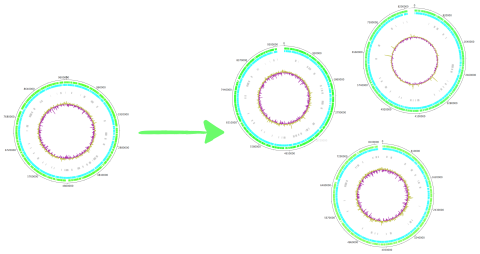
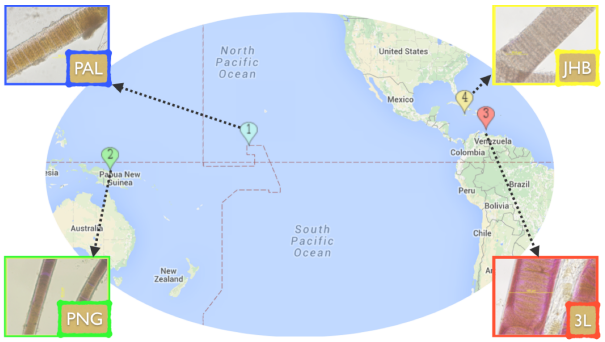
SIO-BUG MEETING

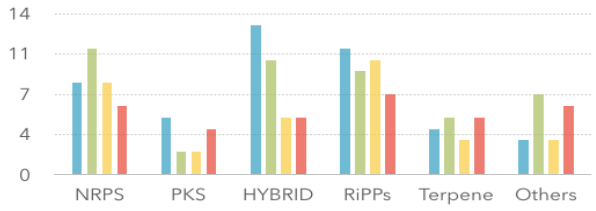
Obtaining genomes from (mini)metagenomes: assembly, binning and scaffolding

TIAGO LEÃO

GERWICK LAB







INTRO => Elevator talk

Importance:

The Nobel Prize in Physiology or Medicine 2015

William C. Campbell | Satoshi Ōmura | Youyou Tu

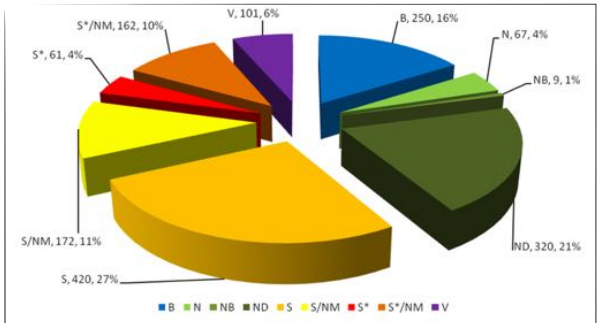
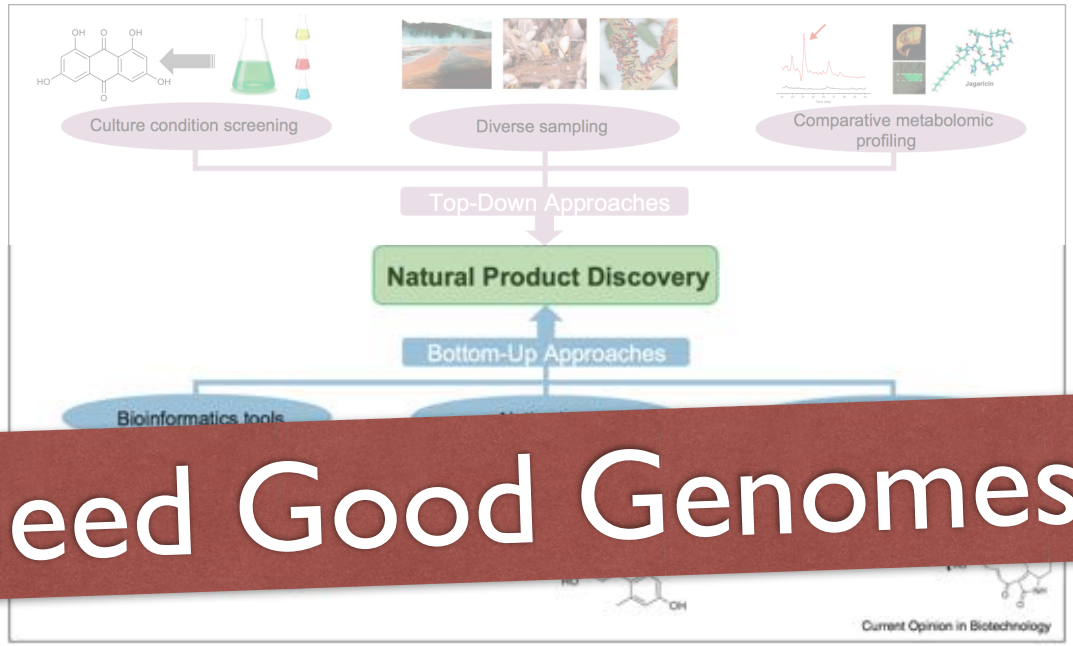
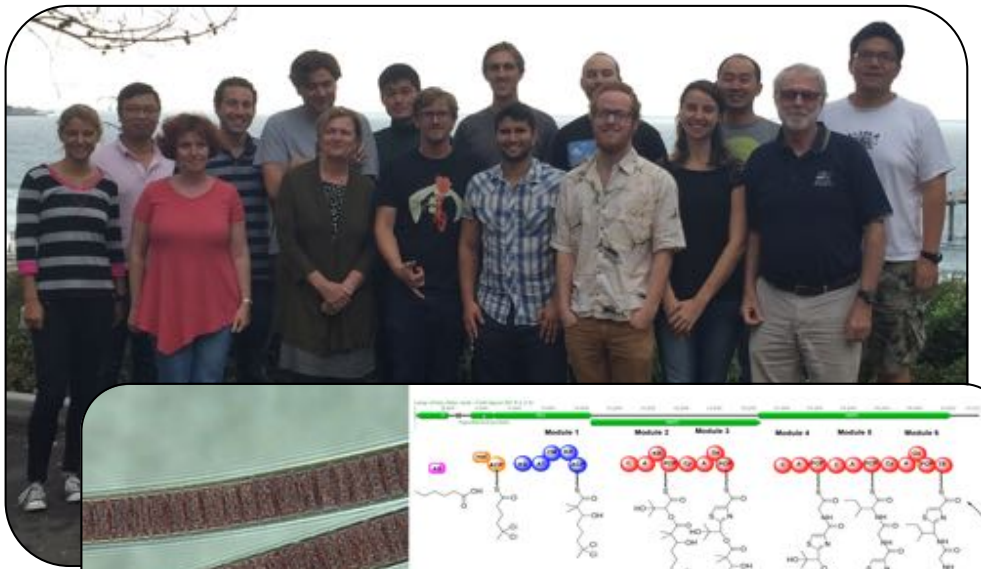


Figure 1. All new approved drugs 1981–2014; $n = 1562$.

Methods:



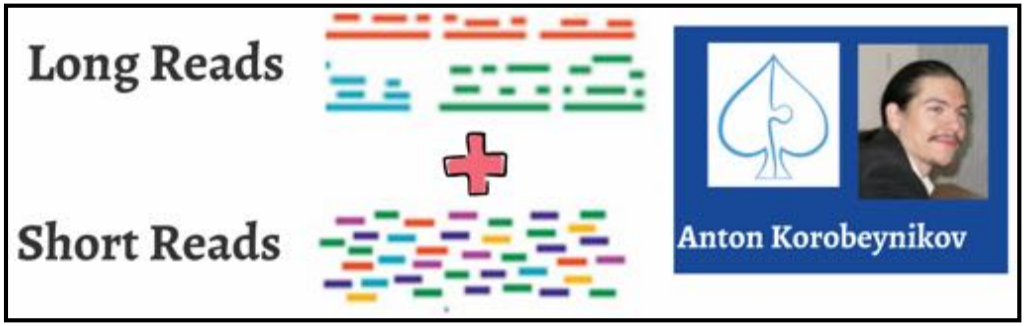
Need Good Genomes!



