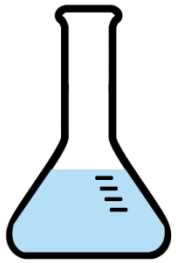
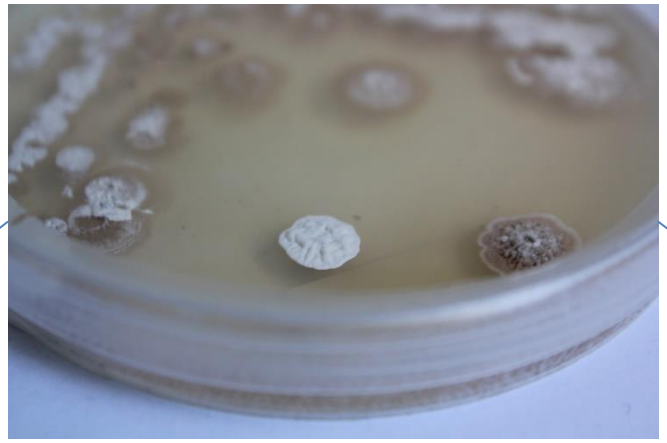
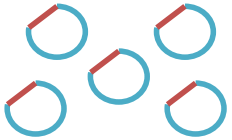
SEQUENCING AND ANALYZING RARE MARINE ACTINOMYCETE GENOMES FOR THEIR NATURAL PRODUCT POTENTIAL

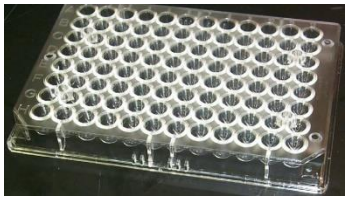
Michelle Schorn
PhD Candidate, 4th year, Moore Lab
SIO BUG Meeting
4/11/16

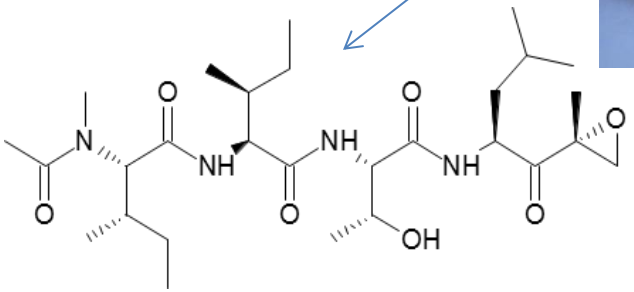
Microbial Natural Products: Connecting Genes and Molecules

