

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

Michelle Schorn

PhD Candidate, 4th year, Moore Lab

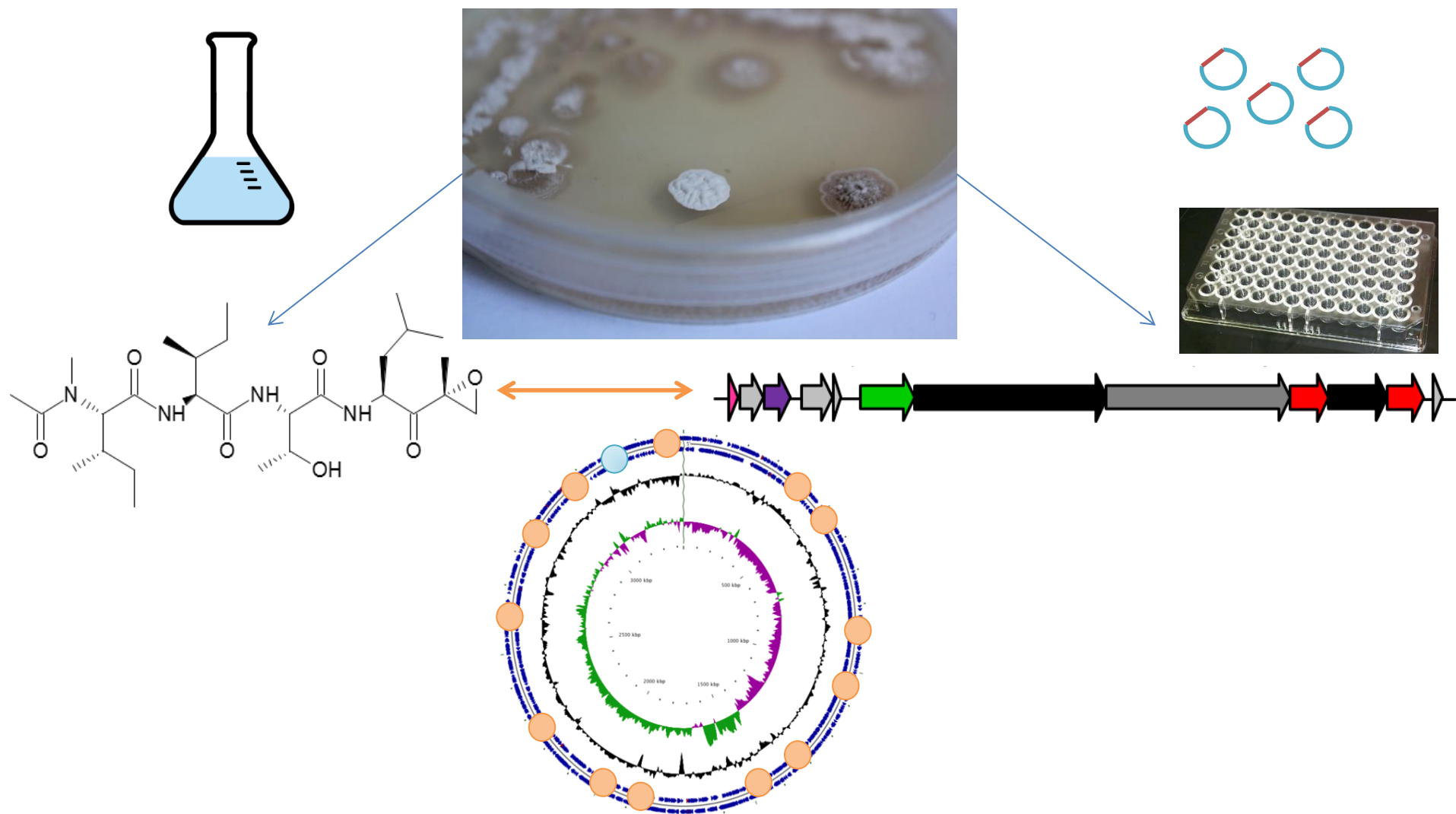
SIO BUG Meeting

4/11/16





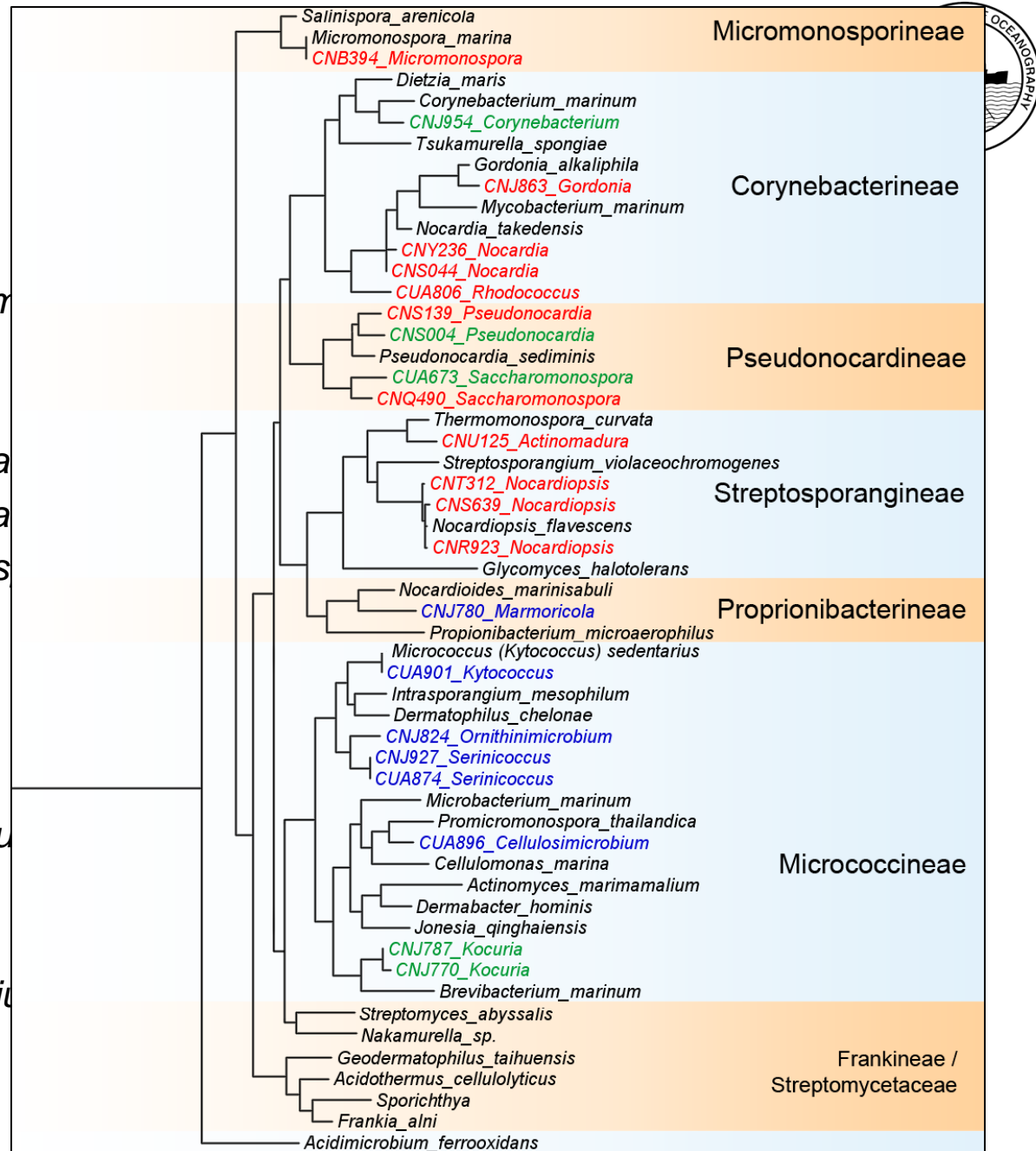
Microbial Natural Products: Connecting Genes and Molecules



Rare Strains

- Strains I sequenced:

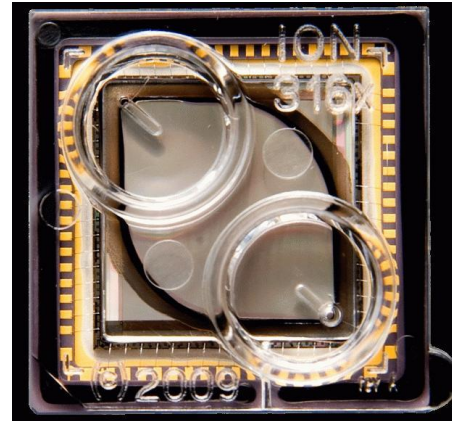
- CNJ954 *Corynebacterium*
- CNJ863 *Gordonia*
- CUA806 *Rhodococcus*
- CNS139 *Pseudonocardia*
- CNS004 *Pseudonocardia*
- CUA673 *Saccharomonos*
- CNU125 *Actinomadura*
- CNR923 *Nocardiopsis*
- CNJ780 *Marmoricola*
- CUA901 *Kytococcus*
- CNJ824 *Ornithinimicrobiu*
- CNJ927 *Serinicoccus*
- CUA874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