D) Post-Acquisition

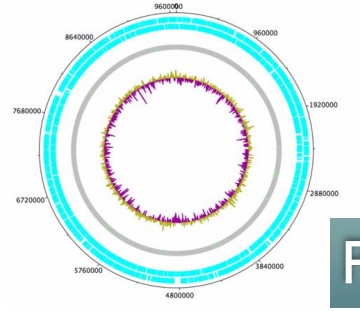
PAL

PacBio Sequence
+ Previous MiSeq



55 Contigs

Binning

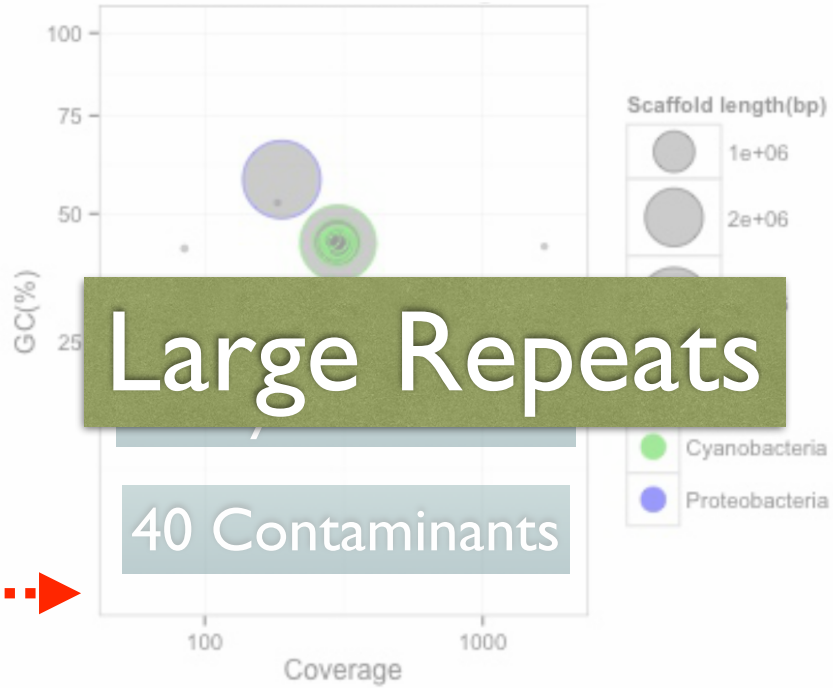


Plasmid (35.5 kb)

Chromosome (9.67 Mb)

Scaffolding w/
Long Reads

SSPACE



Large Repeats

40 Contaminants

A) *De Novo* Assembly

PAL

PacBio Sequence
+ Previous MiSeq

Long Reads

Short Reads

Anton Korobeynikov

55 Contigs

Which sequencing method is the best?

SIMPLEST ANSWER: We'll never know

Depends on: Sample & Goal

- -Complexity
- -Abundance
- -Overall Size
- -Repeats
- -[…]

- -Specific Genome
- -Completness
- (Reference?)
- -Genomic Target
- -[…]

A) *De Novo* Assembly

Many Papers

Which assembler is the best?

Long Read

Short Read

OPEN ACCESS Freely available online

PLoS one

A Practical Comparison of *De Novo* Genome Assembly Software Tools for Next-Generation Sequencing Technologies

Wenyu Zhang, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, Bairong Shen*

Center for Systems Biology, Soochow University, Suzhou, Jiangsu, China

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 30 no. 19 2014, pages 2709–2716
doi:10.1093/bioinformatics/btu391

Genome analysis

Advance Access publication June 14, 2014

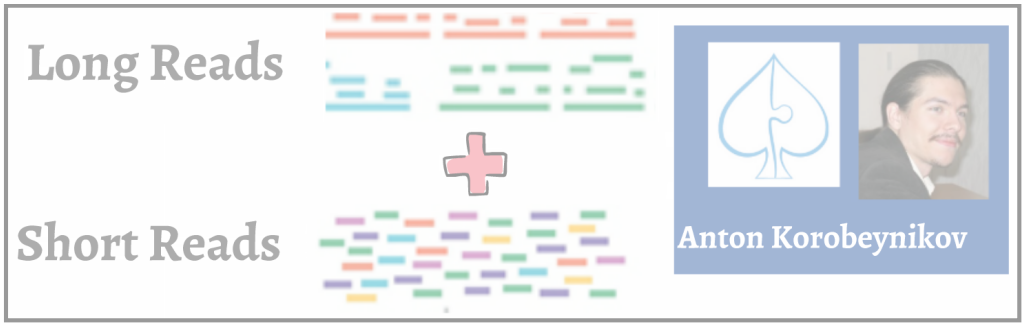
Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences

Sagar M. Utturkar[1], Dawn M. Klingeman[2], Miriam L. Land[2], Christopher W. Schadt[1,2], Mitchel J. Doktycz[1,2], Dale A. Pelletier[1,2] and Steven D. Brown[1,2,*]

[1]Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37919, USA and [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

A) *De Novo* Assembly

PacBio Sequence
+ Previous MiSeq



55 Contigs

Which assembler is the best?