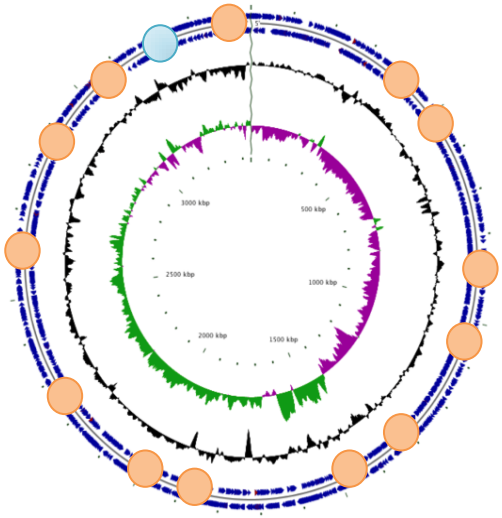






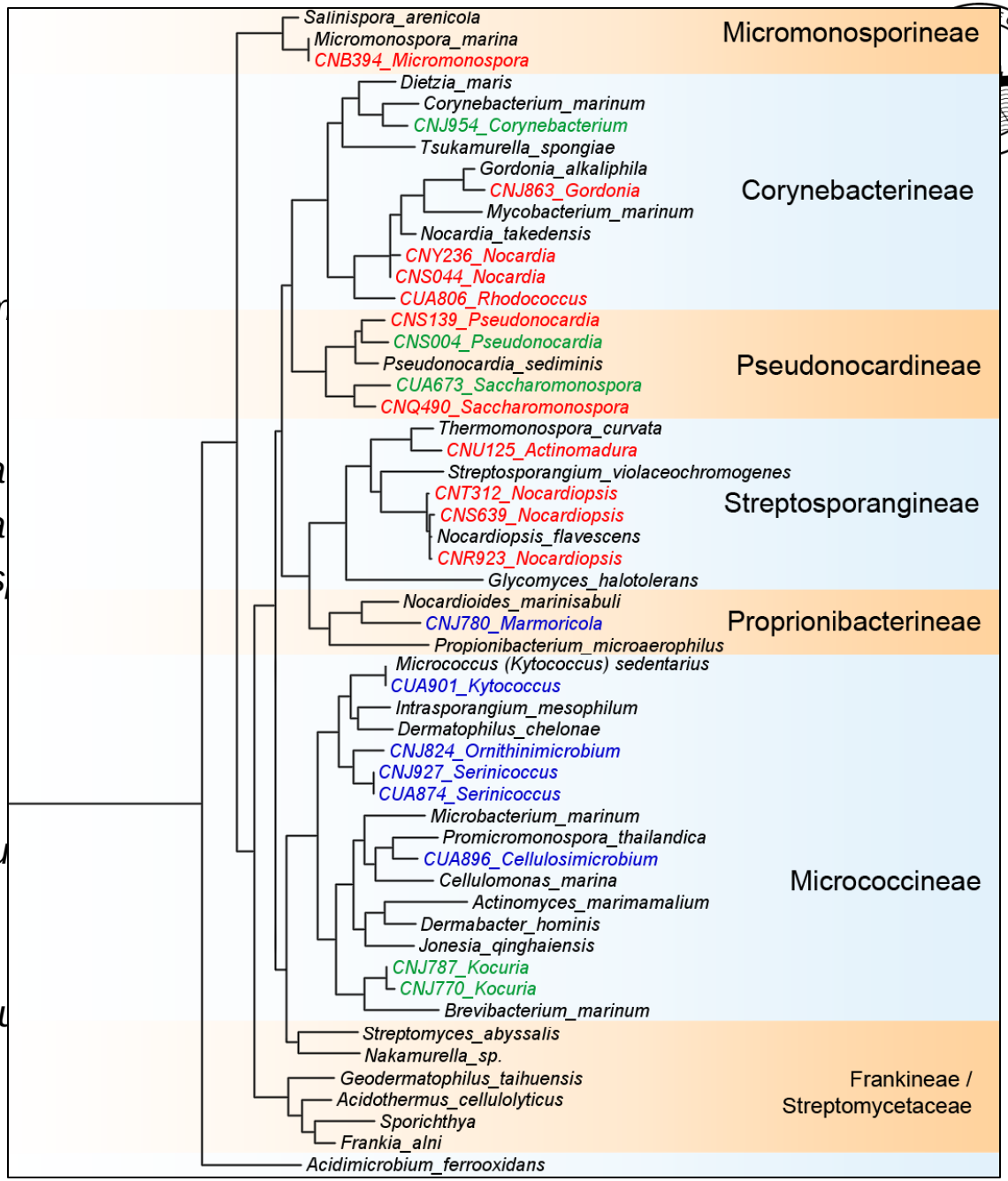






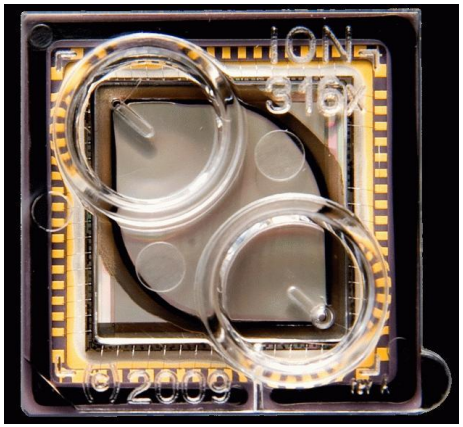
Rare Strains

- Strains I sequenced:
  - CNJ954 *Corynebacterium*
  - CNJ863 *Gordonia*
  - CUA806 *Rhodococcus*
  - CNS139 *Pseudonocardia*
  - CNS004 *Pseudonocardia*
  - CUA673 *Saccharomonos*
  - CNU125 *Actinomadura*
  - CNR923 *Nocardiopsis*
  - CNJ780 *Marmoricola*
  - CUA901 *Kytococcus*
  - CNJ824 *Ornithinimicrobiu*
  - CNJ927 *Serinicoccus*
  - CUA874 *Serinicoccus*
  - CUA896 *Cellulosimicrobiu*
  - CNJ770 *Kocuria*



Ion Torrent Sequencing

Ion Torrent PGM
400bp libraries



318v2 semiconductor sequencing chip
1 chip per genome
1-1.8 Gb data per run

Genome Assembly

- 4 complete clusters
- From 2/15 strains

- 51 complete clusters
- From 12/15 strains

SPAdes Assembly

- SPAdes Genome Assembler
- Developed for single cell genomics, but works well for uneven coverage seen in in high GC genome sequencing

