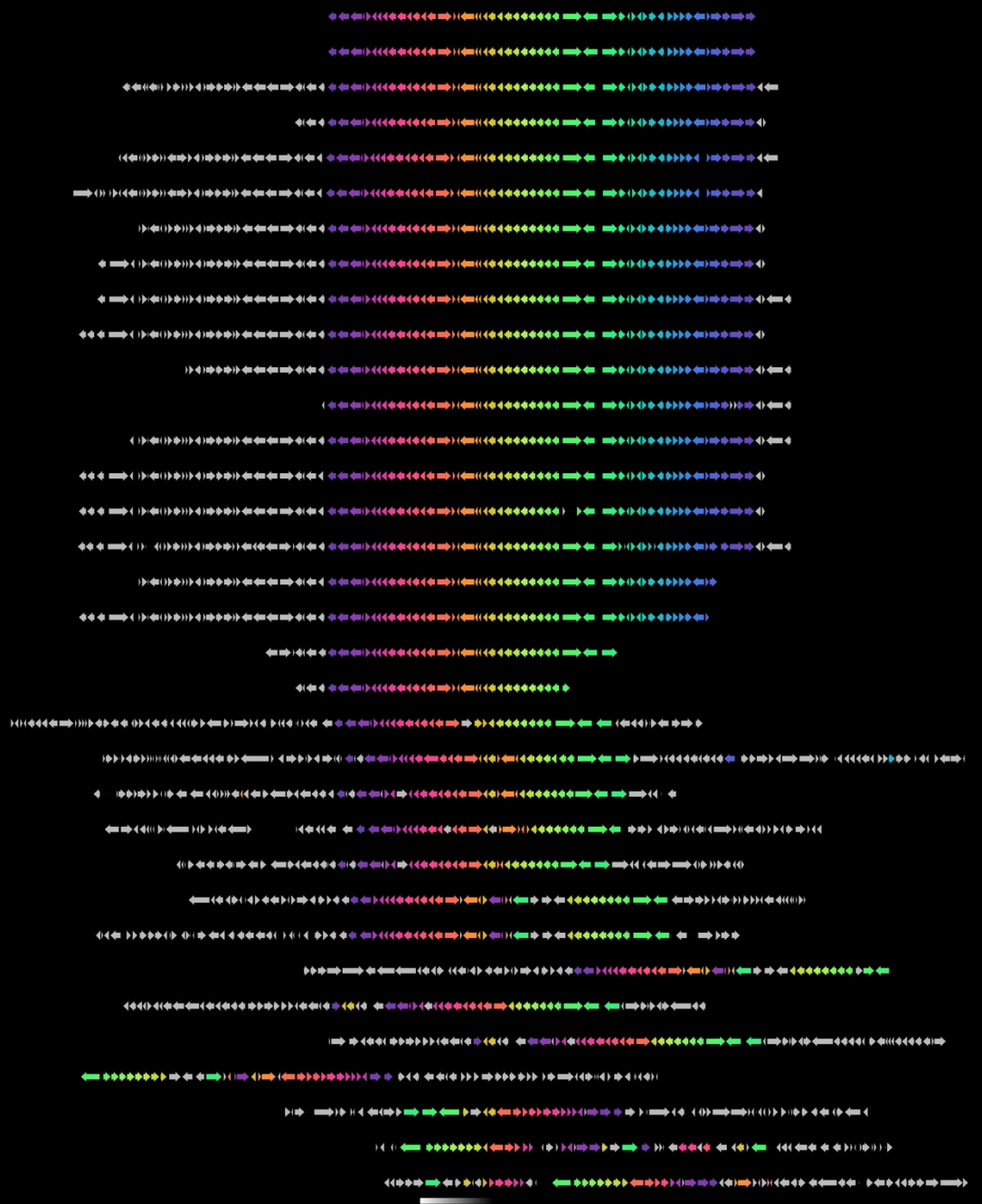
Outdegree of input protein domains

Protein	Locus/Descripton	Unique HMM models	Mean	Min	25%	50%	75%	Max	Sum
3aUcLZlofJ196v15TfvlMaxGpnFc6rMs	CAA60468.1 | no-locus-tag | pteridine-dependent dioxygenase	1	221	221	221	221	221	221	221
mZKlUBeQ7BRLGL8XlsYiYKslIxOgngr2	CAA60470.1 | no-locus-tag | methyltransferase	2	6735	6670	6702	6735	6768	6801	13471
s9HRwBkT10uC6uRfVw6cjuCmnc4BXWpC	CAA60452.1 | no-locus-tag | ABC-transporter	1	17624	17624	17624	17624	17624	17624	17624
KFQSeRRgS631GJRB_Qcou2BEmtH6bz0w	CAA60460.1 | no-locus-tag | polyketide synthase	3	6436	960	1371	1782	9174	16567	19309
sRl0ScCQMVcCyERmnvBlQ6H4s1MztBmm	CAA60475.1 | no-locus-tag | None	1	30410	30410	30410	30410	30410	30410	30410
a0IUrl7zPvLuJxqCD6vtU4pPpIW6um_R	CAA60459.1 | no-locus-tag | polyketide synthase	59	206147	20625	131559	172651	239874	685782	12162718

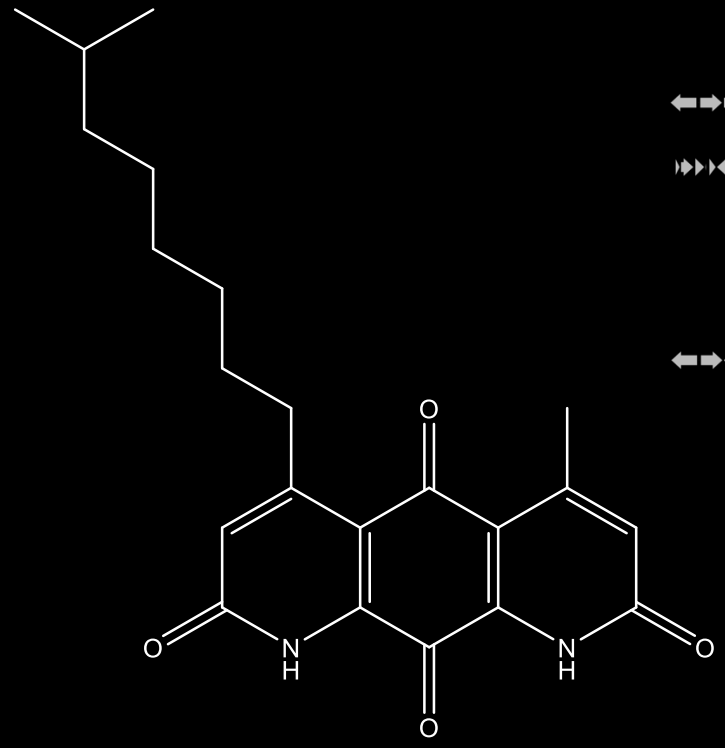
```
                INFO    Searching database for proteins with similar domain content and all of the genomes those are found in              hmmer.py:244

2024-02-26 12:04:53 INFO    Initial search returned 209,038 proteins, found in 27,879 genomes                                           hmmer.py:268
                INFO    Starting with matches across 27879 genomes                                                                         base.py:197
                INFO    Filtering on assembly_uid where unique hits >= 0.4                                                                 base.py:565
                INFO    Counting unique hits per assembly_uid                                                                              base.py:549
                INFO    14965 assemblies, 80149 nucleotide sequences had assembly_uids with >= 3 unique query hits                         base.py:574
                INFO    Filtering on nucleotide_uid where unique hits >= 0.4                                                               base.py:565
                INFO    Counting unique hits per nucleotide_uid                                                                            base.py:549
2024-02-26 12:04:54 INFO    4848 assemblies, 4907 nucleotide sequences had nucleotide_uids with >= 3 unique query hits                     base.py:574
                INFO    Grouping protein hits if less than 20000 bp apart                                                                  base.py:600
                INFO    Sorting genes by start position                                                                                    base.py:584
2024-02-26 12:04:55 INFO    Pulling data from the database for 28 putative BGCs                                                            base.py:334
                INFO    Time to fill: 0 seconds                                                                                            base.py:351
2024-02-26 12:04:56 INFO    Start: Creating links                                                                                          base.py:458
                INFO    Finding reciprocal best hits; protein similarity via Diamond BLASTp                                                base.py:468
2024-02-26 12:05:07 INFO    Finish: Creating links; 279 links produced                                                                     base.py:498
                INFO    Start: Assigning target BGC proteins to input BGC groups                                                           base.py:505
                INFO    Finish: Assigning target BGC proteins to input BGC groups                                                          base.py:515
                INFO    Writing clustermap.js output to: /home/chase/Downloads/clinker/plot/data.json                                  serialize.py:143
                INFO    Creating clustermap.js clusters                                                                                 serialize.py:40
                INFO    Creating clustermap.js links                                                                                   serialize.py:120
                INFO    Creating clustermap.js links                                                                                    serialize.py:95
```

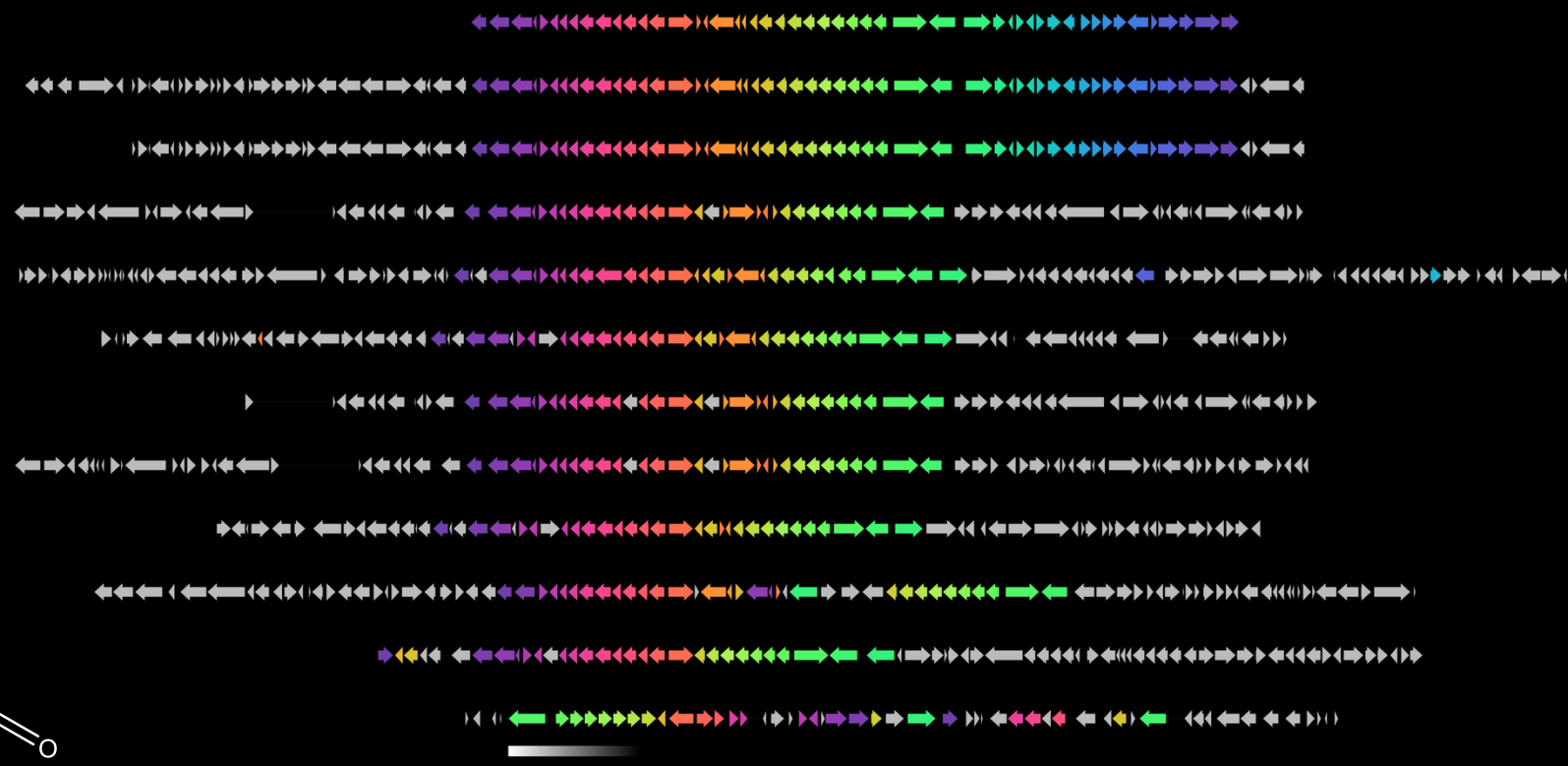
Searching for BGC0001848
in >300k genomes



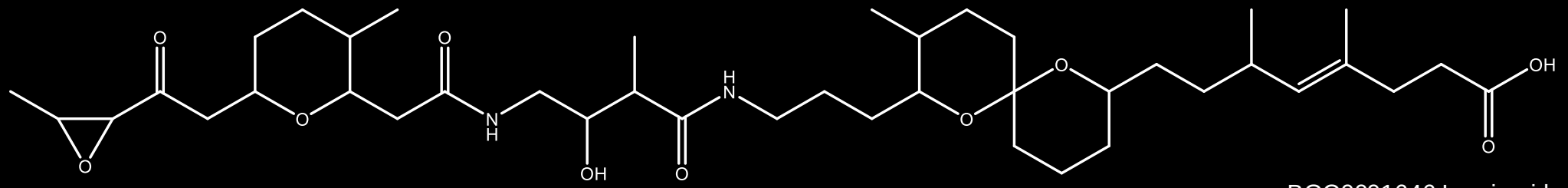


Searching for BGC0001848
in >300k genomes
filter for organisms in culture collections