I'm happy with SPAdes

Now there is metaSPAdes

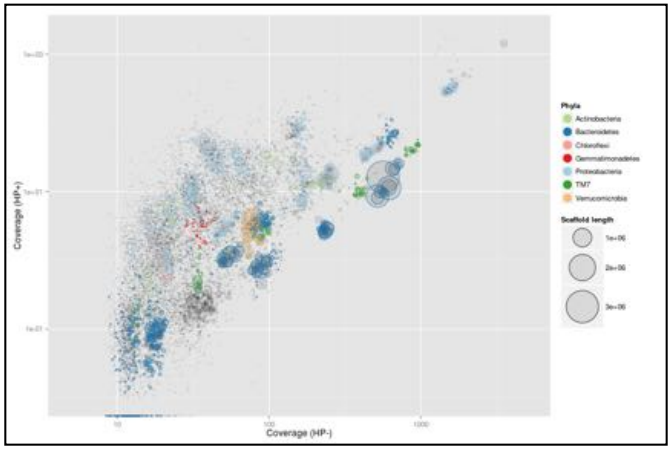
Also tried ~~MIRA~~
and ~~Geneious~~

PS: By experience, try many different k-mers

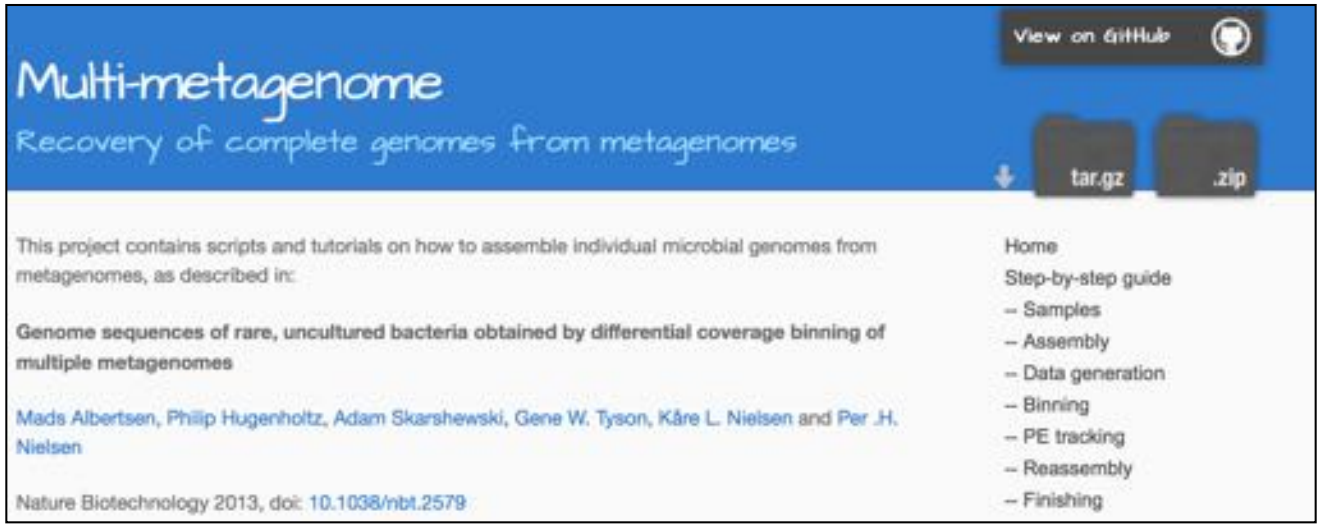
Different k-mer size can assemble better different genomes (from metagenomes)

B) Binning Assembled Contigs

- -GC**
- -Coverage***
- -Tetranucleotide***
- -Taxonomy*

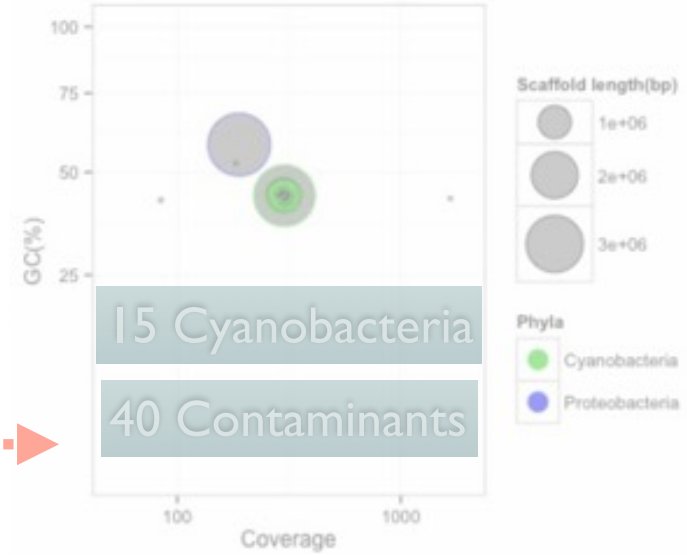


Albertsen *et al.*, 2013

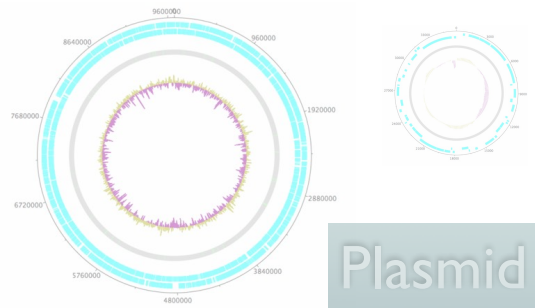


55 Contigs

Binning



C) Scaffolding

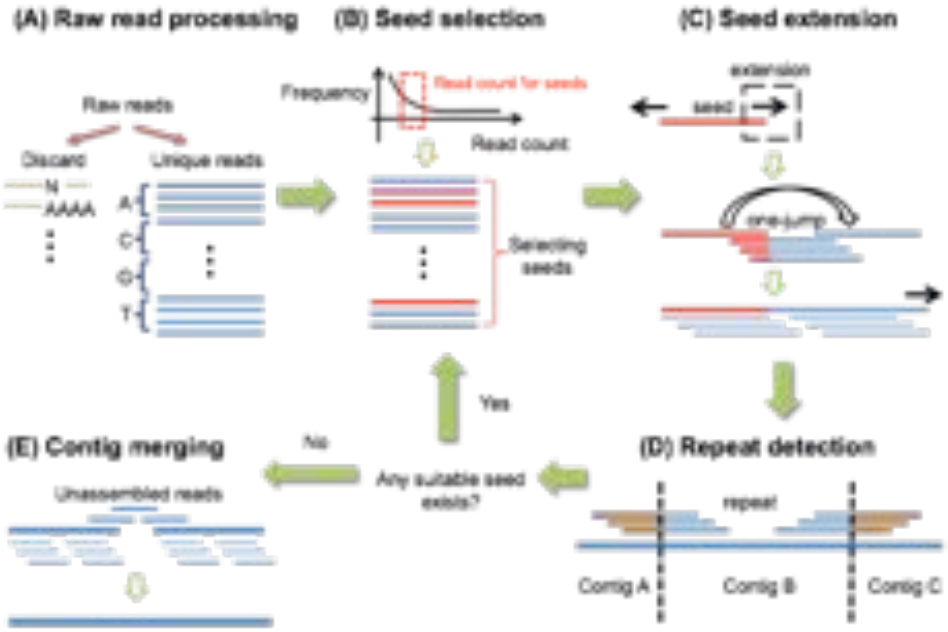


Plasmid (35.5 kb)

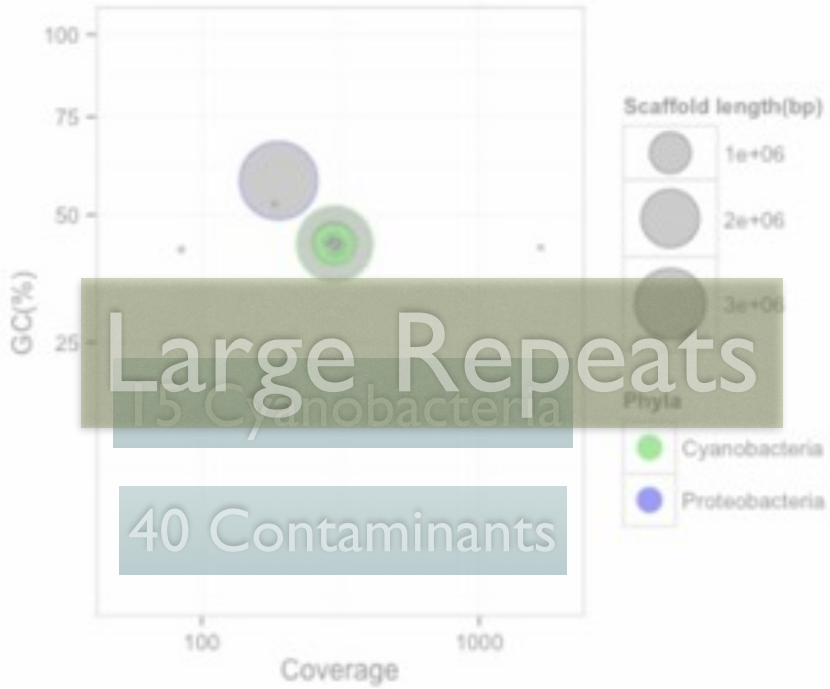
Chromosome (9.67 Mb)

I tried: geneious AlignGraph mauve Multiple Genome Alignment

SSPACE Short (or long)



Scaffolding w/ Long Reads → SSPACE

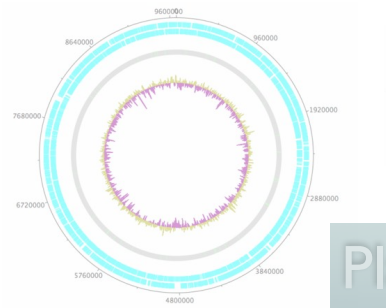


Large Repeats

15 Cyanobacteria

40 Contaminants

C) Scaffolding



Plasmid (35.5 kb)