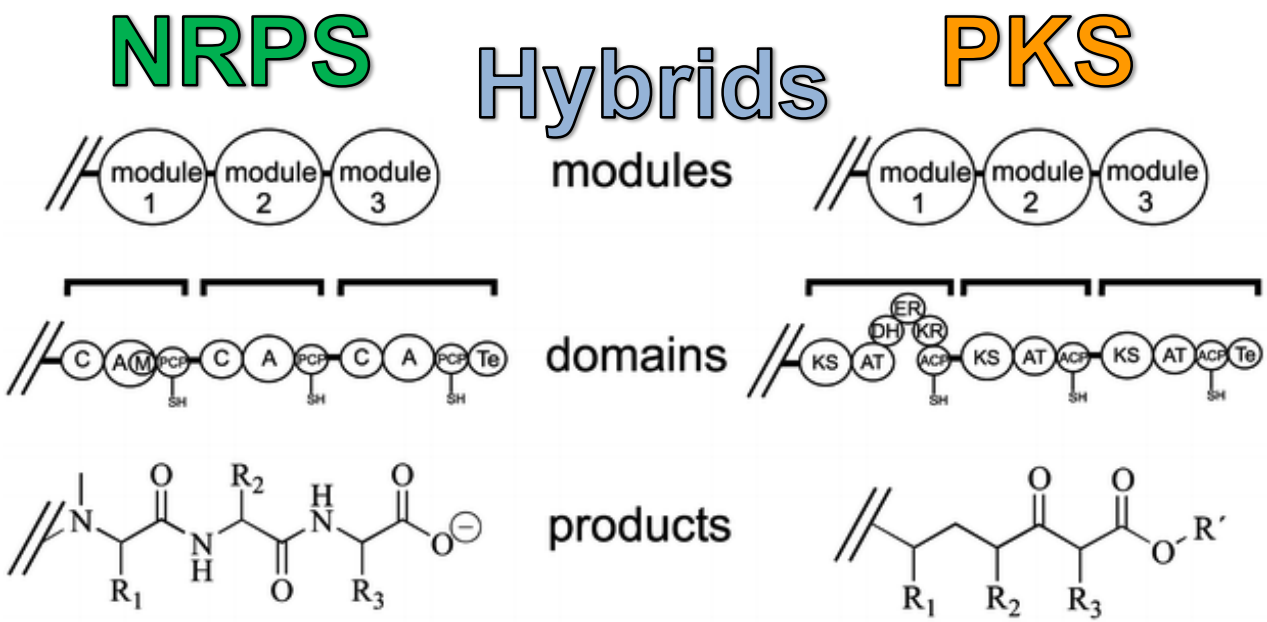
Strain	Genus	Contigs >1 kb	Max contig	Avg contig	GC%	Size (Mb)
CNJ780	*Marmoricola*	36	453,291	13,303	73.1%	4.39
CUA874	*Serinicoccus*	39	464,901	15,770	72.0%	3.56
CNJ954	*Corynebacterium*	45	433,327	83,858	65.2%	3.78
CNJ927	*Serinicoccus*	52	409,679	66,116	72.1%	3.48
CNJ824	*Ornithinimicrobium*	62	450,180	11,267	72.9%	3.50
CUA806	*Rhodococcus*	67	424,519	86,550	63.9%	6.33
CUA896	*Cellulosimicrobium*	79	73,310	10,472	74.9%	5.35
CUA673	*Saccharomonospora*	85	283,169	63,777	70.0%	5.46
CNJ863	*Gordonia*	94	499,033	57,433	67.3%	5.47
CNR923	*Nocardiopsis*	153	230,239	36,276	71.0%	5.66
CNS004	*Pseudonocardia*	156	550,412	58,994	72.6%	9.26
CNJ770	*Kocuria*	166	225,243	24,837	71.9%	4.73
CNS139	*Pseudonocardia*	250	224045	28,501	74.2%	7.34
CUA901	*Kytococcus*	461	301,417	6,995	71.0%	3.61
CNU125	*Actinomadura*	603	114,639	14,065	72.6%	10.44

Bankevich A, *et al.* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 19(5):455-477.

Genome Annotation & Curation

antiSMASH

antibiotics & Secondary Metabolite Analysis SHell



Terpene

RiPPs

Weber T, *et al.* (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 43(W1):W237-243.

