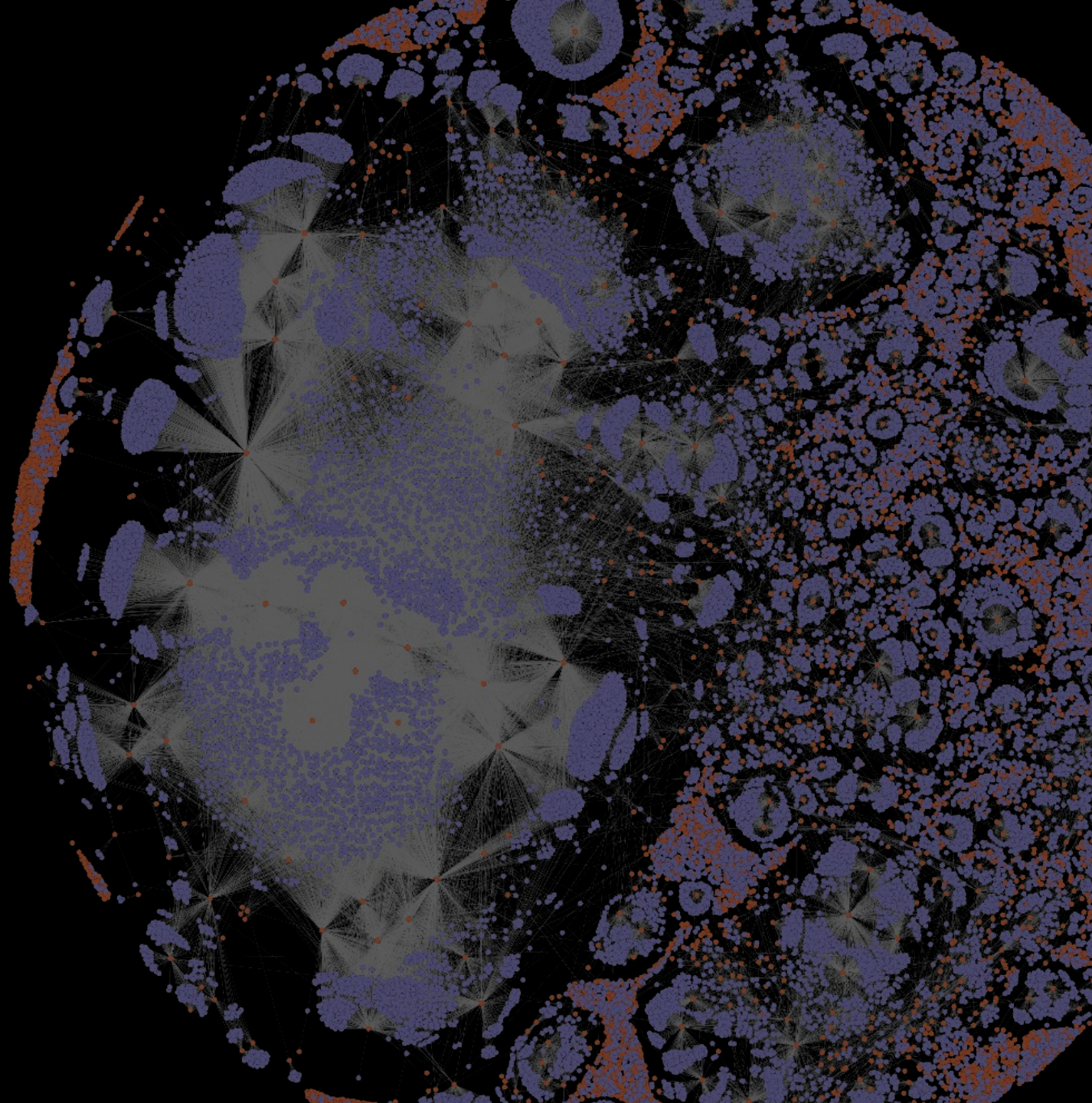
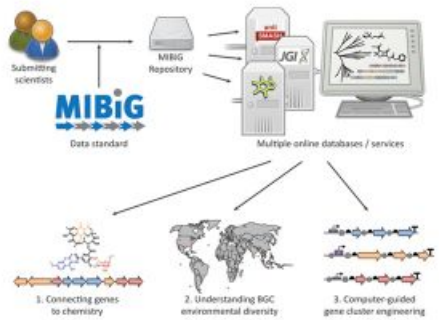
SocialGene: Large Scale Knowledge Graphs for Microbial Based Drug Discovery



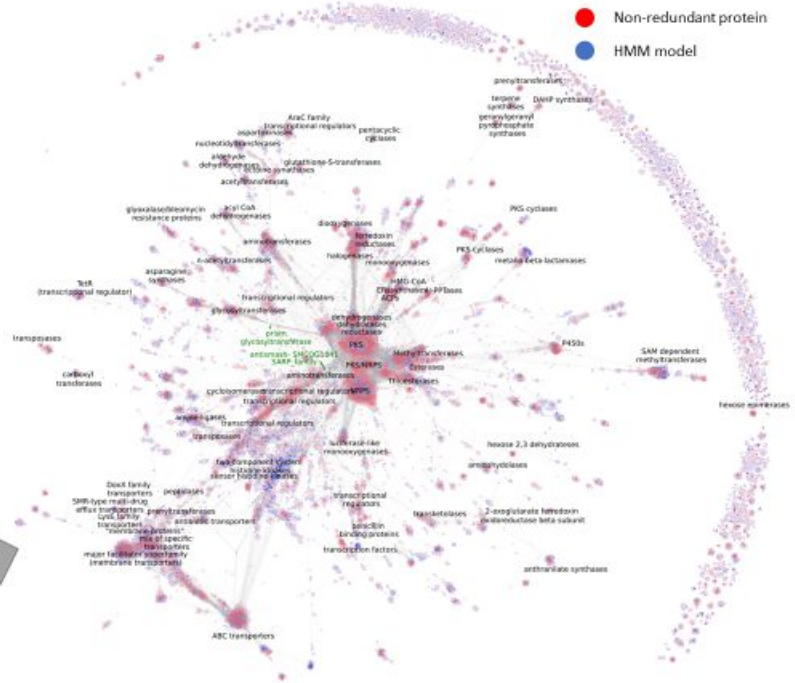
Chase Clark
Postdoctoral Research Associate
Computation and Informatics in
Biology and Medicine Training Program
Jason Kwan Lab
School of Pharmacy
University of Wisconsin-Madison

SocialGene

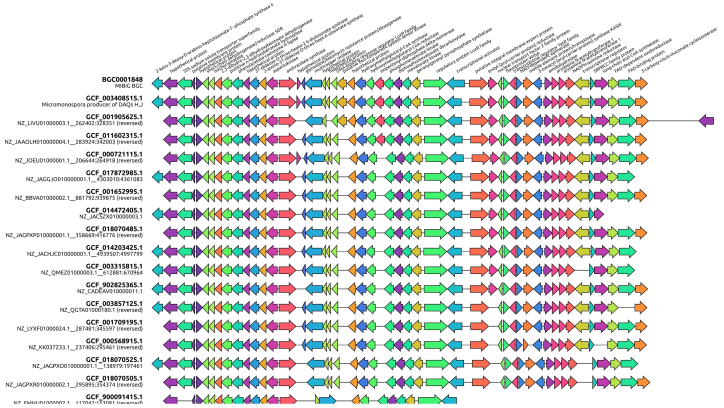
Knowledge graphs for drug discovery
Functional characterization of proteins and searching for similar BGCs



SocialGene

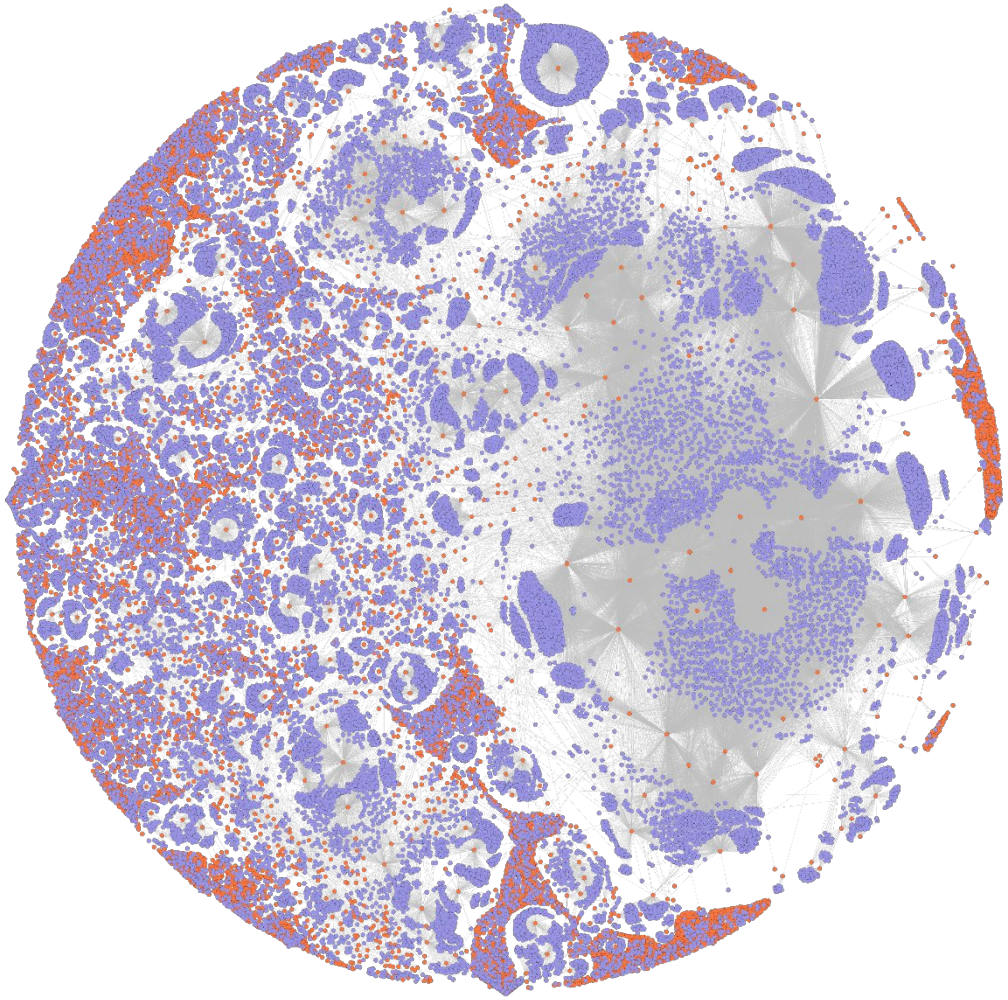


Input BGC → Search 200k genomes



SocialGene

Knowledge graphs for drug discovery
Functional characterization of proteins
and searching for similar BGCs



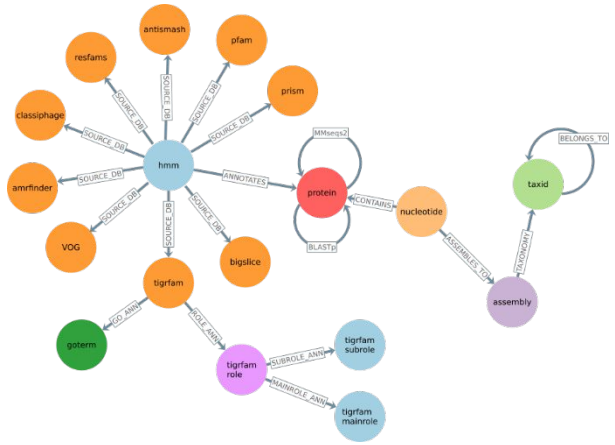
python™

nextflow

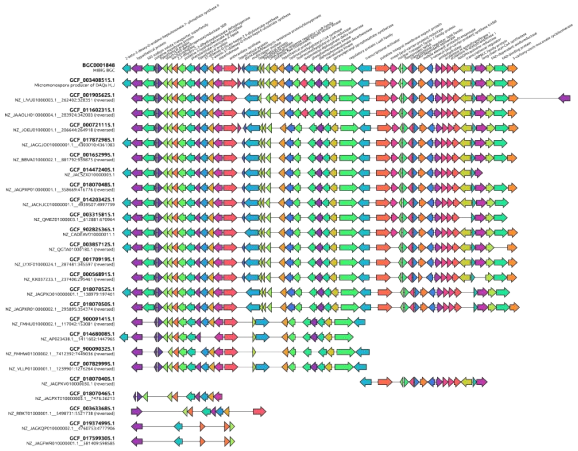
django

neo4j

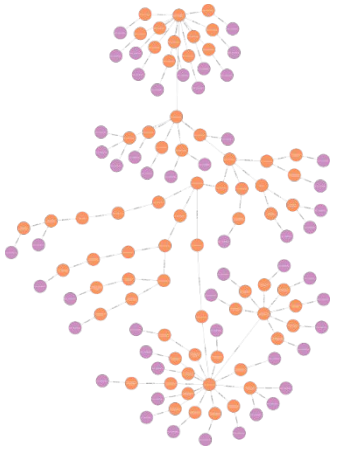
SocialGene



Reproducibly create Graph Databases of up to 100s of thousands of genomes



Search for remotely-homologous proteins and biosynthetic gene clusters



Large-scale phylogenetic and distribution patterns of enzyme occurrence for targeted discovery of enzyme function

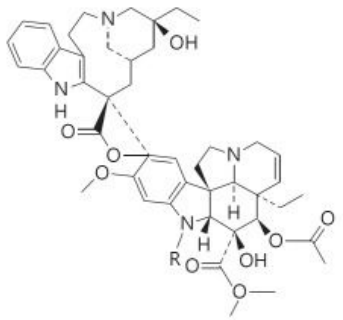
“Drugs”
“Natural Products”
“Specialized Metabolites”

“Drugs”
“Natural Products”
“Specialized Metabolites”

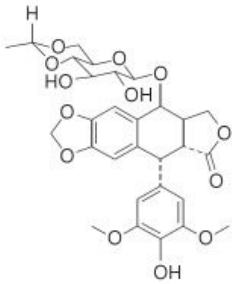
Over 60% of current anticancer drugs are
derived in one way or another from natural sources

Cragg, Gordon M, and John M Pezzuto. “Natural Products as a Vital Source for the Discovery of Cancer Chemotherapeutic and Chemopreventive Agents.” *Medical principles and practice : international journal of the Kuwait University, Health Science Centre* vol. 25 Suppl 2,Suppl 2 (2016): 41-59. doi:10.1159/000443404

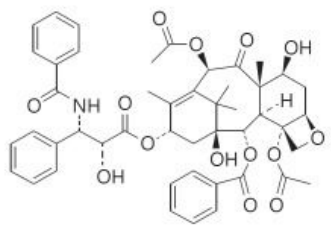
“Natural Products”



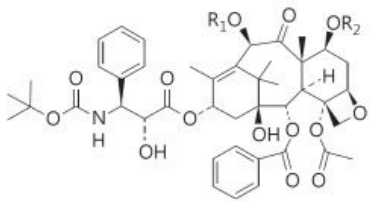
1 VBL R = CH_3
2 VCR R = CHO



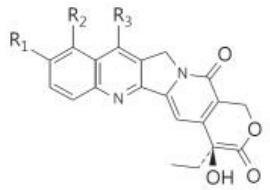
3 Etoposide



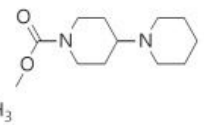
4 Paclitaxel (Taxol™)



5 Docetaxel (Taxotere™) R_1 = R_2 = H
6 Cabazitaxel R_1 = R_2 = CH_3



7 CPT R_1 = R_2 = R_3 = H
8 Topotecan R_1 = OH; R_2 = $CH_2NH(CH_3)_2$; R_3 = H



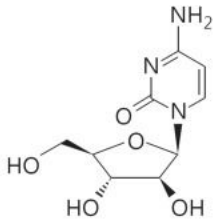
9 Irinotecan R_1 =
R_2 = H
R_3 = CH_2CH_3

10 Belotecan R_1 = R_2 = H; R_3 = $(CH_2)_2NHCH(CH_3)_2$
11 Positecan R_1 = R_2 = H; R_3 = $(CH_2)_2Si(CH_3)_3$
12 SN-38 R_1 = OH; R_2 = H; R_3 = CH_2CH_3

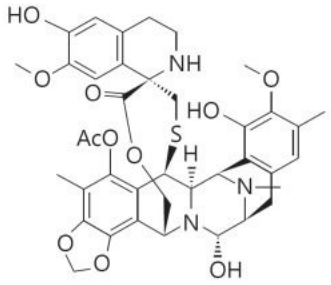
Natural products from plants

- Cragg, Gordon M, and John M Pezzuto. “Natural Products as a Vital Source for the Discovery of Cancer Chemotherapeutic and Chemopreventive Agents.” *Medical principles and practice : international journal of the Kuwait University, Health Science Centre* vol. 25 Suppl 2,Suppl 2 (2016): 41-59. doi:10.1159/000443404
- Cragg, G M et al. “The taxol supply crisis. New NCI policies for handling the large-scale production of novel natural product anticancer and anti-HIV agents.” *Journal of natural products* vol. 56,10 (1993): 1657-68. doi:10.1021/np50100a001
- Stierle, A et al. “The search for a taxol-producing microorganism among the endophytic fungi of the Pacific yew, Taxus brevifolia.” *Journal of natural products* vol. 58,9 (1995): 1315-24. doi:10.1021/np50123a002

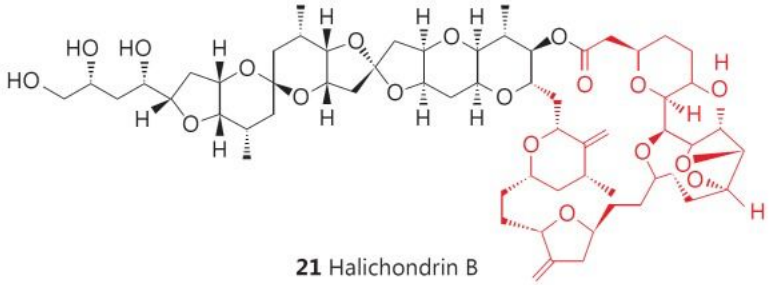
“Natural Products”



19 Cytarabine



20 Trabectedin (ET743; Yondelis®)

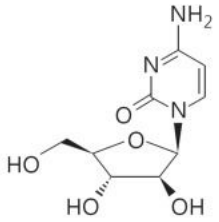


21 Halichondrin B

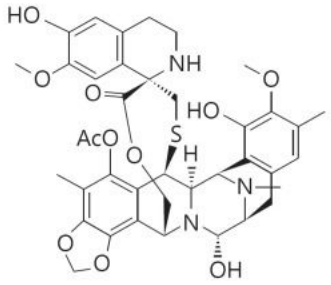
Natural products from marine organisms

Cragg, Gordon M, and John M Pezzuto. “Natural Products as a Vital Source for the Discovery of Cancer Chemotherapeutic and Chemopreventive Agents.” *Medical principles and practice : international journal of the Kuwait University, Health Science Centre* vol. 25 Suppl 2,Suppl 2 (2016): 41-59. doi:10.1159/000443404

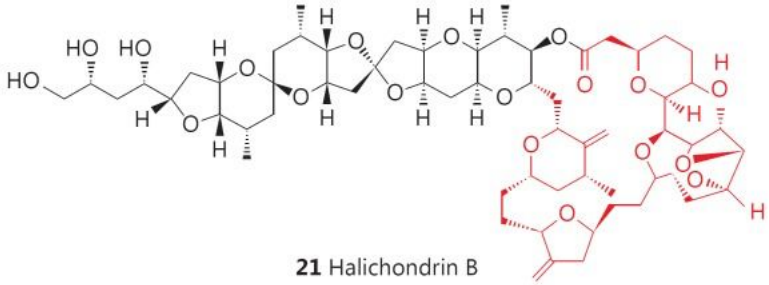
“Natural Products”



19 Cytarabine



20 Trabectedin (ET743; Yondelis®)



21 Halichondrin B

Supply issue can be a huge problem

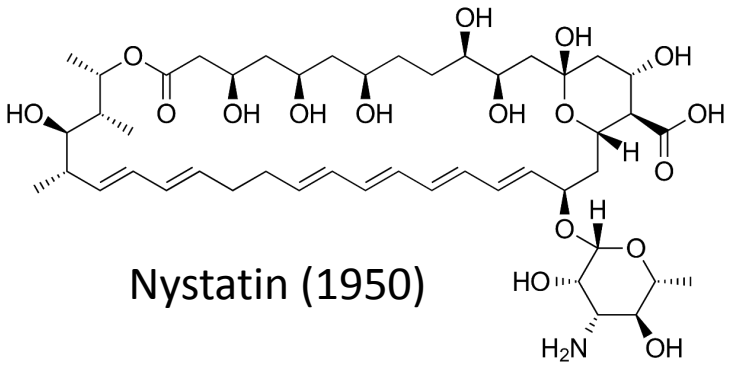
Cragg, Gordon M, and John M Pezzuto. “Natural Products as a Vital Source for the Discovery of Cancer Chemotherapeutic and Chemopreventive Agents.” *Medical principles and practice : international journal of the Kuwait University, Health Science Centre* vol. 25 Suppl 2,Suppl 2 (2016): 41-59. doi:10.1159/000443404

“Natural Products”

Dr. Elizabeth Lee Hazen (left), microbiologist
Dr. Rachel Fuller Brown (right), chemist

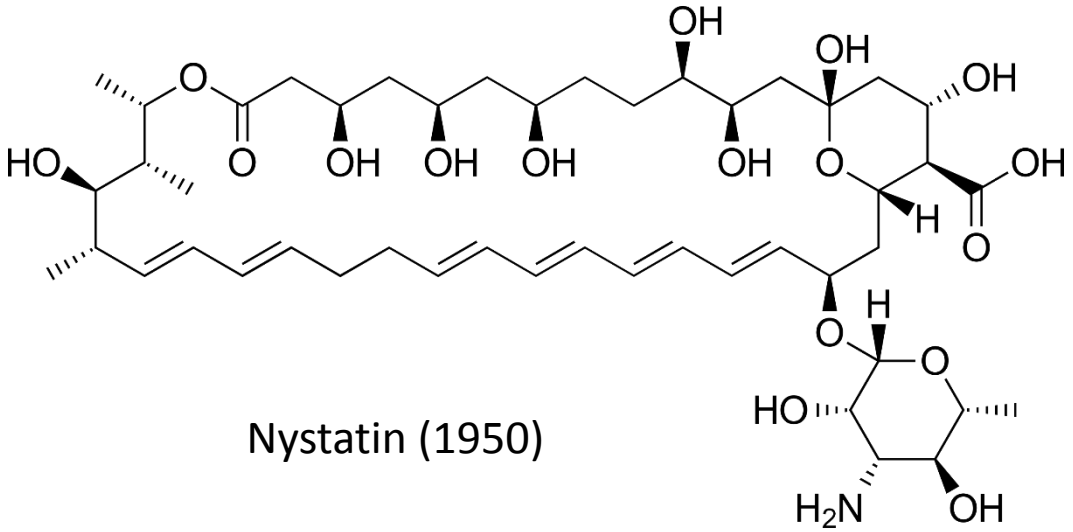
Natural products from bacteria and fungi

Streptomyces noursei



Nystatin (1950)

Dr. Elizabeth Lee Hazen (left), microbiologist
Dr. Rachel Fuller Brown (right), chemist



Nystatin (1950)

Streptomyces noursei

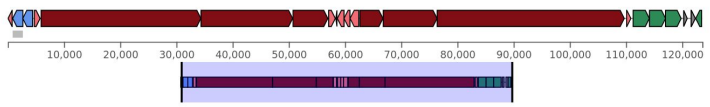
MIBiG Repository of Known Biosynthetic Gene Clusters

Download Contact

BGC0000115: nystatin A1 biosynthetic gene cluster from *Streptomyces noursei ATCC 11455*

Location: 1 - 123,580 nt. (total: 123,580 nt).
This entry is originally from NCBI GenBank AF263912.1.

Download region SVG | Download Cluster GenBank file | View antiSMASH-generated output



Legend:

core biosynthetic genes | additional biosynthetic genes | transport-related genes | regulatory genes | other genes | resistance | TTA codons

reset view | zoom to selection

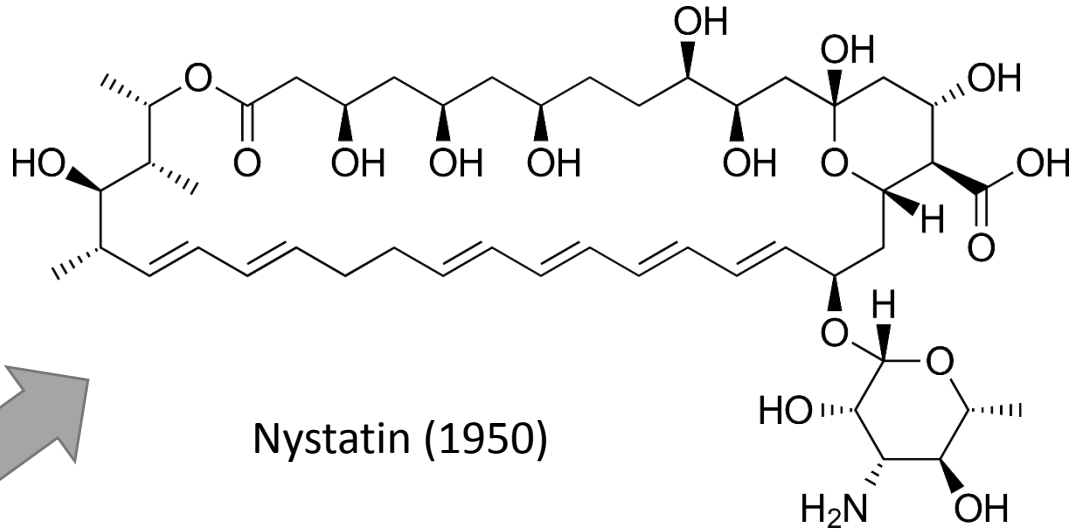
Gene details

Select a gene to view the details available for it

General | Compounds | Genes | Polyketide | Saccharide | History | NRPS/PKS domains | KnownClusterBlast

General information about the BGC

MIBiG accession	BGC0000115
Short description	nystatin A1 biosynthetic gene cluster from *Streptomyces noursei ATCC 11455*
Status	Minimal annotation: no Completeness: complete
Biosynthetic class(es)	Polyketide (Polyene) Saccharide (hybrid/tailoring)
Loci	NCBI GenBank: AF263912.1
Compounds	nystatin A1
Species	*Streptomyces noursei ATCC 11455* [taxonomy]
References	Biosynthesis of the polyene antifungal antibiotic nystatin in Streptomyces noursei ATCC 11455: analysis of the gene cluster and deduction of the biosynthetic pathway. Brautaset T et al., Chem Biol (2000) PMID:10873841



Nystatin (1950)

Streptomyces noursei

MISSING TRANSITION

We’ve barely scratched the surface…

› PLoS Biol. 2021 Nov 9;19(11):e3001421. doi: 10.1371/journal.pbio.3001421.
eCollection 2021 Nov.

Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences

Grace A Blackwell [1 2], Martin Hunt [1 3], Kerri M Malone [1], Leandro Lima [1], Gal Horesh [2], Blaise T F Alako [1], Nicholas R Thomson [2 4], Zamin Iqbal [1]

Affiliations + expand

PMID: 34752446 PMCID: PMC8577725 DOI: 10.1371/journal.pbio.3001421

“639,981 high-quality genomes emphasised the uneven species composition of the ENA/public databases, with just 20 of the total 2,336 species making up 90% of the genomes.”

We've barely scratched the surface…

› PLoS Biol. 2021 Nov 9;19(11):e3001421. doi: 10.1371/journal.pbio.3001421. eCollection 2021 Nov.

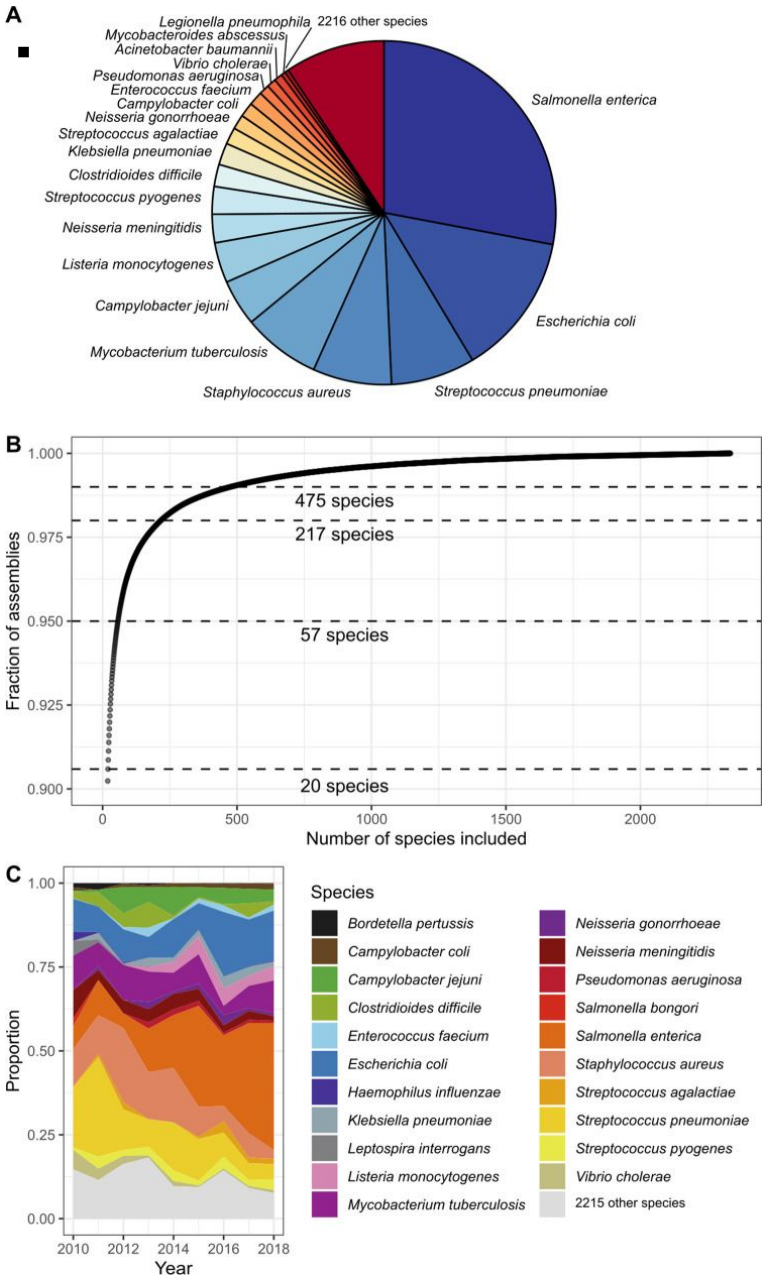
Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences

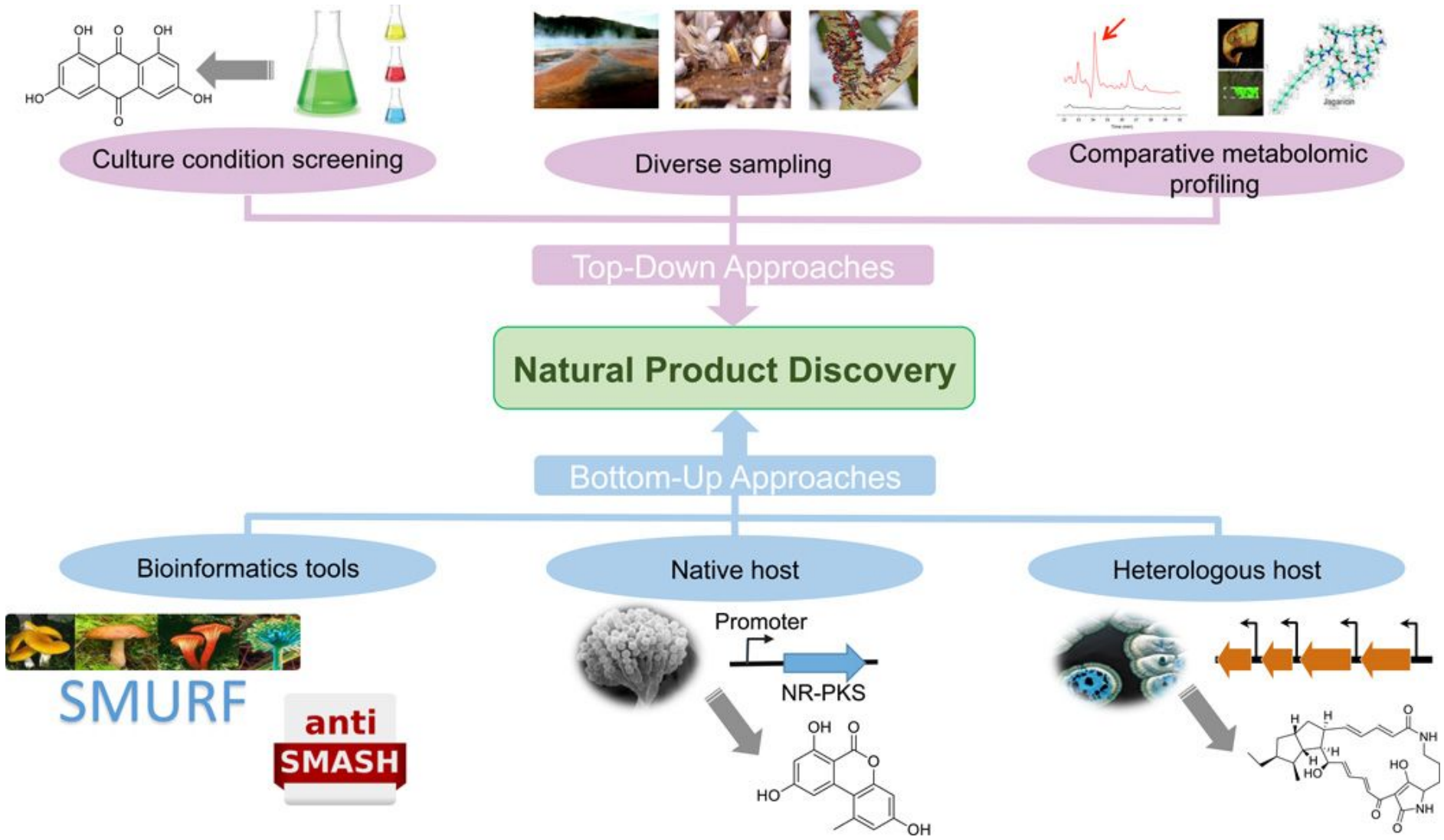
Grace A Blackwell [1 2], Martin Hunt [1 3], Kerri M Malone [1], Leandro Lima [1], Gal Horesh [2], Blaise T F Alako [1], Nicholas R Thomson [2 4], Zamin Iqbal [1]

Affiliations + expand

PMID: 34752446 PMCID: PMC8577725 DOI: 10.1371/journal.pbio.3001421

"639,981 high-quality genomes emphasised the uneven species composition of the ENA/public databases, with just 20 of the total 2,336 species making up 90% of the genomes."





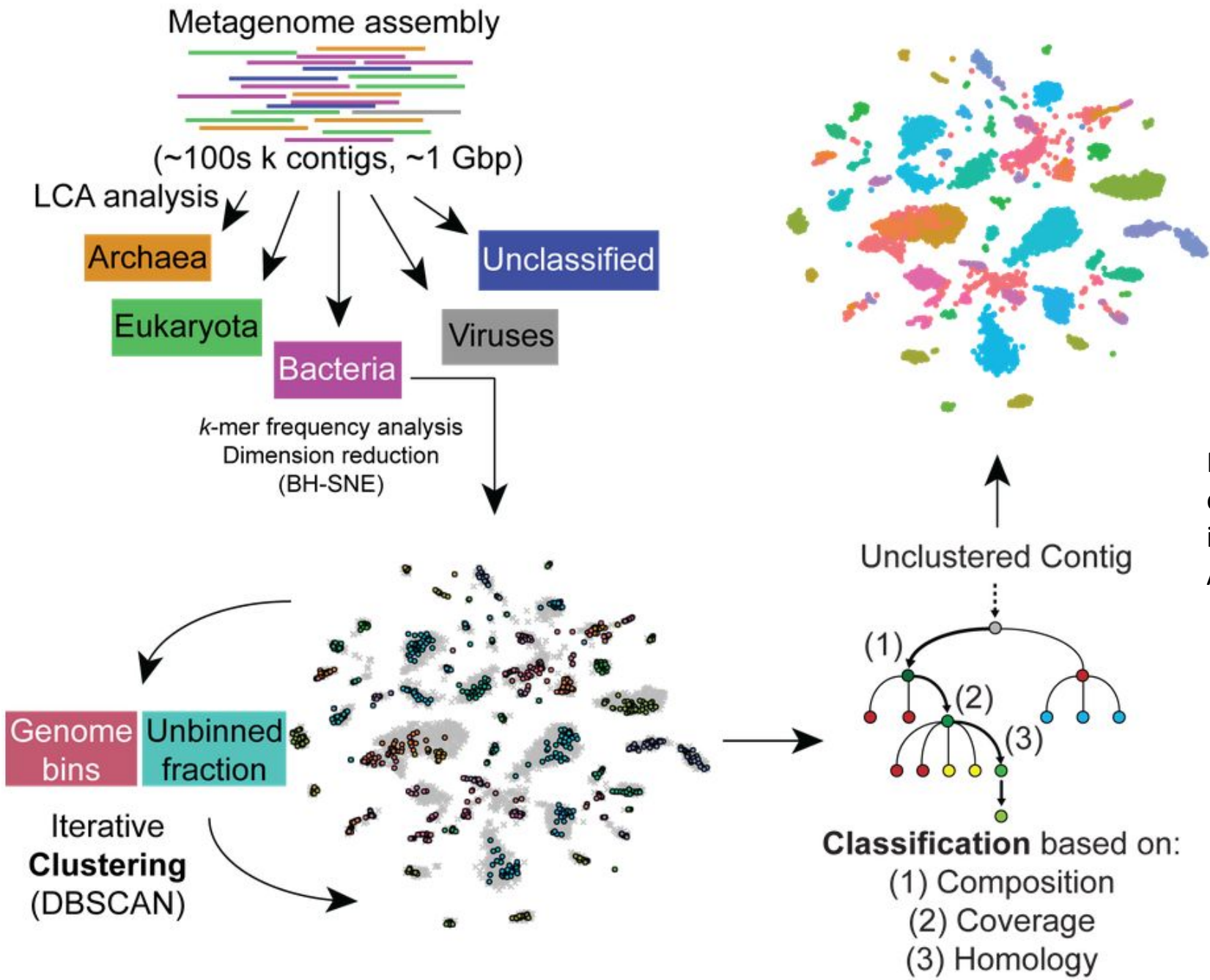
Luo, Yunzi et al. “Recent advances in natural product discovery.” *Current opinion in biotechnology* vol. 30 (2014): 230-7. doi:10.1016/j.copbio.2014.09.002

- Bacteria and fungi produce chemical compounds that can by repurposed as drugs
- We’ve barely genome sequenced bacterial/fungal diversity

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

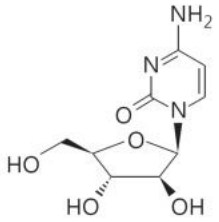
Postdoctoral Research

Bioinformatics, Metagenomics

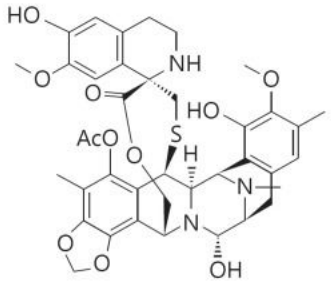


Miller IJ, et al. Autometa: automated extraction of microbial genomes from individual shotgun metagenomes. Nucleic Acids Res. 2019 Jun 4;47(10):e57.

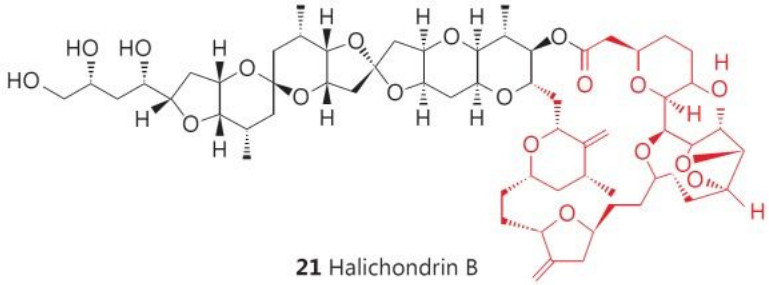
“Natural Products”



19 Cytarabine



20 Trabectedin (ET743; Yondelis®)



21 Halichondrin B

Supply issue can be a huge problem

Cragg, Gordon M, and John M Pezzuto. “Natural Products as a Vital Source for the Discovery of Cancer Chemotherapeutic and Chemopreventive Agents.” *Medical principles and practice : international journal of the Kuwait University, Health Science Centre* vol. 25 Suppl 2,Suppl 2 (2016): 41-59. doi:10.1159/000443404

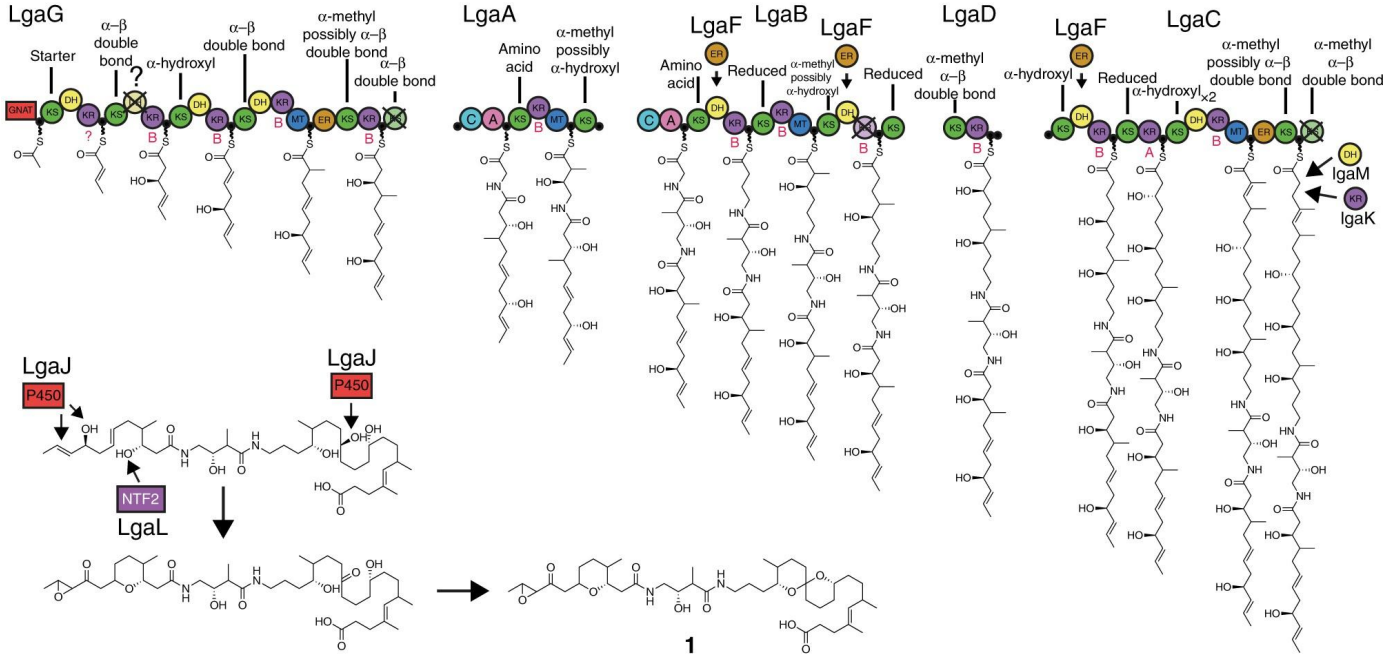
Metagenomics revealed the biosynthetic gene cluster for lagriamide

PHOTO BY LAURA FLÓREZ

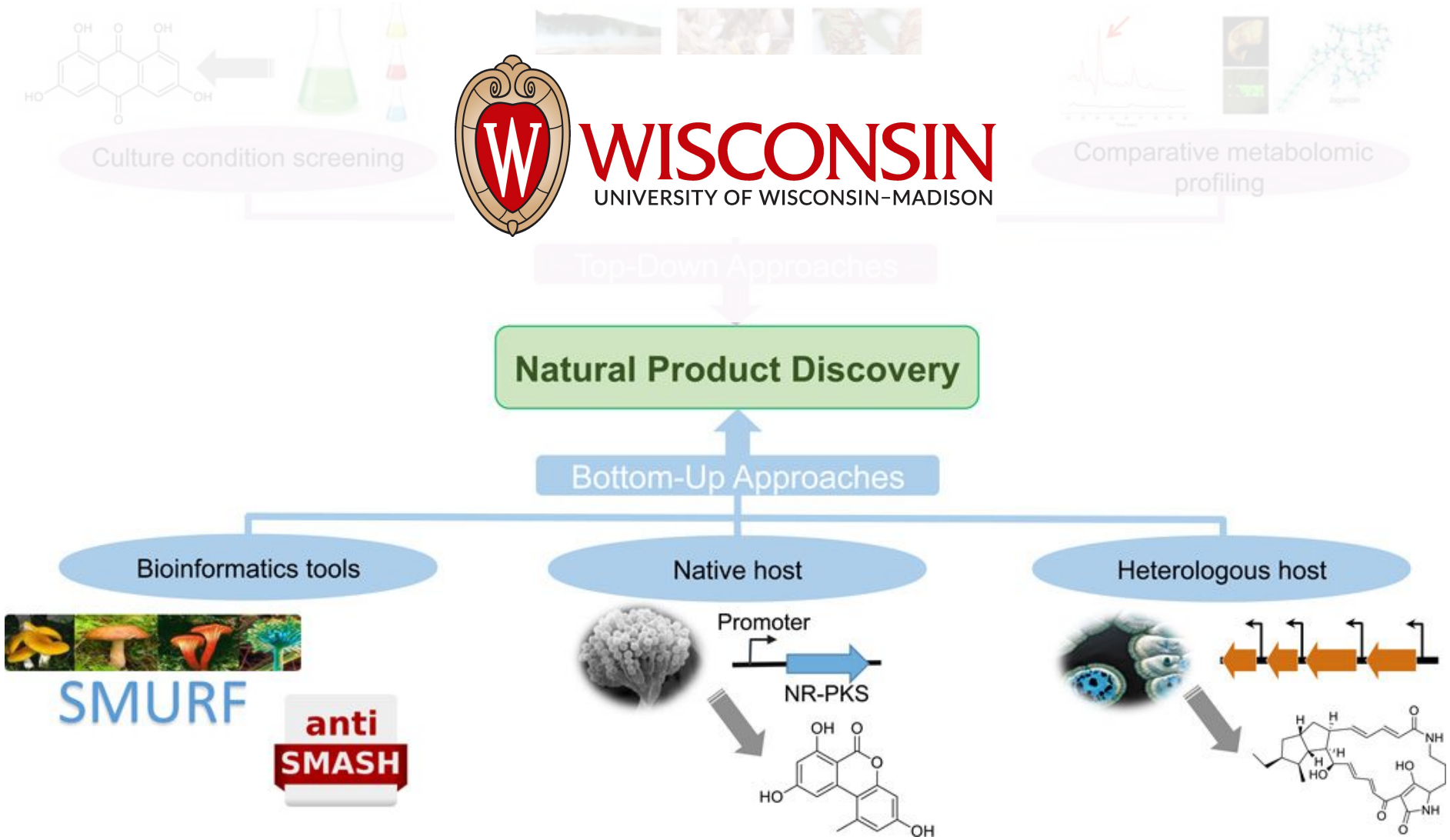
Metagenomics

Uncultured symbiont

Burkholderia gladioli Lv-StB



Waterworth SC, Flórez LV, Rees ER, Hertweck C, Kaltenpoth M, Kwan JC. Horizontal Gene Transfer to a Defensive Symbiont with a Reduced Genome in a Multipartite Beetle Microbiome. mBio. 2020 Feb 25;11(1):e02430-19.



Luo, Yunzi et al. “Recent advances in natural product discovery.” *Current opinion in biotechnology* vol. 30 (2014): 230-7. doi:10.1016/j.copbio.2014.09.002

What do we need?

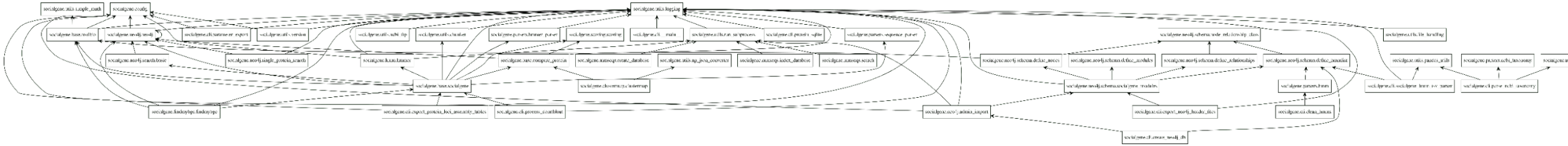
- Collect/process/standardize genomes and the proteins they encode
  - Protein identifiers aren’t constant across databases
    - ID conversion files are huge, online API rate-limited, doesn’t help with in-house proteins
- All-vs-All sequence comparison too slow for real-time search, results too large to store
- Downloading annotated genomes doesn’t because we can’t annotate in-house sequences in the same way

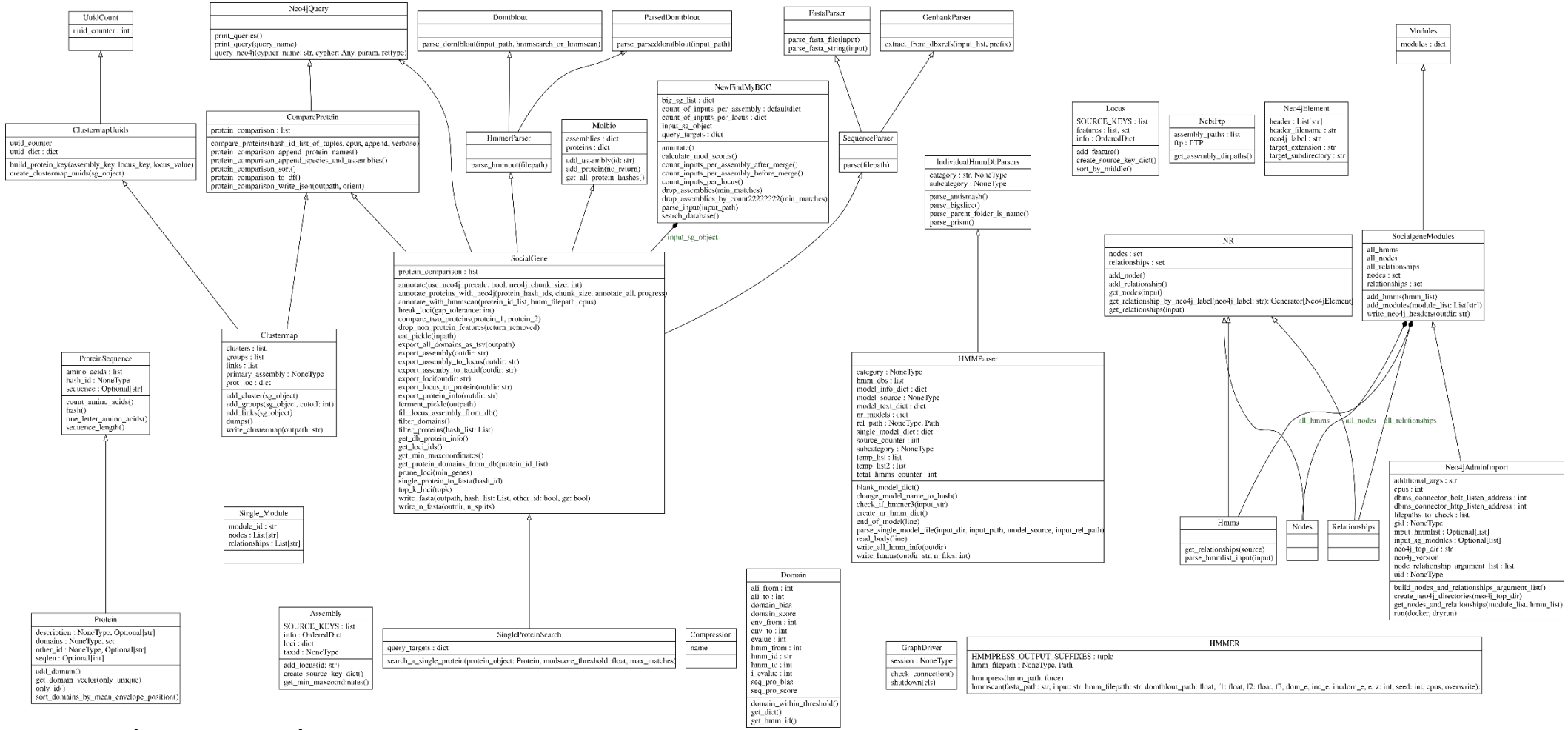
A few challenges, of many

Problems	Solutions
Protein identifiers aren't constant across databases	Hash protein sequences and use the hash as the identifier *(sgpy and crabhash)*
All-vs-All sequence comparison too slow for real-time search, results too large to store	Annotate proteins based on functional domains, use domains as entry-point for search and similarity *(HMMER + many, many models, sgnf)*
Storing/data-retrieval / machine learning	Graph database *(through sgpy, sgnf)*
Doing all steps reproducibly and "easily"	*(sgpy, sgnf)*

sgpy Public

Python





https://github.com/socialgene/sgpy

Input genome(s)
1-????