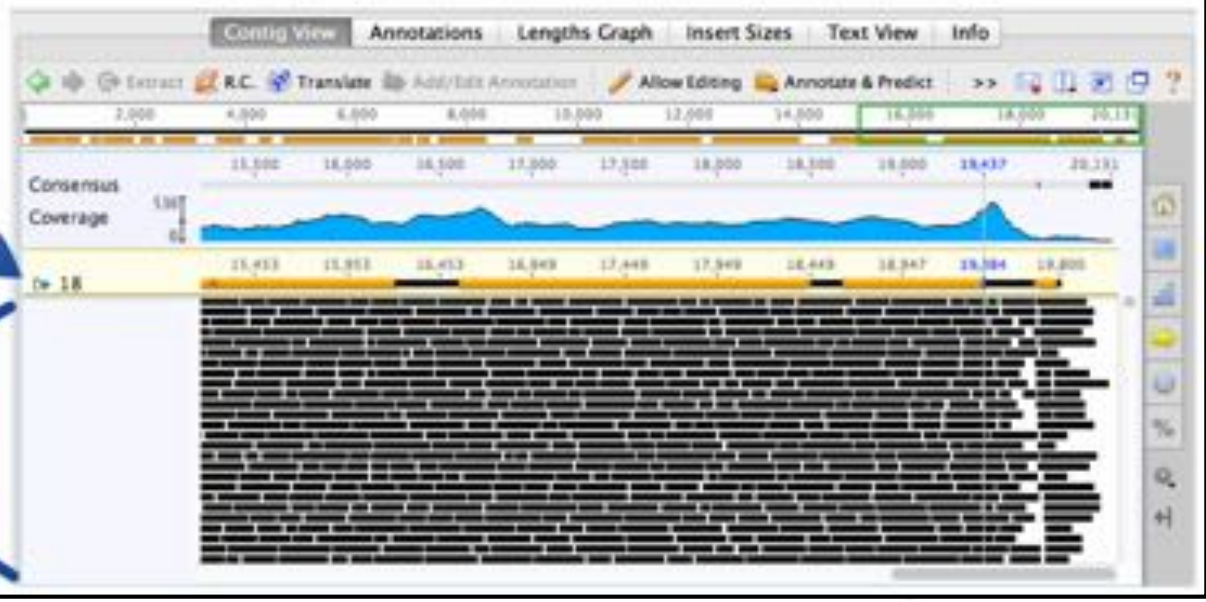
Chromosome (9.67 Mb)

I tried: geneious AlignGraph mauve Multiple Genome Alignment

SSPACE Short (or long)

Scaffolding w/ Long Reads ⇢ SSPACE

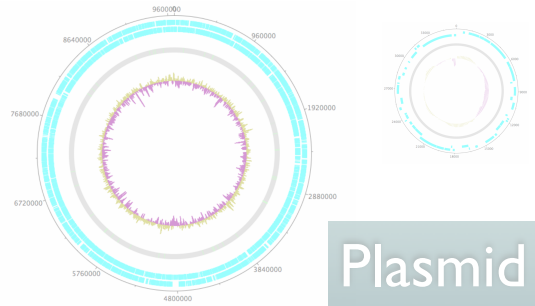
Trimming helps



Contig With ORFs

Paired Reads

C) Scaffolding



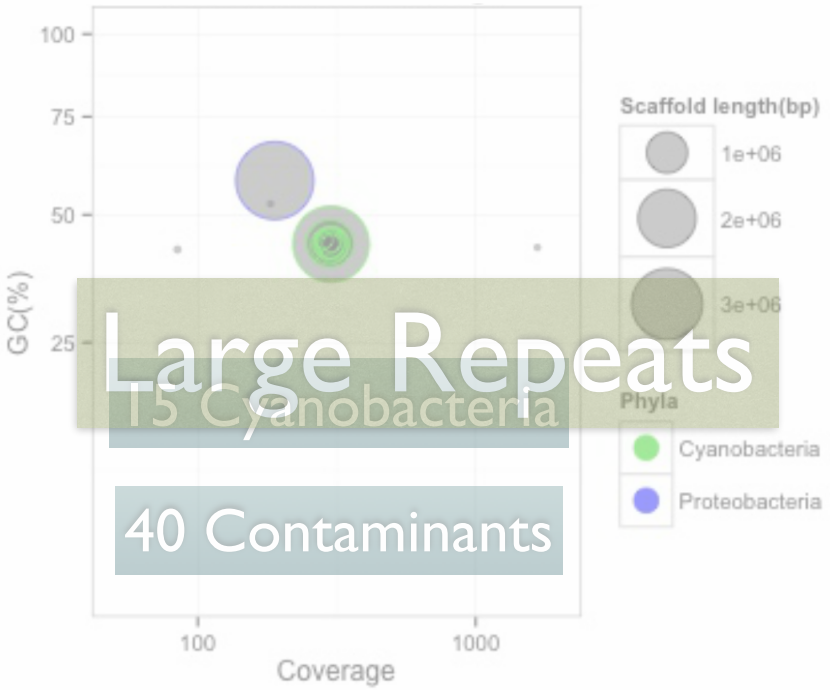
Plasmid (35.5 kb)

Chromosome (9.67 Mb)

I tried: geneious AlignGraph mauve Multiple Genome Alignment

SSPACE Short (or long)

Scaffolding w/ Long Reads → SSPACE



Large Repeats

15 Cyanobacteria

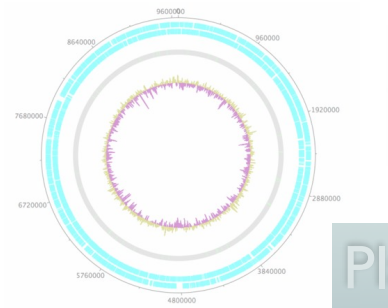
40 Contaminants

C.b) Gap Filling

PBSuite
Software for Long-Rea[illegible]equencing Data from PacBio
Brought to you by: acenglis[illegible]cm

geneious
Good coverage (>60x)
Alignments and Consensus

D) Post-Acquisition



Plasmid (35.5 kb)

Chromosome (9.67 Mb)

How complete is your genome/draft?

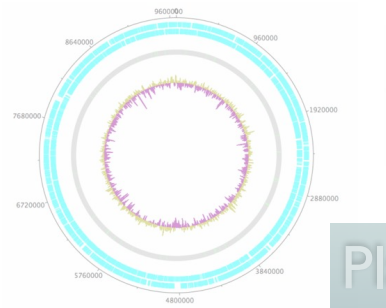
i. Select a set of genomes

JGI JOINT GENOME INSTITUTE

ii. Abundance Profile (COGs and Gene Count)

iii. COGs that are exactly one copy in 95% of all strains (Single Copy Genes)

D) Post-Acquisition

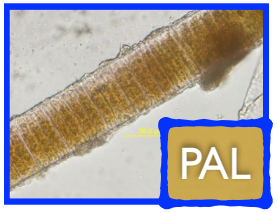


Plasmid (35.5 kb)

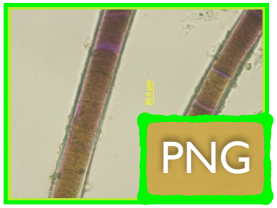
Chromosome (9.67 Mb)

iv. Compare the count of Single Copy Genes

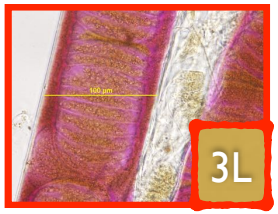
208 Housekeeping Genes from Cyanobacteria