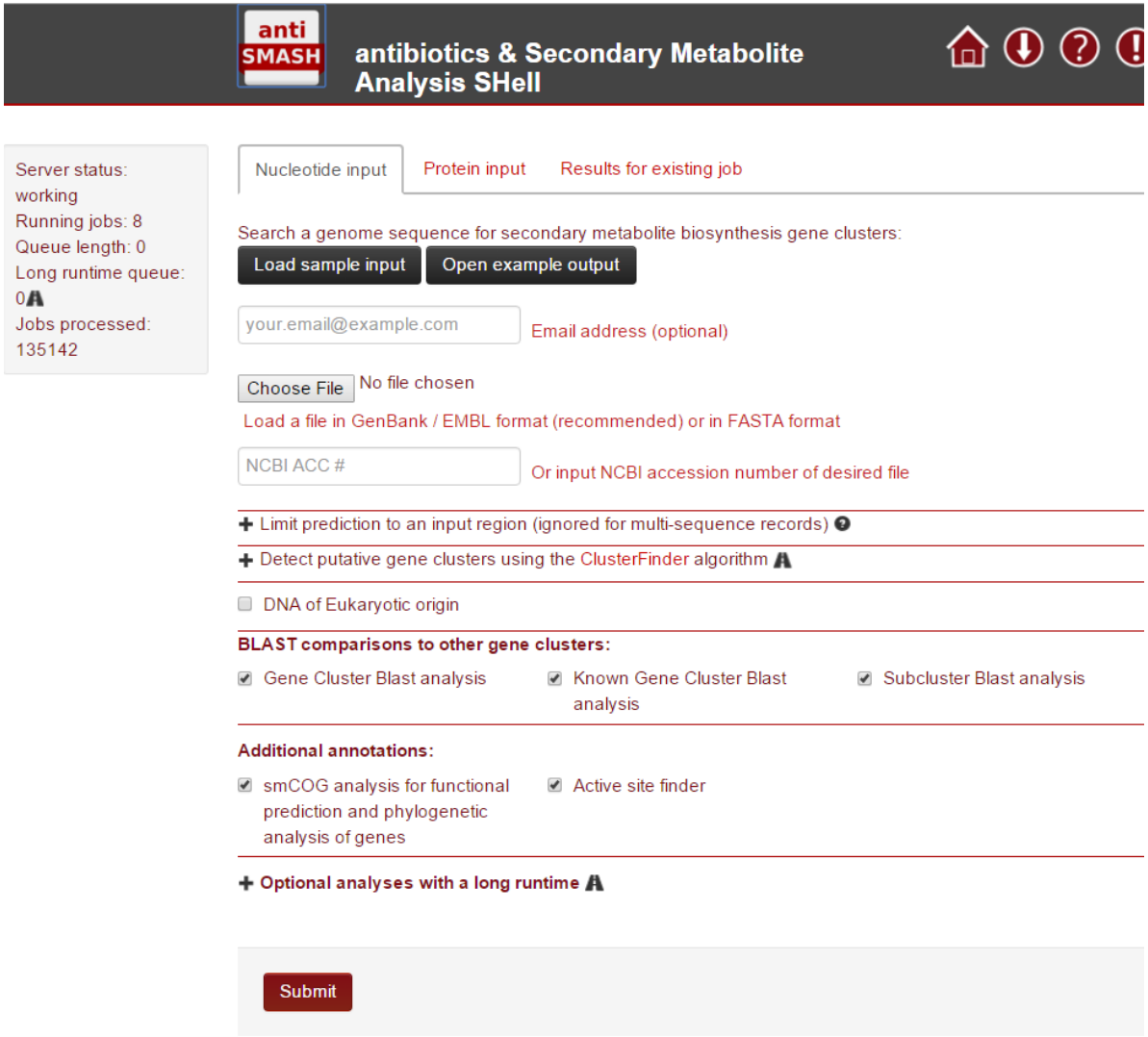
Bioinformatic Tools- antiSMASH

- antibiotics and Secondary Metabolites Analysis Shell (antiSMASH) allows the rapid genome-wide identification, annotation and analysis of secondary metabolite biosynthesis gene clusters

- Submit your genome to the antiSMASH website, and it will **locate and annotate predictable secondary metabolite gene clusters**, such as NRPS, PKS, Terpene, Lantipeptide, Bacteriocin, Siderophore, etc.
- Can make some rudimentary structural predictions

Kai Blin *et al*., (2013). Nucleic Acids Research doi: 10.1093/nar/gkt449.
Marnix H. Medema *et al*., (2011) Nucleic Acids Research 39: W339-W346.

Bioinformatic Tools- antiSMASH



Kai Blin *et al.*, (2013). Nucleic Acids Research doi: 10.1093/nar/gkt449.
Marnix H. Medema *et al.*, (2011) Nucleic Acids Research 39: W339-W346.

Bioinformatic Tools- antiSMASH

Select Gene Cluster:

Overview 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35

Identified secondary metabolite clusters

Cluster	Type	From	To
The following clusters are from record T333DRAFT_sc...1:			
Cluster 1	Nrps	1	30644
The following clusters are from record T333DRAFT_sc...3:			
Cluster 2	Bacteriocin	1	1593
The following clusters are from record T333DRAFT_sc...1:			
Cluster 3	T1pks	294194	340610
Cluster 4	T1pks	574473	619755
Cluster 5	Other	833699	877637
Cluster 6	Nrps	1110887	1156243
Cluster 7	Nrps-terpene	1209176	1307215
Cluster 8	Nrps	1298496	1348148
Cluster 9	Nrps	1666701	1720560
The following clusters are from record T333DRAFT_sc...2:			
Cluster 10	T1pks	171145	257190
Cluster 11	T1pks	330021	374406
Cluster 12	Bacteriocin	977673	988473
Cluster 13	Nrps	985809	1062894
Cluster 14	Nrps	1044655	1150138
The following clusters are from record T333DRAFT_sc...3:			
Cluster 15	Terpene	105733	126833
Cluster 16	Butyrolactone	545848	556834
Cluster 17	Nrps	707717	754932
Cluster 18	Nrps	767318	823661
Cluster 19	T3pks	806727	847872
Cluster 20	Nrps	882251	936926
The following clusters are from record T333DRAFT_sc...4:			
Cluster 21	Other	11792	55280
Cluster 22	Nrps-t1pks	108302	199912
Cluster 23	T1pks-transatpks	591160	664596
The following clusters are from record T333DRAFT_sc...5:			
Cluster 24	Terpene-t1pks-butyrolactone	169594	219884
Cluster 25	Butyrolactone	355409	366404
Cluster 26	Nrps	531146	588281
The following clusters are from record T333DRAFT_sc...6:			
Cluster 27	Ectoine	52096	62491
Cluster 28	T1pks	308081	366308
The following clusters are from record T333DRAFT_sc...7:			
Cluster 29	Nrps	57764	111416
Cluster 30	Nrps	113544	177981
Cluster 31	Nrps	200011	258187
Cluster 32	T3pks	276355	317524
Cluster 33	T1pks	343668	390102
Cluster 34	Nrps	407275	455971
The following clusters are from record T333DRAFT_sc...9:			
Cluster 35	Terpene	12202	33848