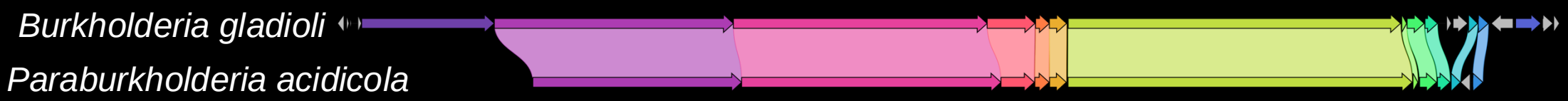




Find isolates containing metagenome-discovered BGCs



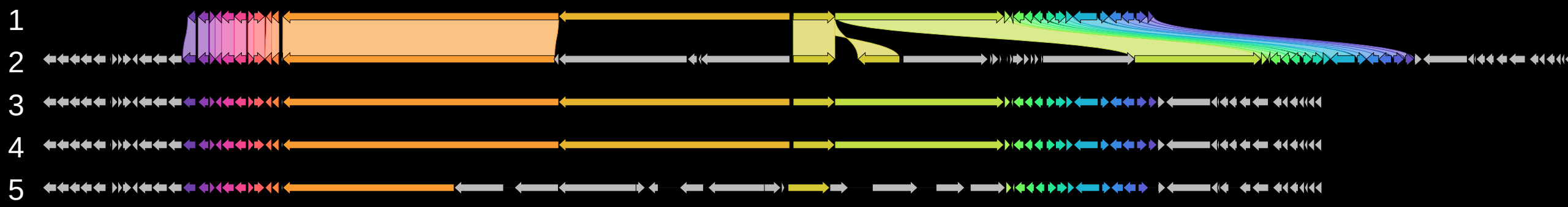
BGC0001646 Lagriamide



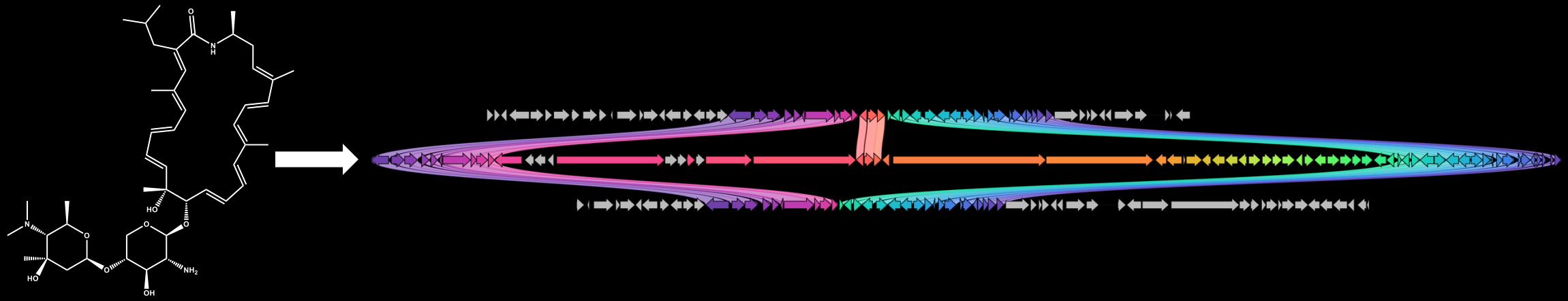
Fergusson, C. H., Saulog, J., Paulo, B. S., Wilson, D. M., Liu, D. Y., Morehouse, N. J., Waterworth, S., Barkei, J., Gray, C. A., Kwan, J. C., Eustaquio, A. S., and Linington, R. (2023) Discovery of the Polyketide Lagriamide B by Integrated Genome Mining, Isotopic Labeling, and Untargeted Metabolomics. ChemRxiv.

Search 30k Actinos for Rapamycin BGCs

Four Genome Assemblies of Streptomyces rapamycinicus NRRL 5491



Find… ??



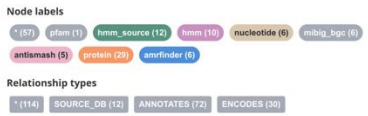
BGC0001522: auroramycin

Application

Can we target specific BGCs functionality?

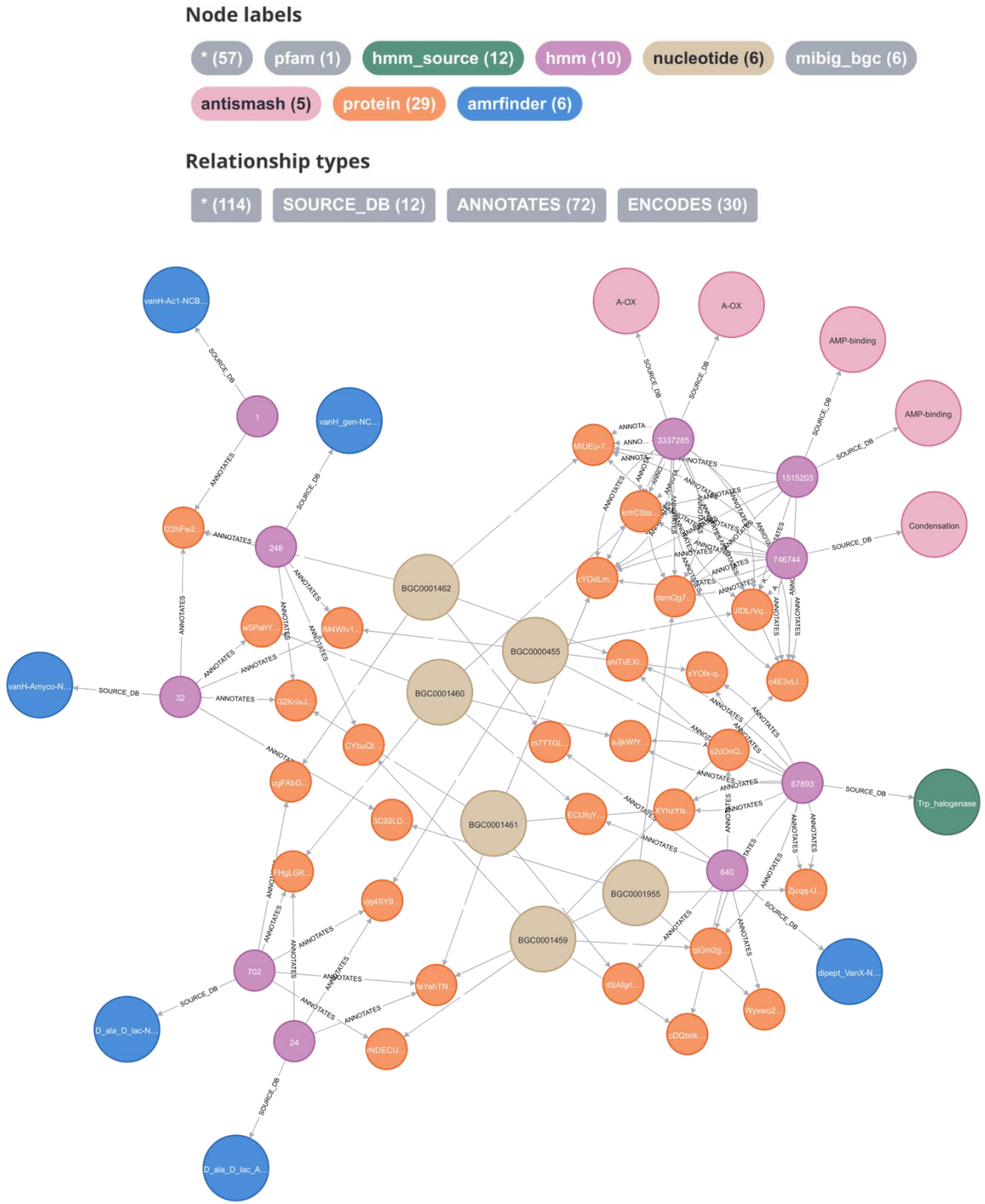
Find sequences with a halogenase, near an NRPS, near an antibiotic resistance gene, all on the same strand of DNA

```
:auto MATCH z1=(n:pfam {name:"Trp_halogenase"})←[:SOURCE_DB]-(h1:hmm)
MATCH z2=(h1)-[:ANNOTATES]-(:protein)←[e1:ENCODES]-(n1:mibig_bgc)
CALL {
    WITH n1, e1
    MATCH z3=(an1:antismash)←[:SOURCE_DB]-(:hmm)-[:ANNOTATES]→(p1:protein)←[e2:ENCODES]-(n1)
    MATCH z4=(an2:antismash)←[:SOURCE_DB]-(:hmm)-[:ANNOTATES]→(p1)
    WHERE an1.name ="Condensation"
        AND an2.name in ["AMP-binding", "A-OX"] AND abs(e1.start - e2.start) < 10000 AND e1.strand = e2.strand
    MATCH z5=(:amrfinder)←[:SOURCE_DB]-(:hmm)-[:ANNOTATES]→(p2:protein)←[e3:ENCODES]-(n1)
    WHERE abs(e1.start - e3.start) < 50000 AND e1.strand = e3.strand
    return z3, z4, z5
} in transactions of 1 rows
RETURN z1, z2, z3, z4, z5
```



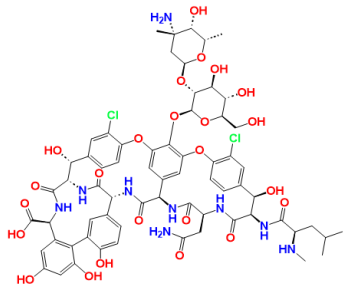
Find sequences with a halogenase, near an NRPS, near an antibiotic resistance gene, all on the same strand of DNA

```
:auto MATCH z1=(n:pfam {name:"Trp_halogenase"})<-[:SOURCE_DB]-(h1:hmm)
MATCH z2=(h1)-[:ANNOTATES]-(:protein)<-[e1:ENCODES]-(n1:mibig_bgc)
CALL {
    WITH n1, e1
    MATCH z3=(an1:antismash)<-[:SOURCE_DB]-(:hmm)-[:ANNOTATES]->(p1:protein)<-[e2:ENCODES]-(n1)
    MATCH z4=(an2:antismash)<-[:SOURCE_DB]-(:hmm)-[:ANNOTATES]->(p1)
    WHERE an1.name ="Condensation"
        AND an2.name in ["AMP-binding", "A-OX"] AND abs(e1.start - e2.start) < 10000 AND e1.strand = e2.strand
    MATCH z5=(:amrfinder)<-[:SOURCE_DB]-(:hmm)-[:ANNOTATES]->(p2:protein)<-[e3:ENCODES]-(n1)
    WHERE abs(e1.start - e3.start) < 50000 AND e1.strand = e3.strand
    return z3, z4, z5
} in transactions of 1 rows
RETURN z1, z2, z3, z4, z5
```

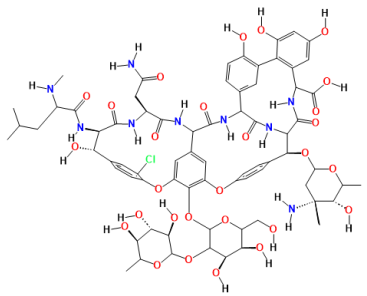


Find sequences with a halogenase, near an NRPS, near an antibiotic resistance gene, all on the same strand of DNA

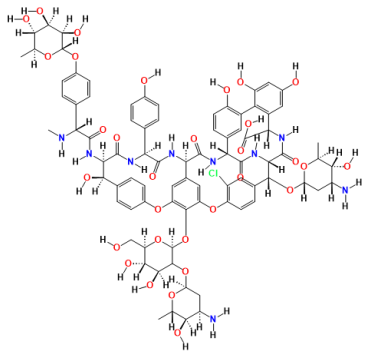
```
:auto MATCH z1=(n:pfam {name:"Trp_halogenase"})<-[:SOURCE_DB]-(h1:hmm)
MATCH z2=(h1)-[:ANNOTATES]-(:protein)<-[e1:ENCODES]-(n1:mibig_bgc)
CALL {
    WITH n1, e1
    MATCH z3=(an1:antismash)<-[:SOURCE_DB]-(:hmm)-[:ANNOTATES]->(p1:protein)<-[e2:ENCODES]-(n1)
    MATCH z4=(an2:antismash)<-[:SOURCE_DB]-(:hmm)-[:ANNOTATES]->(p1)
    WHERE an1.name ="Condensation"
        AND an2.name in ["AMP-binding", "A-OX"] AND abs(e1.start - e2.start) < 10000 AND e1.strand = e2.strand
    MATCH z5=(:amrfinder)<-[:SOURCE_DB]-(:hmm)-[:ANNOTATES]->(p2:protein)<-[e3:ENCODES]-(n1)
    WHERE abs(e1.start - e3.start) < 50000 AND e1.strand = e3.strand
    return z3, z4, z5
} in transactions of 1 rows
RETURN z1, z2, z3, z4, z5
```