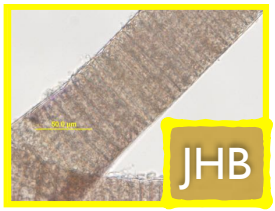
99.52%



98.10%



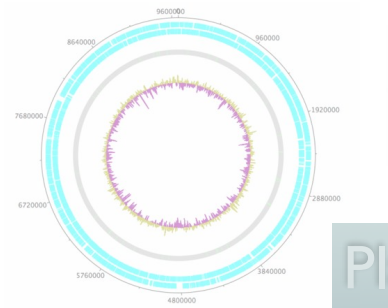
97.13%



100%

D) Post-Acquisition

JGI
JOINT GENOME INSTITUTE

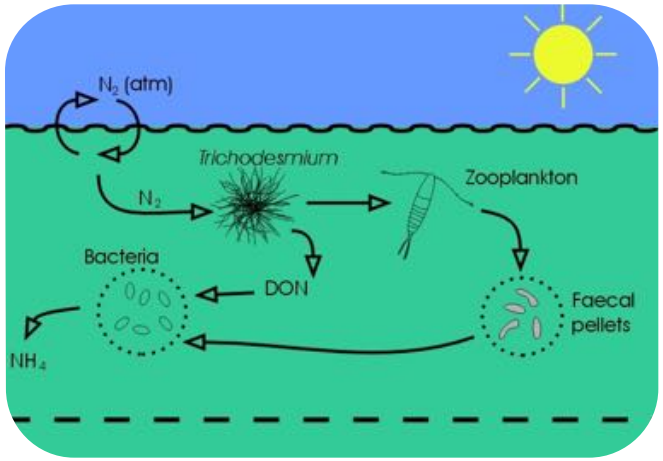


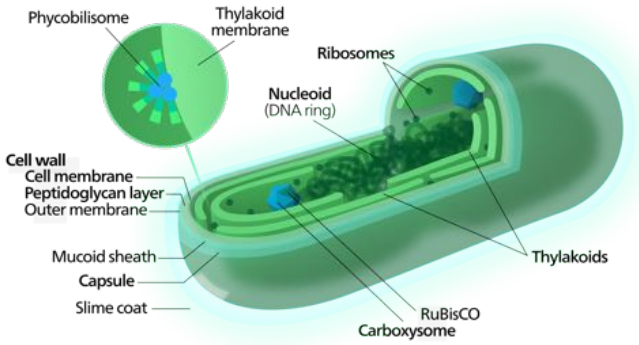
Plasmid (35.5 kb)

Chromosome (9.67 Mb)

-Metabolism and Biosynthesis

Phenotypes/Metabolism from Pathway Assertion	
Metabolism	Prototrophic (L-alanine prototroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Prototrophic (L-aspartate prototroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Prototrophic (L-glutamate prototroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (L-phenylalanine auxotroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (L-tyrosine auxotroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (L-tryptophan auxotroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (L-histidine auxotroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Prototrophic (Glycine prototroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Prototrophic (L-asparagine prototroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Prototrophic (L-glutamine prototroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (L-isoleucine auxotroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (L-leucine auxotroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (L-serine auxotroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (L-valine auxotroph) (IMG_PIPELINE; 2015-09-25)
Metabolism	(Non-selenocysteine synthesizer) (IMG_PIPELINE; 2015-09-25)
Metabolism	(Non-biotin synthesizer) (IMG_PIPELINE; 2015-09-25)
Metabolism	Auxotroph (Incomplete Coenzyme A biosynthesis) (IMG_PIPELINE; 2015-09-25)

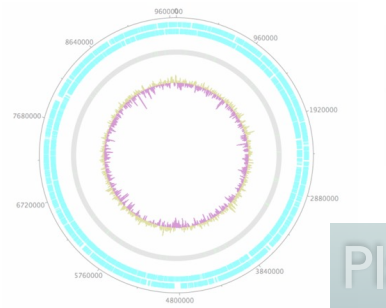




http://www.soes.soton.ac.uk/staff/tt/nf/structure/intro.htm

https://en.wikipedia.org/wiki/Thylakoid

D) Post-Acquisition



Plasmid (35.5 kb)

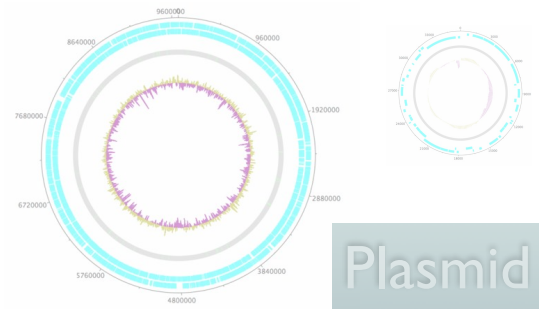
Chromosome (9.67 Mb)

-COGs Profile and Comparison

		Reference	Draft Genomes				
Category	Category Description	PAL	D-rank	PNG	D-rank	JL	D-rank	JHB
B	Chromatin structure and dynamics	2	0*	2	0*	2	0*	2
C	Energy production and conversion	157	0.03	155	0.05	151	0.02	171
D	Cell cycle control, cell division, chromosome partitioning	43	0.01	40	0.05	35	0.03	39
E	Amino acid transport and metabolism	245	0.05	237	0.03	236	0.02	252
F	Nucleotide transport and metabolism	69	0.01	62	0.01	63	0.03	61
G	Carbohydrate transport and metabolism	141	0.05	155	0.04	137	0.04	153
H	Coenzyme transport and metabolism	210	0.03	206	0.01	194	0.01	210
I	Lipid transport and metabolism	115	0.01	96	0.05	92	0.08	123
J	Translation, ribosomal structure and biogenesis	194	0*	190	0*	188	0*	194
K	Transcription	114	0.03	112	0.09	101	0.08	113
L	Replication, recombination and repair	104	0.09	107	0*	91	0.03	91
M	Cell wall/membrane/envelope biogenesis	296	0.09	285	0.23	293	0.16	326
N	Cell motility	69	0.09	63	0.12	59	0.07	69
O	Posttranslational modification, protein turnover, chaperones	163	0.04	149	0.01	145	0.04	169
P	Inorganic ion transport and metabolism	153	0.04	148	0.01	141	0.02	153
Q	Secondary metabolites biosynthesis, transport and catabolism	171	0.58	118	0.57	112	0.08	161
R	General function prediction only	480	0.02	425	0.04	409	0.03	465
S	Function unknown	211	0.01	203	0.01	200	0.04	226
T	Signal transduction mechanisms	266	0.21	256	0.05	237	0.03	269
U	Intracellular trafficking, secretion, and vesicular transport	62	0.27	39	0.31	42	0.2	49
V	Defense mechanisms	135	0.03	125	0.16	115	0.08	135
W	Extracellular structures	17	0*	16	0*	19	0*	18
X	Mobilome: prophages, transposons	294	0.73	183	0.49	226	0.48	235
B-X	Total	3711		3372		3288		3684