Bioinformatic Tools- antiSMASH

Gene cluster description

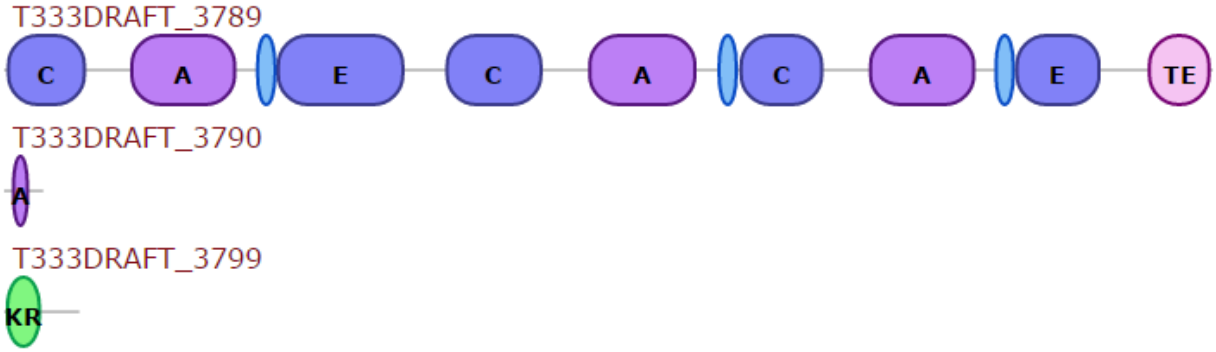
Gene Cluster 20. Type = nrps. Location: 882251 - 936926 nt. Click on genes for more information. Download cluster GenBank file

Show pHMM detection rules used

Legend:

biosynthetic genes | transport-related genes | regulatory genes | other genes

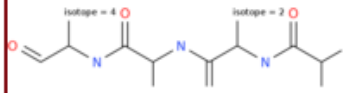
Detailed annotation



Homologous gene clusters

All hits | Download graphic

Predicted core structure



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

Prediction details

Monomers prediction:
(nrp) + (nrp-nrp-nrp)

T333DRAFT_3789
- NRPSPredictor2 SVM: N/A
- Stachelhaus code: orn
- Minowa: trp
- consensus: nrp

- NRPSPredictor2 SVM: gly,ala
- Stachelhaus code: thr
- Minowa: trp
- consensus: nrp

- NRPSPredictor2 SVM: gly,ala
- Stachelhaus code: thr
- Minowa: phe
- consensus: nrp

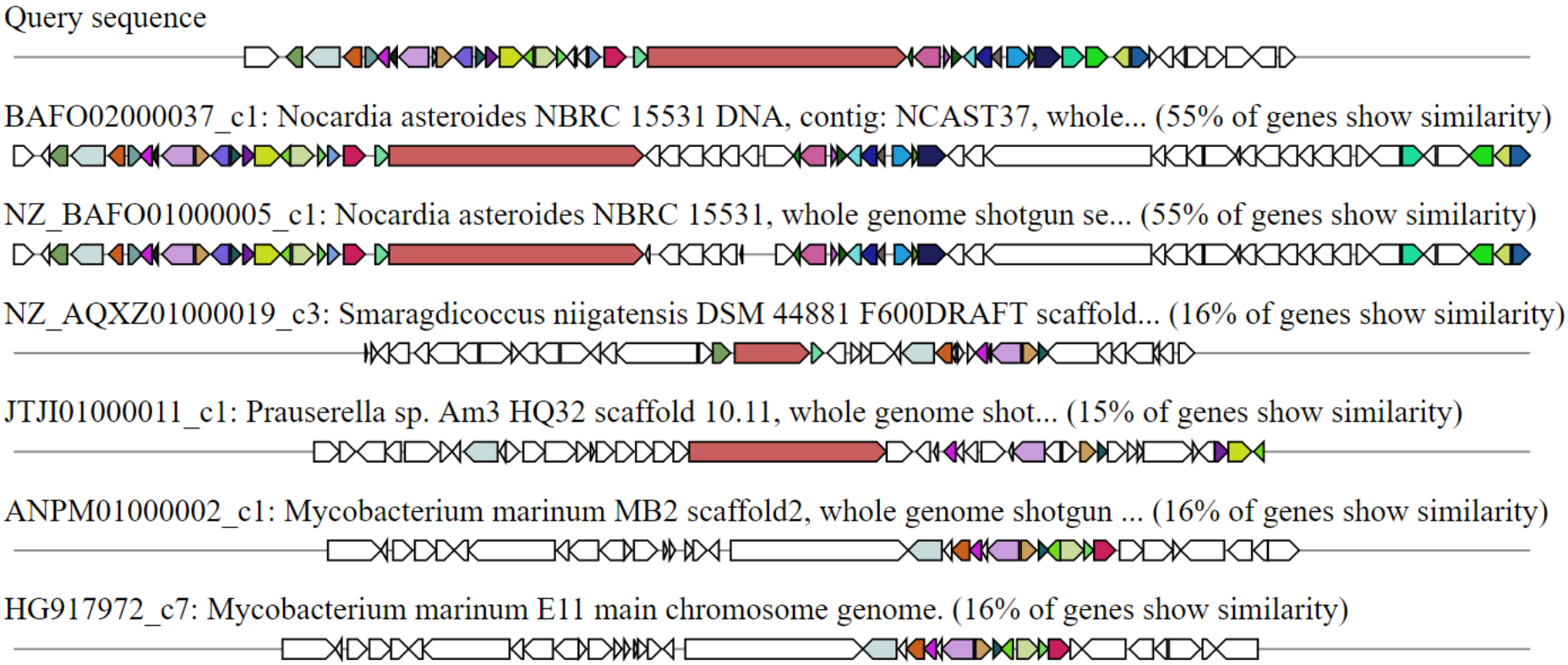
T333DRAFT_3790
- NRPSPredictor2 SVM: N/A
- Stachelhaus code: N/A
- Minowa: pip
- consensus: nrp