Ion Torrent Sequencing

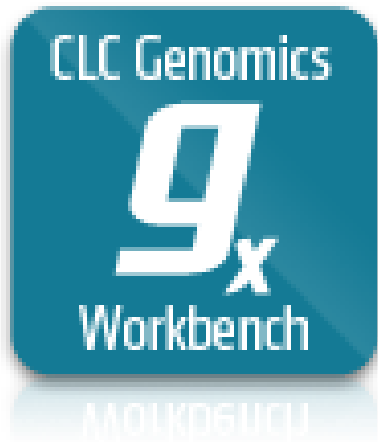
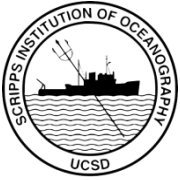


Ion Torrent PGM
400bp libraries



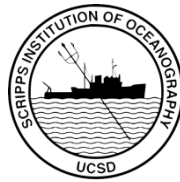
318v2 semiconductor sequencing chip
1 chip per genome
1-1.8 Gb data per run

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 high GC genome sequencing

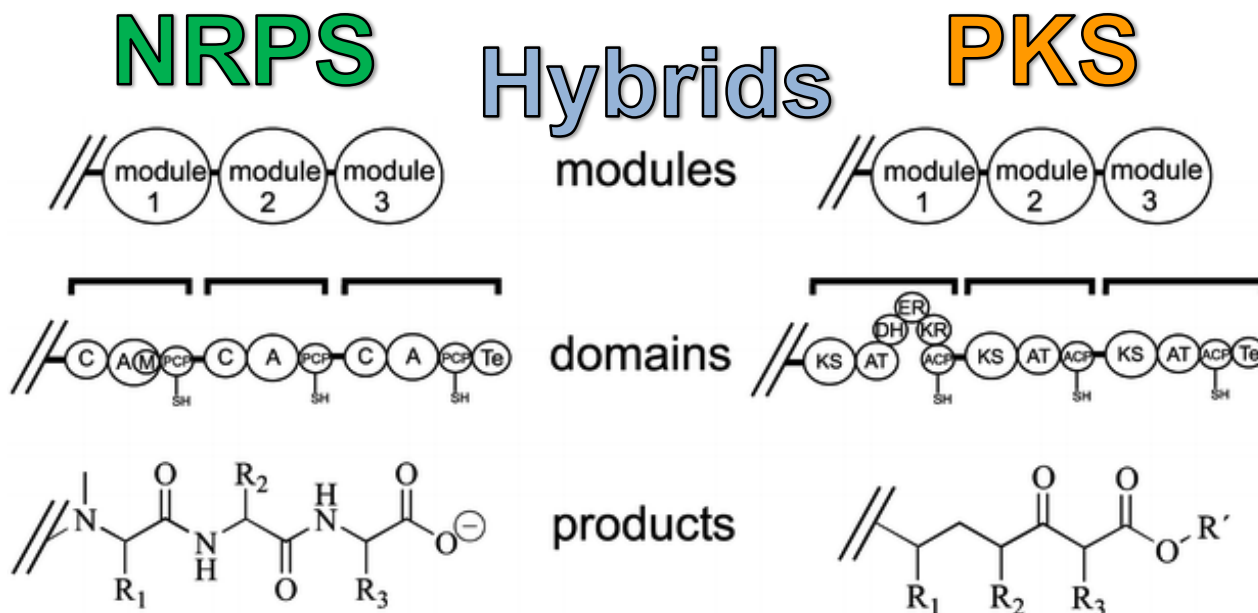
Strain	Genus	Contigs >1 kb	Max contig	Avg contig	GC%	Size (Mb)
CNJ780	<i>Marmoricola</i>	36	453,291	13,303	73.1%	4.39
CUA874	<i>Serinicoccus</i>	39	464,901	15,770	72.0%	3.56
CNJ954	<i>Corynebacterium</i>	45	433,327	83,858	65.2%	3.78
CNJ927	<i>Serinicoccus</i>	52	409,679	66,116	72.1%	3.48
CNJ824	<i>Ornithinimicrobium</i>	62	450,180	11,267	72.9%	3.50
CUA806	<i>Rhodococcus</i>	67	424,519	86,550	63.9%	6.33
CUA896	<i>Cellulosimicrobium</i>	79	73,310	10,472	74.9%	5.35
CUA673	<i>Saccharomonospora</i>	85	283,169	63,777	70.0%	5.46
CNJ863	<i>Gordonia</i>	94	499,033	57,433	67.3%	5.47
CNR923	<i>Nocardiosis</i>	153	230,239	36,276	71.0%	5.66
CNS004	<i>Pseudonocardia</i>	156	550,412	58,994	72.6%	9.26
CNJ770	<i>Kocuria</i>	166	225,243	24,837	71.9%	4.73
CNS139	<i>Pseudonocardia</i>	250	224,045	28,501	74.2%	7.34
CUA901	<i>Kytococcus</i>	461	301,417	6,995	71.0%	3.61
CNU125	<i>Actinomadura</i>	603	114,639	14,065	72.6%	10.44