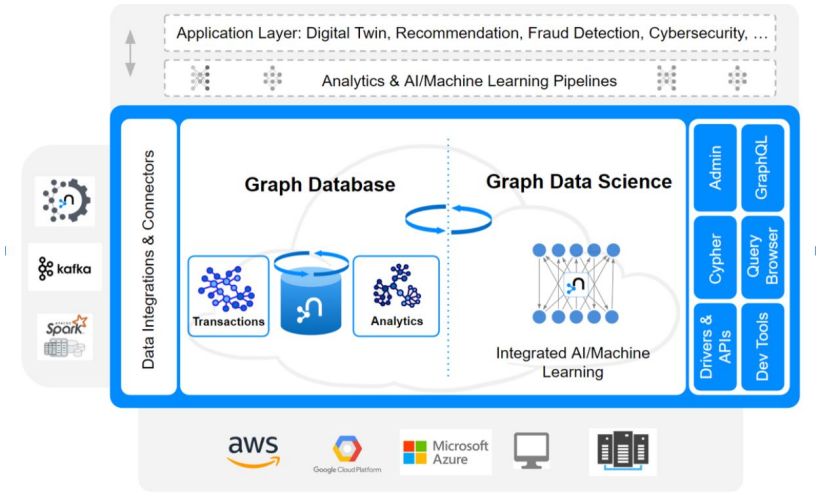
nextflow

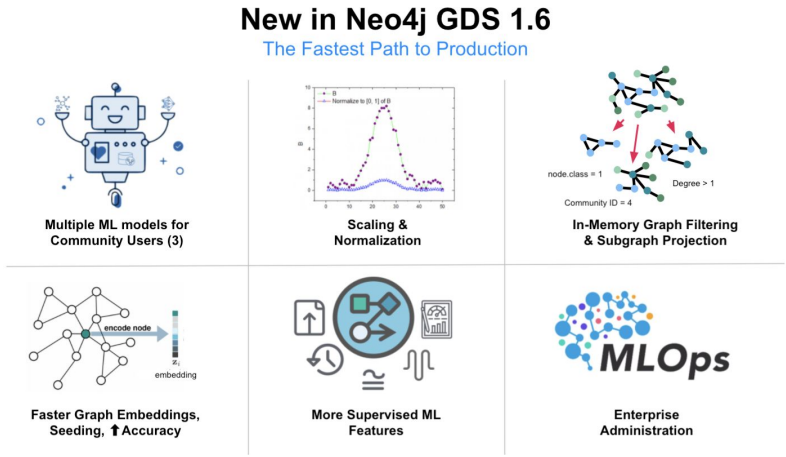
Reproducibly run multiple scripts/software in parallel, across different machines

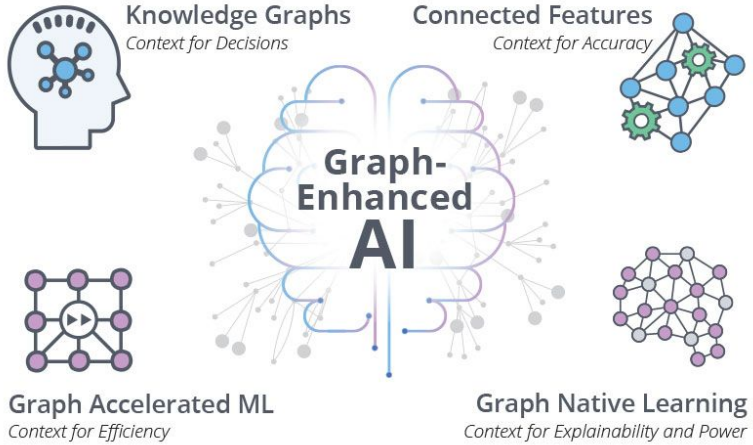
neo4j

Graph database

Creating knowledge graphs with neo4j

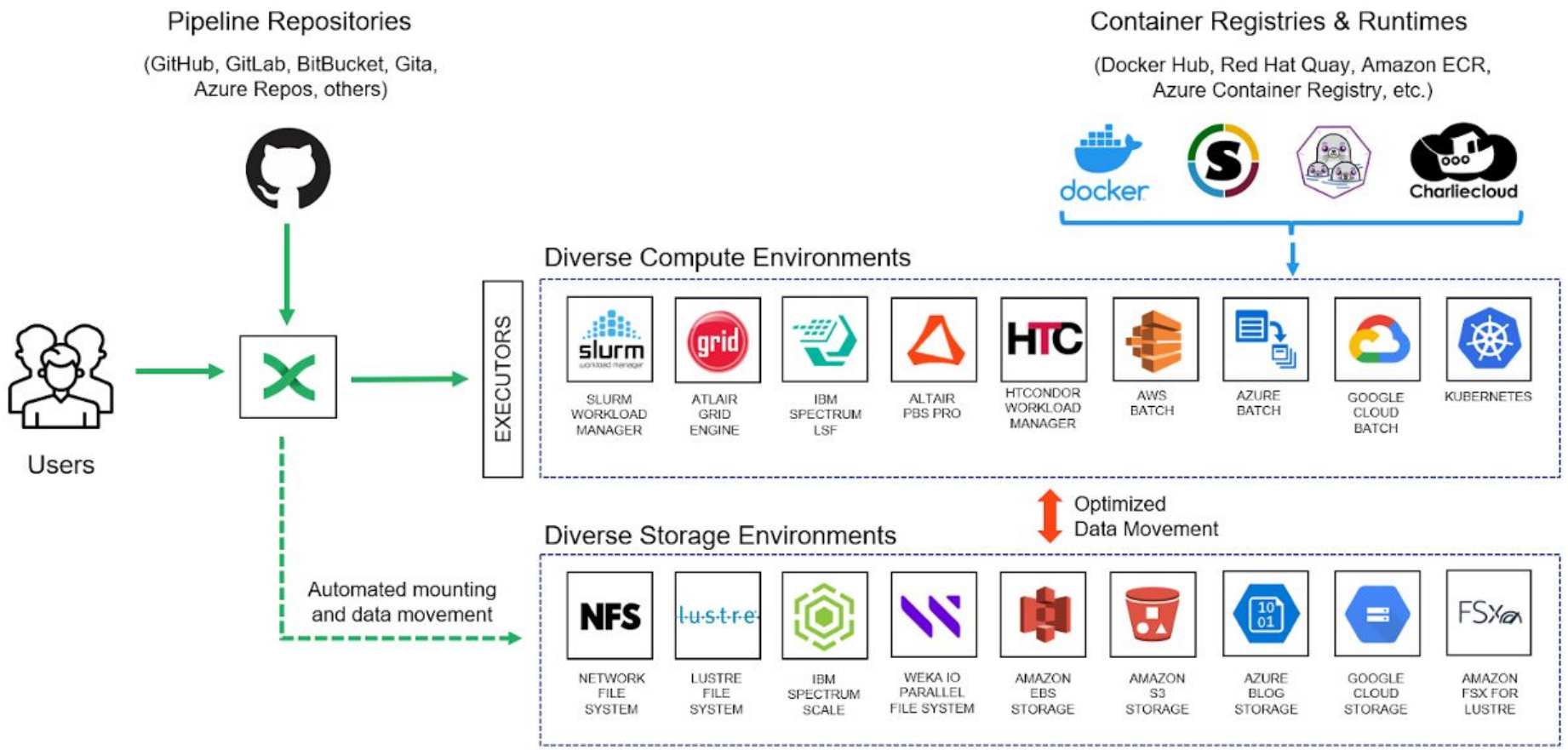






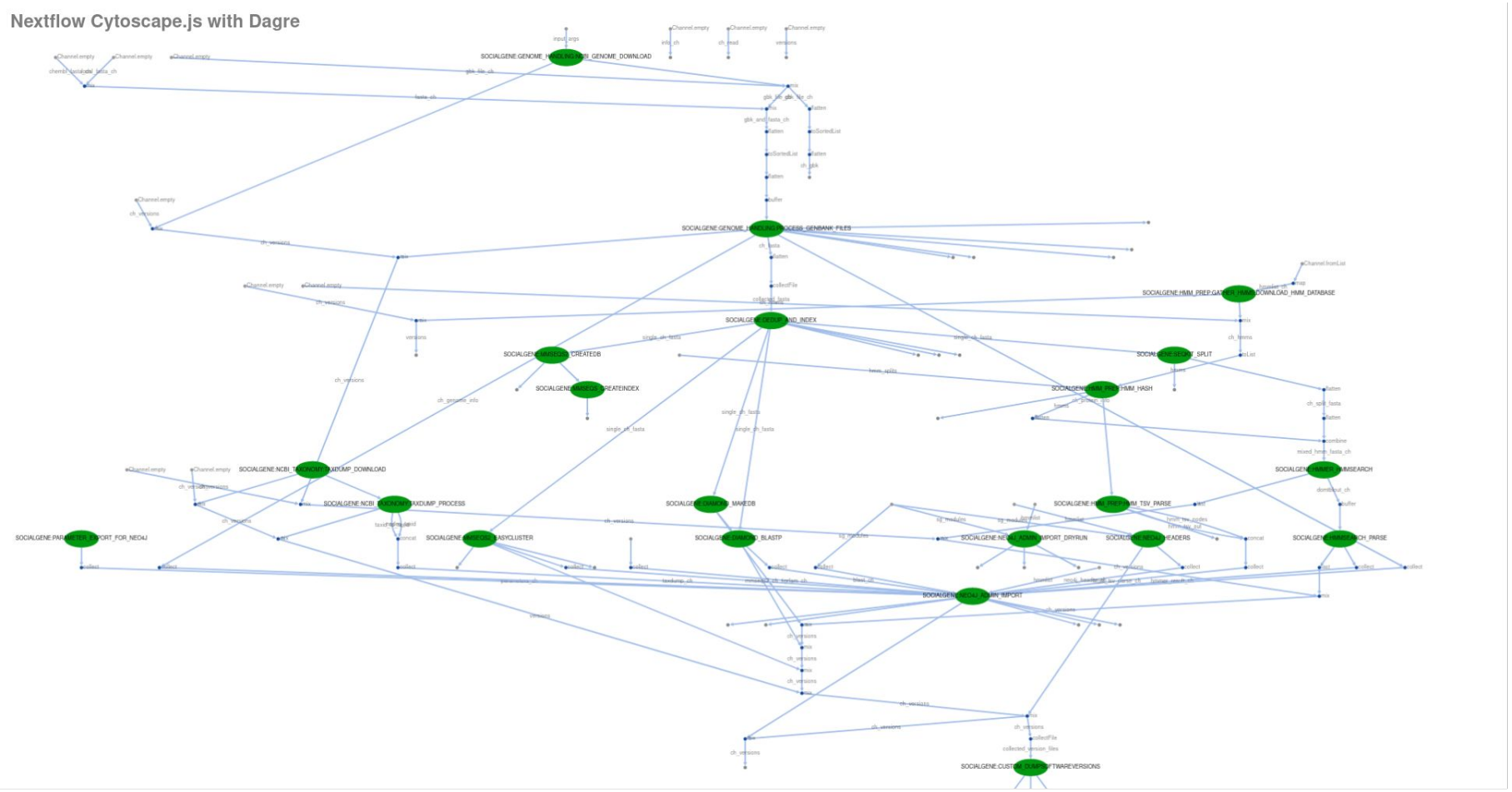
https://workflows.community/stories/2022/09/28/nextflow/

Simplifying complicated bioinformatics workflows with nextflow

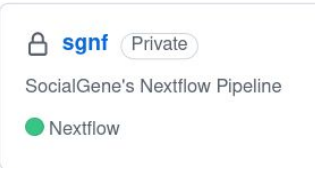


https://workflows.community/stories/2022/09/28/nextflow

Why Nextflow?
It’s complicated



```
(base) chase@titan:~/Documents/github/kwan_lab/socialgene/sgnf$
```



https://github.com/socialgene/sgnf

Sh#$ happens…


Chase Clark, PhD
@ChasingMicrobes
[99%] 5913 of 5914, failed: 1
Bummer
GIF ALT
3:22 PM · Mar 3, 2023 · 55 Views
View Tweet analytics

Sh#$ happens…


Chase Clark, PhD @ChasingMicrobes · 1m
[100%] 5914 of 5914, cached: 5913 ✓
GIF ALT
2

Nextflow workflow report

[trusting_monod] *(resumed run)*

Workflow execution completed successfully!

Run times

28-Feb-2023 12:09:31 - 28-Feb-2023 12:10:53 (duration: **1m 23s**)

37 succeeded | 5 cached

Nextflow command

```
nextflow run . -profile ultraquickstart,docker --outdir /home/chase/Documents/socialgene_data/example --outdir_download_cache /home/chase
/Documents/socialgene_data/cache -resume
```

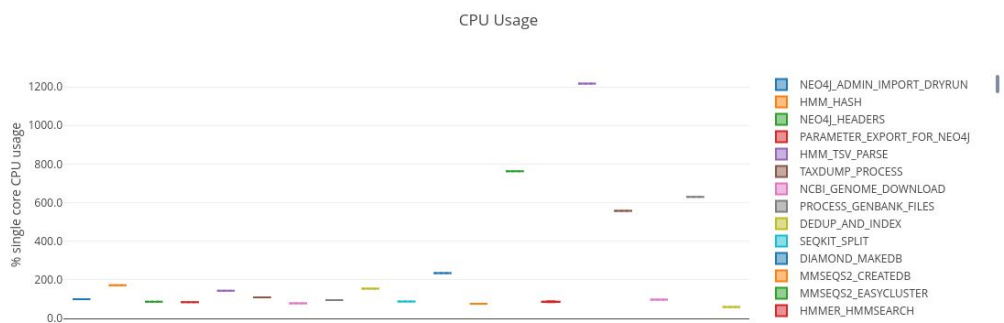
CPU-Hours	0.3 (1.2% cached)
Launch directory	/home/chase/Documents/github/kwan_lab/socialgene/sgnf
Work directory	/home/chase/Documents/github/kwan_lab/socialgene/sgnf/work
Project directory	/home/chase/Documents/github/kwan_lab/socialgene/sgnf
Script name	main.nf
Script ID	a28244b78ec8421acc43ed824425d888
Workflow session	0106669a-7144-4423-82df-a6bd864f98fd
Workflow profile	ultraquickstart,docker
Workflow container	chasemc2/socialgene-nf:0.0.1
Container engine	docker
Nextflow version	version 22.10.6, build 5843 (23-01-2023 23:20 UTC)

Resource Usage

These plots give an overview of the distribution of resource usage for each process.

CPU

Raw Usage | % Allocated



Nextflow workflow report

[trusting_monod] *(resumed run)*

Workflow execution completed successfully!

Run times

28-Feb-2023 12:09:31 - 28-Feb-2023 12:10:53 (duration: **1m 23s**)

37 succeeded | 5 cached

Nextflow command

```
nextflow run . -profile ultraquickstart,docker --outdir /home/chase/Documents/socialgene_data/example --outdir_download_cache /home/chase
/Documents/socialgene_data/cache -resume
```

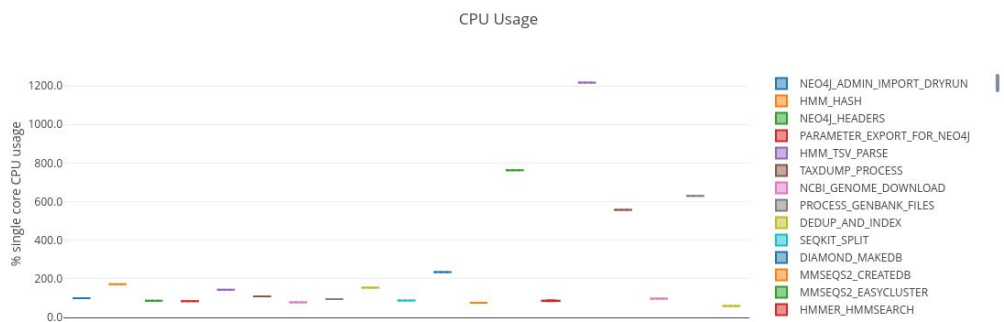
CPU-Hours	0.3 (1.2% cached)
Launch directory	/home/chase/Documents/github/kwan_lab/socialgene/sgnf
Work directory	/home/chase/Documents/github/kwan_lab/socialgene/sgnf/work
Project directory	/home/chase/Documents/github/kwan_lab/socialgene/sgnf
Script name	main.nf
Script ID	a28244b78ec8421acc43ed824425d888
Workflow session	0106669a-7144-4423-82df-a6bd864f98fd
Workflow profile	ultraquickstart,docker
Workflow container	chasemc2/socialgene-nf:0.0.1
Container engine	docker
Nextflow version	version 22.10.6, build 5843 (23-01-2023 23:20 UTC)

Resource Usage

These plots give an overview of the distribution of resource usage for each process.

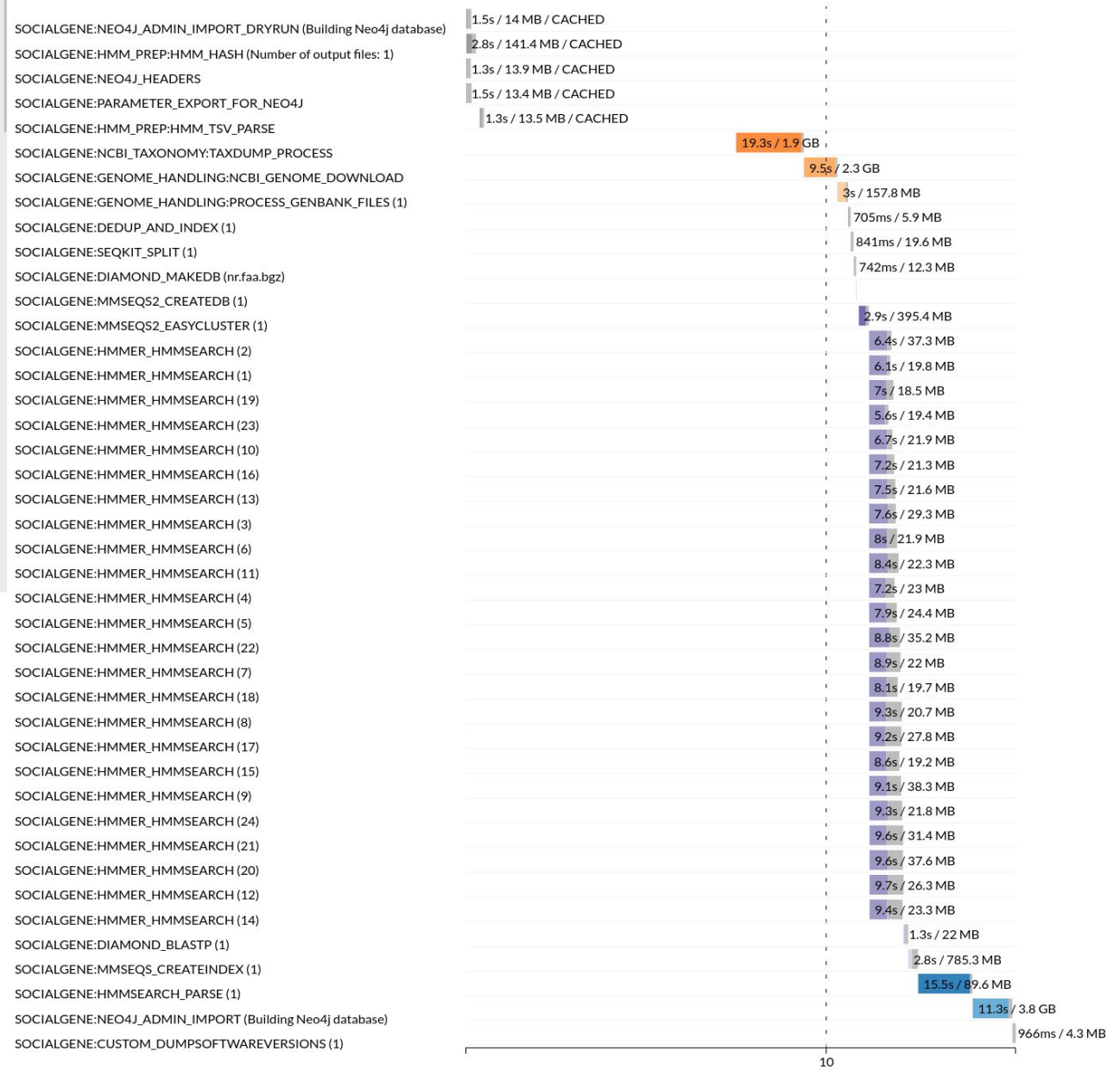
CPU

Raw Usage | % Allocated

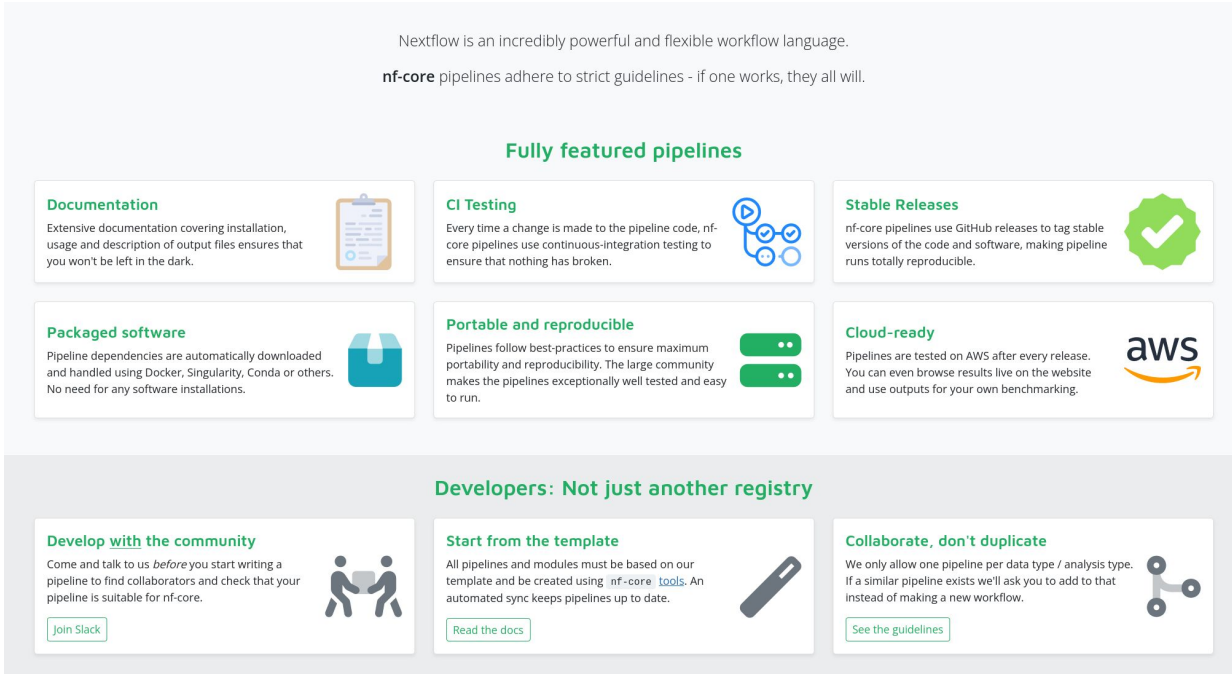


Processes execution timeline

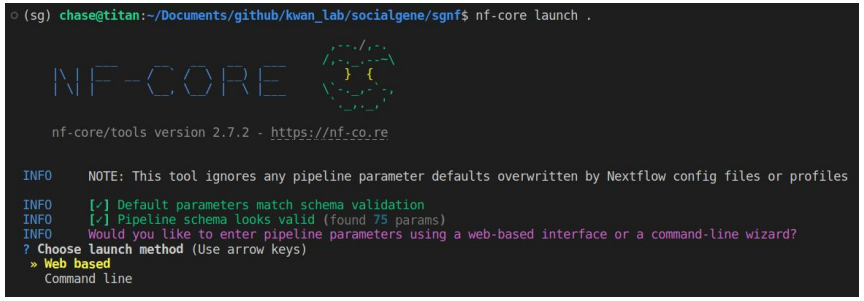
Launch time: 28 Feb 2023 12:08
Elapsed time: 1m 23s
Legend: job wall time / memory usage (RAM)