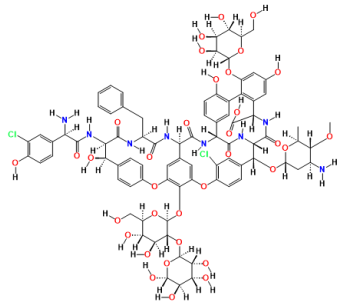
Vancomycin



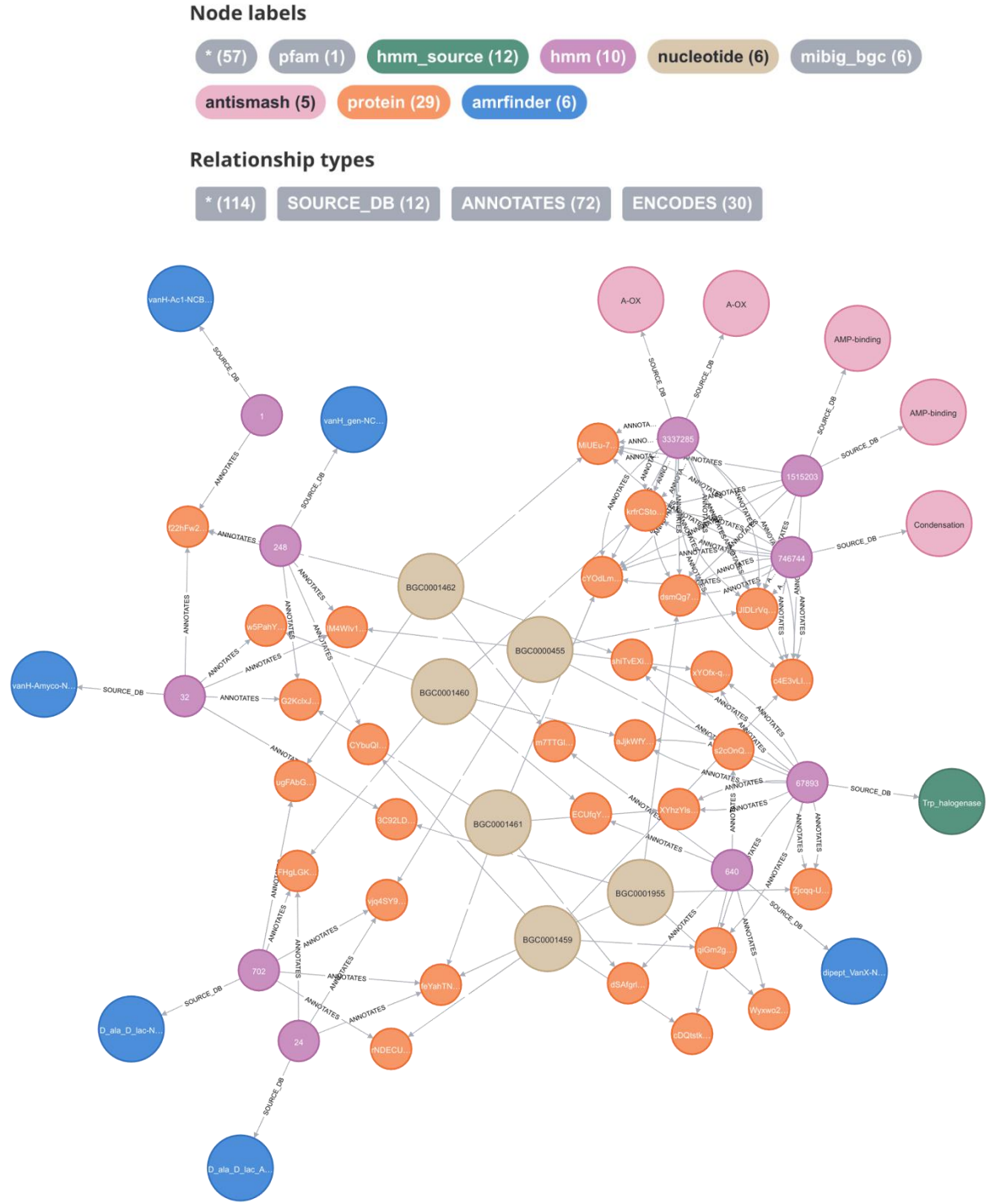
Decaplanin



Avoparcin

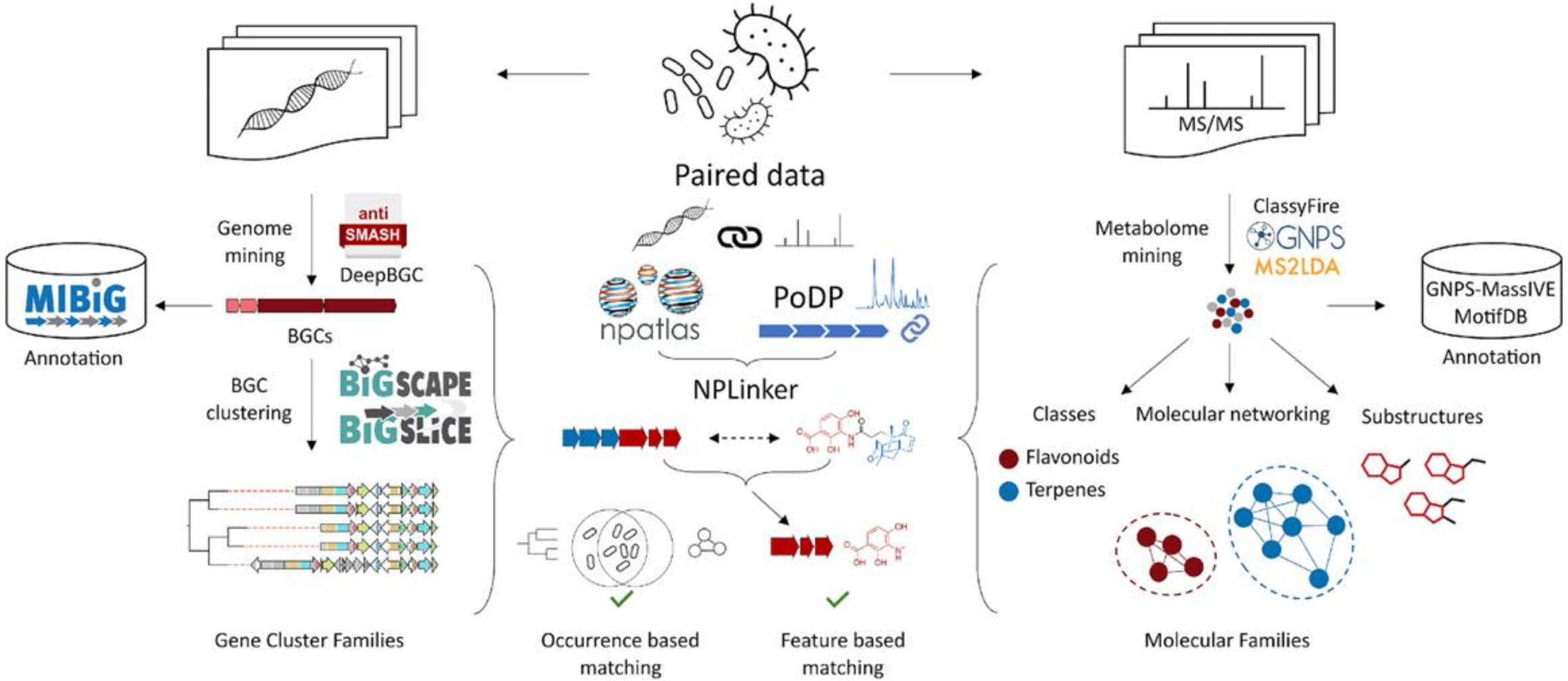


Keratinimicin A



Application

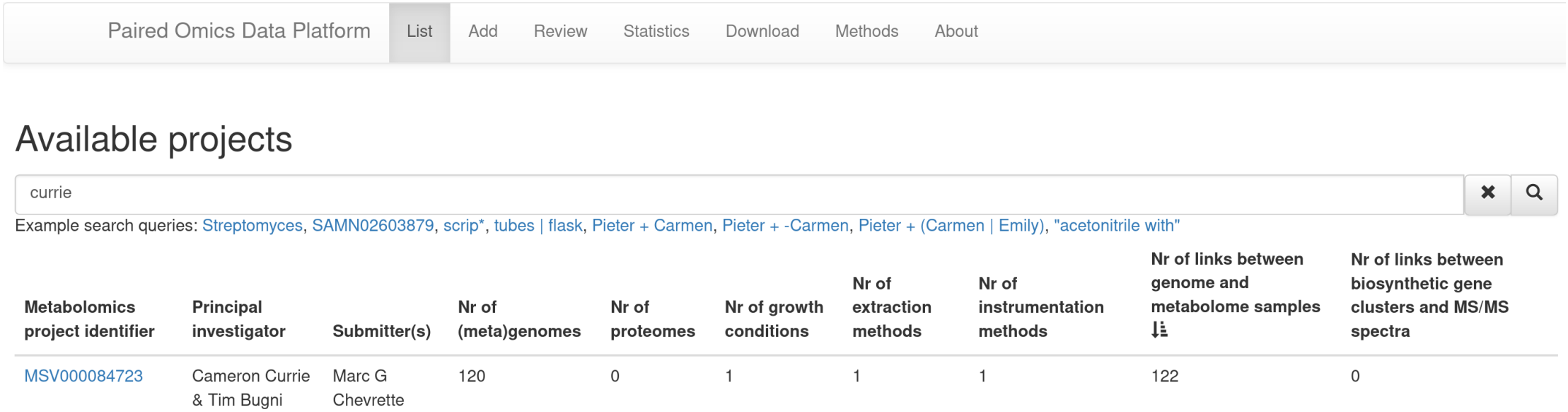
Paired Omics



Louwen JJR, van der Hooft JJJ. Comprehensive Large-Scale Integrative Analysis of Omics Data To Accelerate Specialized Metabolite Discovery. mSystems. 2021 Aug 31;6(4):e0072621. doi: 10.1128/mSystems.00726-21. Epub 2021 Aug 24

SocialGene

All 30,000+ Actinobacteria in RefSeq



pairedomicsdata.bioinformatics.nl

GNPS

Metabolomics project identifier	Principal investigator	Submitter(s)	Nr of (meta)genomes	Nr of proteomes	Nr of growth conditions	Nr of extraction methods	Nr of instrumentation methods	Nr of links between genome and metabolome samples	Nr of links between biosynthetic gene clusters and MS/MS spectra
MSV000084723	Cameron Currie & Tim Bugni	Marc G Chevrette	120	0	1	1	1	122	0

```
(sgpy) chase@titan:~/Downloads$ sg_import_gnps --gnps_dirpath ProteoSAFe-METABOLOMICS-SNETS-V2-927974dc-view_all_clusters_withID_beta --map_path genome_to_file_map.csv
2024-03-08 08:31:47 INFO     Connected to Neo4j database at bolt://localhost:7687                                neo4j.py:62
2024-03-08 08:31:50 INFO     Created 21930 (:ms2_spectrum) nodes, set 131580 properties                  neo4j_element.py:369
                    INFO     Created 63 (:gnps_library_spectrum) nodes, set 1633 properties              neo4j_element.py:369
                    INFO     Created 0 (:gnps_library_spectrum) nodes, set 0 properties                  neo4j_element.py:369
                    INFO     Created 1214 (:gnps_cluster) nodes, set 25695 properties                    neo4j_element.py:369
2024-03-08 08:31:52 INFO     Created 122 (:mass_spectrum_file) nodes, set 366 properties                 neo4j_element.py:369
2024-03-08 08:31:55 INFO     20027 relationships created (:ms2_spectrum)-[:CLUSTERS_TO]->(:gnps_cluster)  neo4j_element.py:538
2024-03-08 08:31:59 INFO     21930 relationships created (:mass_spectrum_file)-[:HAS]->(:ms2_spectrum)    neo4j_element.py:538
                    INFO     Created 0 (:gnps_cluster) nodes, set 0 properties                           neo4j_element.py:369
                    INFO     Created 55 (:gnps_library_spectrum) nodes, set 55 properties                neo4j_element.py:369
                    INFO     67 relationships created (:gnps_cluster)-[:LIBRARY_HIT]->(:gnps_library_spectrum)  neo4j_element.py:538
                    INFO     Created 0 (:gnps_cluster) nodes, set 0 properties                           neo4j_element.py:369
                    INFO     Created 0 (:gnps_library_spectrum) nodes, set 0 properties                  neo4j_element.py:369
                    INFO     67 relationships created (:gnps_cluster)-[:LIBRARY_HIT]->(:gnps_library_spectrum)  neo4j_element.py:538
2024-03-08 08:32:00 INFO     Created 0 (:gnps_cluster) nodes, set 0 properties                           neo4j_element.py:369
                    INFO     Created 0 (:gnps_cluster) nodes, set 0 properties                           neo4j_element.py:369
                    INFO     2018 relationships created (:gnps_cluster)-[:MOLECULAR_NETWORK]->(:gnps_cluster)  neo4j_element.py:538
                    INFO     Created 39 (:chemical_compound) nodes, set 1599 properties                  neo4j_element.py:369
                    INFO     46 relationships created (:gnps_library_spectrum)-[:IS_A]->(:chemical_compound)  neo4j_element.py:538
                    INFO     Assemblies in GNPS results found in db: 84 of 84                                parse.py:237
                    INFO     Assemblies in GNPS results not found in db: set()                               parse.py:240
2024-03-08 08:32:01 INFO     86 relationships created (:mass_spectrum_file)-[:ANALYSIS_OF]->(:assembly)  neo4j_element.py:538
                    INFO     GNPS molecular network has been integrated into the SocialGene Neo4j database       cli.py:146
```