D) Post-Acquisition

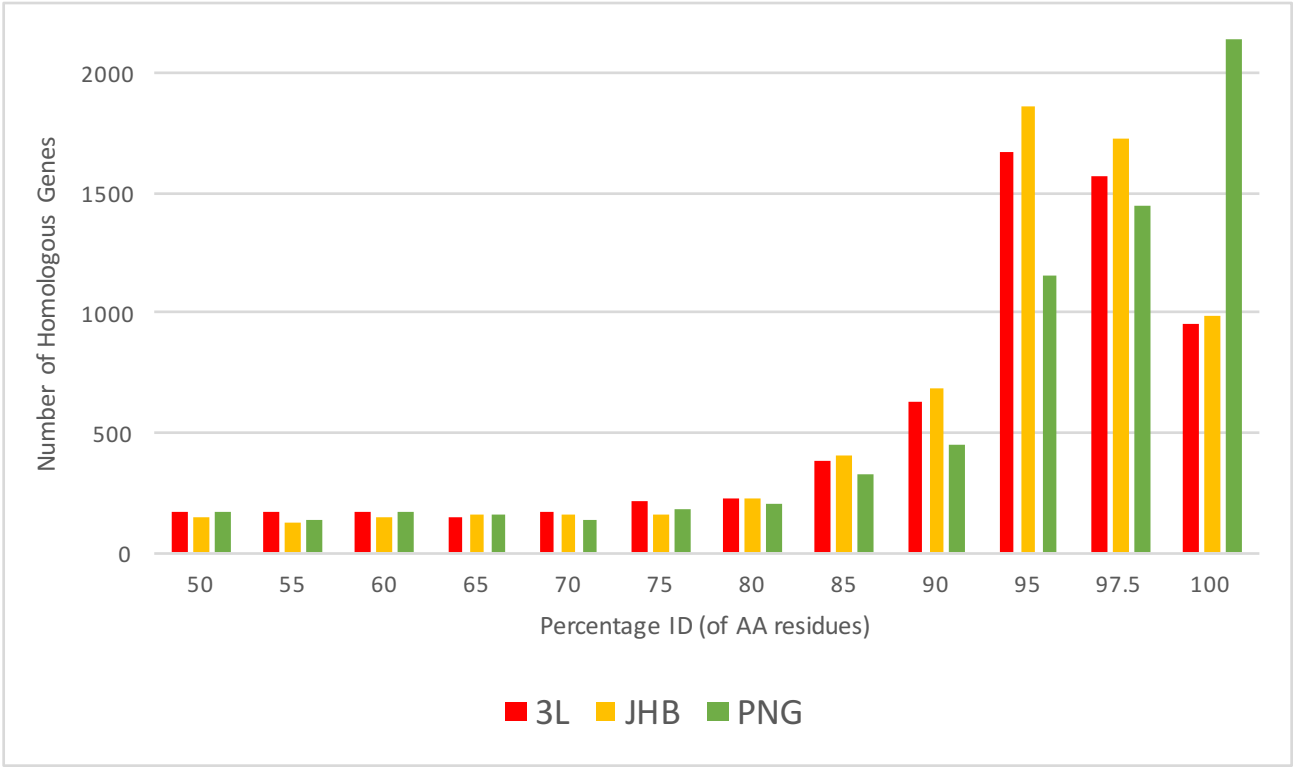
JGI JOINT GENOME INSTITUTE



Plasmid (35.5 kb)

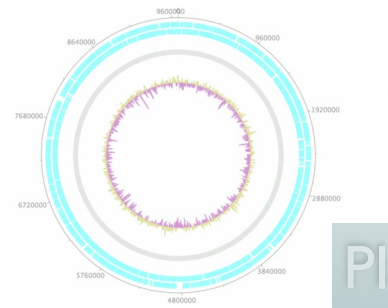
Chromosome (9.67 Mb)

-List of Best Homologs Between Genomes



D) Post-Acquisition

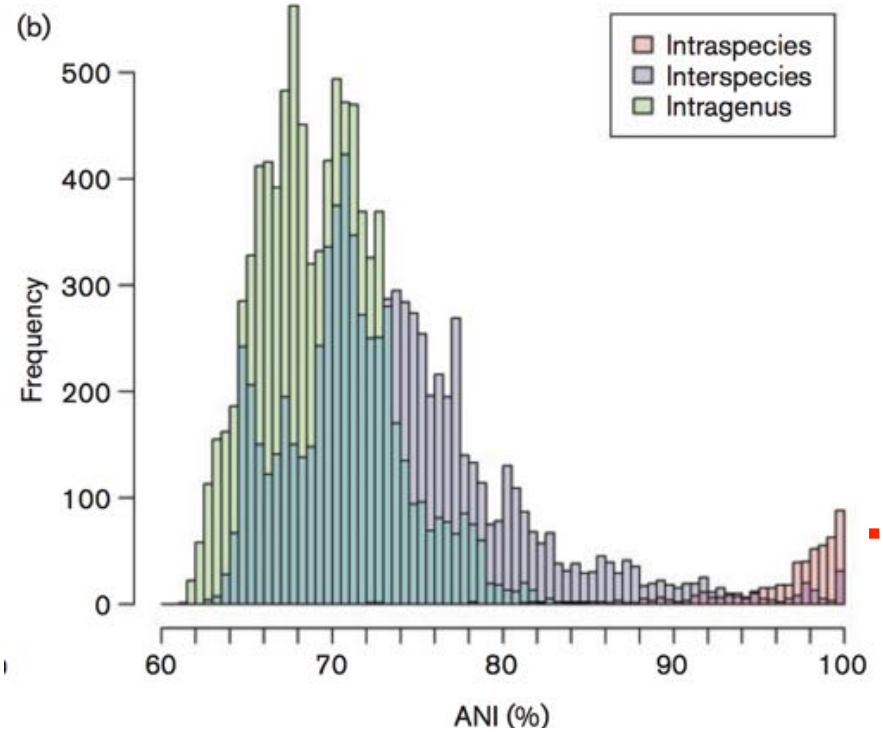
JGI JOINT GENOME INSTITUTE



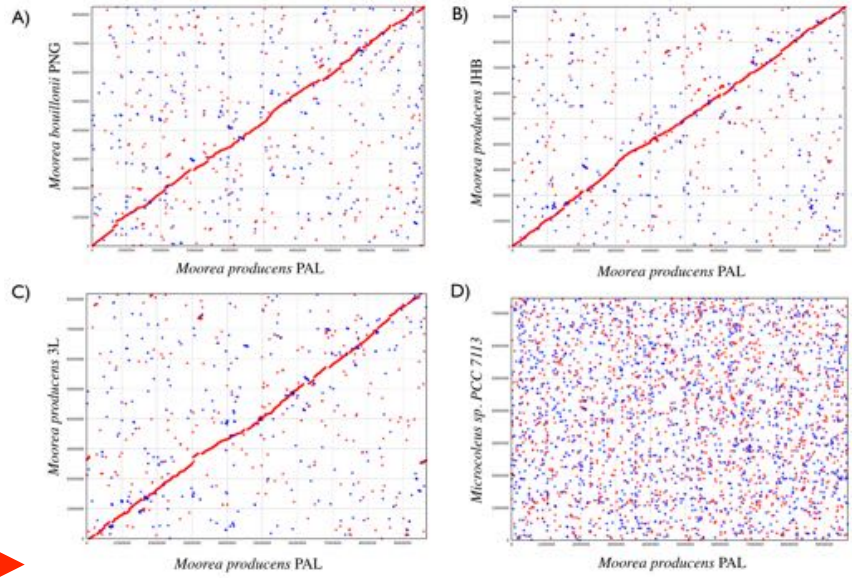
Plasmid (35.5 kb)

Chromosome (9.67 Mb)