Genome Annotation & Curation



antibiotics & Secondary Metabolite Analysis Shell



Terpene

RiPPs









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

Bioinformatic Tools- antiSMASH



**antibiotics & Secondary Metabolite Analysis Shell**

Server status:
working
Running jobs: 8
Queue length: 0
Long runtime queue: 0 
Jobs processed: 135142

Nucleotide input Protein input Results for existing job


Search a genome sequence for secondary metabolite biosynthesis gene clusters:


Email address (optional)

No file chosen

Load a file in GenBank / EMBL format (recommended) or in FASTA format

Or input NCBI accession number of desired file

+ Limit prediction to an input region (ignored for multi-sequence records) 

+ Detect putative gene clusters using the ClusterFinder algorithm 


☐ DNA of Eukaryotic origin

BLAST comparisons to other gene clusters:

☒ Gene Cluster Blast analysis ☒ Known Gene Cluster Blast analysis ☒ Subcluster Blast analysis

Additional annotations:

☒ smCOG analysis for functional prediction and phylogenetic analysis of genes ☒ Active site finder

+ Optional analyses with a long runtime 



Bioinformatic Tools- antiSMASH

Select Gene Cluster:



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

T333DRAFT_3789



T333DRAFT_3790



T333DRAFT_3799



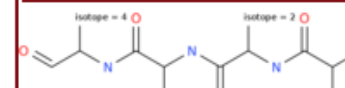
Homologous gene clusters

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

NRSPredictor2 SVM: N/A

Stachelhaus code: orn

Minowa: trp

consensus: nrp

NRSPredictor2 SVM: gly,ala

Stachelhaus code: thr

Minowa: trp

consensus: nrp

NRSPredictor2 SVM: gly,ala

Stachelhaus code: thr

Minowa: phe

consensus: nrp

T333DRAFT_3790

NRSPredictor2 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

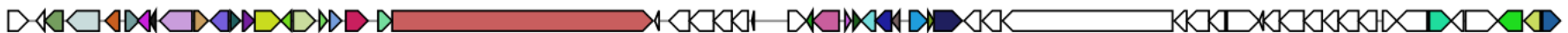
Query sequence



BAFO02000037_c1: *Nocardia asteroides* NBRC 15531 DNA, contig: NCAST37, whole... (55% of genes show similarity)



NZ_BAFO01000005_c1: *Nocardia asteroides* NBRC 15531, whole genome shotgun se... (55% of genes show similarity)



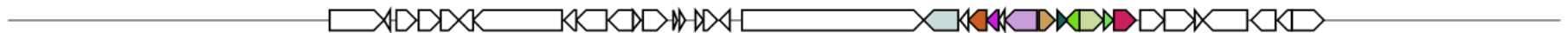
NZ_AQXZ01000019_c3: *Smaragdicoccus niigatensis* DSM 44881 F600DRAFT scaffold... (16% of genes show similarity)



JTJI01000011_c1: *Prauserella* sp. Am3 HQ32 scaffold 10.11, whole genome shot... (15% of genes show similarity)



ANPM01000002_c1: *Mycobacterium marinum* MB2 scaffold2, whole genome shotgun ... (16% of genes show similarity)



HG917972_c7: *Mycobacterium marinum* E11 main chromosome genome. (16% of genes show similarity)



Bioinformatic Tools- antiSMASH

- MultiGeneBlast against known gene clusters

Query sequence



BGC0000392_c1: Mirubactin biosynthetic gene cluster (57% of genes show similarity)



BGC0001185_c1: Bacillibactin biosynthetic gene cluster (14% of genes show similarity)



BGC0000368_c1: Griseobactin biosynthetic gene cluster (28% of genes show similarity)



BGC0000300_c1: Amychelin biosynthetic gene cluster (21% of genes show similarity)



BGC0000309_c1: Bacillibactin biosynthetic gene cluster (21% of genes show similarity)



Genome Annotation & Curation



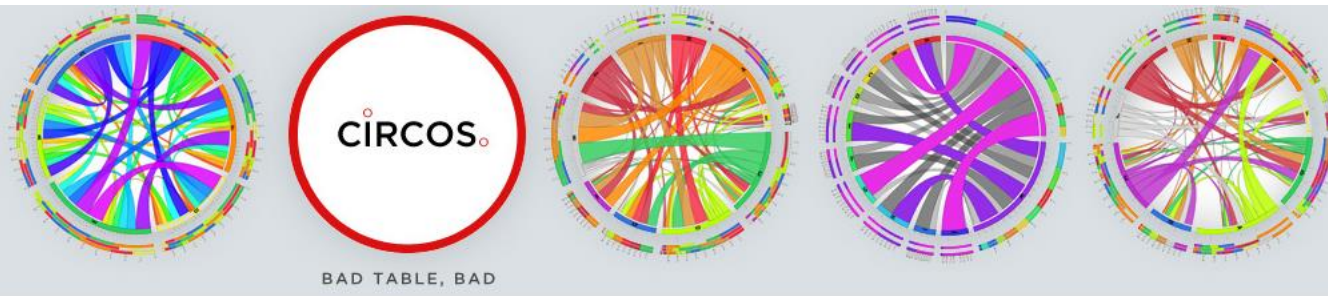
antibiotics & Secondary Metabolite Analysis SHell

NaPDoS

Natural Product Domain Seeker



Bioinformatic Tools- Circos



Circos in beta for 3rd year at 2012. Bioinformatics and Computational Genomics Analysis course by Prof. Patrick Henslin-McCoy, M.D.

visualize




settings

samples

archive

about

0. READ SLOGAN BADGES

1. CHECK DATA FORMAT

Before uploading a data file, check the [samples gallery](#) to make sure that your data format is compatible.

- Your file must be **plain text**.
- Your data values must be **non-negative integers**.
- Data must be space-separated (**one or more** tab or space, which will be collapsed).
- No two rows or columns may have the same name.
- Column and row names must **begin with a letter** (e.g. 'A', 'A0', 'A-0') and can only contain letters, numbers and **_**. No punctuation!
- Maximum row + column total is 150 — if exceeded, rows and columns are limited to 75.
- If you are using order, size and color rows/columns in combination they must appear in that order.

Need help? Post questions to the [Circos Google Group](#).

2A. UPLOAD YOUR FILE

If you are using the size, order or color options below, make sure your input file has the appropriate content (see [samples 5-9](#)).