GNPS npatlas

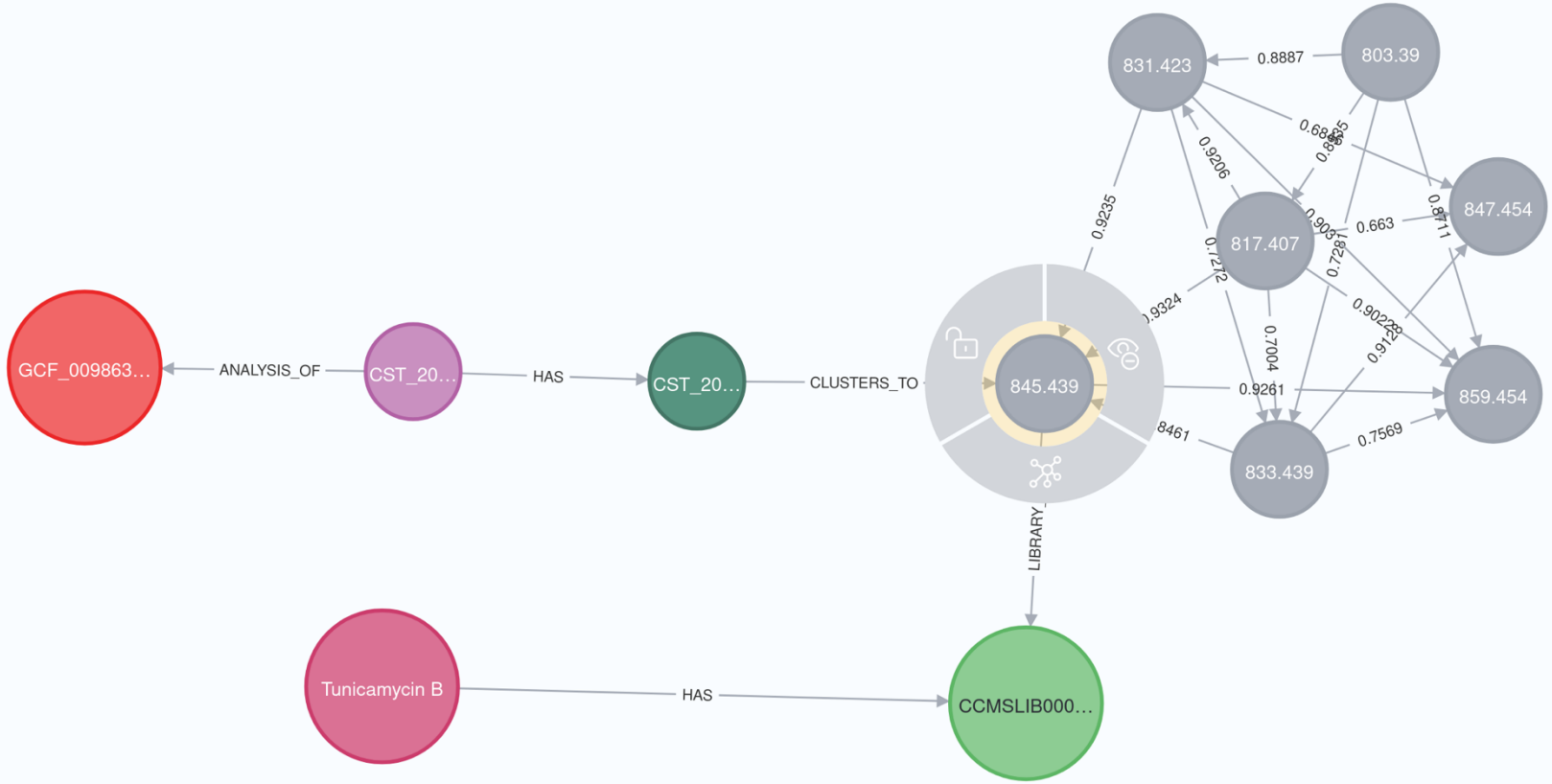
Metabolomics project identifier	Principal investigator	Submitter(s)	Nr of (meta)genomes	Nr of proteomes	Nr of growth conditions	Nr of extraction methods	Nr of instrumentation methods	Nr of links between genome and metabolome samples	Nr of links between biosynthetic gene clusters and MS/MS spectra
MSV000084723	Cameron Currie & Tim Bugni	Marc G Chevrette	120	0	1	1	1	122	0

```
(sgpy) chase@titan:~/Downloads$ sg_import_npatlas --input NPAtlas_download.json
33372 Processing NPAtlas entries... ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:02:30
2024-03-08 08:35:13 INFO     Creating/Merging npatlas nodes in neo4                                  cli.py:73
                    INFO     Connected to Neo4j database at bolt://localhost:7687                    neo4j.py:62
2024-03-08 08:35:16 INFO     Created 33372 (:npatlas) nodes, set 467208 properties                   neo4j_element.py:369
2024-03-08 08:35:18 INFO     Creating/Merging nodes linked to npatlas entries in neo4j               cli.py:83
2024-03-08 08:35:19 INFO     Created 13058 (:publication) nodes, set 78348 properties                neo4j_element.py:369
                    INFO     Created 8 (:taxid) nodes, set 8 properties                              neo4j_element.py:369
2024-03-08 08:35:20 INFO     Created 27196 (:gnps_library_spectrum) nodes, set 27196 properties      neo4j_element.py:369
2024-03-08 08:35:21 INFO     Created 96 (:assembly:mibig) nodes, set 96 properties                   neo4j_element.py:369
                    INFO     Created 1632 (:classyfire) nodes, set 1632 properties                   neo4j_element.py:369
                    INFO     Created 509 (:npclassifier_class) nodes, set 509 properties             neo4j_element.py:369
                    INFO     Created 7 (:npclassifier_pathway) nodes, set 7 properties               neo4j_element.py:369
                    INFO     Created 76 (:npclassifier_superclass) nodes, set 76 properties          neo4j_element.py:369
2024-03-08 08:35:28 INFO     Created 33339 (:chemical_compound) nodes, set 1366899 properties        neo4j_element.py:369
                    INFO     Created 911 (:chebi) nodes, set 1822 properties                         neo4j_element.py:369
                    INFO     Linking npatlas entries and related nodes in neo4j                      cli.py:87
2024-03-08 08:35:31 INFO     33372 relationships created (:npatlas)-[:HAS]->(:publication)           neo4j_element.py:538
2024-03-08 08:35:33 INFO     31481 relationships created (:taxid)-[:PRODUCES]->(:npatlas)            neo4j_element.py:538
2024-03-08 08:35:36 INFO     31887 relationships created (:npatlas)-[:HAS]->(:gnps_library_spectrum) neo4j_element.py:538
                    INFO     2511 relationships created (:assembly:mibig)-[:PRODUCES]->(:npatlas)    neo4j_element.py:538
2024-03-08 08:35:38 INFO     32888 relationships created (:npatlas)-[:LOWEST_CLASS]->(:classyfire)   neo4j_element.py:538
2024-03-08 08:35:40 INFO     32888 relationships created (:npatlas)-[:DIRECT_PARENT]->(:classyfire)  neo4j_element.py:538
2024-03-08 08:35:41 INFO     10694 relationships created (:npatlas)-[:INTERMEDIATE_NODES]->(:classyfire) neo4j_element.py:538
2024-03-08 08:36:15 INFO     444892 relationships created (:npatlas)-[:ALTERNATIVE_PARENTS]->(:classyfire) neo4j_element.py:538
2024-03-08 08:36:18 INFO     31400 relationships created (:npatlas)-[:IS_A]->(:npclassifier_class)   neo4j_element.py:538
2024-03-08 08:36:21 INFO     34822 relationships created (:npatlas)-[:IS_A]->(:npclassifier_pathway) neo4j_element.py:538
2024-03-08 08:36:23 INFO     28759 relationships created (:npatlas)-[:IS_A]->(:npclassifier_superclass) neo4j_element.py:538
2024-03-08 08:36:26 INFO     33371 relationships created (:npatlas)-[:IS_A]->(:chemical_compound)     neo4j_element.py:538
2024-03-08 08:37:37 INFO     801979 relationships created (:npatlas)-[:IS_A]->(:chebi)               neo4j_element.py:538
```

GNPS npatlas

```
MATCH z1=(:assembly)<-[r:ANALYSIS_OF]-(msf:mass_spectrum_file)-[:HAS]->(ms2:ms2_spectrum)-[:CLUSTERS_TO]->(gc:gnps_cluster)-[:LIBRARY_HIT]->(:gnps_library_spectrum)-[HAS]-(np:npatlas)
MATCH (np)<-[:PRODUCES]-(:taxid {name: "Streptomyces"})
MATCH z2=(gc)-[:MOLECULAR_NETWORK*1..2]-(:gnps_cluster)
RETURN z1, z2 limit 20
```

Graph
Table
Text
Code

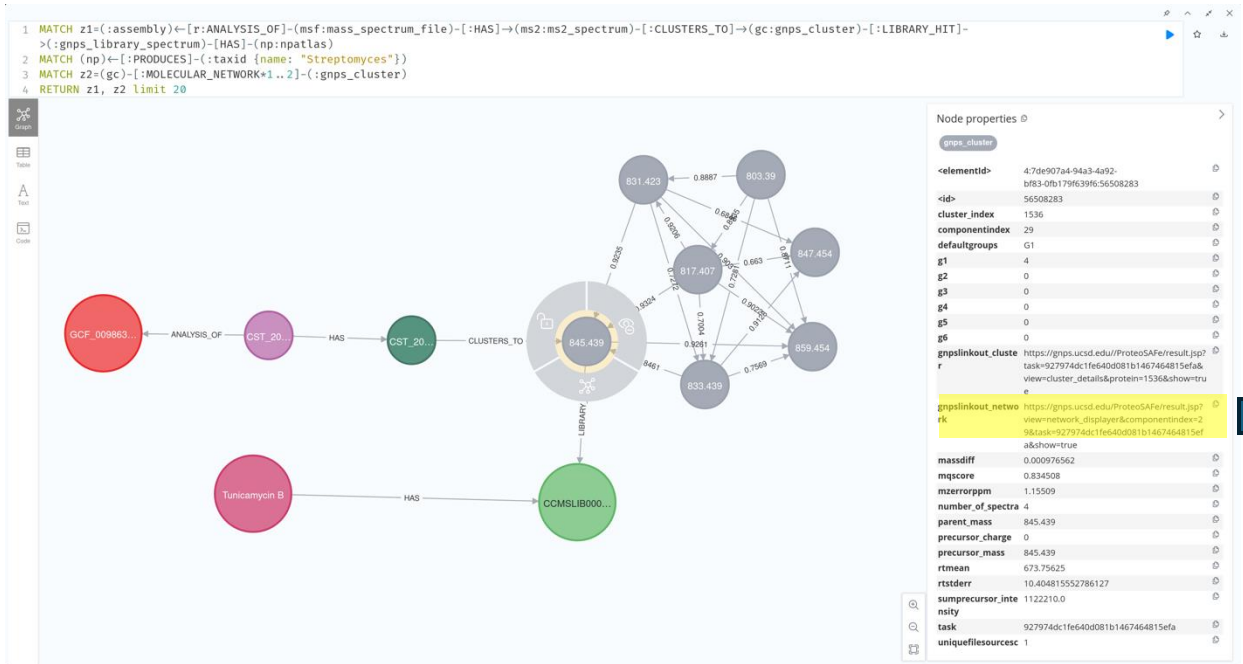


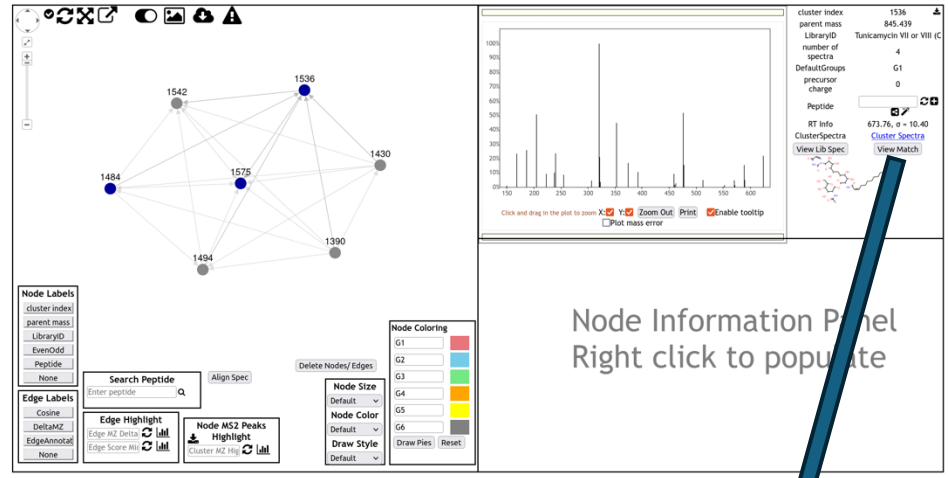
Node properties

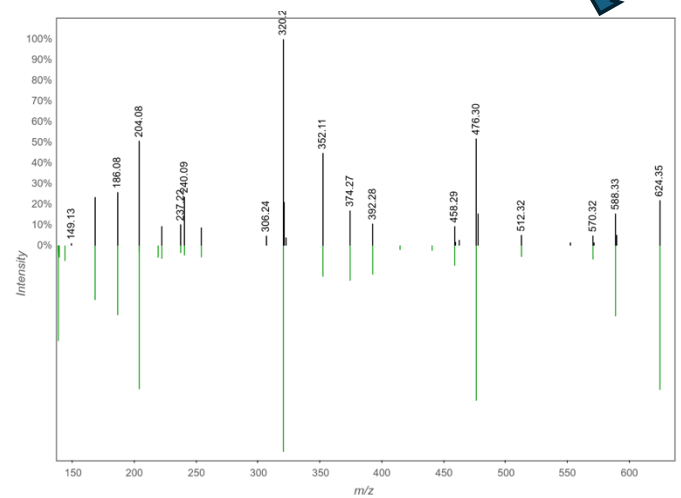
gnps_cluster

<elementId>	4:7de907a4-94a3-4a92-bf83-0fb179f639f6:56508283
<id>	56508283
cluster_index	1536
componentindex	29
defaultgroups	G1
g1	4
g2	0
g3	0
g4	0
g5	0
g6	0
gnpslinkout_cluster	https://gnps.ucsd.edu//ProteoSAFe/result.jsp?task=927974dc1fe640d081b1467464815efa&view=cluster_details&protein=1536&show=true
gnpslinkout_network	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?view=network_displayer&componentindex=29&task=927974dc1fe640d081b1467464815efa&show=true
massdiff	0.000976562
mqscore	0.834508
mzerrorppm	1.15509
number_of_spectra	4
parent_mass	845.439
precursor_charge	0
precursor_mass	845.439
rtmean	673.75625
rtstderr	10.404815552786127
sumprecursor_intensity	1122210.0
task	927974dc1fe640d081b1467464815efa
uniquefilesourcesc	1