Database cross-links

Look up in NORINE database

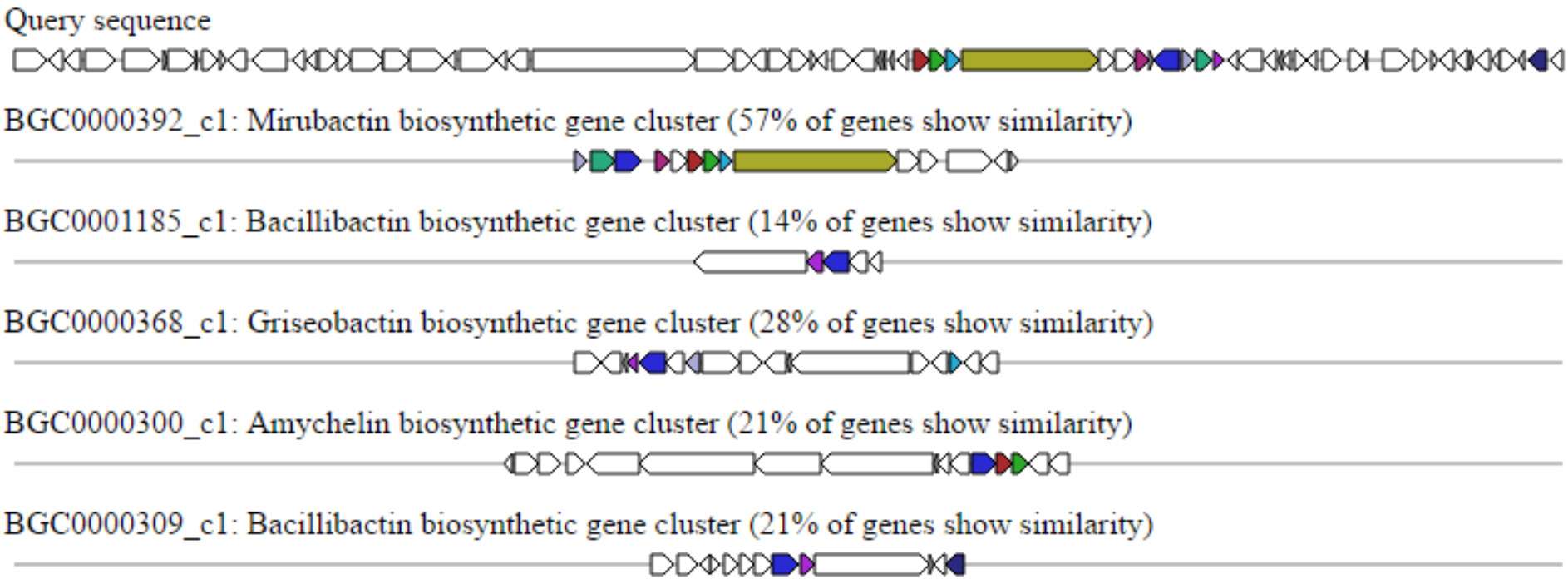
Bioinformatic Tools- antiSMASH

- MultiGeneBlast against all genomes



Bioinformatic Tools- antiSMASH

- MultiGeneBlast against known gene clusters



Genome Annotation & Curation

antiSMASH

antibiotics & Secondary Metabolite Analysis SHell

NaPDoS

Natural Product Domain Seeker

JGI

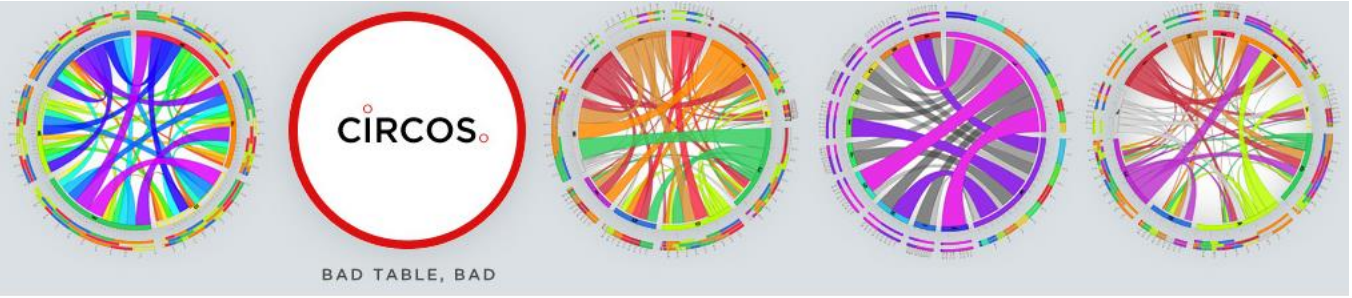
JOINT GENOME INSTITUTE

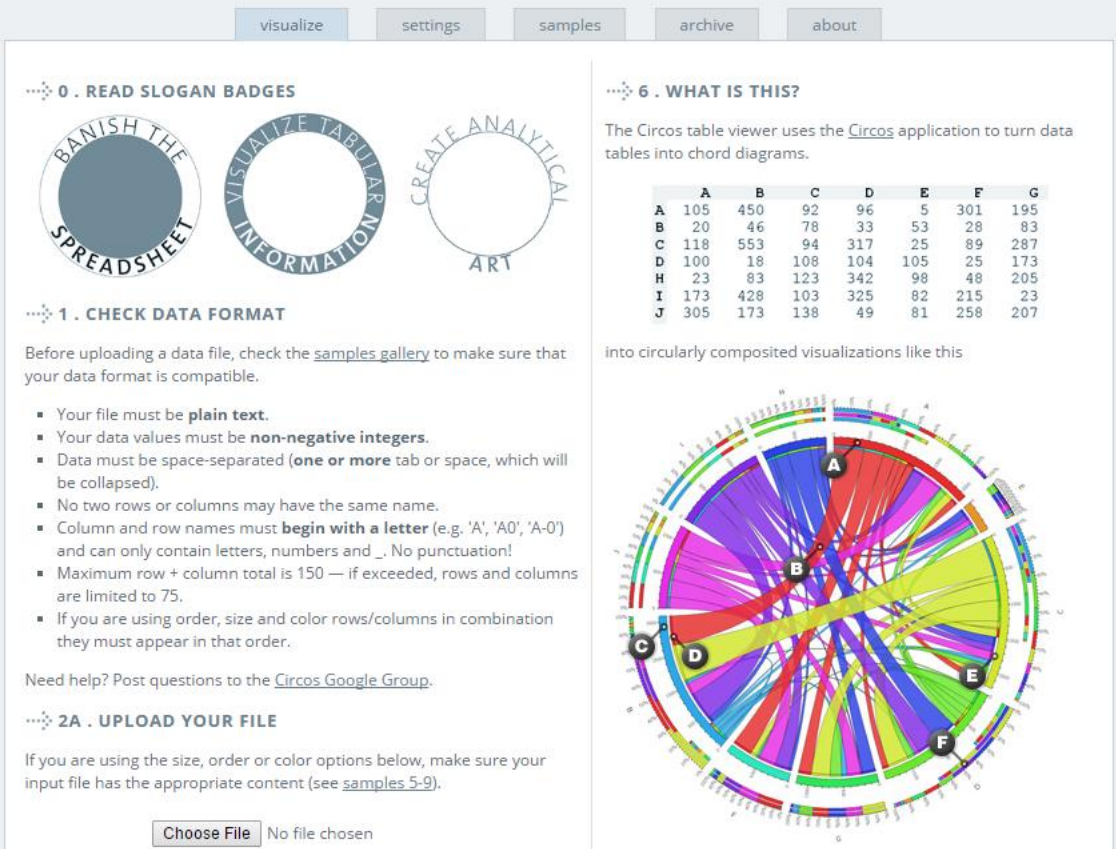
DEPARTMENT OF ENERGY

img/er

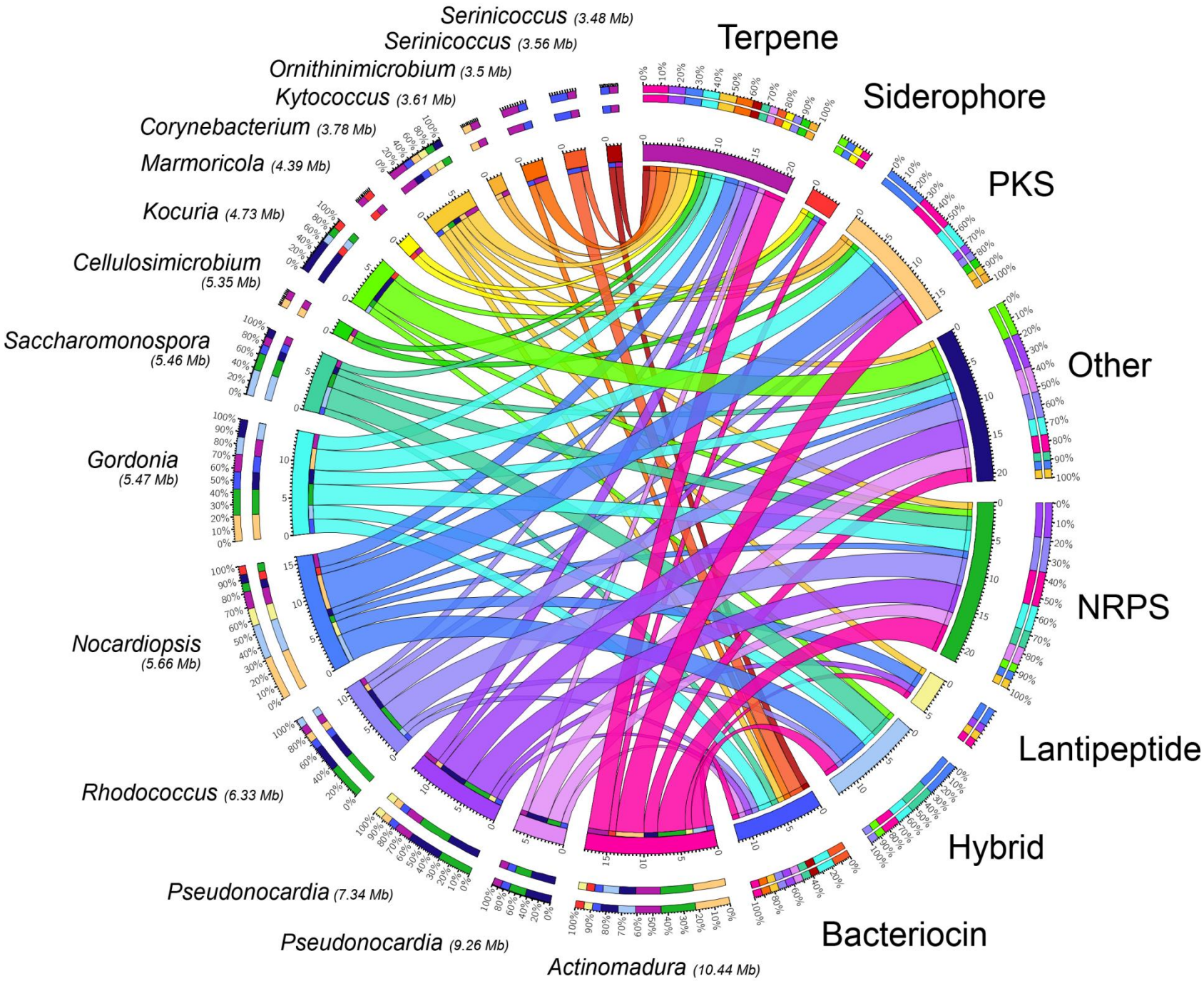