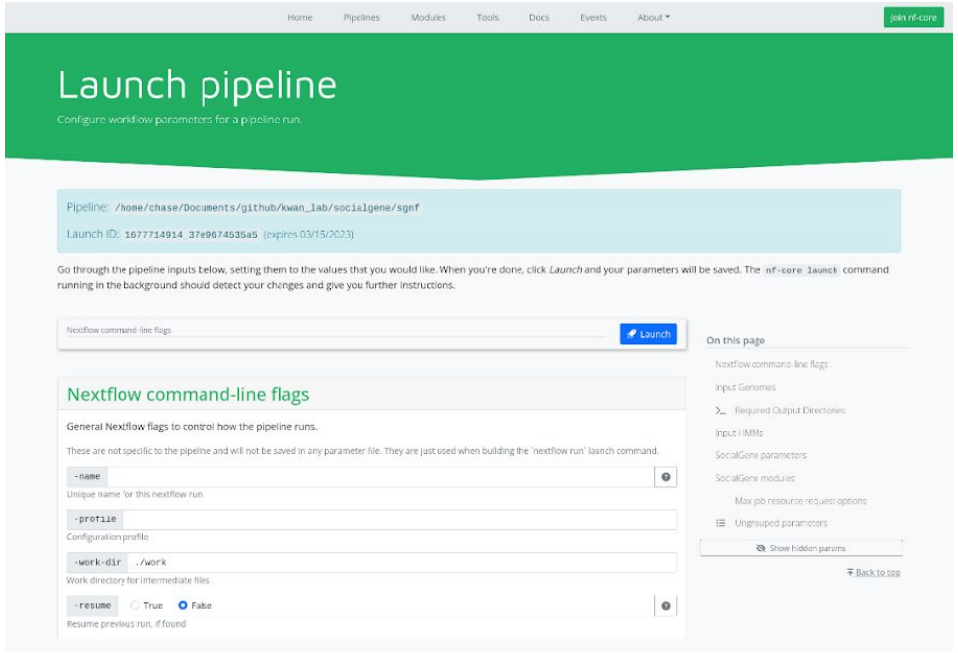


Standardizing the pipeline with nf-core

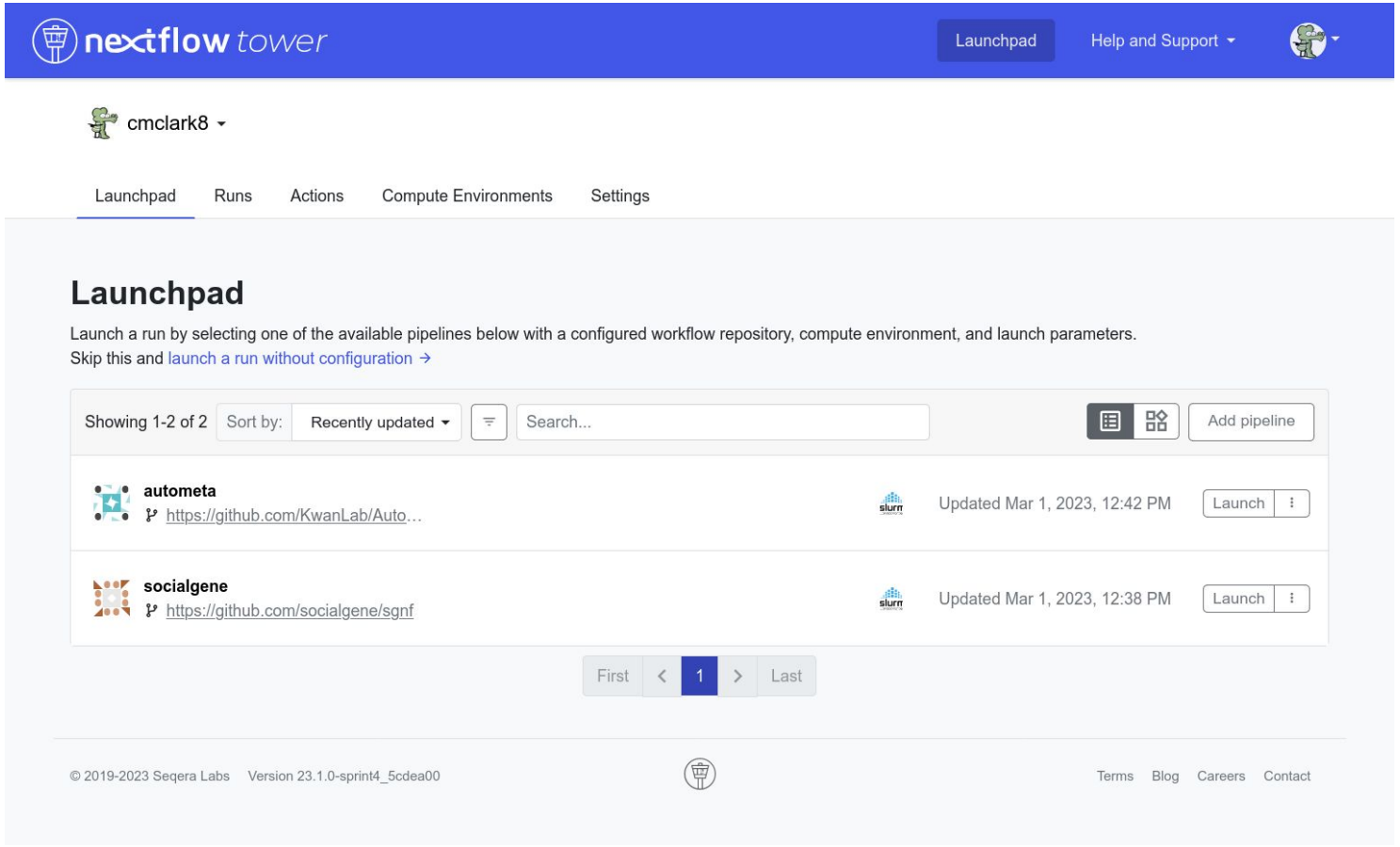


Making the pipeline more accessible with nf-core

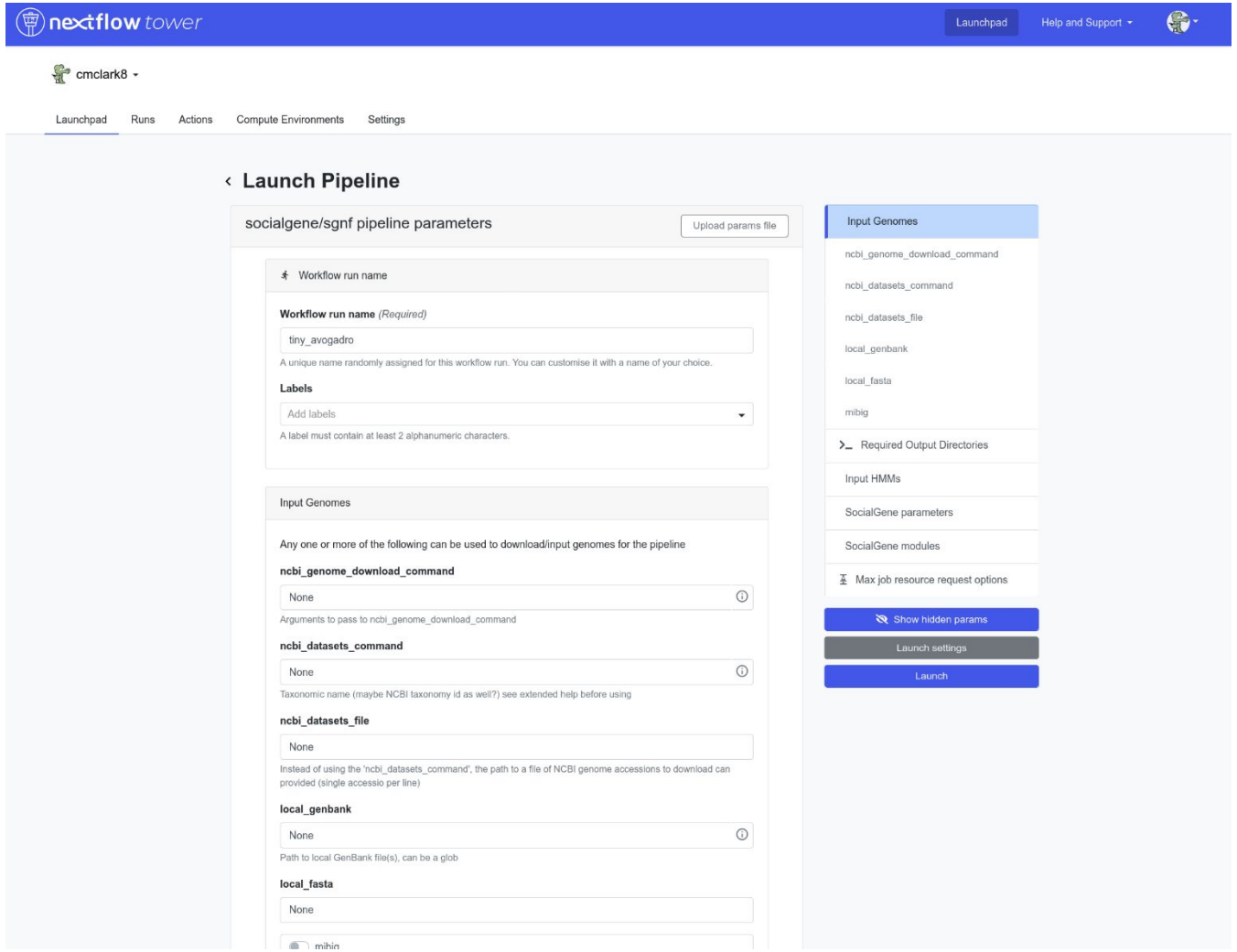




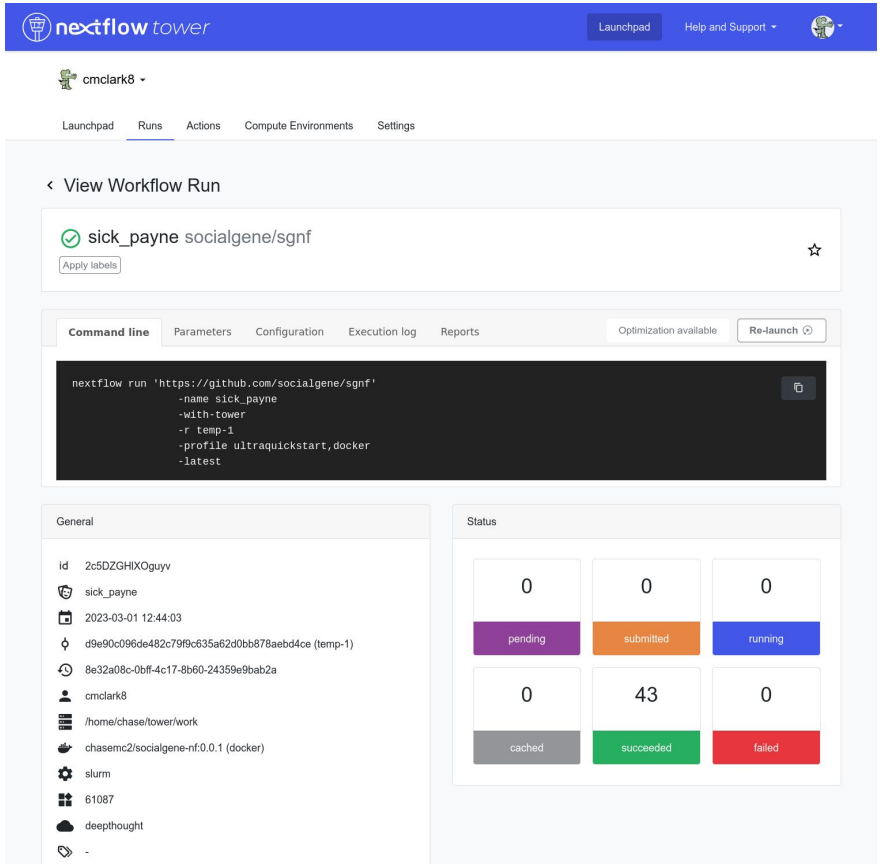
Making the pipeline more accessible with nextflow tower

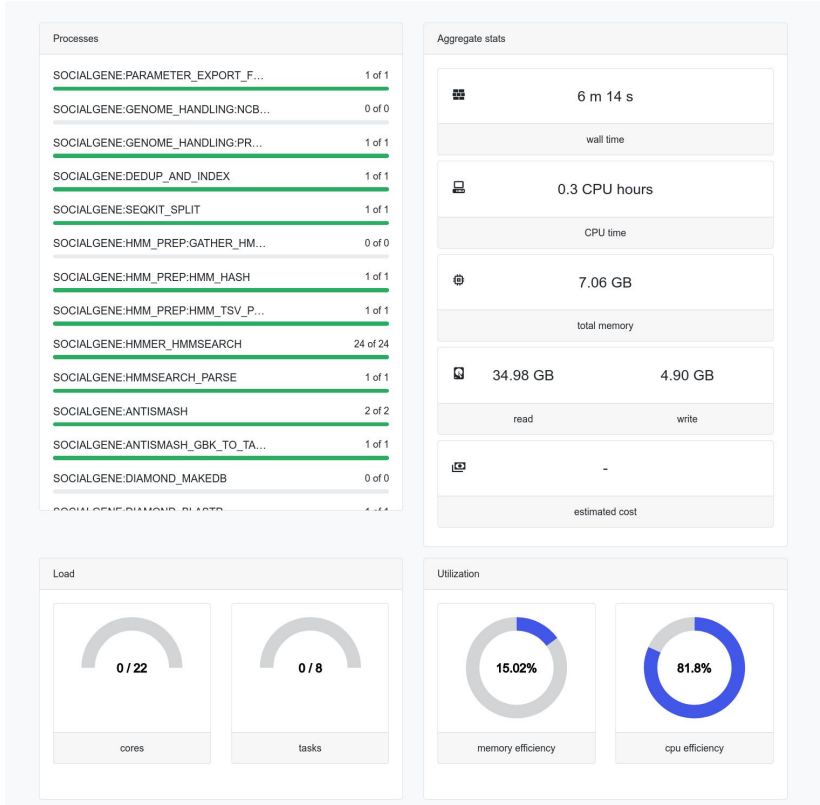


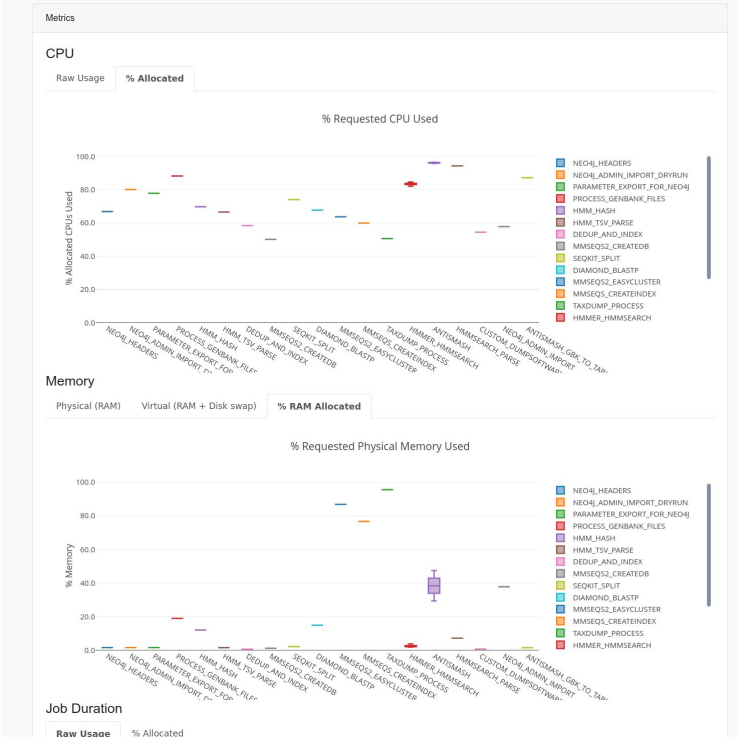
Making the pipeline more accessible with nextflow tower



Making the workflow more accessible with nextflow tower

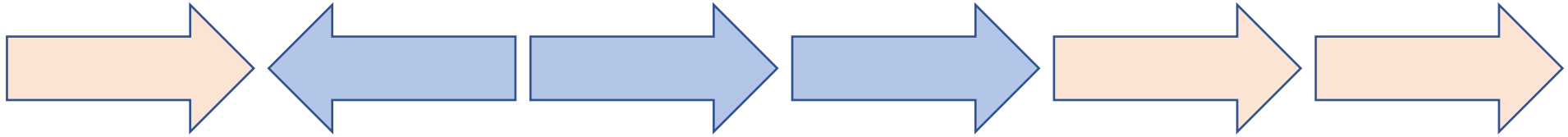




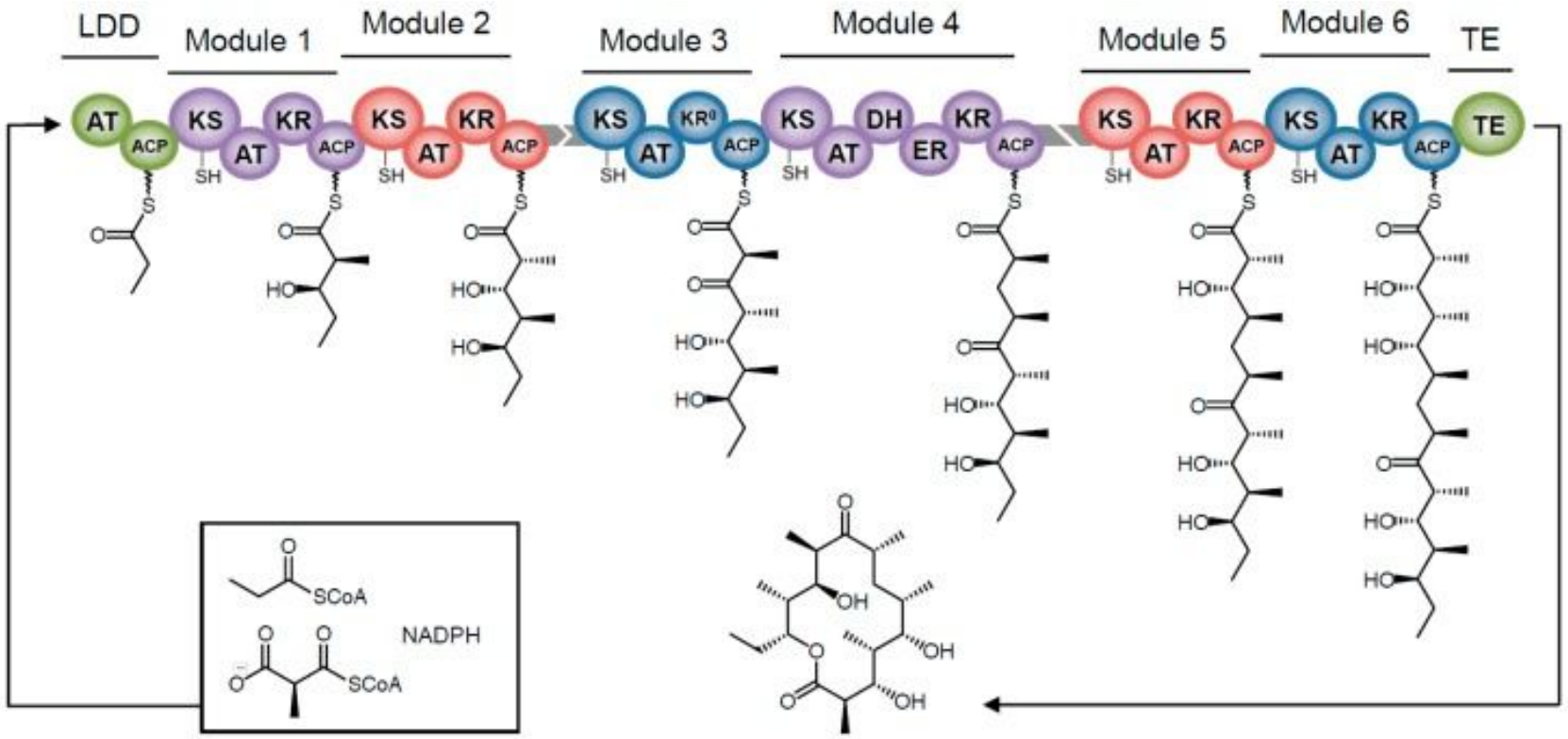


Molecular Biology in < 1 minute

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			

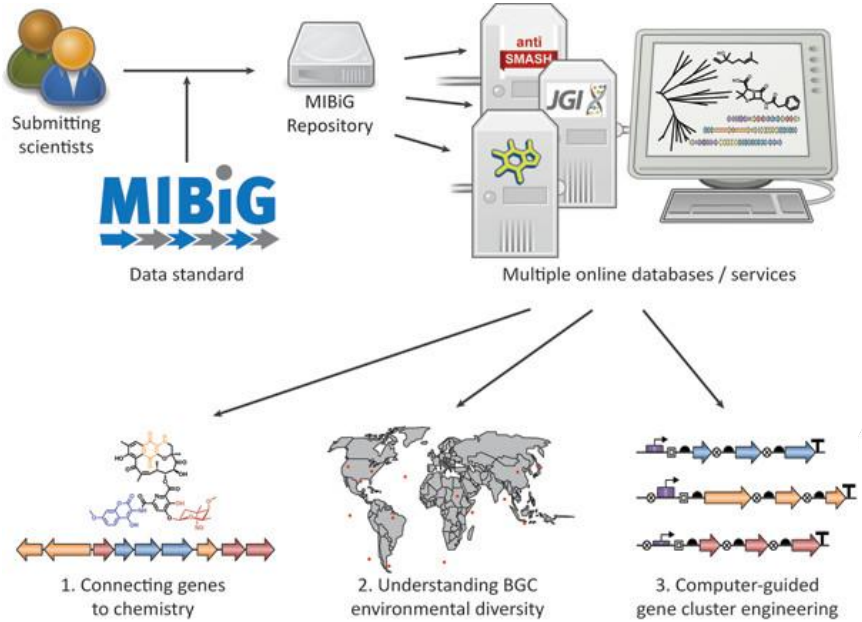
Polyketide synthases are large, multidomain proteins
“Beads on a String”



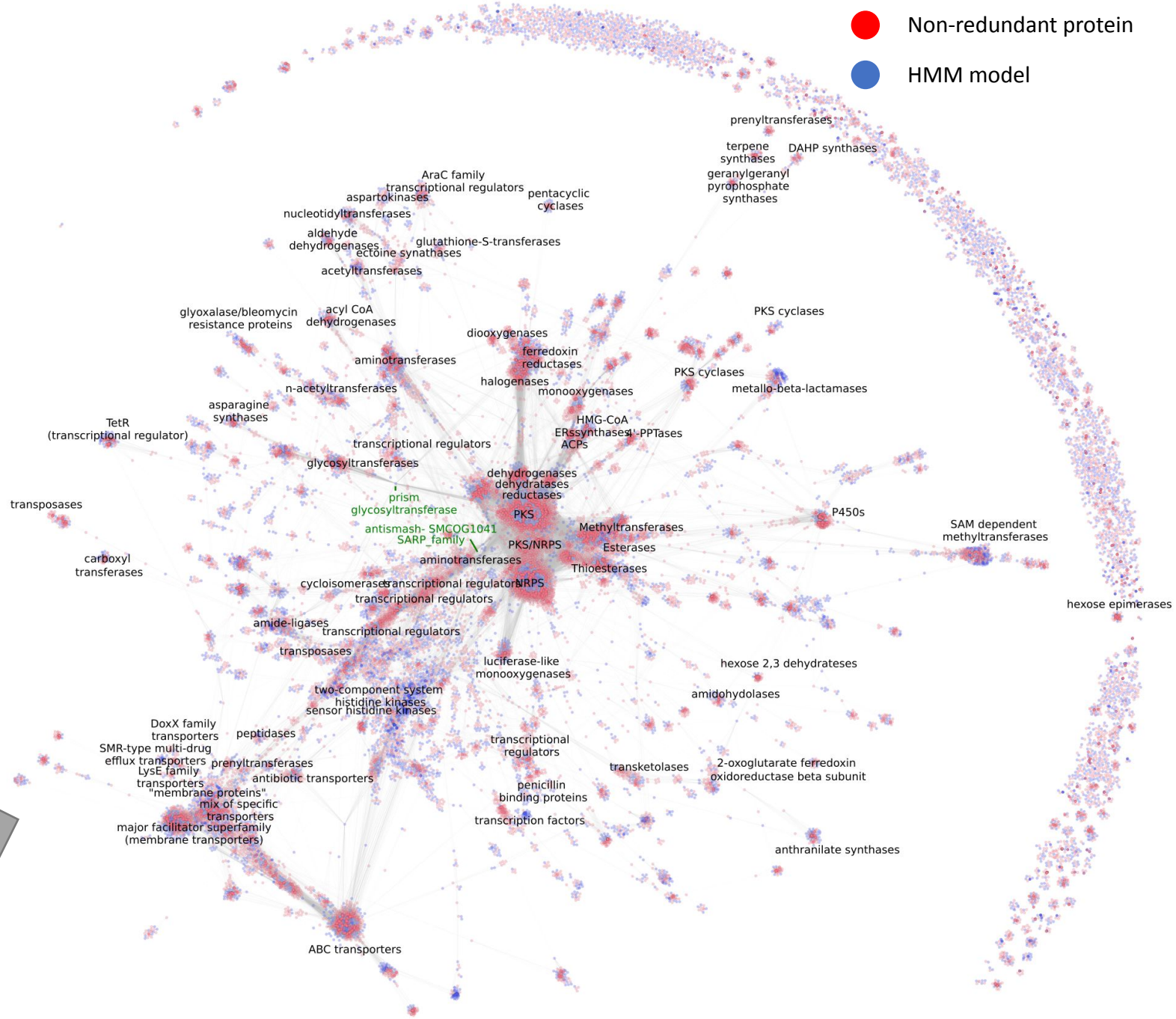
Bayly, Carmen L, and Vikramaditya G Yadav. “Towards Precision Engineering of Canonical Polyketide Synthase Domains: Recent Advances and Future Prospects.” *Molecules (Basel, Switzerland)*vol. 22,2 235. 5 Feb. 2017, doi:10.3390/molecules22020235