Choose File

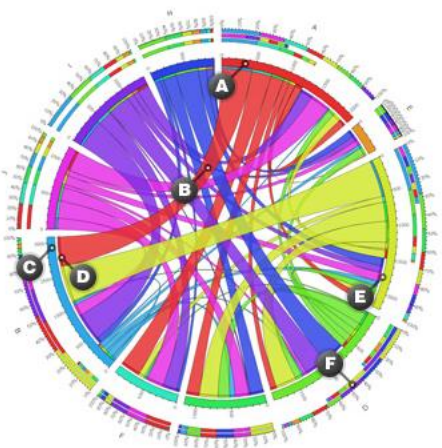
No file chosen

6. WHAT IS THIS?

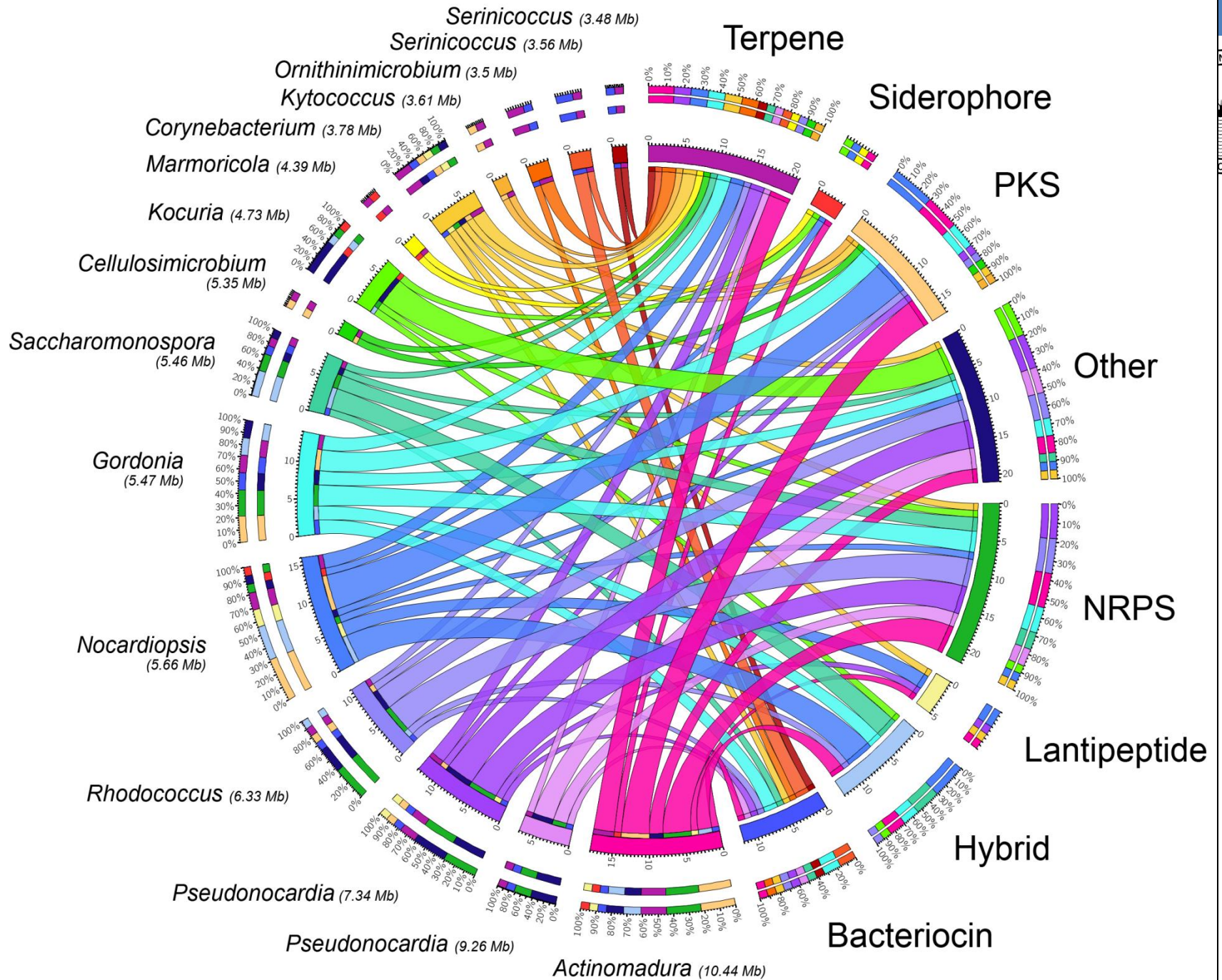
The Circos table viewer uses the [Circos](#) application to turn data tables into chord diagrams.

	A	B	C	D	E	F	G
A	105	450	92	96	5	301	195
B	20	46	78	33	53	28	83
C	118	553	94	317	25	89	287
D	100	18	108	104	105	25	173
H	23	83	123	342	98	48	205
I	173	428	103	325	82	215	23
J	305	173	138	49	81	258	207

into circularly composited visualizations like this



- 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



Thank you!

Moore lab

Dr. Bradley Moore

Ziemert lab

Dr. Nadine Ziemert
Mohammad Alanjary

Jensen lab

Dr. Paul Jensen
Nastassia Patin

Dr. Bill Fenical

Allen Lab

Dr. Sheila Podell

Dr. Anton Korobeynikov

Ion Torrent

Dr. Tommie Lincecum
Kristen Aguinaldo



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