-Average Nucleotide Identity



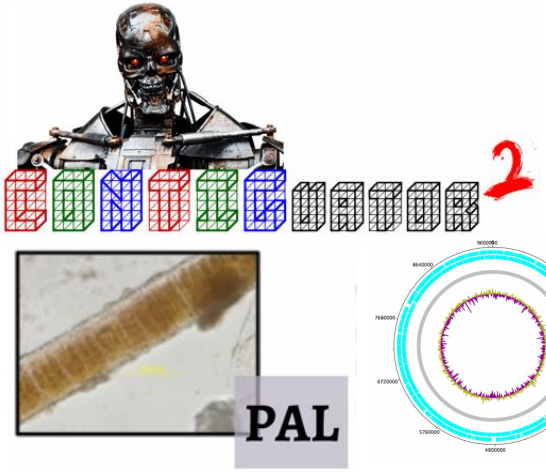
-Synteny Plots



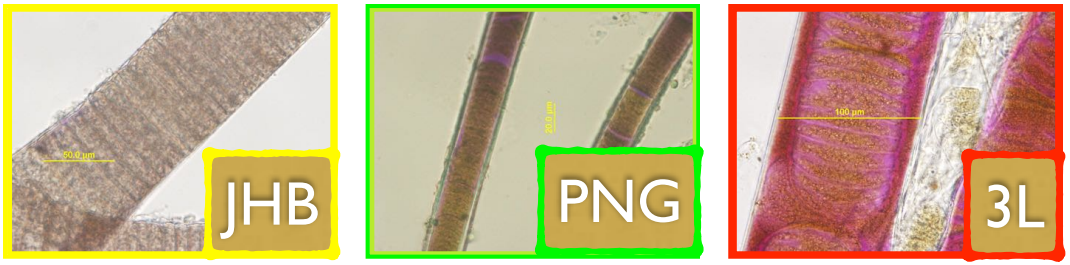
Find the Best Reference Genome

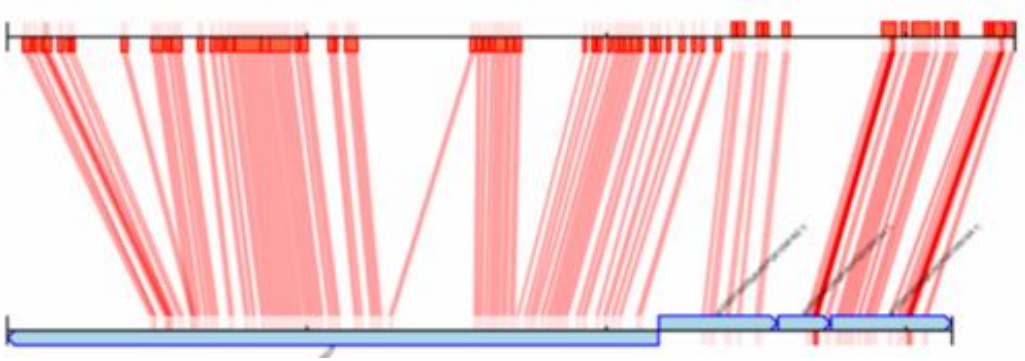
D) Post-Acquisition

How to use my new genome?



Reference Assembly

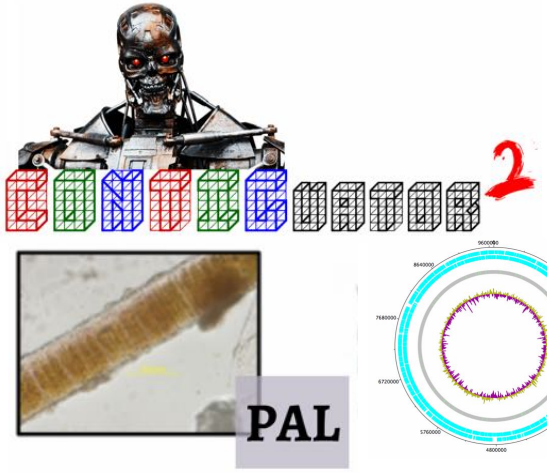




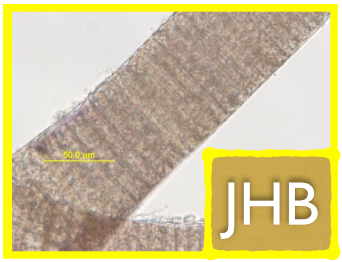
Be careful on your comparisons after reference assembly
Make sure to do not discard **important(?)** unmapped contigs

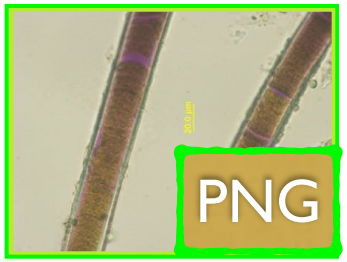
D) Post-Acquisition

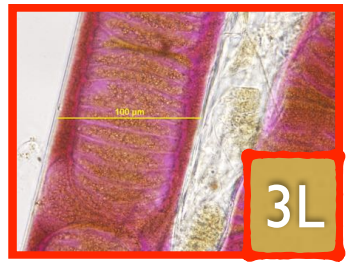
How to use my new genome?



Reference Assembly





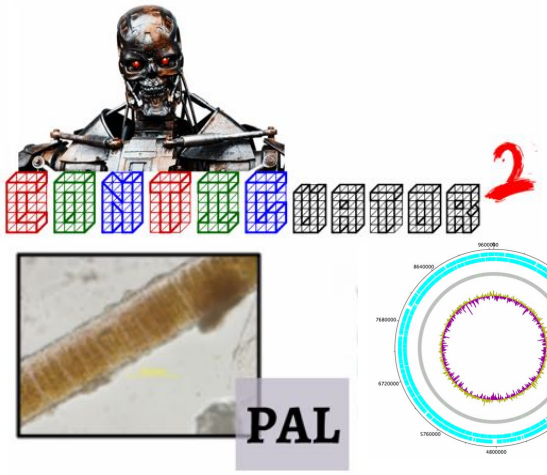


Genome	Sequencing Method	Avg. Read Lengths	Current Size	# Contigs	GC Content	N50
M. producens PAL	**Illumina MiSeq + PacBio**	**300 bp PE + 10 kb**	**9.67 Mb**	**1**	**43.53%**	—
M. producens JHB	Illumina HiSeq	100bp PE	9.64 Mb	2435	43.65%	18,820
M. bouillonii PNG	Illumina HiSeq + PacBio	100bp PE	8.61 Mb	913	43.67%	135,461
M. producens 3L	Sanger + 454 (MDA)	600-800bp	8.38 Mb	287	43.68%	66,028

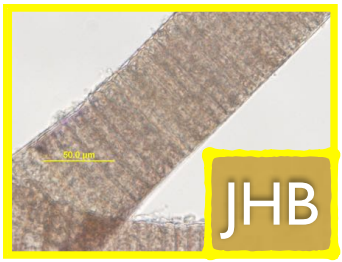
Before

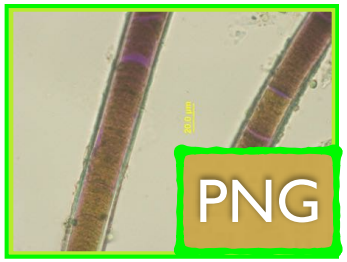
D) Post-Acquisition

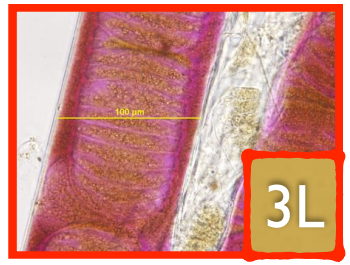
How to use my new genome?