Vignette 1

Mapping MIBiG

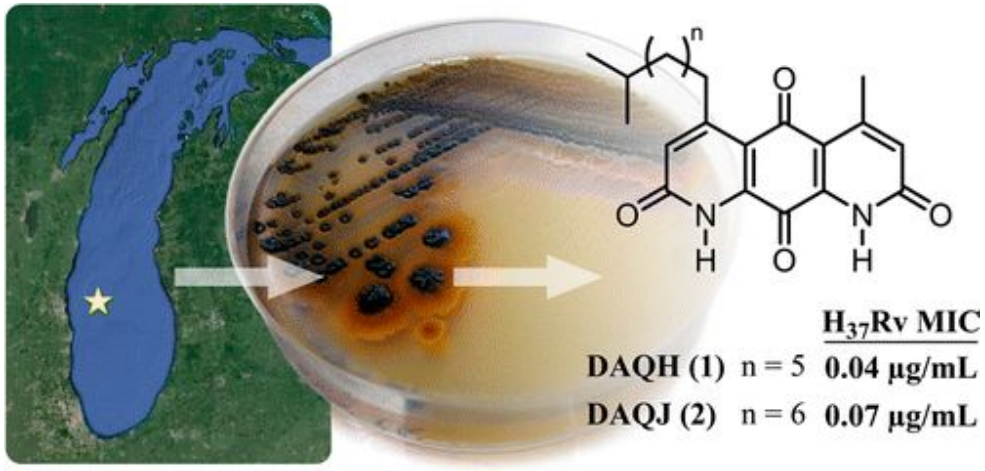

Submitting scientists
MIBiG Repository
anti SMASH
JGI
MIBiG
Data standard
Multiple online databases / services
1. Connecting genes to chemistry
2. Understanding BGC environmental diversity
3. Computer-guided gene cluster engineering

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 synthases
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
prenyltransferases
transketolases
2-oxoglutarate ferredoxin oxidoreductase beta subunit
LysE family transporters
antibiotic transporters
penicillin binding proteins
"membrane proteins" mix of specific transporters
transcription factors
major facilitator superfamily (membrane transporters)
anthranilate synthases
ABC transporters

Vignette 2

Looking for diazaquinomycin analogs



Mullowney, Michael W et al. “Diaza-anthracene Antibiotics from a Freshwater-Derived Actinomycete with Selective Antibacterial Activity toward *Mycobacterium tuberculosis*.” *ACS infectious diseases* vol. 1,4 (2015): 168-174. doi:10.1021/acsinfecdis.5b00005

RETURN TO ISSUE ‹ PREV ARTICLE NEXT ›

Diazaquinomycin Biosynthetic Gene Clusters from Marine and Freshwater Actinomycetes

Jana Braesel, Jung-Ho Lee, Benoit Arnould, Brian T. Murphy, and Alessandra S. Eustáquio*

Cite this: *J. Nat. Prod.* 2019, 82, 4, 937–946
Publication Date: March 21, 2019
https://doi.org/10.1021/acs.jnatprod.8b01028
Copyright © 2019 American Chemical Society and American Society of Pharmacognosy
RIGHTS & PERMISSIONS

Article Views 1685 | Altmetric 8 | Citations 8
LEARN ABOUT THESE METRICS

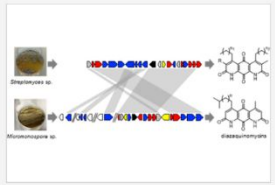
Share Add to Export

Read Online | PDF (2 MB) | Supporting Info (1) » | SUBJECTS: Peptides and proteins, Bacteria, Genetics,


Journal of Natural Products

Abstract

Tuberculosis is an infectious disease of global concern. Members of the diazaquinomycin (DAQ) class of natural products have shown potent and selective activity against drug-resistant *Mycobacterium tuberculosis*. However, poor solubility has prevented further development of this compound class. Understanding DAQ biosynthesis may provide a viable route for the generation of derivatives with improved properties. We have sequenced the genomes of two actinomycete bacteria that produce distinct DAQ derivatives. While software tools for automated biosynthetic gene cluster (BGC) prediction failed to detect DAQ BGCs, comparative genomics using MAUVE alignment led to the identification of putative BGCs in the marine *Streptomyces* sp. F001 and in the freshwater *Micromonospora* sp. B006. Deletion of the identified *daq* BGC in strain B006 using CRISPR-Cas9 genome editing abolished DAQ production, providing experimental evidence for BGC assignment. A complete model for DAQ biosynthesis is proposed based on the genes identified. Insufficient knowledge of natural product biosynthesis is one of the major challenges of productive genome mining approaches. The results reported here fill a gap in knowledge regarding the genetic basis for the biosynthesis of DAQ antibiotics. Moreover, identification of the *daq* BGC shall enable future generations of improved derivatives using biosynthetic methods.



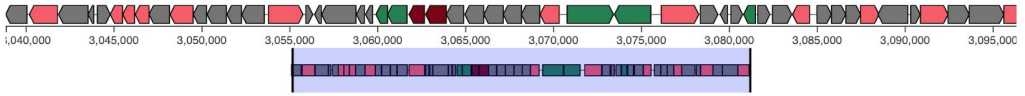
mibig.secondarymetabolites.org

MIBiG Repository of Known Biosynthetic Gene Clusters — Download — Contact

BGC0001848: diazaquinomycin H biosynthetic gene cluster from *Micromonospora sp. B006*

Location: 3,038,865 - 3,097,062 nt. (total: 58,198 nt).
This entry is originally from NCBI GenBank CP030865.1.

Download Cluster GenBank file | View antiSMASH-generated output



Gene details

Select a gene to view the details available for it

Legend:

core biosynthetic genes | additional biosynthetic genes | transport-related genes | regulatory genes | resistance genes | other genes

reset view | zoom to selection

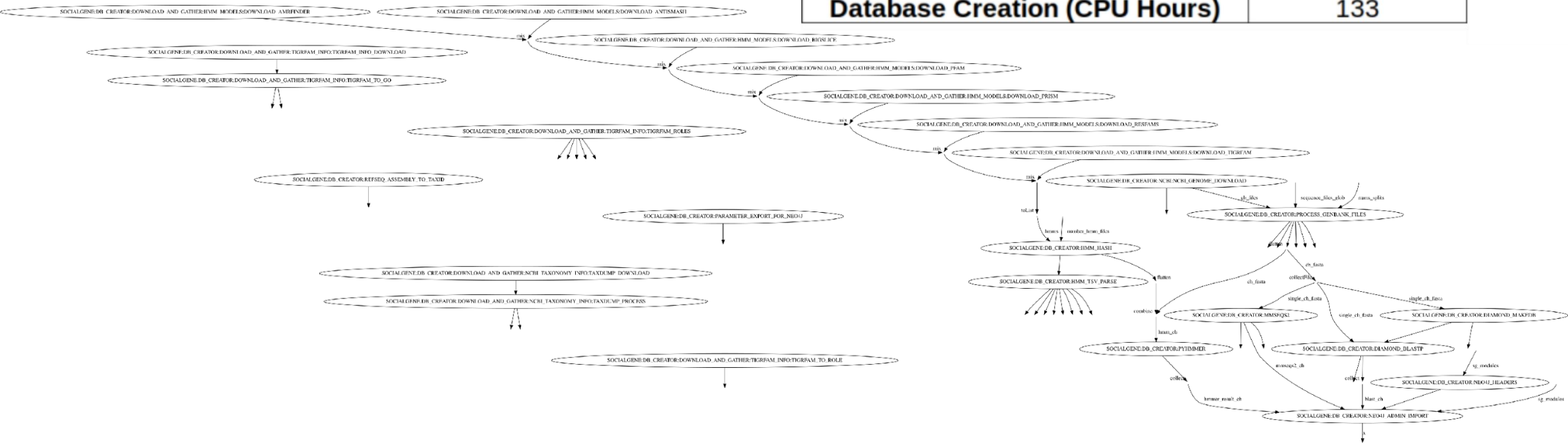
General | Compounds | Genes | History | KnownClusterBlast | NRPS/PKS domains

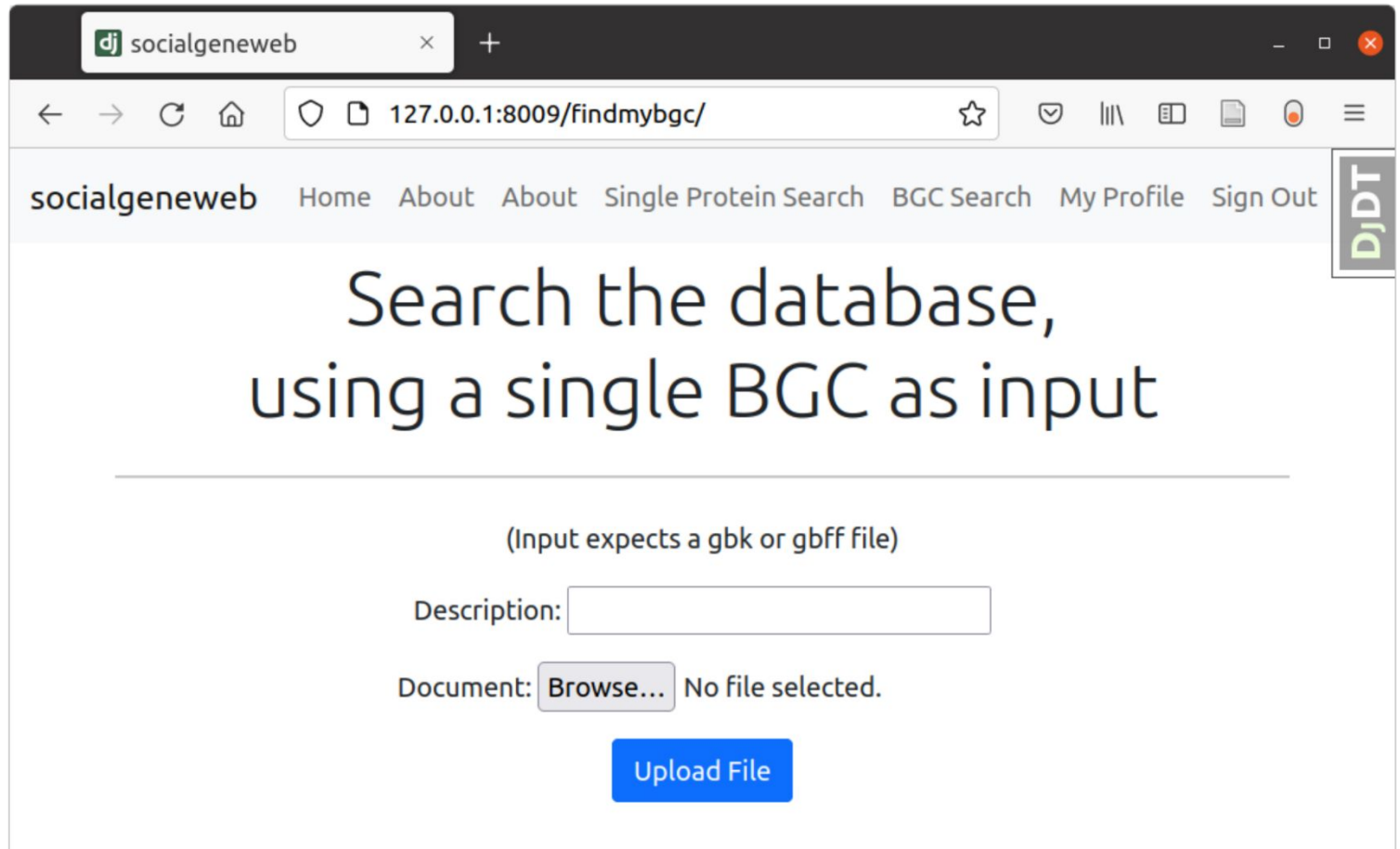
List of genes involved in compound(s) production

Identifiers	Position	Product	Functions	Evidence	Extra
MicB006_2892 AXO35173.1	3038865 - 3040082 (-)	2-keto-3-deoxy-D-arabino-heptulosonate-7- phosphate synthase II	Precursor biosynthesis	Sequence-based prediction	copy AA seq copy Nt seq
MicB006_2893 AXO35174.1	3040216 - 3041814 (-)	hypothetical protein			copy AA seq copy Nt seq
MicB006_2894 AXO35175.1	3041874 - 3043553 (-)	SSS sodium solute transporter superfamily			copy AA seq copy Nt seq
MicB006_2895 AXO35176.1	3043550 - 3043885 (-)	putative membrane protein			copy AA seq copy Nt seq
MicB006_2896 AXO35177.1	3044080 - 3044820 (+)	hypothetical protein			copy AA seq copy Nt seq
MicB006_2897 AXO35178.1	3044813 - 3045514 (-)	short-chain dehydrogenase/reductase SDR			copy AA seq copy Nt seq
MicB006_2898 AXO35179.1	3045526 - 3046200 (-)	isochorismatase	Precursor biosynthesis	Sequence-based prediction	copy AA seq copy Nt seq
MicB006_2899 AXO35180.1	3046233 - 3047018 (-)	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase	Precursor biosynthesis	Sequence-based prediction	copy AA seq copy Nt seq
MicB006_2900 AXO35181.1	3047018 - 3048187 (-)	putative n-hydroxybenzoate hydroxylase			copy AA seq copy Nt seq
MicB006_2901 AXO35182.1	3048219 - 3049535 (-)	phenylacetate-coenzyme A ligase			copy AA seq copy Nt seq
MicB006_2902 AXO35183.1	3049532 - 3050293 (-)	hypothetical protein			copy AA seq copy Nt seq
MicB006_2903 AXO35184.1	3050319 - 3051422 (-)	37-dideoxy-D-threo-hepto-2 6-diulosonate synthase			copy AA seq copy Nt seq
MicB006_2904 AXO35185.1	3051454 - 3052278 (-)	2-amino-37-dideoxy-D-threo-hept-6-ulosonate synthase			copy AA seq copy Nt seq
MicB006_2905 AXO35186.1	3052294 - 3053577 (-)	aspartokinase			copy AA seq copy Nt seq
MicB006_2906 AXO35187.1	3053813 - 3055777 (+)	anthranilate synthase	Precursor biosynthesis	Sequence-based prediction	copy AA seq copy Nt seq
MicB006_2907 AXO35188.1	3055937 - 3056335 (+)	hypothetical protein			copy AA seq copy Nt seq
MicB006_2908 AXO35189.1	3056436 - 3056825 (-)	glyoxalase/bleomycin resistance protein/dioxygenase			copy AA seq copy Nt seq
MicB006_2909 AXO35190.1	3056836 - 3058797 (-)	multicopper oxidase			copy AA seq copy Nt seq
MicB006_2910 AXO35191.1	3058818 - 3059273 (-)	hypothetical protein			copy AA seq copy Nt seq
MicB006_2911 AXO35192.1	3059338 - 3059733 (-)	hypothetical protein			copy AA seq copy Nt seq
MicB006_2912	3059926 - 3060588 (-)	DNA-binding response regulator LuxR family			copy AA seq

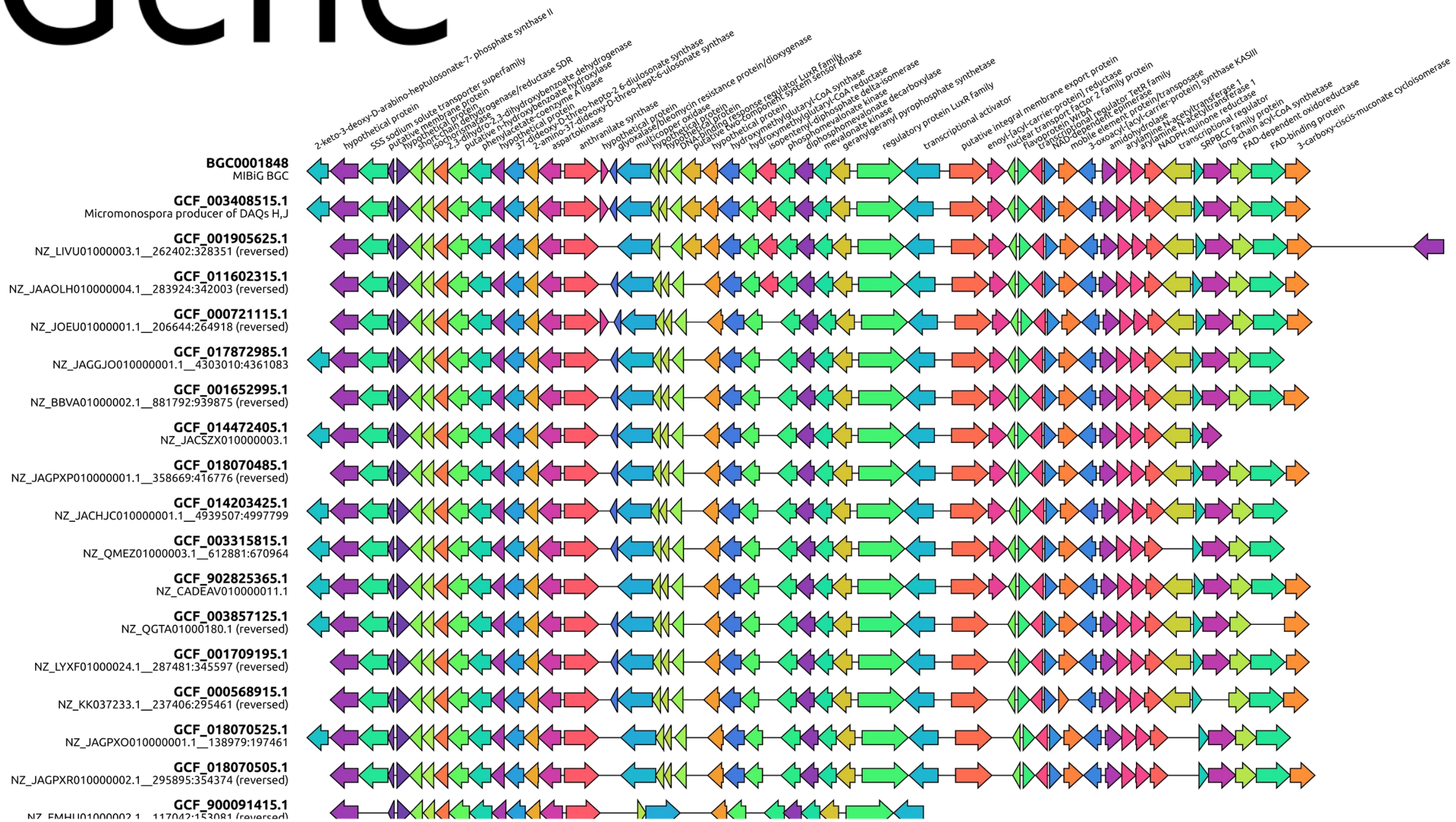
SocialGene

	Micromonospora
Genomes	226
Contigs/Scaffolds	34,856
Proteins	1,373,641
Non-Redundant Proteins	1,076,690
HMM Annotations	10,656,754
MMSEQS2	1,009,056
Bidirectional BLASTP (45.3 GB)	420,788,981
Database Creation (Real Duration)	4h 30m
Database Creation (CPU Hours)	133




socialgeneweb
127.0.0.1:8009/findmybgc/
socialgeneweb
Home
About
About
Single Protein Search
BGC Search
My Profile
Sign Out
DjDT
Search the database,
using a single BGC as input
(Input expects a gbk or gbff file)
Description:
Document:
Browse...
No file selected.
Upload File

SocialGene

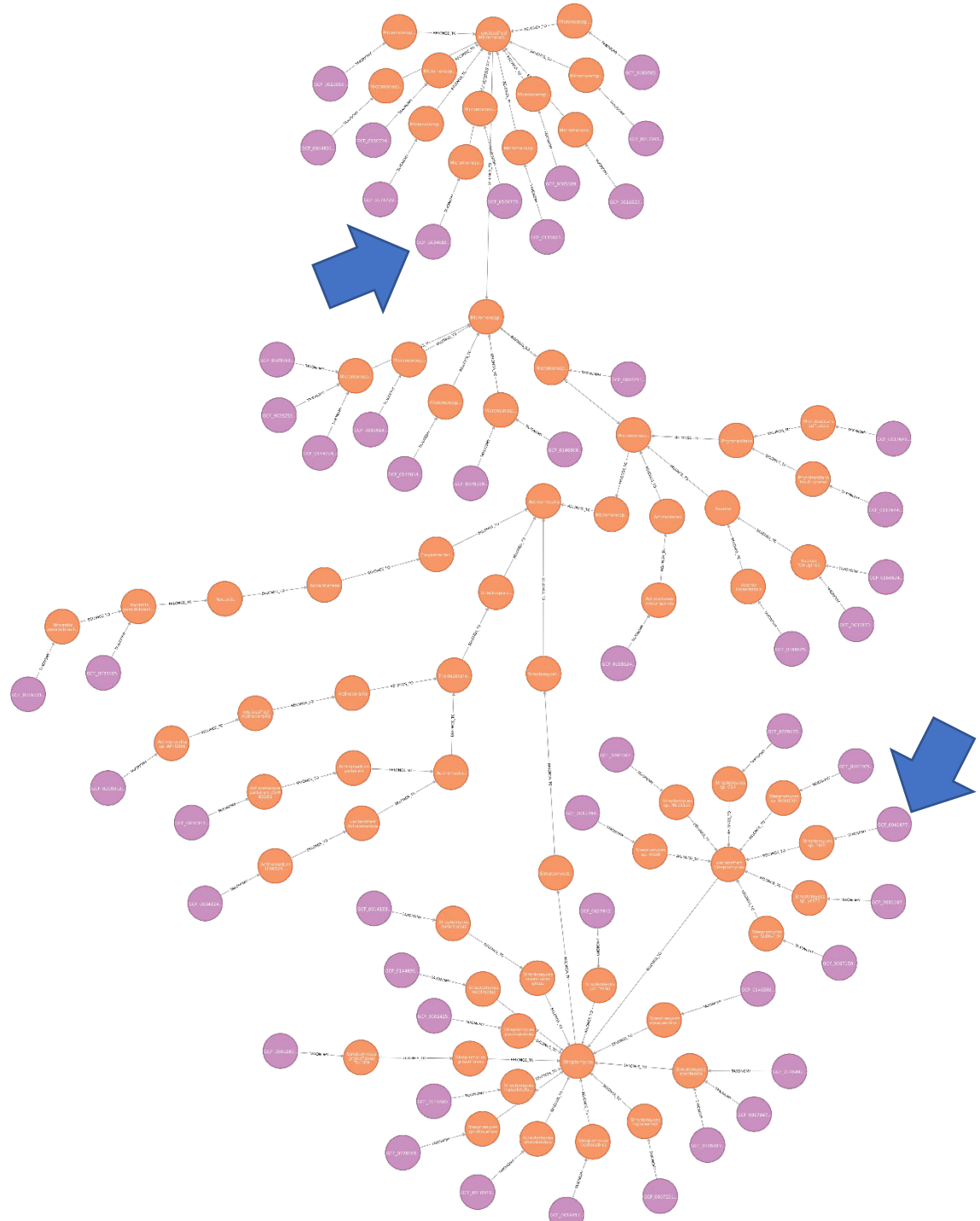

2-keto-3-deoxy-D-arabino-heptulosonate-7- phosphate synthase II
hypothetical protein
SSS sodium solute transporter superfamily
putative membrane protein
hypothetical protein
short-chain dehydrogenase/reductase SDR
isochorismatase
2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase
putative n-hydroxybenzoate hydroxylase
phenylacetate-coenzyme A ligase
hypothetical protein
37-dideoxy-D-threo-hepto-2 6-diulosonate synthase
2-amino-37-dideoxy-D-threo-hept-6-ulosonate synthase
aspartokinase
anthranilate synthase
hypothetical protein
glyoxalase/bleomycin resistance protein/dioxygenase
multicopper oxidase
hypothetical protein
hypothetical protein
DNA-binding response regulator LuxR family
putative two-component system sensor kinase
hypothetical protein
hydroxymethylglutaryl-CoA synthase
hydroxymethylglutaryl-CoA reductase
isopentenyl-diphosphate delta-isomerase
phosphomevalonate kinase
diphosphomevalonate decarboxylase
mevalonate kinase
geranylgeranyl pyrophosphate synthetase
regulatory protein LuxR family
transcriptional activator
putative integral membrane export protein
enoyl-[acyl-carrier-protein] reductase
nuclear transport factor 2 family protein
flavoprotein WrbA
transcriptional regulator TetR family
NAD-dependent epimerase
mobile element protein/transposase
3-oxoacyl-[acyl-carrier-protein] synthase KASIII
amidohydrolase
arylamine N-acetyltransferase 1
arylamine N-acetyltransferase 1
NADPH:quinone reductase
transcriptional regulator
SRPBCC family protein
long-chain acyl-CoA synthetase
FAD-dependent oxidoreductase
FAD-binding protein
3-carboxy-ciscis-muconate cycloisomerase
BGC0001848
MIBiG BGC
GCF_003408515.1
Micromonospora producer of DAQs H,J
GCF_001905625.1
NZ_LIVU01000003.1__262402:328351 (reversed)
GCF_011602315.1
NZ_JAAOLH010000004.1__283924:342003 (reversed)
GCF_000721115.1
NZ_JOEU01000001.1__206644:264918 (reversed)
GCF_017872985.1
NZ_JAGGJO010000001.1__4303010:4361083
GCF_001652995.1
NZ_BBVA01000002.1__881792:939875 (reversed)
GCF_014472405.1
NZ_JACSZX010000003.1
GCF_018070485.1
NZ_JAGPXP010000001.1__358669:416776 (reversed)
GCF_014203425.1
NZ_JACHJC010000001.1__4939507:4997799
GCF_003315815.1
NZ_QMEZ01000003.1__612881:670964
GCF_902825365.1
NZ_CADEAV010000011.1
GCF_003857125.1
NZ_QGTA01000180.1 (reversed)
GCF_001709195.1
NZ_LYXF01000024.1__287481:345597 (reversed)
GCF_000568915.1
NZ_KK037233.1__237406:295461 (reversed)
GCF_018070525.1
NZ_JAGPXO010000001.1__138979:197461
GCF_018070505.1
NZ_JAGPXR010000002.1__295895:354374 (reversed)
GCF_900091415.1