SocialGene







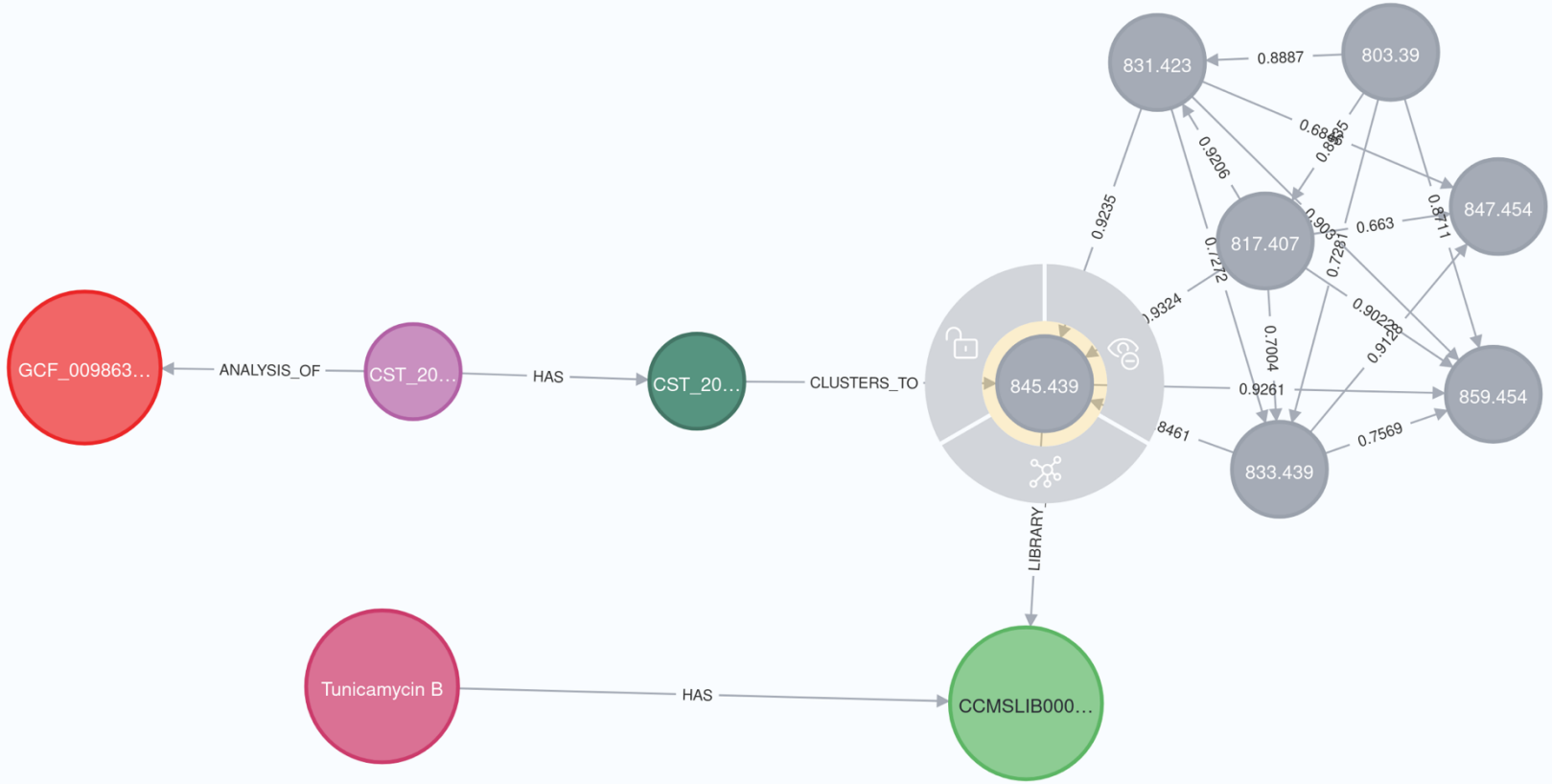
SocialGene

GNPS

npatlas

```
MATCH z1=(:assembly)<-[r:ANALYSIS_OF]-(msf:mass_spectrum_file)-[:HAS]->(ms2:ms2_spectrum)-[:CLUSTERS_TO]->(gc:gnps_cluster)-[:LIBRARY_HIT]->(:gnps_library_spectrum)-[HAS]-(np:npatlas)
MATCH (np)<-[:PRODUCES]-(:taxid {name: "Streptomyces"})
MATCH z2=(gc)-[:MOLECULAR_NETWORK*1..2]-(:gnps_cluster)
RETURN z1, z2 limit 20
```

Graph
Table
Text
Code



Node properties

gnps_cluster

<elementId>	4:7de907a4-94a3-4a92-bf83-0fb179f639f6:56508283
<id>	56508283
cluster_index	1536
componentindex	29
defaultgroups	G1
g1	4
g2	0
g3	0
g4	0
g5	0
g6	0
gnpslinkout_cluster	https://gnps.ucsd.edu//ProteoSAFe/result.jsp?task=927974dc1fe640d081b1467464815efa&view=cluster_details&protein=1536&show=true
gnpslinkout_network	https://gnps.ucsd.edu/ProteoSAFe/result.jsp?view=network_displayer&componentindex=29&task=927974dc1fe640d081b1467464815efa&show=true
massdiff	0.000976562
mqscore	0.834508
mzerrorppm	1.15509
number_of_spectra	4
parent_mass	845.439
precursor_charge	0
precursor_mass	845.439
rtmean	673.75625
rtstderr	10.404815552786127
sumprecursor_intensity	1122210.0
task	927974dc1fe640d081b1467464815efa
uniquefilesourcesc	1

SocialGene

Application

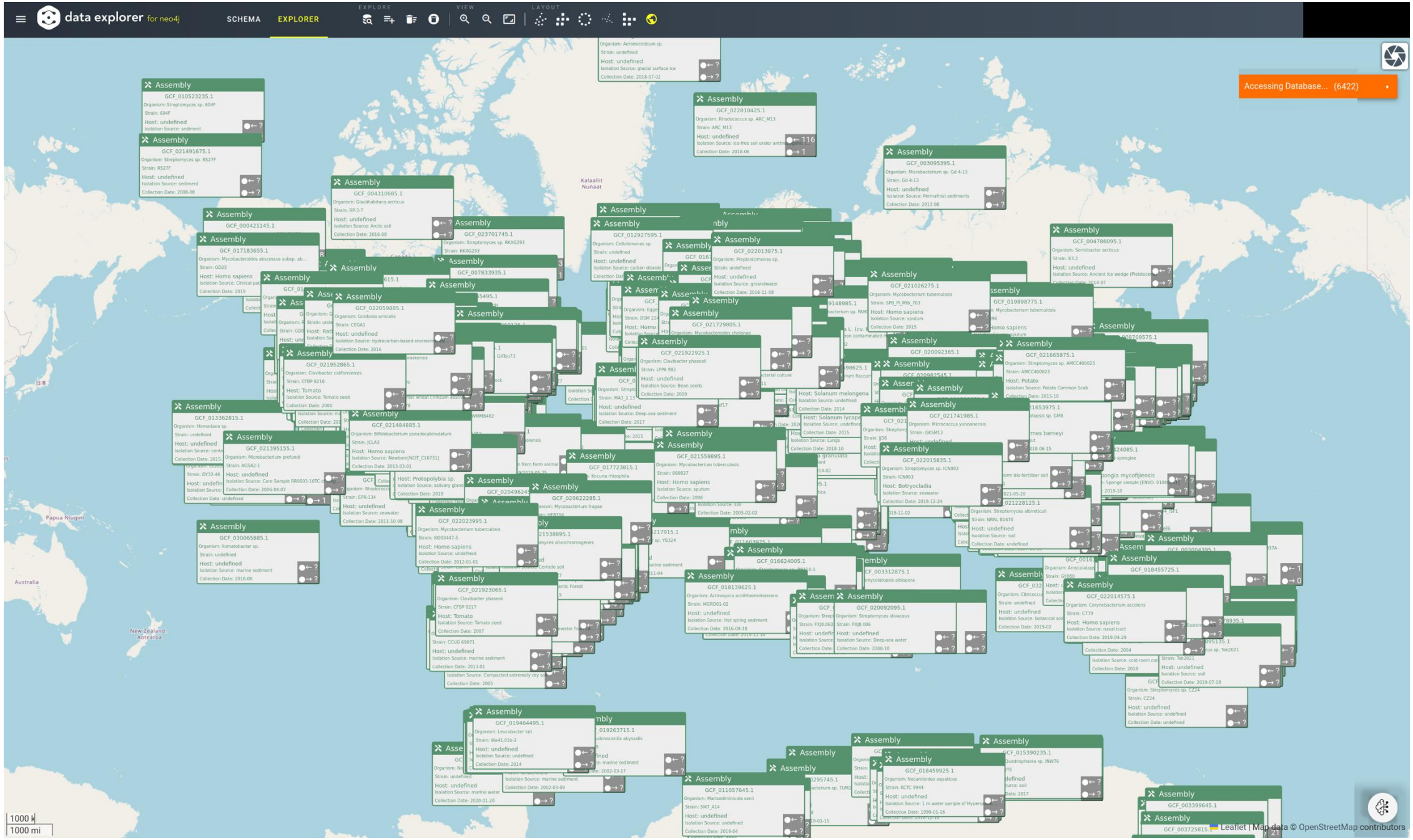
Where in the world are my BGCs?

Pull and store latitude and longitude in-database

```
1 MATCH (n:assembly) where  n.lat_lon is not null
2 WITH n, split(n.lat_lon, ' ') AS parts
3 WITH n, toFloat(parts[0]) AS latitude, parts[1] AS lat_direction, toFloat(parts[2]) AS longitude, parts[3] AS lon_direction
4 WITH n,
5   CASE lat_direction WHEN 'N' THEN latitude WHEN 'S' THEN -latitude ELSE null END AS latitude_value,
6   CASE lon_direction WHEN 'E' THEN longitude WHEN 'W' THEN -longitude ELSE null END AS longitude_value
7 SET n.geolocation= point({latitude: latitude_value, longitude: longitude_value})
```

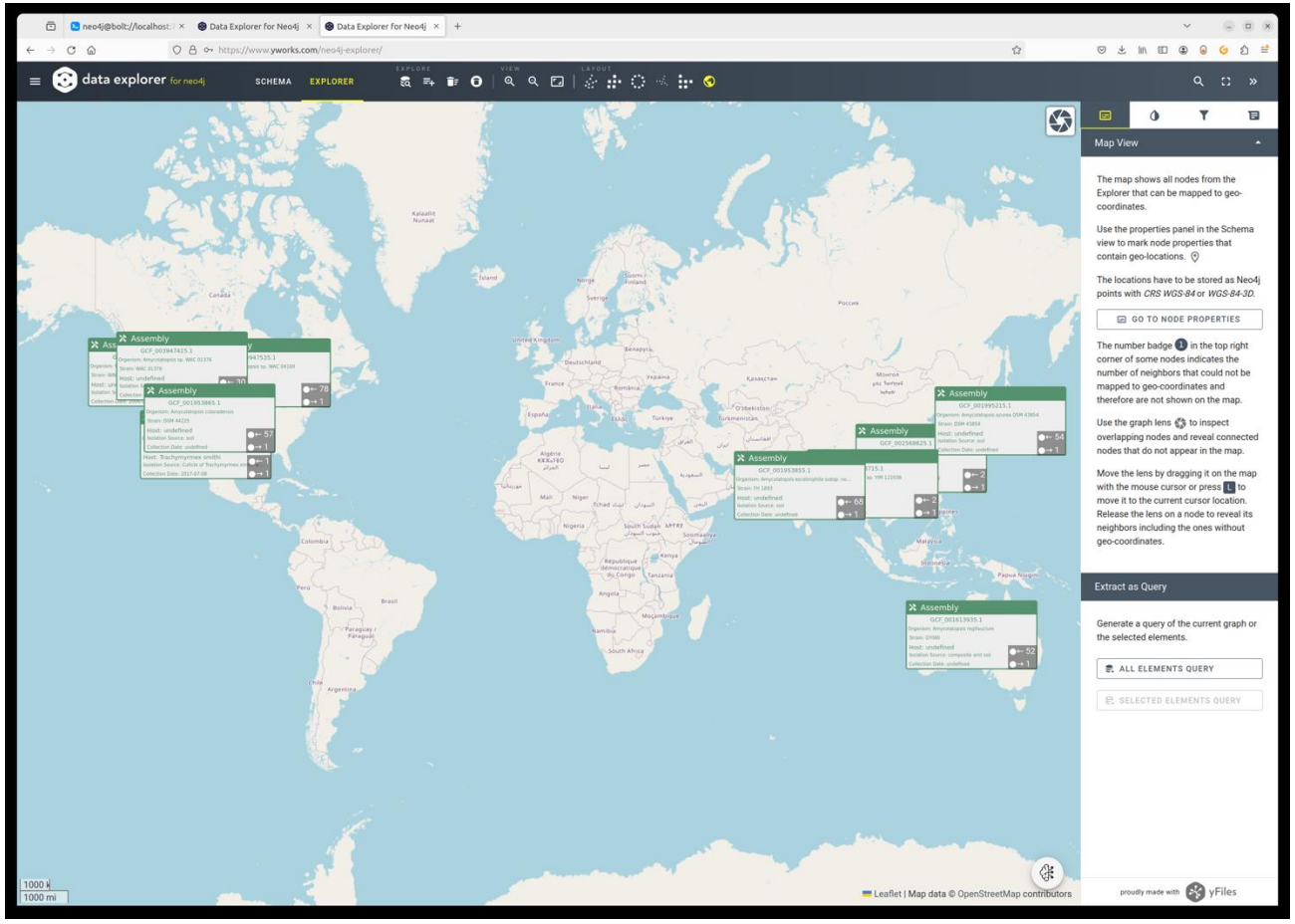
Table

Set 88699 properties, completed after 553 ms.


data explorer for neo4j
SCHEMA
EXPLORER
EXPLORE
VIEW
LAYOUT
Accessing Database... (6422)
Assembly
GCF_010523235.1
Organism: Streptomyces sp. 604F
Strain: 604F
Host: undefined
Isolation Source: sediment
Assembly
GCF_021491675.1
Organism: Streptomyces sp. RS27F
Strain: RS27F
Host: undefined
Isolation Source: sediment
Collection Date: 2008-08
Assembly
GCF_022810425.1
Organism: Rhodococcus sp. ARC_M13
Strain: ARC_M13
Host: undefined
Collection Date: 2018-06
Assembly
GCF_003095395.1
Organism: Microbacterium sp. Gd 4-13
Strain: Gd 4-13
Host: undefined
Isolation Source: Permafrost sediments
Collection Date: 2013-08
Assembly
GCF_004310685.1
Organism: Glaciihabitans arcticus
Strain: RP-3-7
Host: undefined
Isolation Source: Arctic soil
Collection Date: 2016-08
Assembly
GCF_000421145.1
Assembly
GCF_017183655.1
Organism: Mycobacteroides abscessus subsp. ab...
Strain: GD25
Host: Homo sapiens
Isolation Source: Clinical pa
Collection Date: 2019
Assembly
GCF_023701745.1
Organism: Streptomyces sp. RKAG293
Strain: RKAG293
Assembly
GCF_007833935.1
Assembly
GCF_022059885.1
Organism: Gordonia amicalis
Strain: CEGA1
Host: undefined
Isolation Source: hydrocarbon-based environm
Collection Date: 2016
Assembly
GCF_021952865.1
Organism: Clavibacter californiensis
Strain: CFBP 8216
Host: Tomato
Isolation Source: Tomato seed
Collection Date: 2000
Assembly
GCF_013362815.1
Organism: Hamadaea sp.
Strain: undefined
Host: undefined
Assembly
GCF_021395155.1
Organism: Microbacterium profundi
Strain: AGSA2-1
Host: undefined
Isolation Source: Core Sample RR0603-10TC
Collection Date: 2006-04-07
Assembly
GCF_021484885.1
Organism: Bifidobacterium pseudocatenulatum
Strain: JCLA3
Host: Homo sapiens
Isolation Source: Newborn[NCIT_C16731]
Collection Date: 2013-03-01
Assembly
GCF_030065885.1
Organism: Ilumatobacter sp.
Strain: undefined
Host: undefined
Isolation Source: marine sediment
Collection Date: 2018-08
Assembly
GCF_012927595.1
Organism: Cellulomonas sp.
Strain: undefined
Host: undefined
Isolation Source: carbon dioxide
Assembly
GCF_022013875.1
Organism: Propionicimonas sp.
Strain: undefined
Host: undefined
Isolation Source: groundwater
Collection Date: 2016-11-08
Assembly
GCF_021729805.1
Organism: Mycobacteroides chelonae
Assembly
GCF_021922925.1
Organism: Clavibacter phaseoli
Strain: LPPA 982
Host: undefined
Isolation Source: Bean seeds
Collection Date: 2009
Assembly
GCF_021026275.1
Organism: Mycobacterium tuberculosis
Strain: 5PB_PI_Mtb_703
Host: Homo sapiens
Isolation Source: sputum
Collection Date: 2015
Assembly
GCF_004786095.1
Organism: Serinibacter arcticus
Strain: K3-2
Host: undefined
Isolation Source: Ancient ice wedge (Pleistoce
Assembly
GCF_019898775.1
Organism: Mycobacterium tuberculosis
Host: Homo sapiens
Isolation Source: sputum
Assembly
GCF_020092365.1
Assembly
GCF_021665875.1
Organism: Streptomyces sp. AMCC400023
Strain: AMCC400023
Host: Potato
Isolation Source: Potato Common Scab
Collection Date: 2015-10
Assembly
GCF_021741985.1
Organism: Micrococcus yunnanensis
Strain: GKSM13
Assembly
GCF_022015835.1
Organism: Streptomyces sp. ICN903
Strain: ICN903
Host: Botryocladia
Isolation Source: seawater
Collection Date: 2018-12-24
Assembly
GCF_021559895.1
Organism: Mycobacterium tuberculosis
Strain: 060827
Host: Homo sapiens
Isolation Source: sputum
Collection Date: 2006
Assembly
GCF_017723815.1
Assembly
GCF_020622285.1
Organism: Mycobacterium fragae
Assembly
GCF_022023995.1
Organism: Mycobacterium tuberculosis
Strain: I0003447.5
Host: Homo sapiens
Isolation Source: undefined
Collection Date: 2012-01-01
Assembly
GCF_021923065.1
Organism: Clavibacter phaseoli
Strain: CFBP 8217
Host: Tomato
Isolation Source: Tomato seed
Collection Date: 2007
Strain: CCUG 69071
Host: undefined
Isolation Source: marine sediment
Collection Date: 2013-01
Isolation Source: Compacted extremely dry so
Collection Date: 2005
Assembly
GCF_016624005.1
Assembly
GCF_018139625.1
Organism: Actinospica acidithermotolerans
Strain: MGRD01-02
Host: undefined
Isolation Source: Hot spring sediment
Collection Date: 2016-08-18
Assembly
GCF_020092095.1
Organism: Streptomyces olivaceus
Strain: FXJ8.006
Host: undefined
Isolation Source: Deep-sea water
Collection Date: 2008-10
Assembly
GCF_003312875.1
Assembly
GCF_018455725.1
Assembly
GCF_022014575.1
Organism: Corynebacterium accolens
Strain: C779
Host: Homo sapiens
Isolation Source: nasal tract
Collection Date: 2019-04-29
Assembly
GCF_019464495.1
Organism: Leucobacter luti
Strain: Wo41.01b-2
Host: undefined
Isolation Source: undefined
Collection Date: 2014
Assembly
GCF_019263715.1
Assembly
GCF_011057645.1
Organism: Marisediminicola senii
Strain: SM7_A14
Host: undefined
Isolation Source: undefined
Collection Date: 2019-04
Assembly
GCF_018459925.1
Organism: Nocardioides aquaticus
Strain: KCTC 9944
Host: undefined
Isolation Source: 1 m water sample of Hypers
Collection Date: 1990-01-16
Assembly
GCF_015390235.1
Organism: Quadrisphaera sp. INWT6
Host: undefined
Isolation Source: soil
Collection Date: 2017
Assembly
GCF_003399645.1
Assembly
GCF_003725815.1
Kalaallit Nunaat
日本
Papua Niugini
Australia
New Zealand Aotearoa
1000 k
1000 mi
Leaflet | Map data © OpenStreetMap contributors

66 seconds

Geolocate Putative Halogenated NRPS Antibiotics

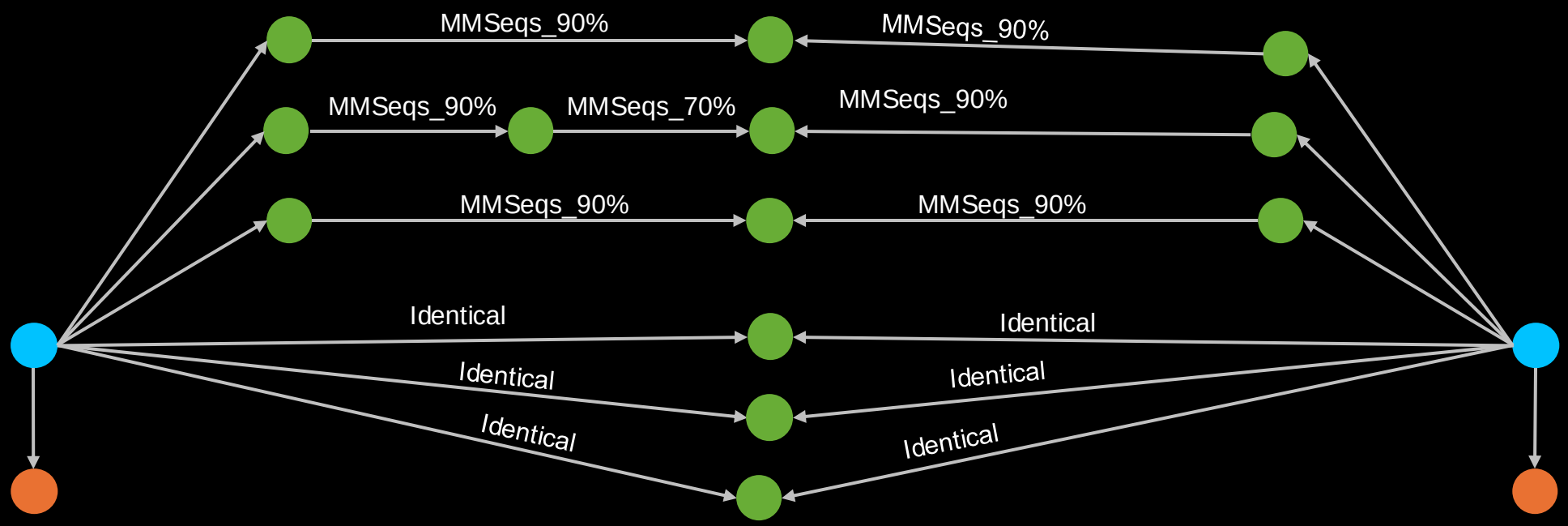



```
MATCH z1=(n:pfam {name:"Trp_halogenase"})←[:SOURCE_DB]-(h1:hmm)
MATCH z2=(h1)-[:ANNOTATES]-(:protein)←[e1:ENCODES]-(n1:nucleotide)
CALL {
    WITH n1, e1
    MATCH z3=(an1:antismash)←[:SOURCE_DB]-(:hmm)-[:ANNOTATES]→(p1:protein)←[e2:ENCODES]-(n1)
    MATCH z4=(an2:antismash)←[:SOURCE_DB]-(:hmm)-[:ANNOTATES]→(p1)
    WHERE an1.name ="Condensation"
        AND an2.name in ["AMP-binding", "A-OX"] AND abs(e1.start - e2.start) < 10000 AND e1.strand = e2.strand
    MATCH z5=(:amrfinder)←[:SOURCE_DB]-(:hmm)-[:ANNOTATES]→(p2:protein)←[e3:ENCODES]-(n1)
    WHERE abs(e1.start - e3.start) < 50000 AND e1.strand = e3.strand
    MATCH (n1)-[:ASSEMBLES_TO]→(a1:assembly)
    WHERE az.lat_lon is not null
    return az
} in transactions of 1 rows
RETURN distinct a1
```

Application

Global analyses

Protein sequence similarity with MMSeqs2 clustering



Genome DNA Sequence Protein

Look in 343,381 genomes
For similar BGCs to
Any of 2,502 MIBiG BGCs
Limit to the
2.1 million sequence regions
predicted by antiSMASH 7

Look in 343,381 genomes
For similar BGCs to
Any of 2,502 MIBiG BGCs
Limit to the
2.1 million sequence regions
predicted by antiSMASH 7

- Genome with match

MIBiG BGC

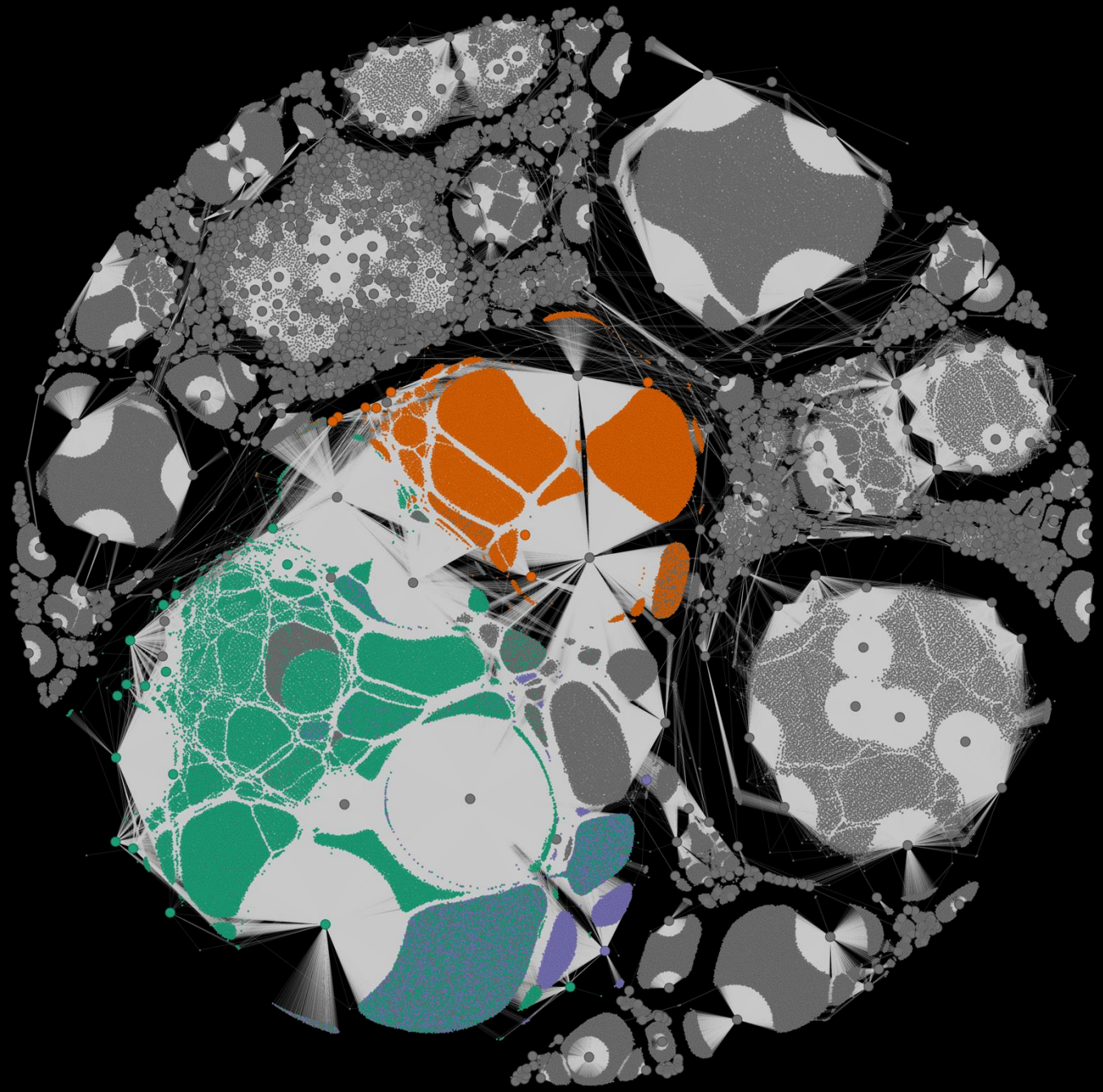
>70% of proteins, >70% AA identity

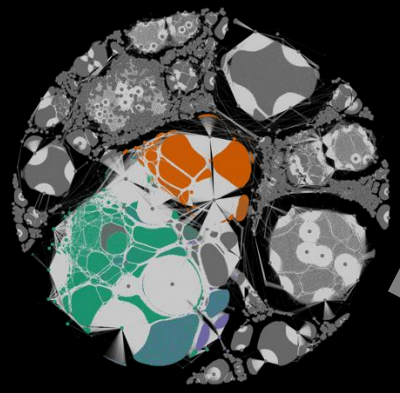
184,571 genomes
1,964 MIBiG BGCs
597,917 relationships
2.5 minutes

Look in 343,381 genomes
For similar BGCs to
Any of 2,502 MIBiG BGCs
Limit to the
2.1 million sequence regions
predicted by antiSMASH 7

- Genome with match
- MIBiG BGC
- *>70% of proteins, >70% AA identity*
- *Klebsiella*
- *Escherichia*
- *Salmonella*

184,571 genomes
1,964 MIBiG BGCs
597,917 relationships
2.5 minutes





How does BGC distribution contribute to pathogenicity?