CIRCOS

Bioinformatic Tools- Circos





- Circos.ca → download, tutorials or click “Circos Online” to use online interface
- mkweb.bcgsc.ca/tableviewer/
- Can change colors in online interface, just have to work it into your table


Serinicoccus (3.48 Mb)
Serinicoccus (3.56 Mb)
Ornithinimicrobium (3.5 Mb)
Kytococcus (3.61 Mb)
Corynebacterium (3.78 Mb)
Marmoricola (4.39 Mb)
Kocuria (4.73 Mb)
Cellulosimicrobium (5.35 Mb)
Saccharomonospora (5.46 Mb)
Gordonia (5.47 Mb)
Nocardiopsis (5.66 Mb)
Rhodococcus (6.33 Mb)
Pseudonocardia (7.34 Mb)
Pseudonocardia (9.26 Mb)
Actinomadura (10.44 Mb)
Terpene
Siderophore
PKS
Other
NRPS
Lantipeptide
Hybrid
Bacteriocin

Thank you!

<u>Moore lab</u>
Dr. Bradley Moore

<u>Ziemert lab</u>
Dr. Nadine Ziemert
Mohammad Alanjary

<u>Jensen lab</u>
Dr. Paul Jensen
Nastassia Patin

Dr. Bill Fenical

<u>Allen Lab</u>
Dr. Sheila Podell

Dr. Anton Korobeynikov

<u>Ion Torrent</u>
Dr. Tommie Lincecum
Kristen Aguinaldo

Funding: Edna Bailey Sussman Foundation, UNICO Foundation, NIH grants, NIH Marine Biotechnology Training grant, Teach@Tübingen Program