	Micromonospora
Genomes	226
Contigs/Scaffolds	34,856
Proteins	1,373,641
Non-Redundant Proteins	1,076,690
HMM Annotations	10,656,754
MMSEQS2	1,009,056
Bidirectional BLASTP (45.3 GB)	420,788,981
Database Creation (Real Duration)	4h 30m
Database Creation (CPU Hours)	133

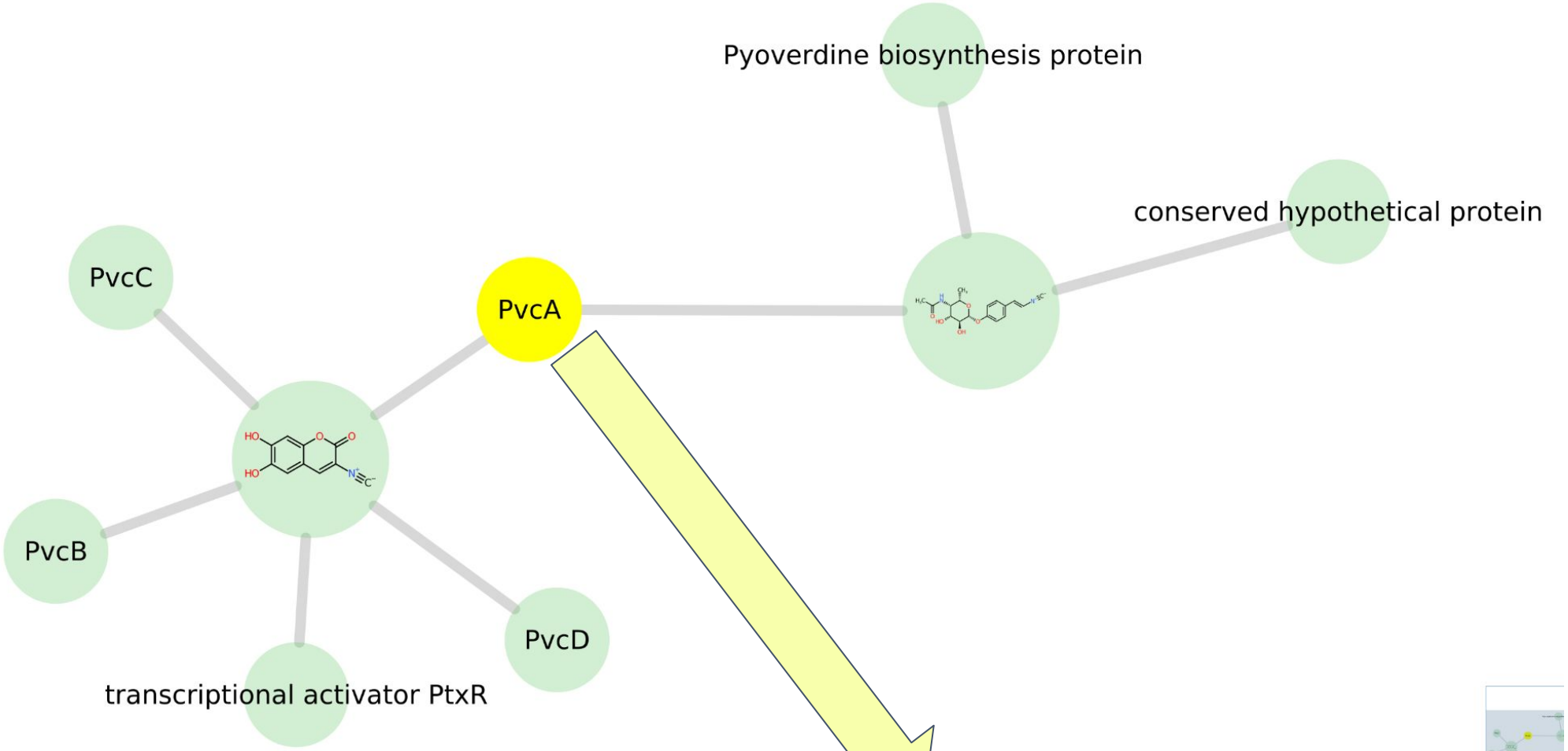
SocialGene

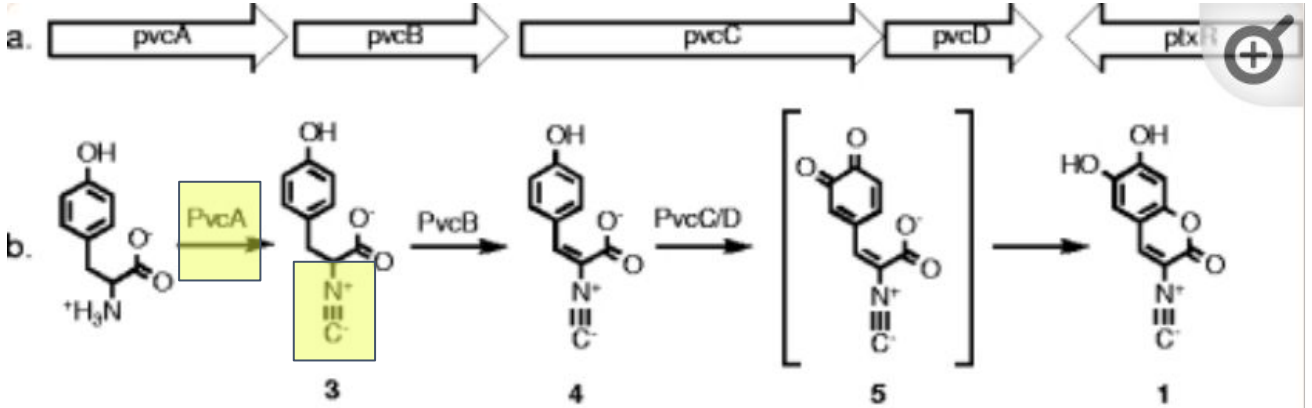
Open Science Grid

	RefSeq
Genomes	266,668
Contigs/Scaffolds	23,941,594
Proteins	188,429,555
HMM models	25,648
HMM annotations	1,403,423,051
MMseqs2 (bug in MMseqs2)	188,327,165
Contigs to Proteins	971,298,319
Species	49,902
Genera	6,460

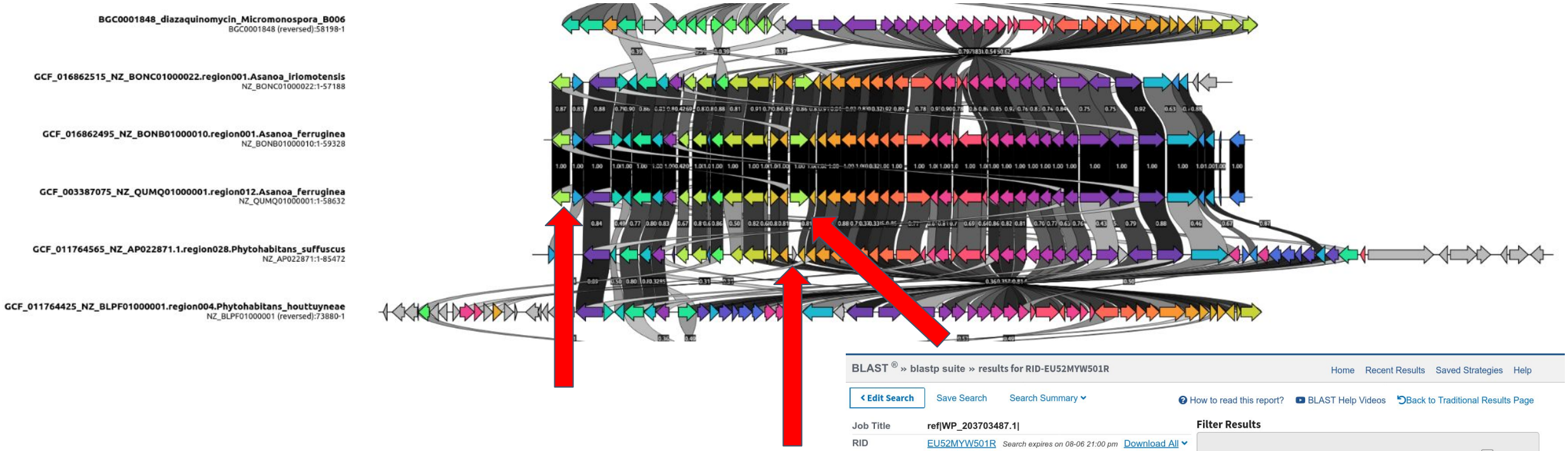


Teasers- What Next?

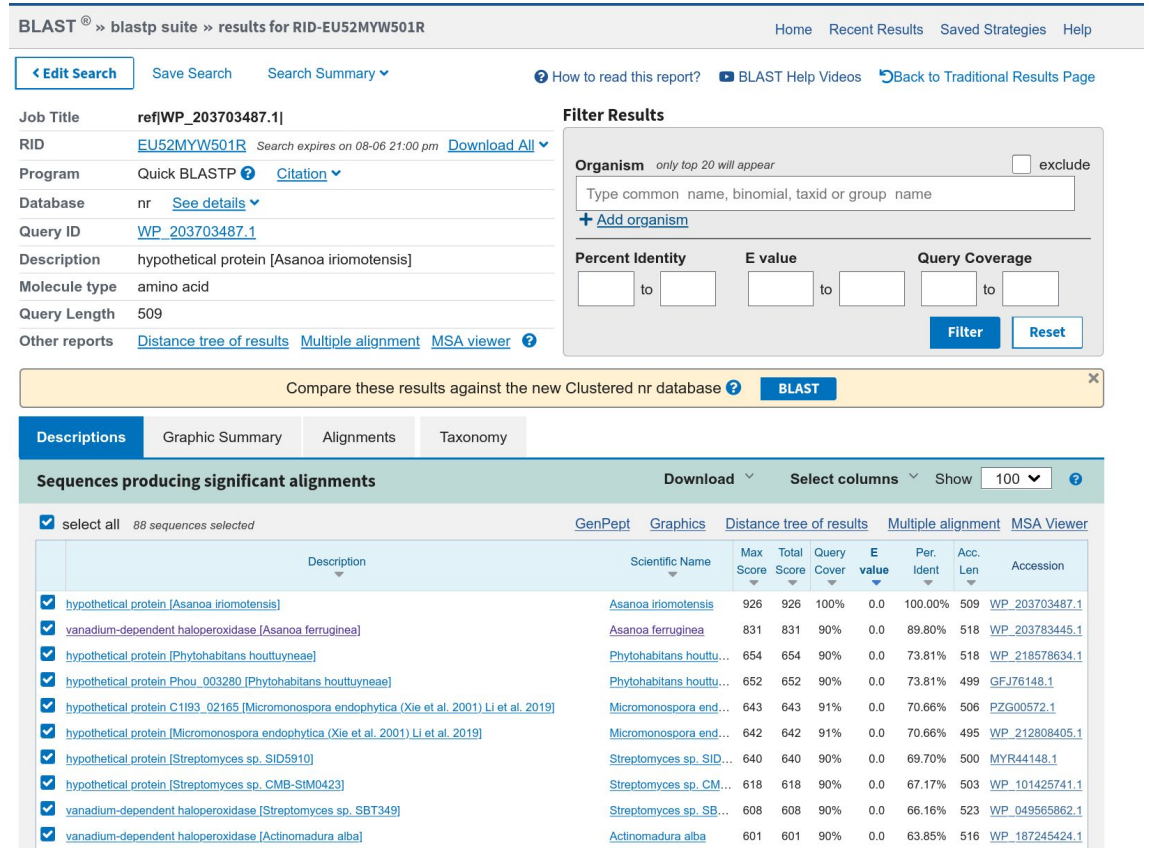




Clarke-Pearson MF, Brady SF. Paerucumarin, a new metabolite produced by the pvc gene cluster from Pseudomonas aeruginosa. J Bacteriol. 2008 Oct;190(20):6927-30.



Vanadium-dependent Haloperoxidases



BLAST® » blastp suite » results for RID-EU52MYW501R

Home | Recent Results | Saved Strategies | Help

‹ Edit Search | Save Search | Search Summary

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	ref|WP_203703487.1|
RID	EU52MYW501R Search expires on 08-06 21:00 pm Download All
Program	Quick BLASTP Citation
Database	nr See details
Query ID	WP_203703487.1
Description	hypothetical protein [Asanoa iriomotensis]
Molecule type	amino acid
Query Length	509
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism only top 20 will appear — exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity — to

E value — to

Query Coverage — to

Filter | Reset

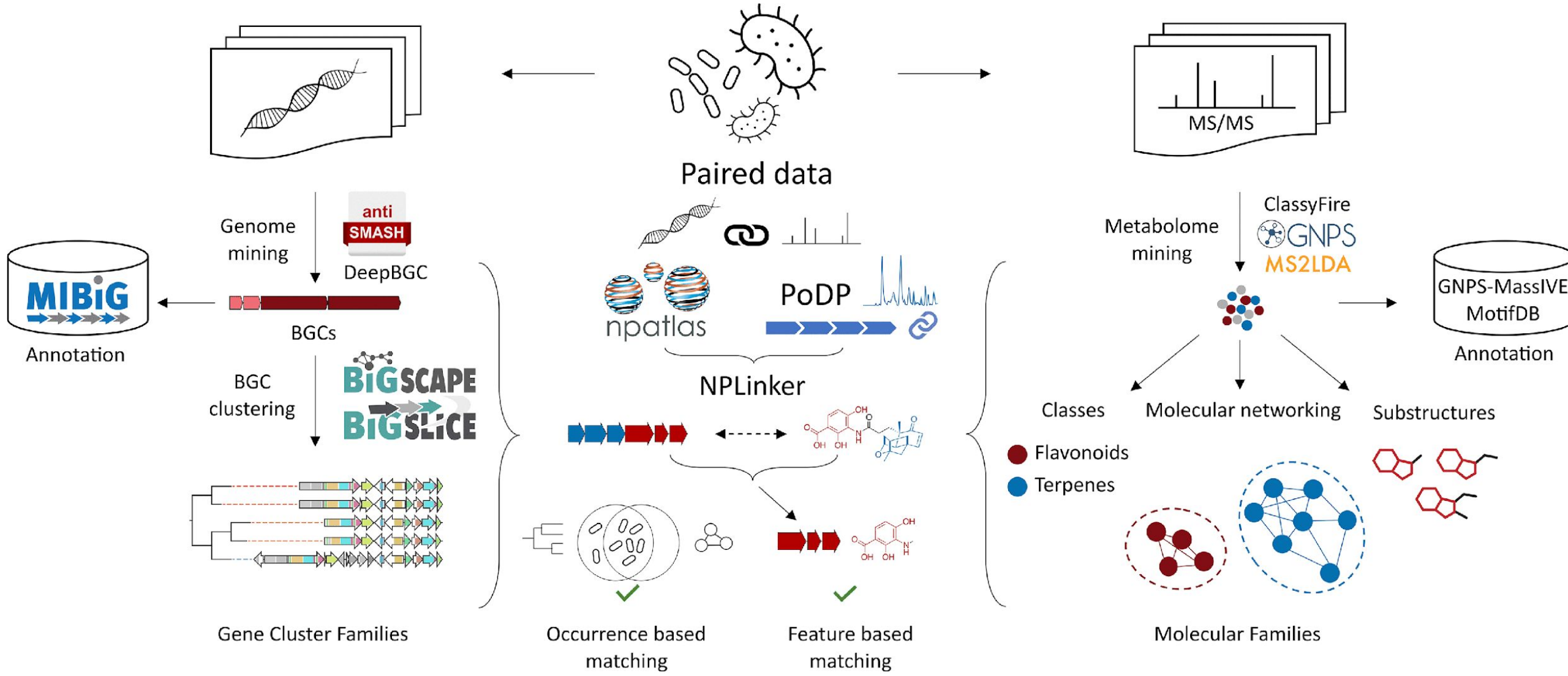
Compare these results against the new Clustered nr database BLAST

Descriptions | Graphic Summary | Alignments | Taxonomy

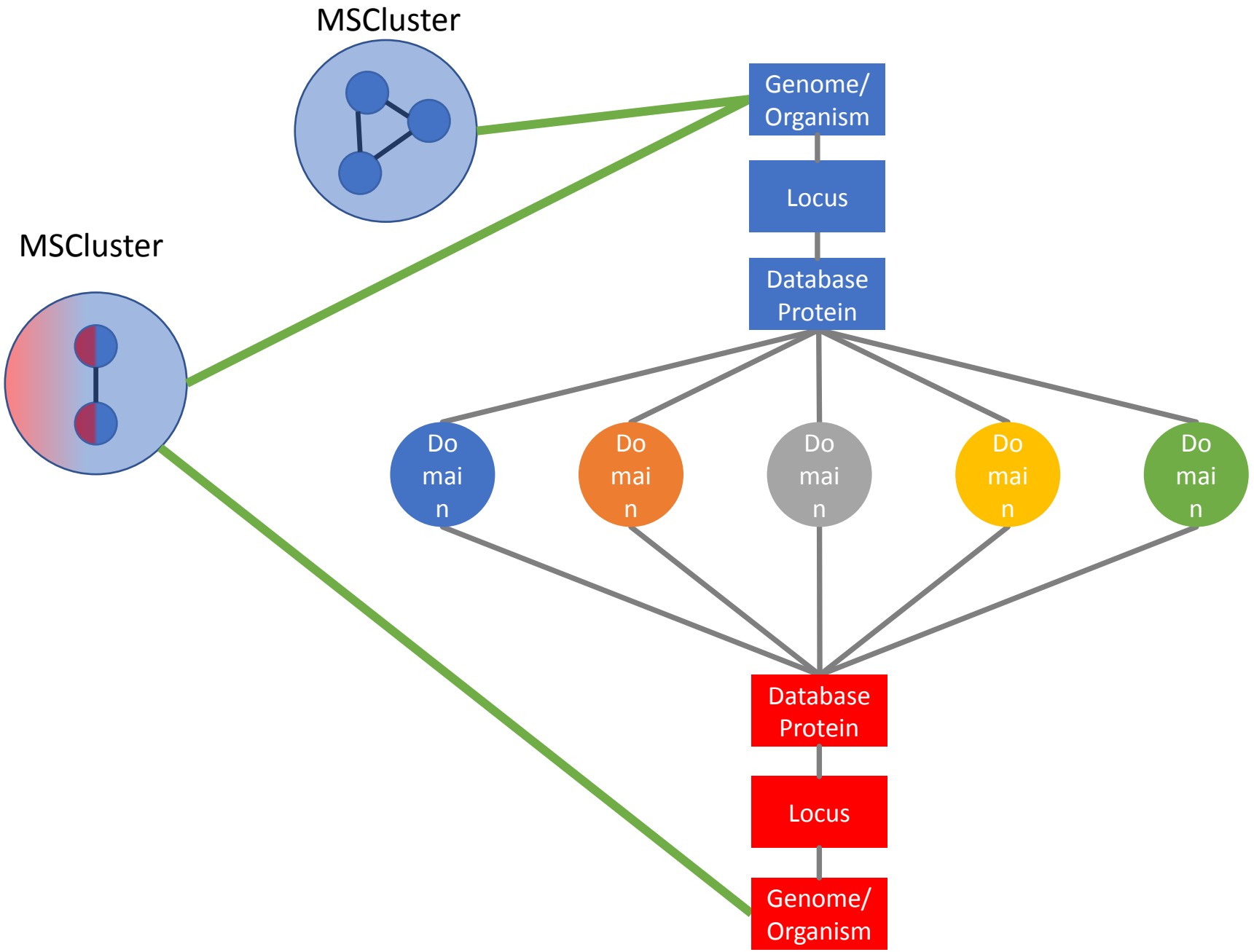
Sequences producing significant alignments — Download | Select columns | Show 100

select all 88 sequences selected — GenPept | Graphics | Distance tree of results | Multiple alignment | MSA Viewer

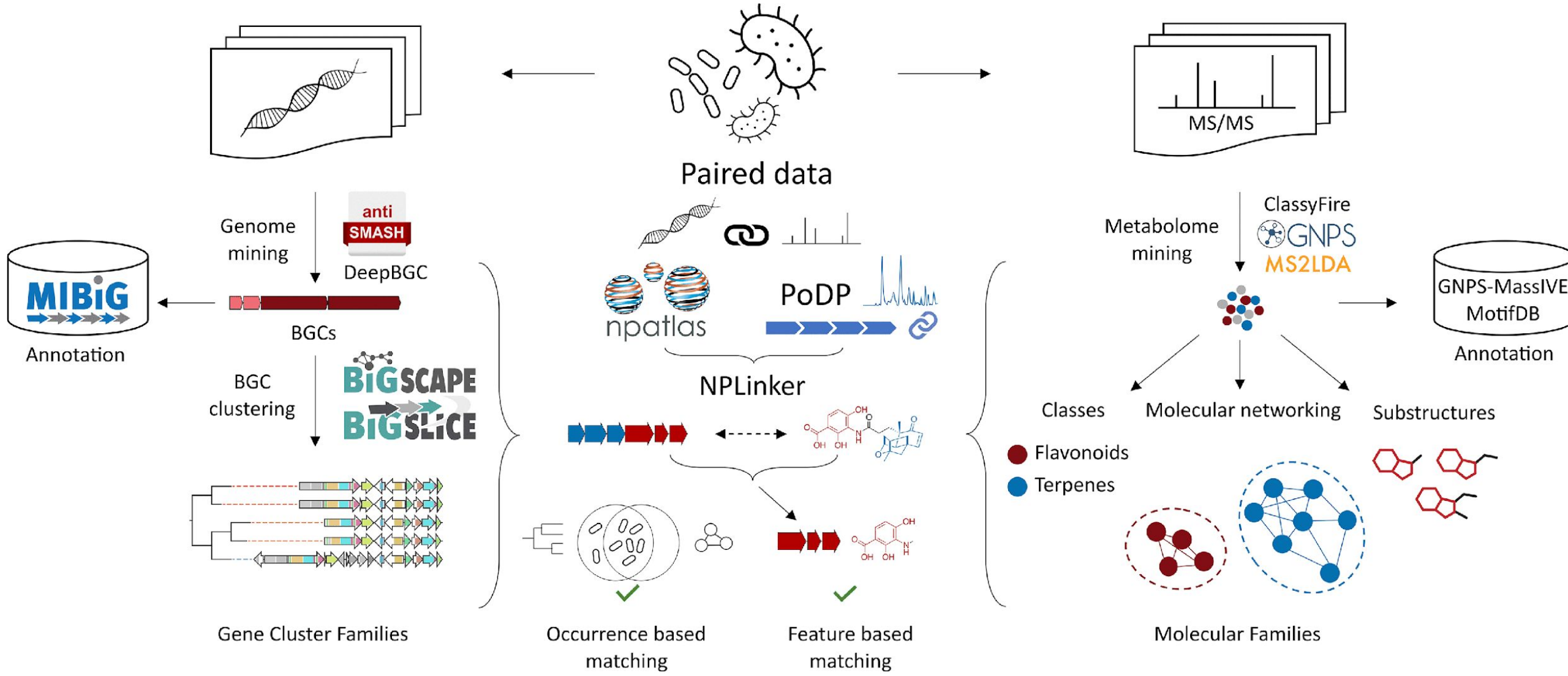
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
hypothetical protein [Asanoa iriomotensis]	Asanoa iriomotensis	926	926	100%	0.0	100.00%	509	WP_203703487.1
vanadium-dependent haloperoxidase [Asanoa ferruginea]	Asanoa ferruginea	831	831	90%	0.0	89.80%	518	WP_203783445.1
hypothetical protein [Phytohabitans houttuyneae]	Phytohabitans houttu…	654	654	90%	0.0	73.81%	518	WP_218578634.1
hypothetical protein Phou_003280 [Phytohabitans houttuyneae]	Phytohabitans houttu…	652	652	90%	0.0	73.81%	499	GFJ76148.1
hypothetical protein C1I93_02165 [Micromonospora endophytica (Xie et al. 2001) Li et al. 2019]	Micromonospora end…	643	643	91%	0.0	70.66%	506	PZG00572.1
hypothetical protein [Micromonospora endophytica (Xie et al. 2001) Li et al. 2019]	Micromonospora end…	642	642	91%	0.0	70.66%	495	WP_212808405.1
hypothetical protein [Streptomyces sp. SID5910]	Streptomyces sp. SID…	640	640	90%	0.0	69.70%	500	MYR44148.1
hypothetical protein [Streptomyces sp. CMB-StM0423]	Streptomyces sp. CM…	618	618	90%	0.0	67.17%	503	WP_101425741.1
vanadium-dependent haloperoxidase [Streptomyces sp. SBT349]	Streptomyces sp. SB…	608	608	90%	0.0	66.16%	523	WP_049565862.1
vanadium-dependent haloperoxidase [Actinomadura alba]	Actinomadura alba	601	601	90%	0.0	63.85%	516	WP_187245424.1



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


MSCluster
MSCluster
Genome/ Organism
Locus
Database Protein
Do mai n
Do mai n
Do mai n
Do mai n
Do mai n
Database Protein
Locus
Genome/ Organism

GNPS



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

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

UIC PHARMACEUTICAL SCIENCES COLLEGE OF PHARMACY

UIC Center for Biomolecular Sciences

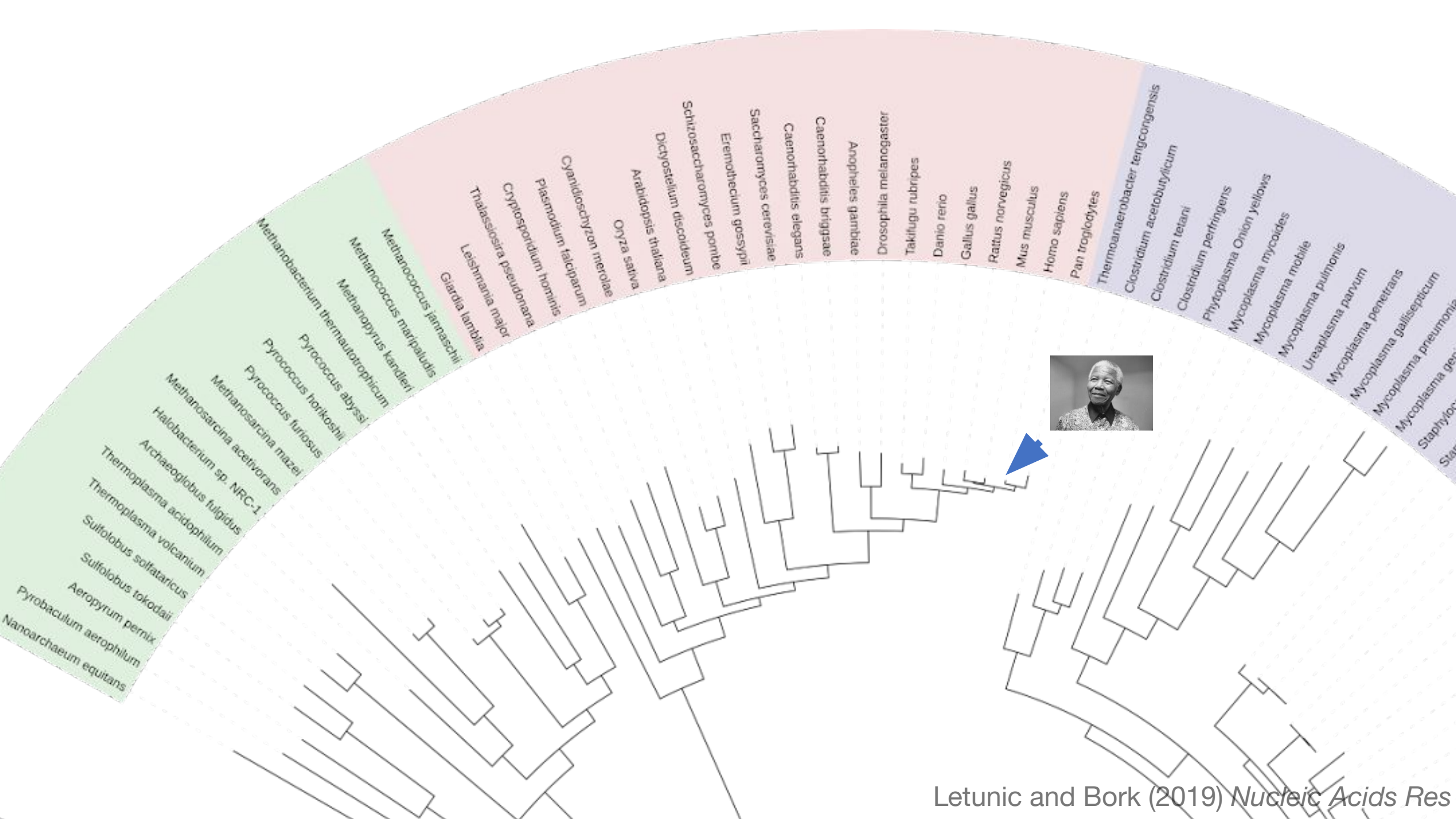
Funding:

National Institute of Health (NIH)

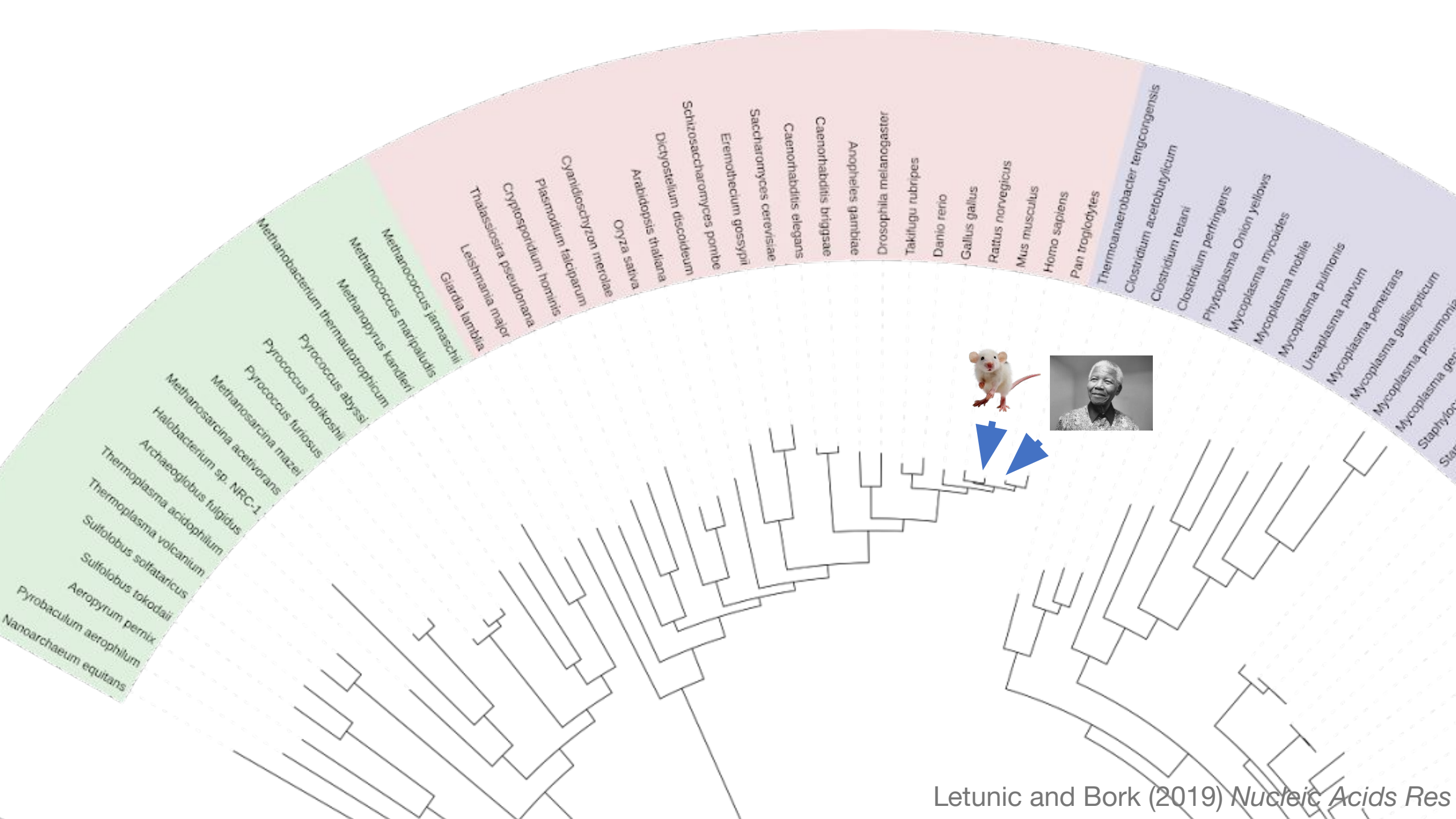
Current:
NLM: 5T15LM007359
NIGMS: R35 GM133776
Past:

NCCIH: F31 AT010419
NIGMS: R01 GM125943

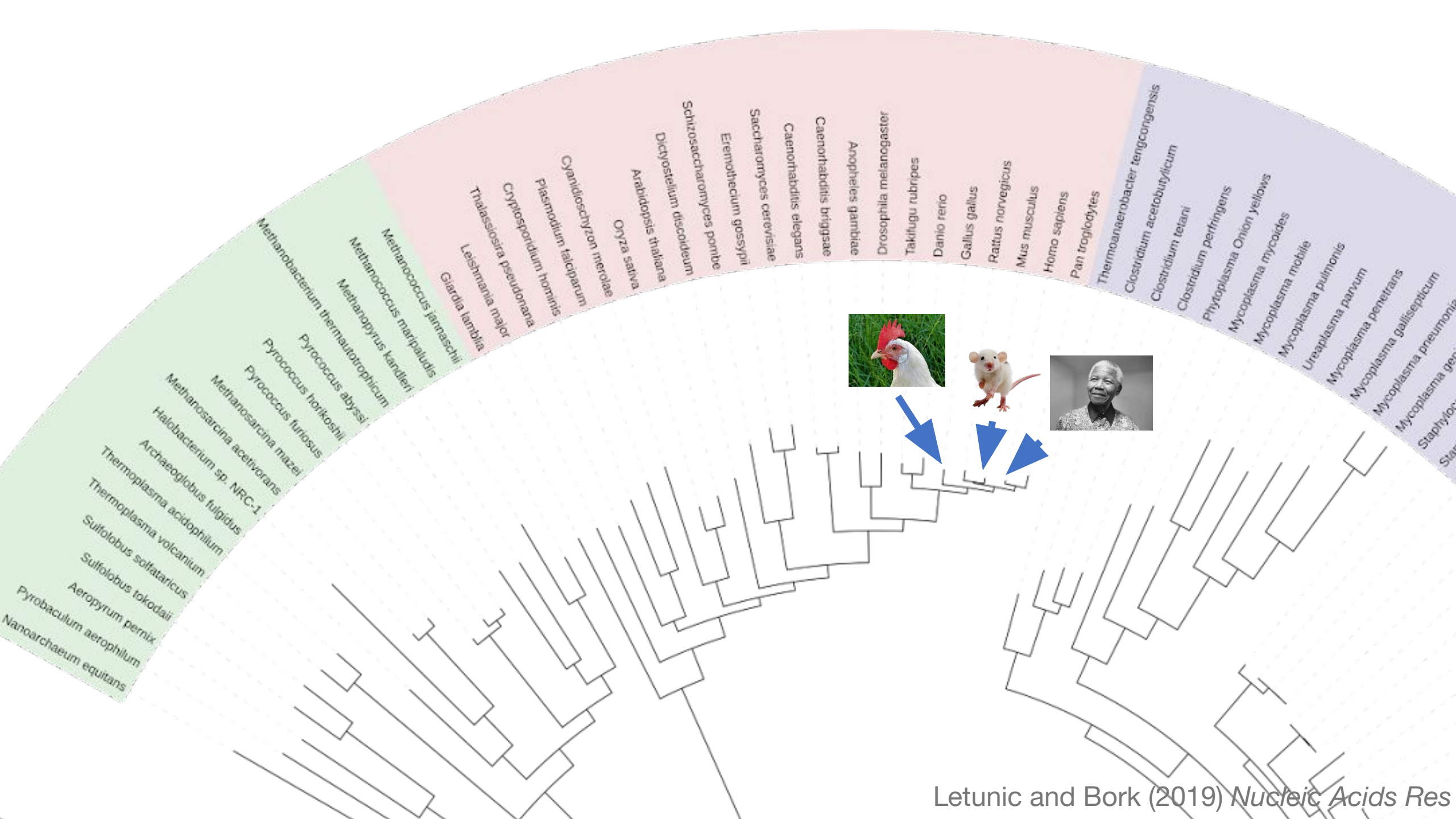
NATIONAL GEOGRAPHIC
Grant CP-044R-17



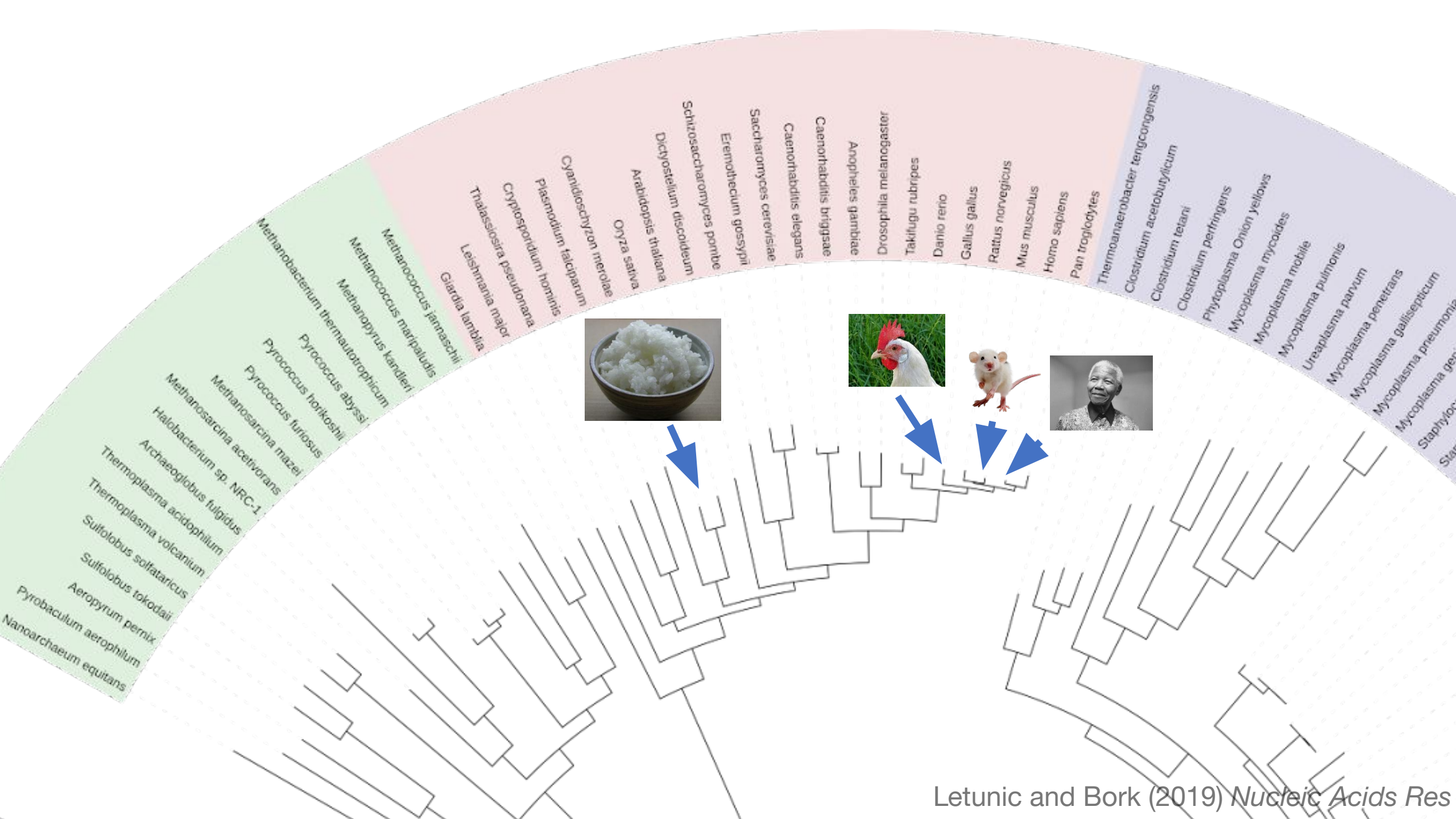
Letunic and Bork (2019) *Nucleic Acids Res*



Letunic and Bork (2019) *Nucleic Acids Res*



Letunic and Bork (2019) *Nucleic Acids Res*

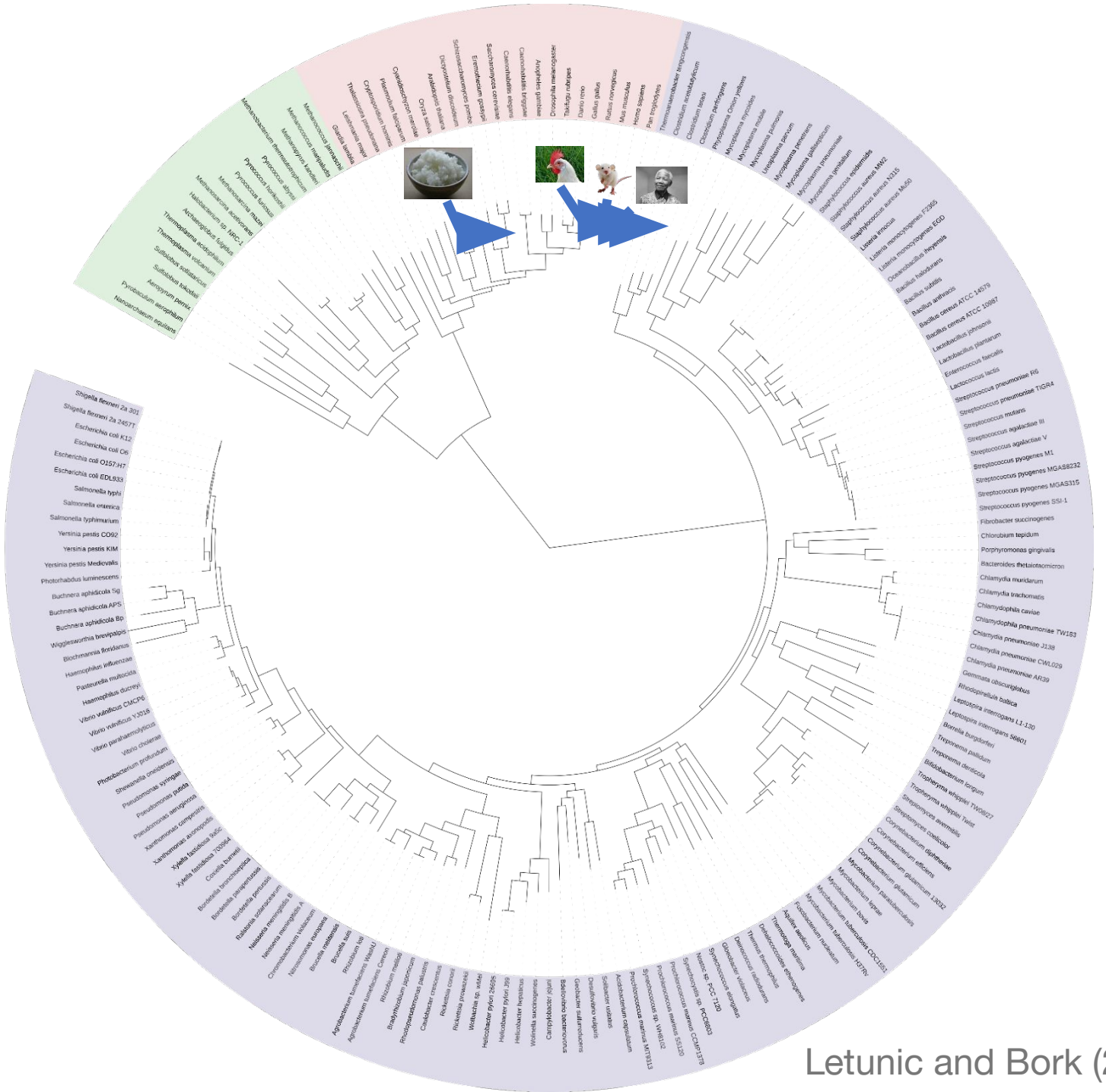


Letunic and Bork (2019) *Nucleic Acids Res*

Tree scale: 1

Colored ranges

- Eukaryota
- Archaea
- Bacteria



Letunic and Bork (2019) *Nucleic Acids Res*

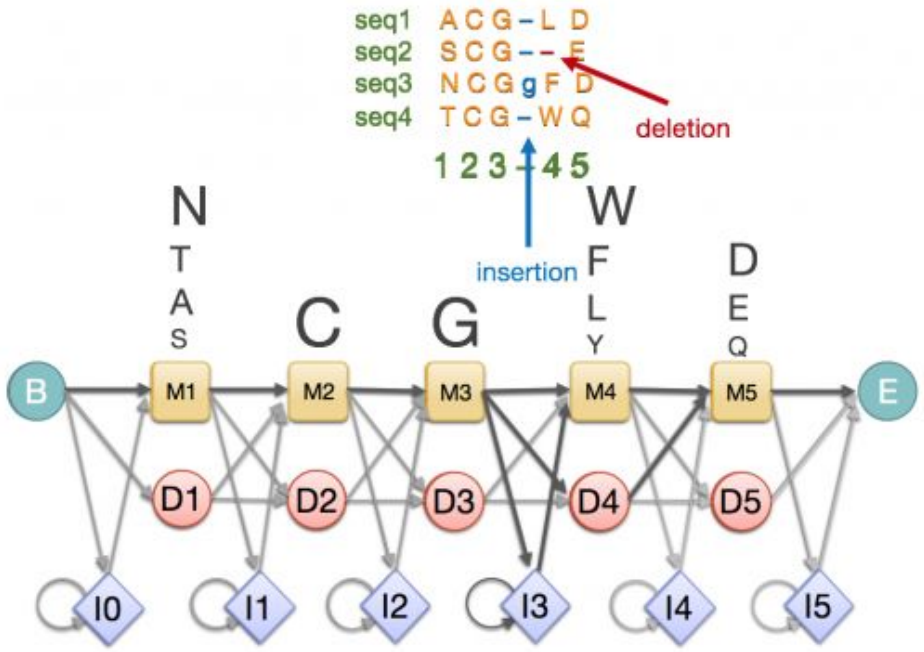
pHMM models are a popular tool for classifying sequences

Start with a multiple sequence alignment

Insertions / deletions can be modelled

Occupancy and amino acid frequency at each position in the alignment are encoded

Profile created



pHMM models are a popular tool for classifying sequences

Input: Query Sequence Set

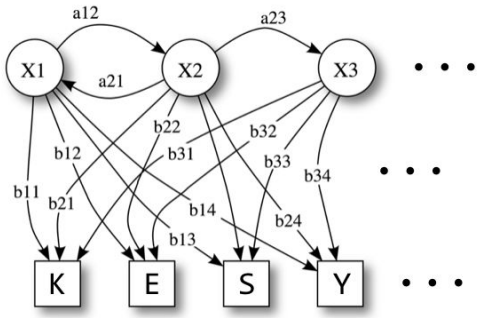
```
...SKEAEYLVKQLNTVME...
...SKEAKYLIQQLDTVMK...
...SKERYAAISMFMK...
...AKEGEYLYSNMLNAVMK...
```

?

Multiple Alignment

```
...SKEAEYLVK-QLNTVME...
...SKEAKYLIQ-QLDTVMK...
...SKERYAA----ISMFMK...
...AKEGEYLYSNMLNAVMK...
```

hmmbuild



HMM Profile

Input: Target Sequence Set

```
...CMSDKPDLSEVETFDKSKLTIQQEKEYNQRS...
...SCALEEHVSKEAEYLVKMLNAVMKVTGSFDP...
...DRSQNPPQSKGCCFVTFYTRKAALEAQNALH...
...KMPKDKERSLNPAAAQRKLDKQKSLKKGKAE...
...
```

hmmsearch

SKEAEYLVKMLNAVMKV

Output: Resulting Match

A Primer for Computational Biology by Shawn T. O'Neil

Artificial Intelligence Approaches to Natural Product Drug Discovery

Nature Reviews Drug Discovery - submitted

(58 authors)

The Threat of Antibiotic Resistance in the United States

Antibiotic resistance—when germs (bacteria, fungi) develop the ability to defeat the antibiotics designed to kill them—is one of the greatest global health challenges of modern time.

New National Estimate*

Each year, antibiotic-resistant bacteria and fungi cause at least an estimated:

2,868,700 infections

35,900 deaths

+

***Clostridioides difficile* is related to antibiotic use and antibiotic resistance:**

223,900 cases

12,800 deaths

New Antibiotic Resistance Threats List

Updated urgent, serious, and concerning threats—totaling 18

5 urgent threats

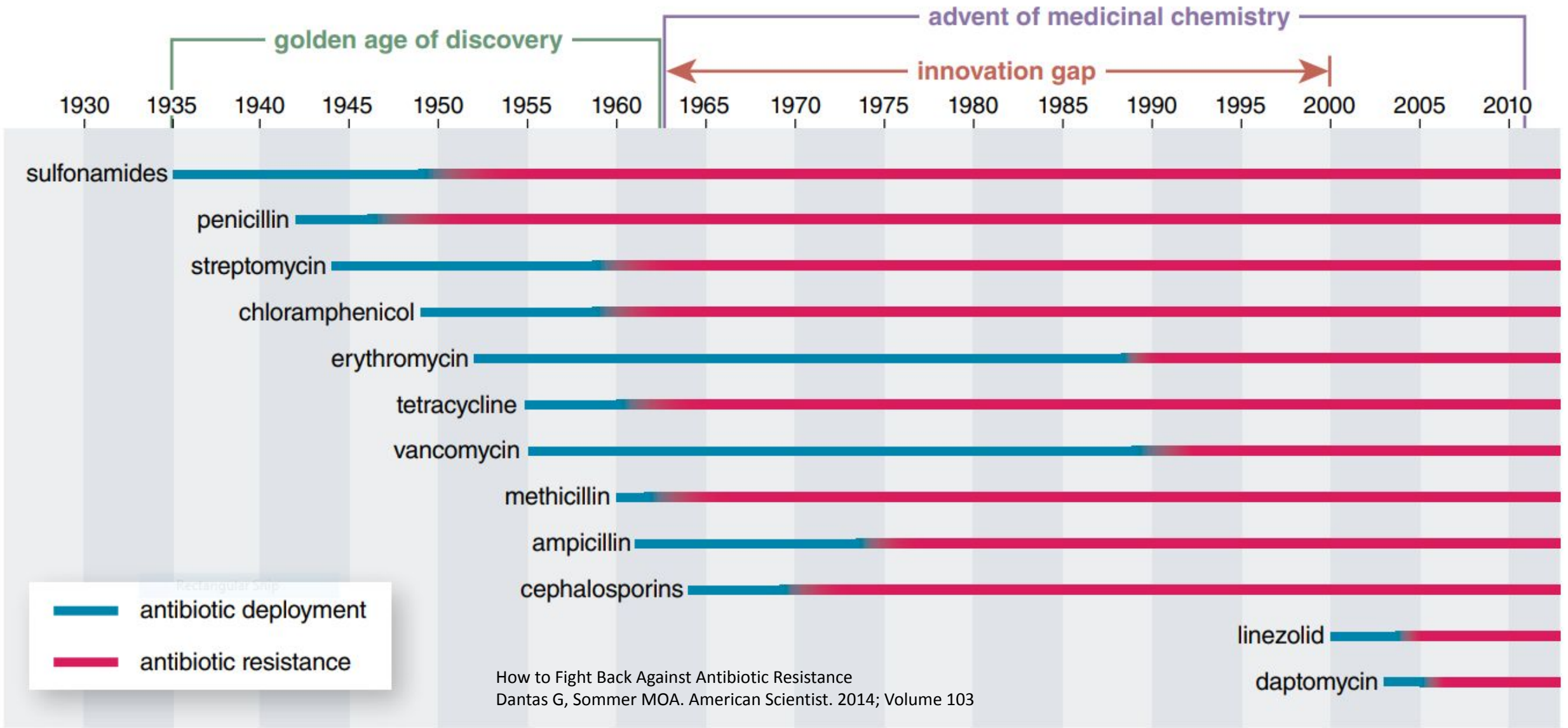
2 new threats

NEW: Watch List with **3** threats

Antibiotic resistance remains a significant One Health problem, affecting humans, animals, and the environment. Data show infection prevention and control is saving lives—especially in hospitals—but threats may undermine this progress without continued aggressive action now.

Learn more: www.cdc.gov/DrugResistance/Biggest-Threats

*National burden reflects de-duplicated infection and death estimates.



How to Fight Back Against Antibiotic Resistance
Dantas G, Sommer MOA. American Scientist. 2014; Volume 103

Combinatorial Chemistry vs Natural Products

70 HTS campaigns
3 million compounds
19 Hits
5 Leads

65 HTS campaigns
2 million compounds
57 Hits
19 Leads

3 million compounds

___ Gram-negative antibiotics with cellular activity

Nat Rev Drug Discov **2007**, *6*, 29
Nat Rev Drug Discov **2015**, *14*, 529
J Med Chem **2014**, *57*, 10144

Slide from Professor Paul J. Hergenrother, UIUC, adapted with permission

Combinatorial Chemistry vs Natural Products

70 HTS campaigns
3 million compounds
19 Hits
5 Leads

65 HTS campaigns
2 million compounds
57 Hits
19 Leads

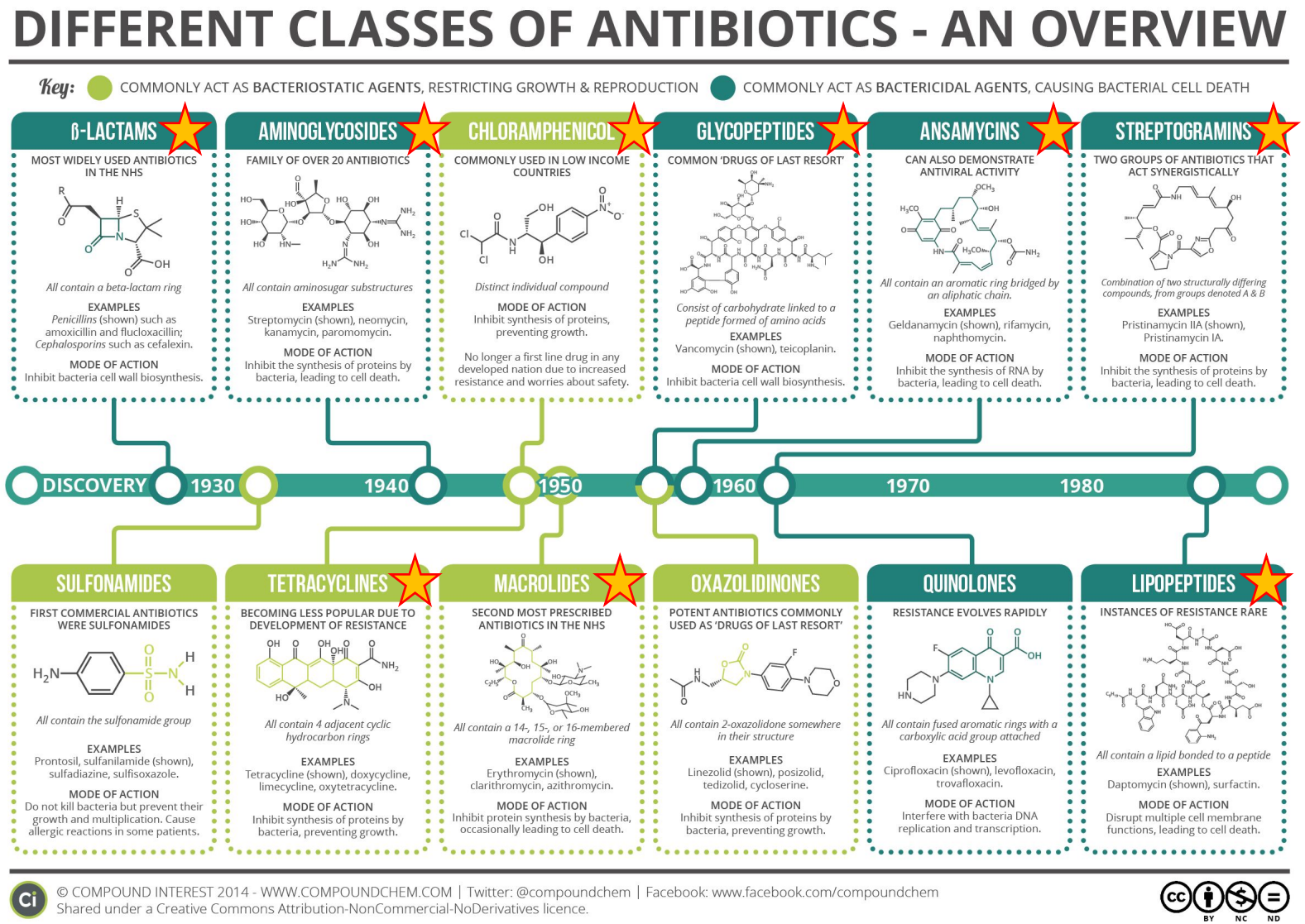
3 million compounds

0 Gram-negative antibiotics with cellular activity

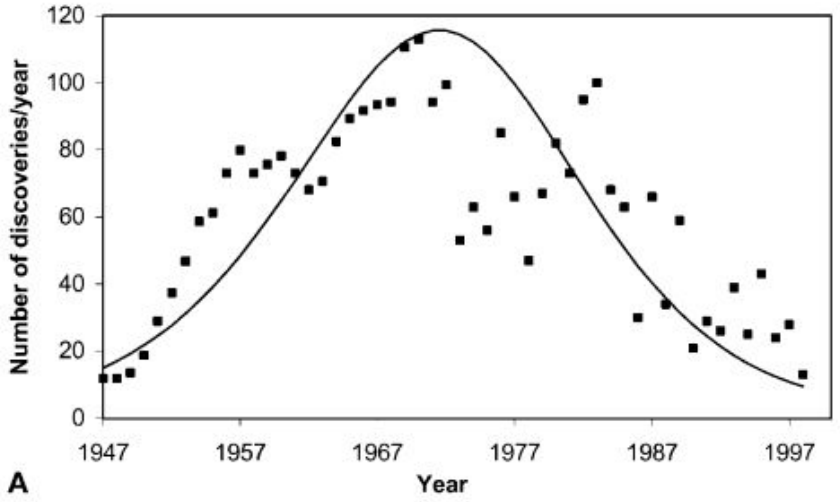
Nat Rev Drug Discov **2007**, *6*, 29
Nat Rev Drug Discov **2015**, *14*, 529
J Med Chem **2014**, *57*, 10144

Slide from Professor Paul J. Hergenrother, UIUC, adapted with permission

Bacteria (and fungi) as sources of antibiotics

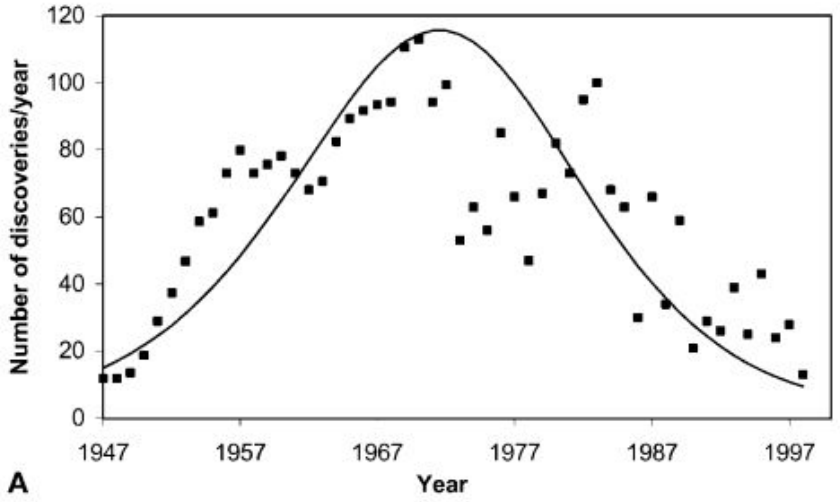


Are there more compounds to find in *Streptomyces?*



Watve MG, Tickoo R, Jog MM, Bhole BD. How many antibiotics are produced by the genus Streptomyces? Archives of Microbiology. 2001. p. 386–390.

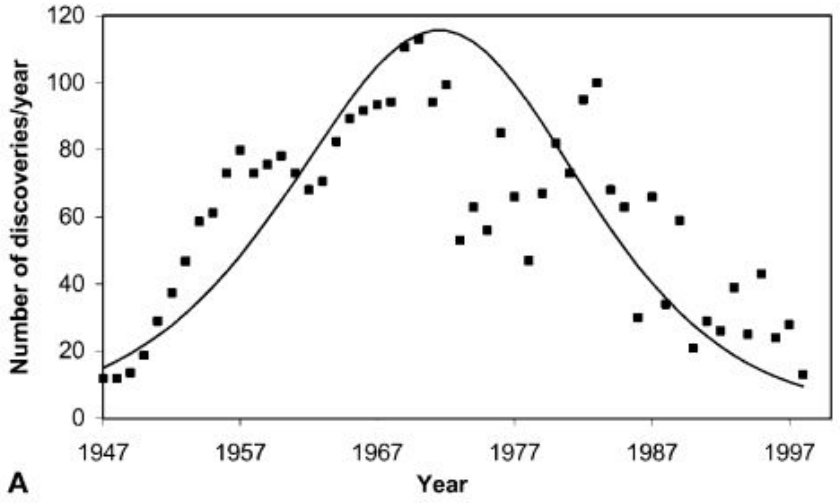
Are there more compounds to find in *Streptomyces?*



"…even if we accept the more conservative estimate, only about 3% of the existing compounds have been reported so far."

Watve MG, Tickoo R, Jog MM, Bhole BD. How many antibiotics are produced by the genus Streptomyces? Archives of Microbiology. 2001. p. 386–390.

Are there more compounds to find in *Streptomyces?*



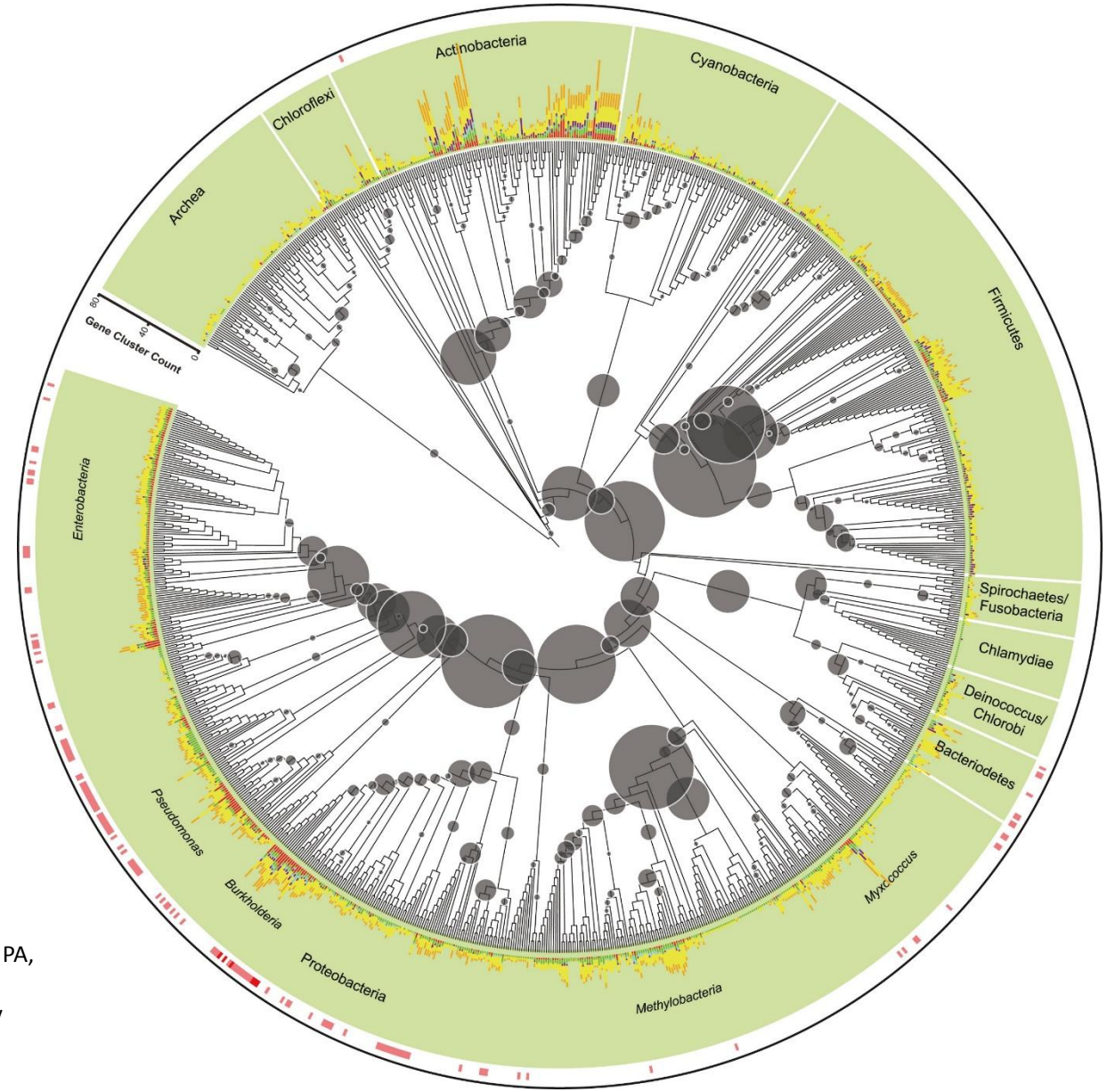
"…even if we accept the more conservative estimate, only about 3% of the existing compounds have been reported so far."

Problems:

1. **Rediscovery of commonly occurring compounds**
2. **Only tested in limited assays**

Watve MG, Tickoo R, Jog MM, Bhole BD. How many antibiotics are produced by the genus Streptomyces? Archives of Microbiology. 2001. p. 386–390.

But… natural products aren't just found in *Streptomyces*

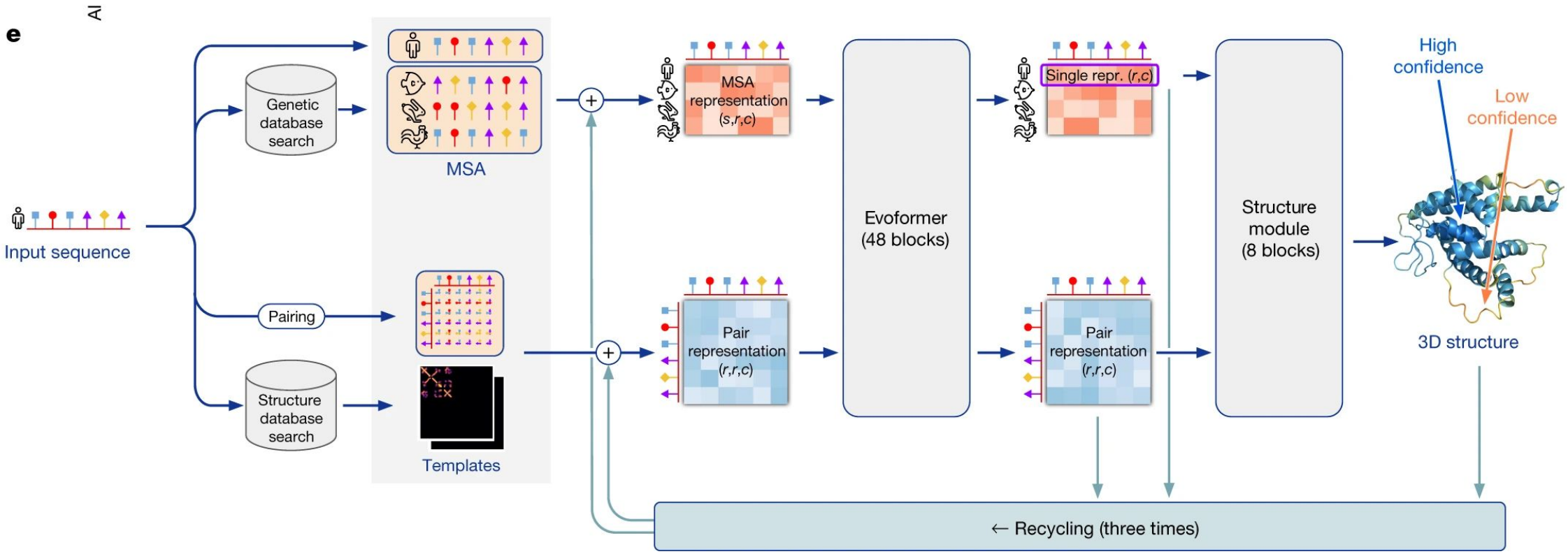


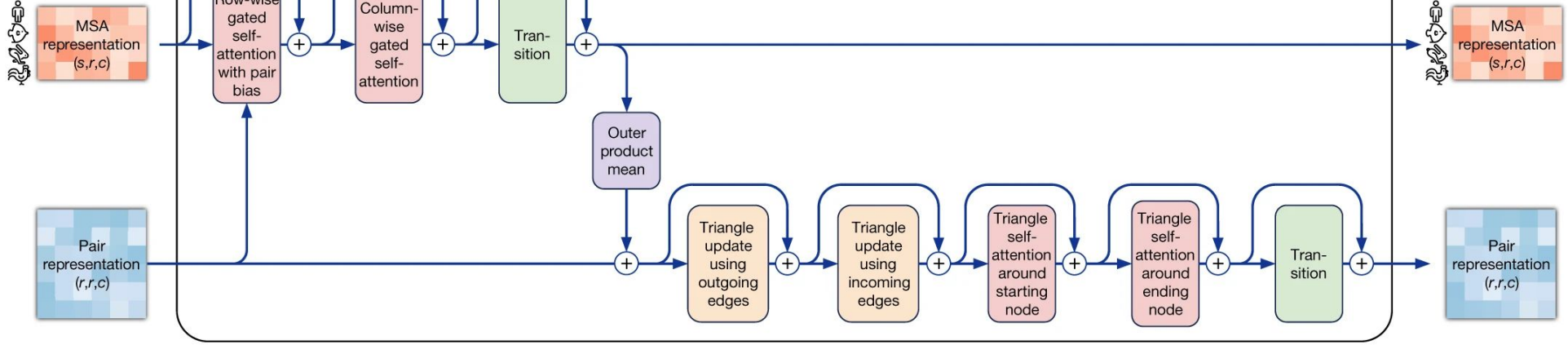
Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. Cell Press; 2014 Jul 17;158(2):412–421.

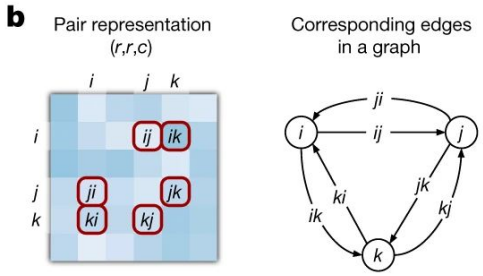
To think about…

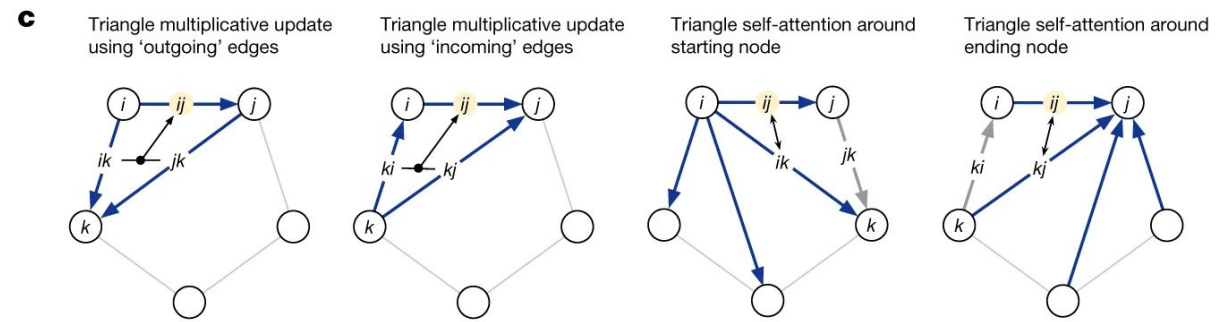
- Where should we look for new bacterial chemistry?
  - New species of bacteria? New phyla?
  - Same species, different location/host organism?
- How to access enough of the chemical compounds?
  - From uncultured organisms?
  - Total synthesis can be slow isn’t always an answer

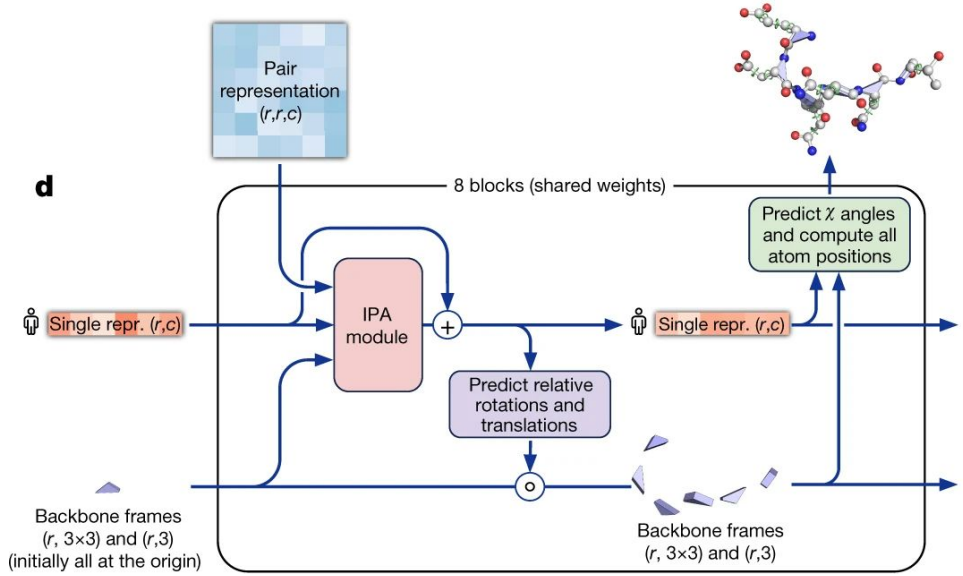
Our methods are scalable to very long pr
domains and domain-packing (see Fig. 1c
2,180-residue protein with no structural h

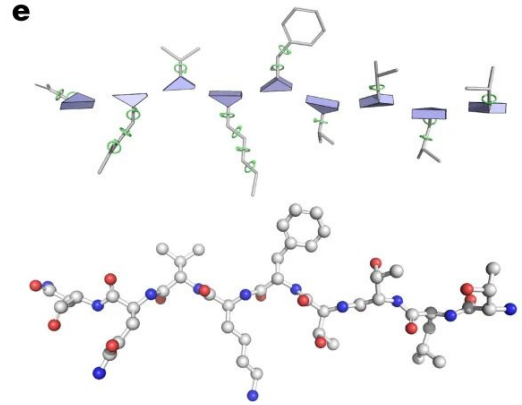


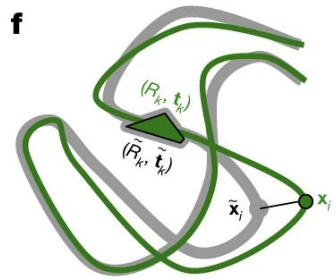

MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
MSA representation (s,r,c)
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
Pair representation (r,r,c)


b
Pair representation (r,r,c)
Corresponding edges in a graph


c
Triangle multiplicative update using 'outgoing' edges
Triangle multiplicative update using 'incoming' edges
Triangle self-attention around starting node
Triangle self-attention around ending node


d
Pair representation (r,r,c)
8 blocks (shared weights)
Predict χ angles and compute all atom positions
IPA module
Single repr. (r,c)
Single repr. (r,c)
Predict relative rotations and translations
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)


e


f