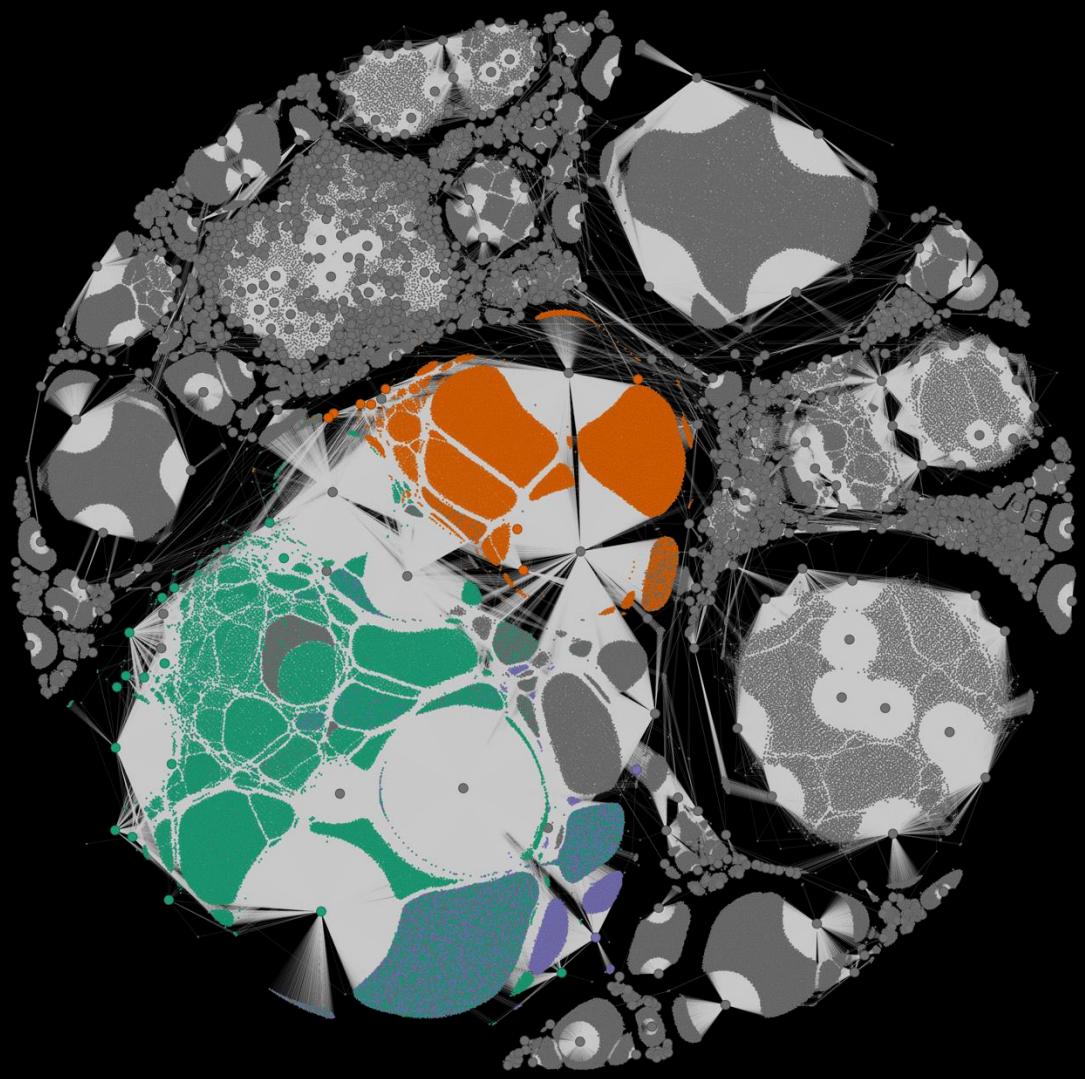
- Genome with match
- MIBiG BGC

- *Klebsiella*
- *Escherichia*
- *Salmonella*

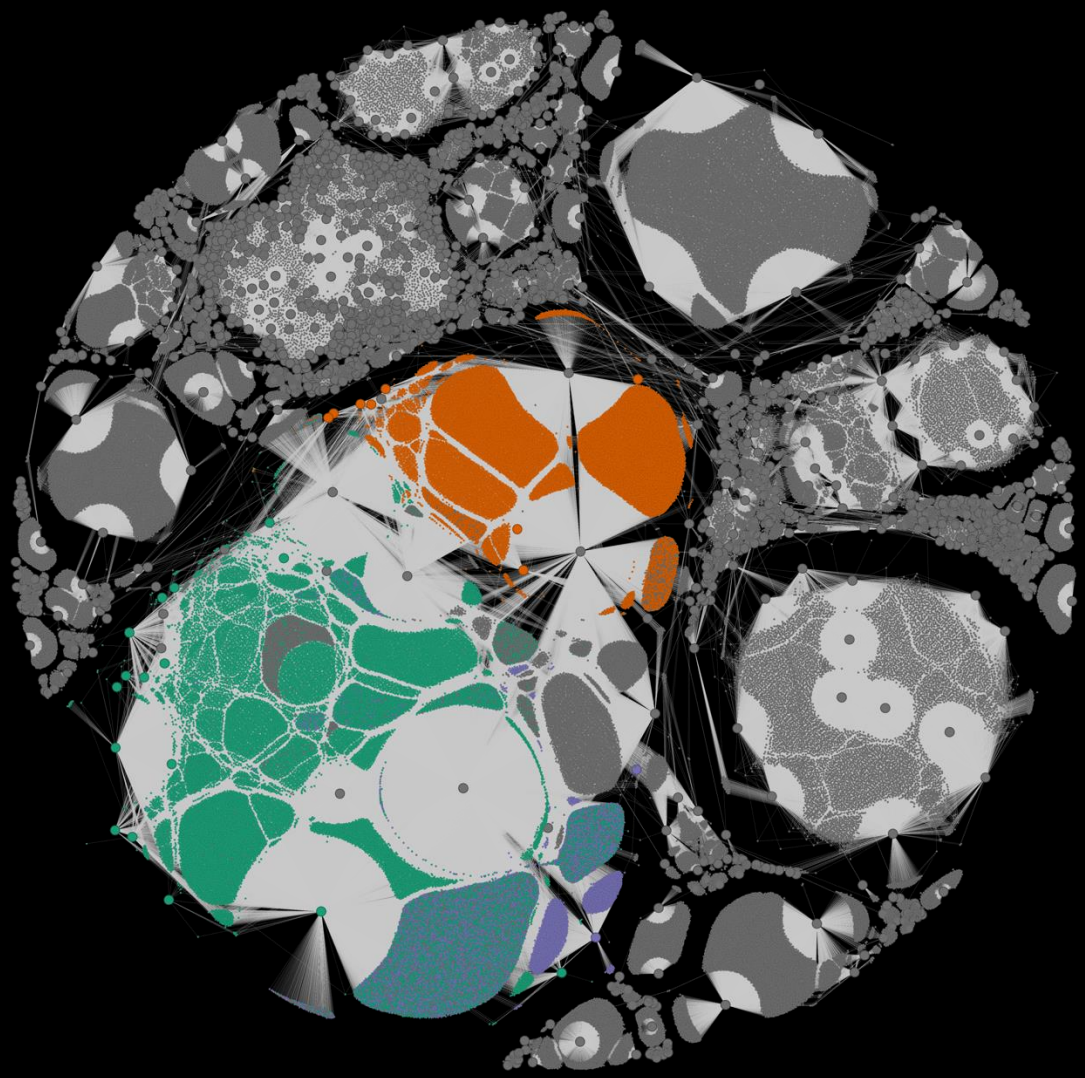
69,880 genomes
42 MIBiG BGCs
203,092 relationships

Weighted: % shared proteins with > 70% identity

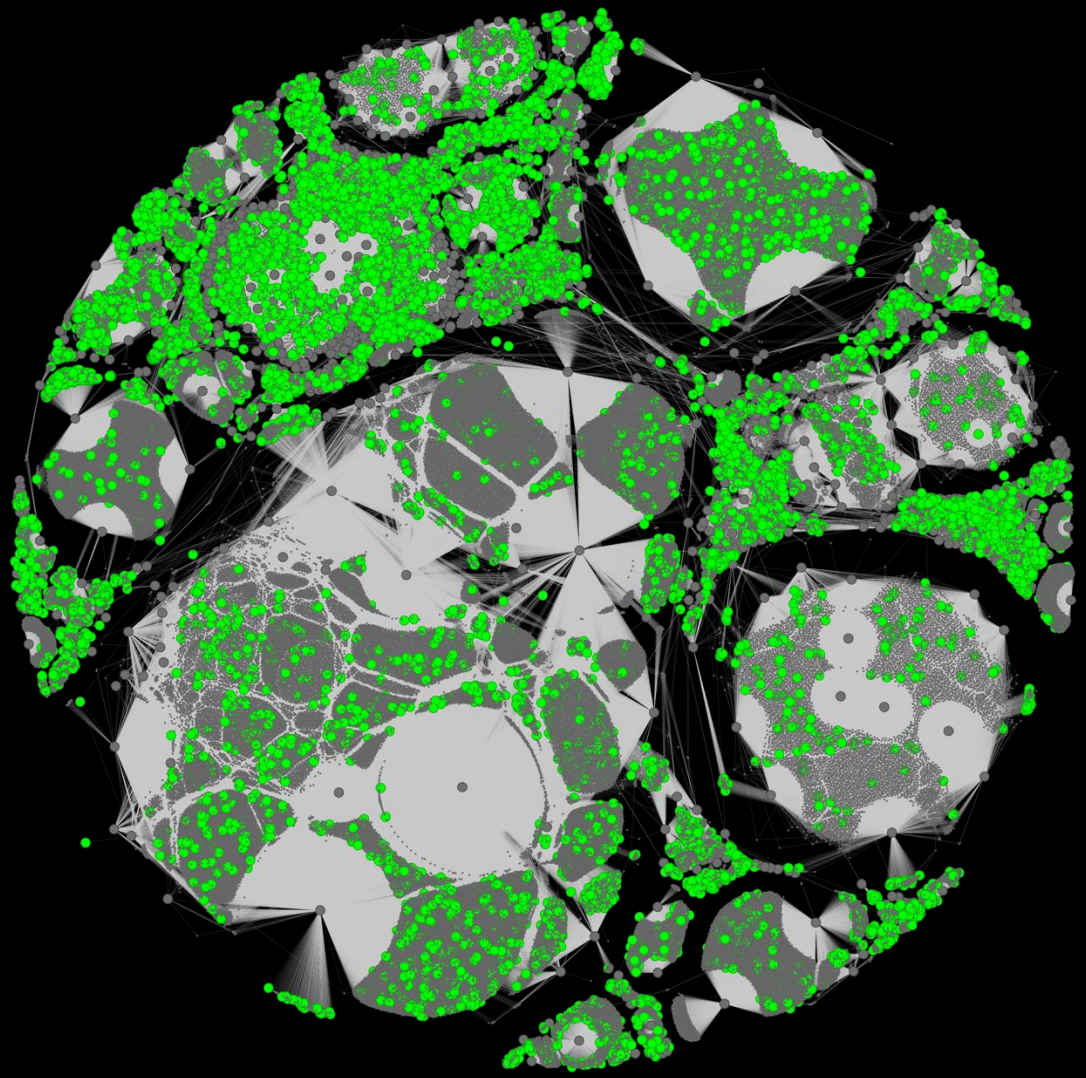
How to test our hypotheses?



8,984 of the 184,571 genomes are in a culture collection



In a culture collection

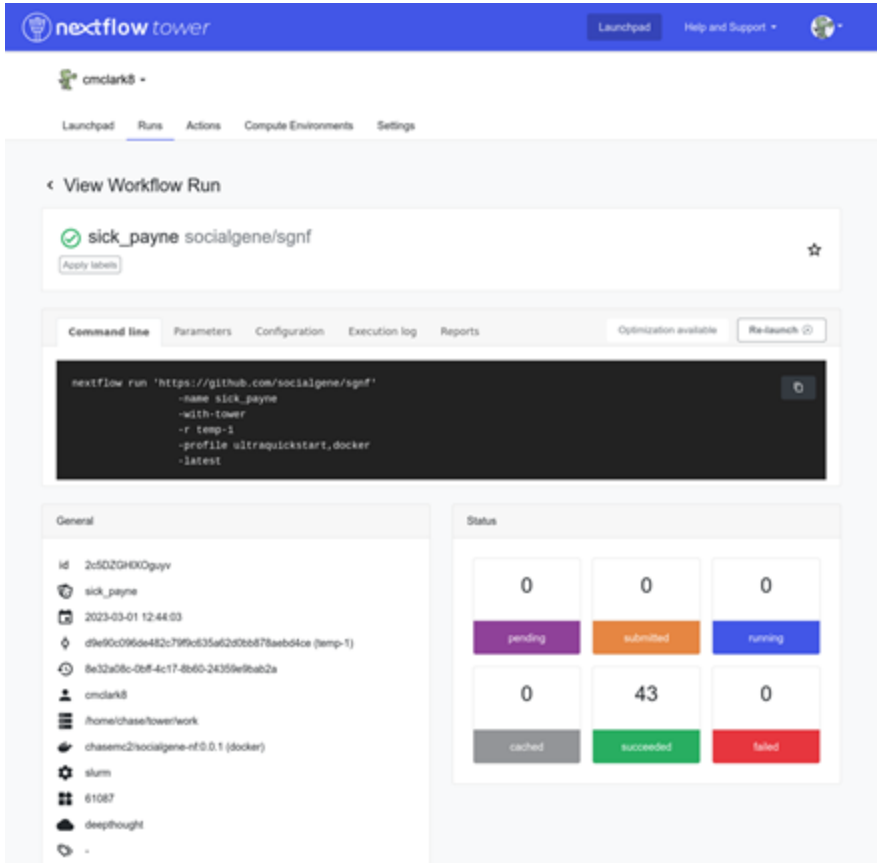


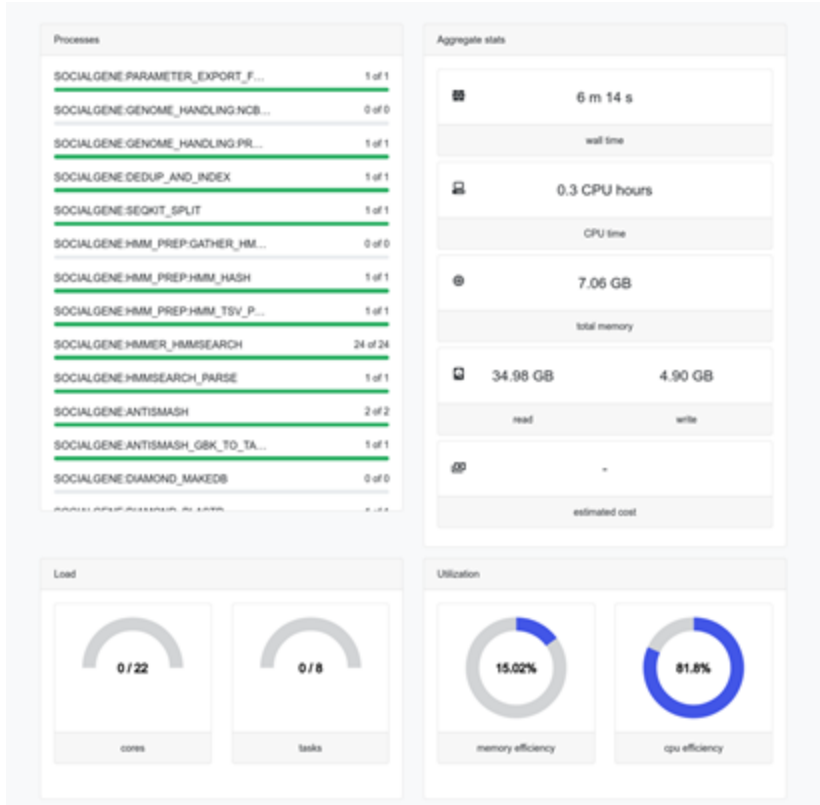
How?

nextflow

docker

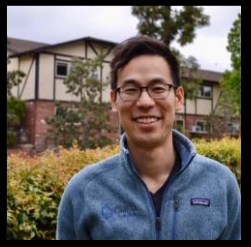
Making the workflow more accessible with nextflow tower





B. Murphy

N. Krull



M. Wang

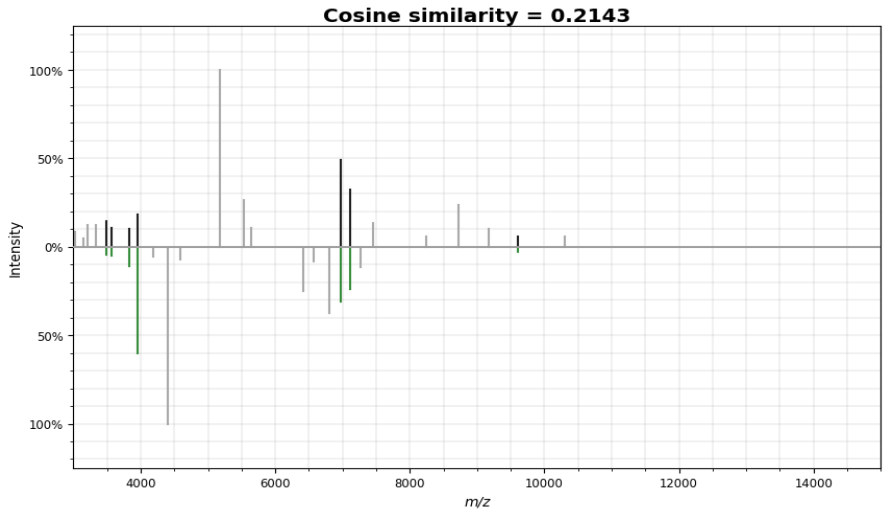
M. Strobel

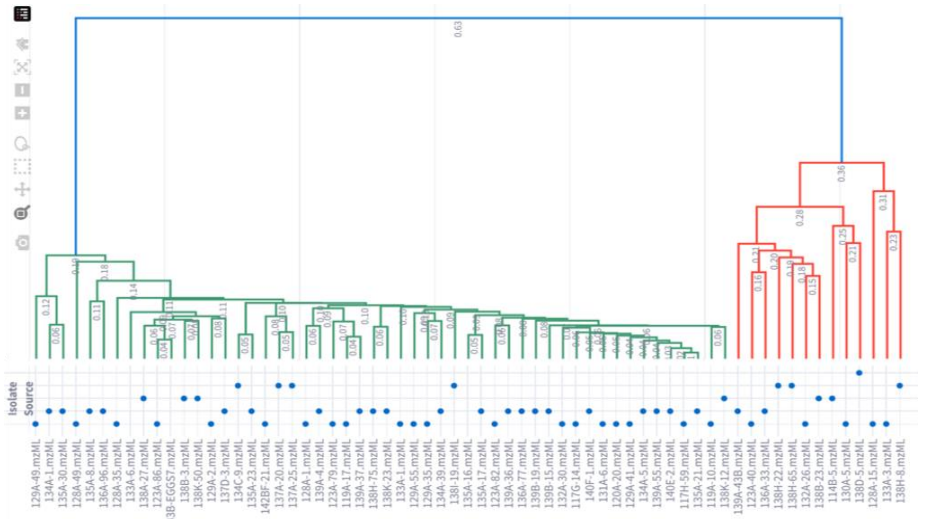
L. Sanchez

R. Shepherd

IDBac

- Available on the web soon!
- *Taxa ID.* Query MS spectrum of ‘unknown’ against a public repository (>4500 seeds)
- Integrate metadata into dendrogram of bacteria

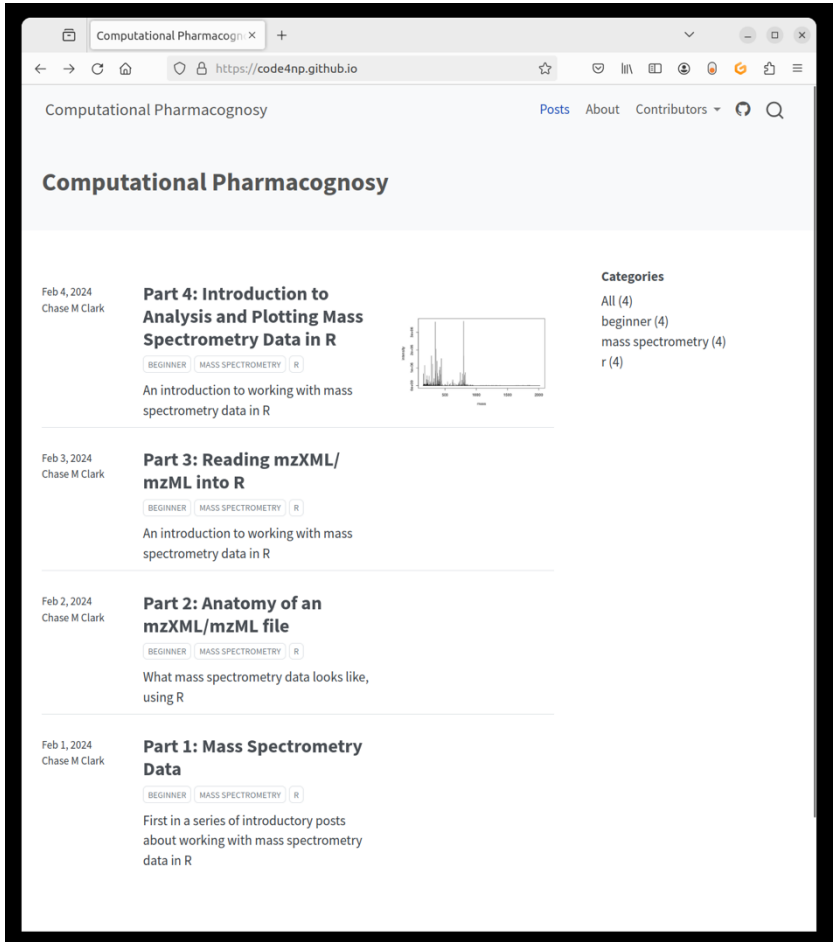


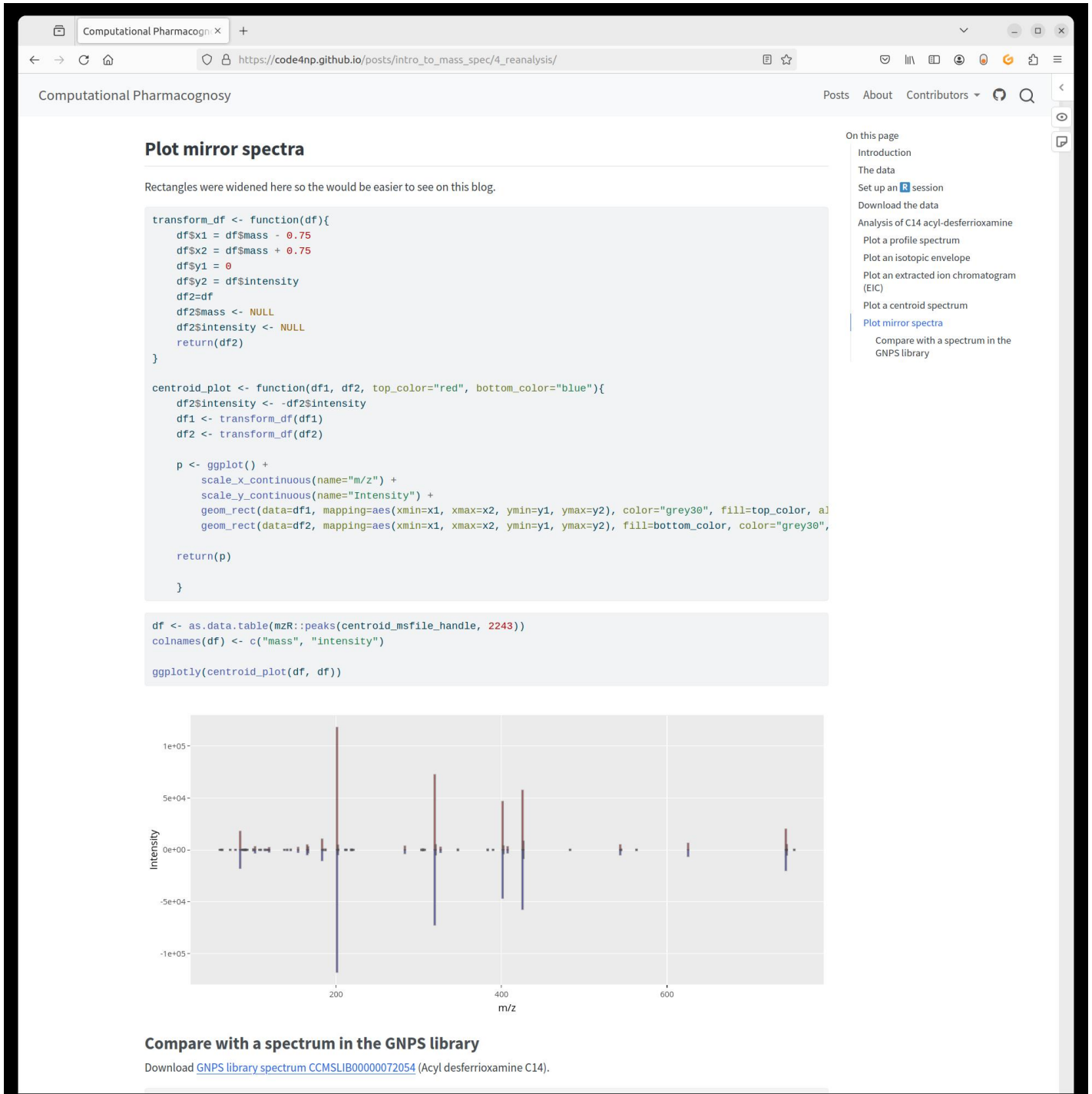


code4np.github.io

Chase Clark

Joseph Egan





socialgene.github.io

CENTER FOR HIGH THROUGHPUT COMPUTING

Open Science Grid

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Funding
NLM: 5T15LM007359
NIGMS: R35GM133776