Reference Assembly





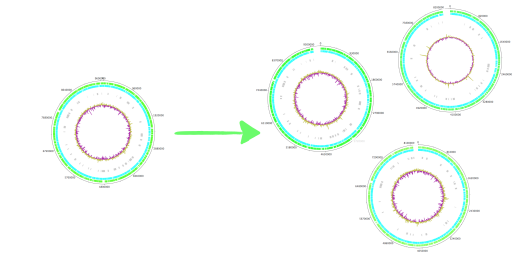


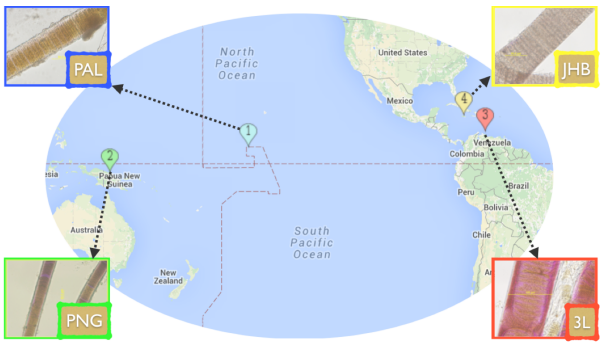
Genome	Sequencing Method	Unmapped Contigs	Scaffold Size	# Contigs	GC Content	N50
M. producens PAL	Illumina MiSeq + PacBio	1 (plasmid)	9.67 Mb	1	43.52%	—
M. producens JHB	Illumina HiSeq	2 (plasmid[s])	9.35 Mb	205	43.67%	—
M. bouillonii PNG	Illumina HiSeq + PacBio	12 (56 kb)	8.23 Mb	291	43.63%	8,262,658
M. producens 3L	Sanger + 454 (MDA)	78 (199 kb)	8.15 Mb	204	43.68%	8,171,464

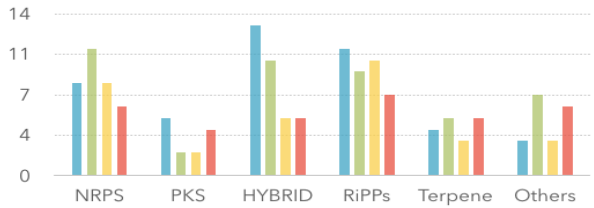
After

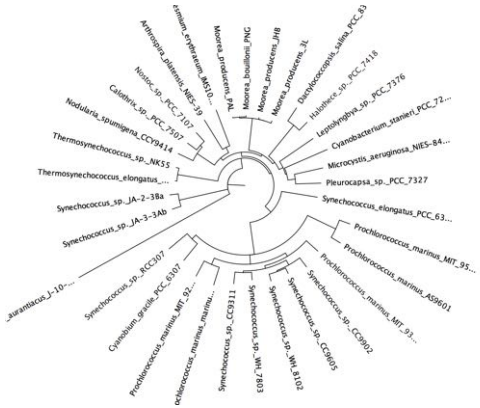
Genome Comparison

First complete genome of a natural product rich filamentous marine cyanobacterium gives insights into a gap in the tree of life: the genus *Moorea* is metabolically and genetically distinct from all known cyanobacteria



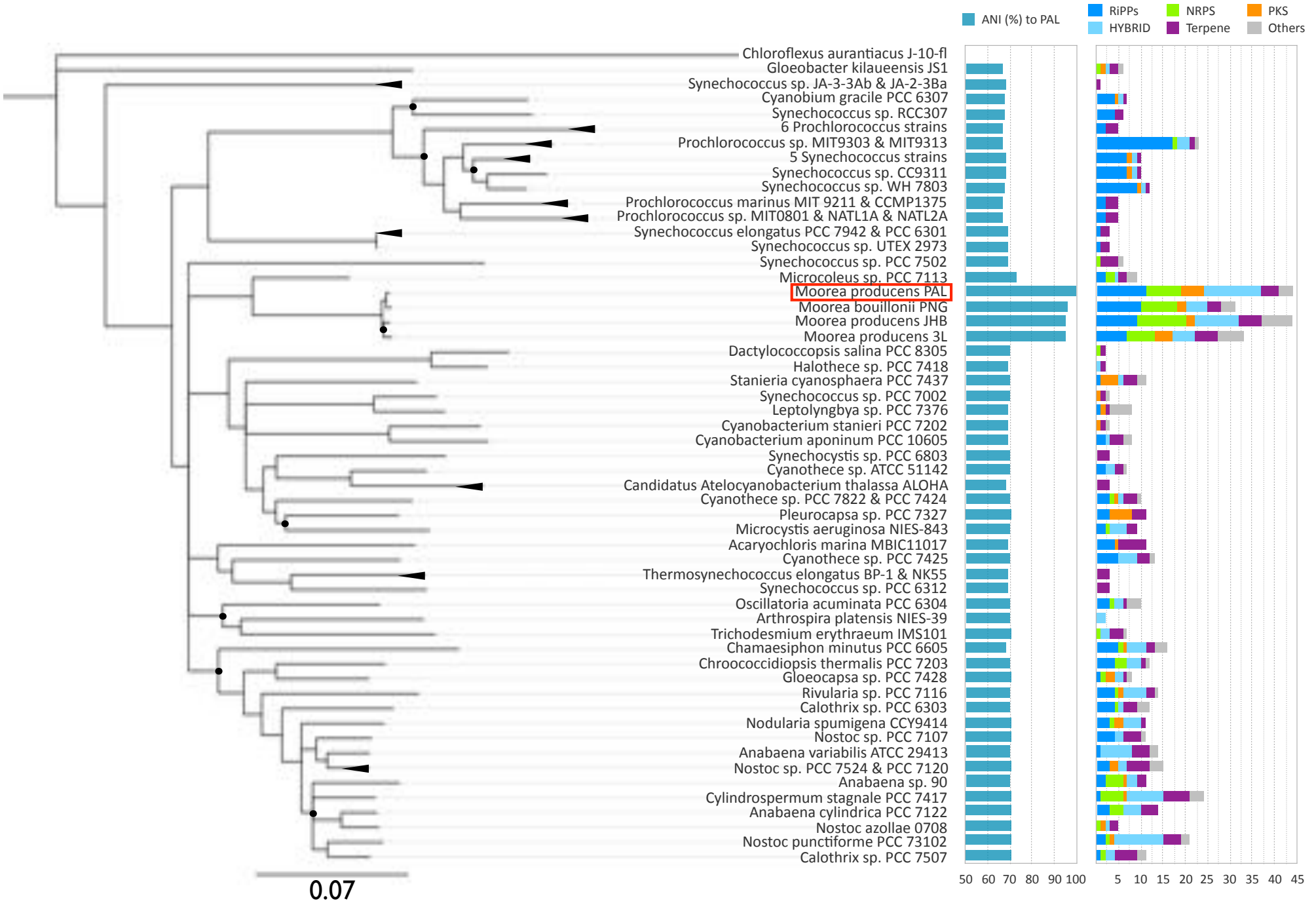






Leao TF, Korobeynikov A, Monroe E, Podell S, Allen E, Gerwick WH, Gerwick L. Under progress. Proposed Journal: PNAS

SOME DATA => Genomes Comparison => Moorea are different from all known cyanobacteria



Dotted bootstraps are higher than 50 and lower than 80, concatenated branches represent multiple sequences with bootstrap 100

Thanks to:

All the Gerwick Lab
(former and current)

Dr William Gerwick
Dr Lena Gerwick
Dr Evgenia Glukhov
Nathan Moss

SCIENCE WITHOUT BORDERS

CAPES

Dr Anton Korobeynikov
Pevzner Lab (St. Peterburg)

Dr Sheila Podell
Allen Lab (SIO)

Dr Luke Thompson
Knight Lab (UCSD)

Dr Tristan Carland
Former Gerwick Lab
Illumina®

Uma pequena mensagem: