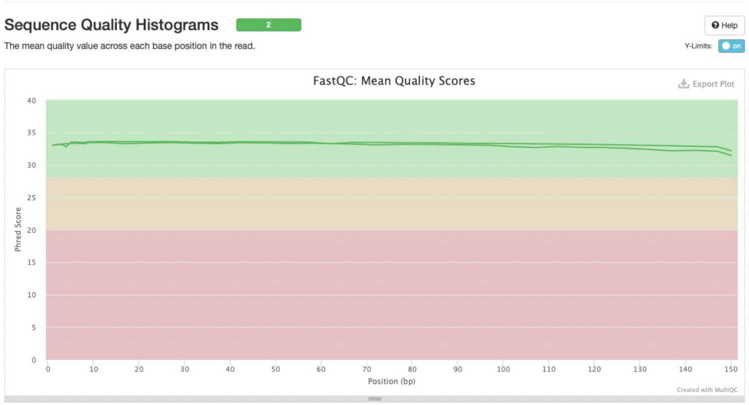
Initial Findings for Neveda Naz, Ph.D.

Adelaide Rhodes, Ph.D. and Jason Laird, M.S.

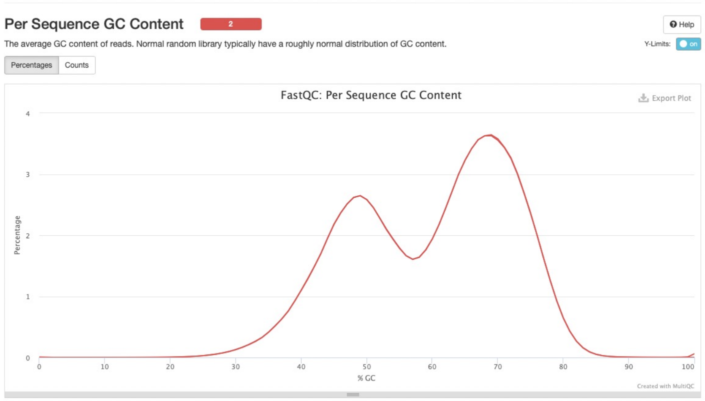
MultiQC on Initial Data



Sequence Quality Histograms

The mean quality value across each base position in the read.

Good quality scores



Per Sequence GC Content

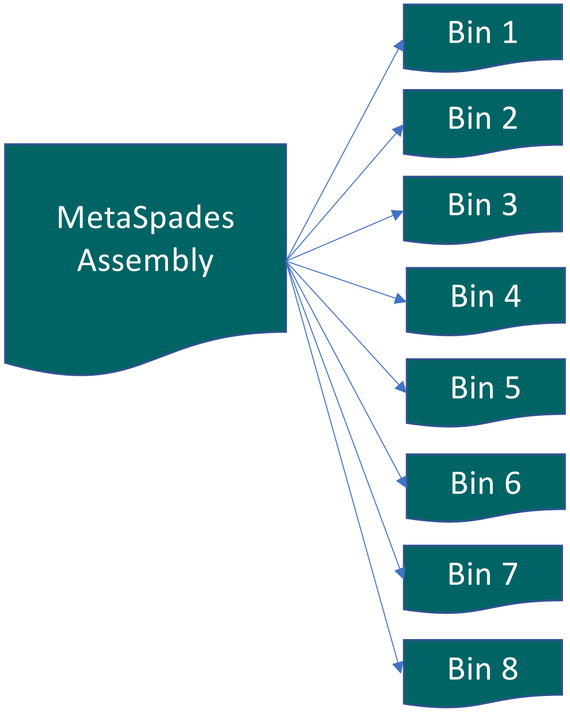
The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

We note two GC peaks, probable contamination

Assembly with MetaSpades

Assembly	contigs
# contigs (>= 0 bp)	25155
# contigs (>= 1000 bp)	4579
# contigs (>= 5000 bp)	1020
# contigs (>= 10000 bp)	564
# contigs (>= 25000 bp)	246
# contigs (>= 50000 bp)	107
Total length (>= 0 bp)	39198854
Total length (>= 1000 bp)	31297820
Total length (>= 5000 bp)	24281365
Total length (>= 10000 bp)	21126250
Total length (>= 25000 bp)	16091790
Total length (>= 50000 bp)	11355212
# contigs	9319
Largest contig	580104
Total length	34578602
GC (%)	64.67
N50	21521
N75	3412
L50	298
L75	1424
# N's per 100 kbp	0.00

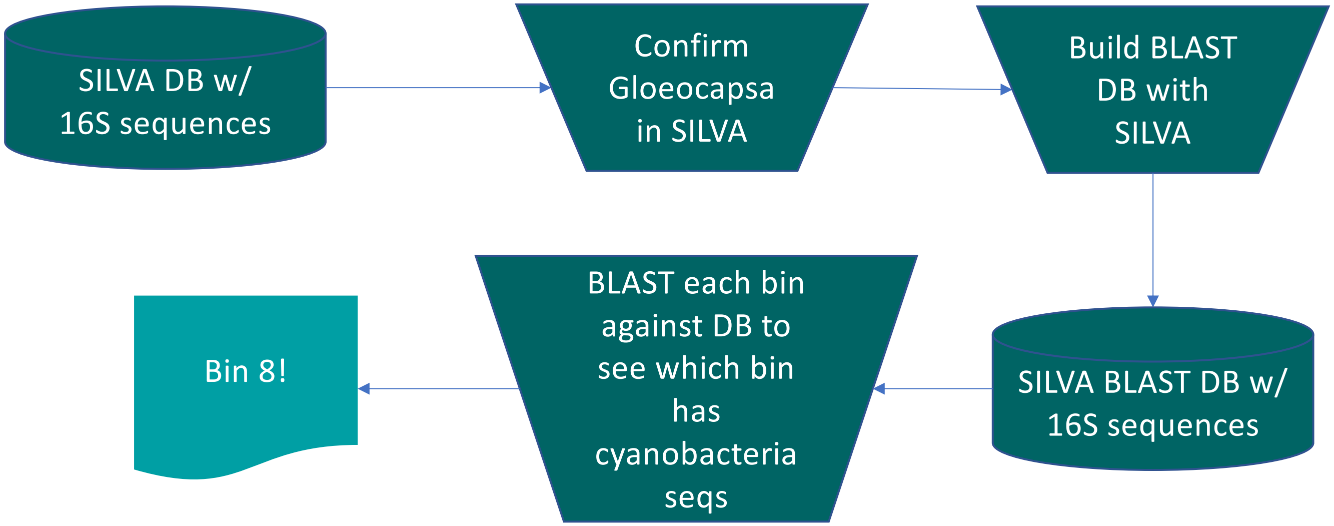
Binning With Metabat2



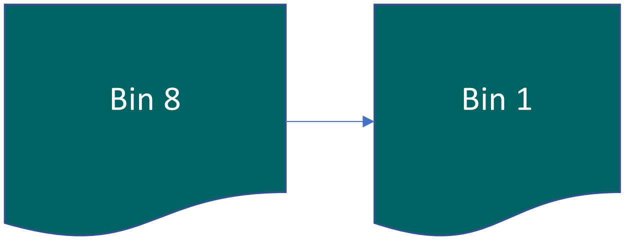
Bin Stats

```
71971 bin.1.fa
79689 bin.2.fa
3760 bin.3.fa
62090 bin.4.fa
19516 bin.5.fa
4061 bin.6.fa
50460 bin.7.fa
82748 bin.8.fa
```

Which Bin now Holds the Cyano Sequences?


SILVA DB w/ 16S sequences
Confirm Gloeocapsa in SILVA
Build BLAST DB with SILVA
SILVA BLAST DB w/ 16S sequences
BLAST each bin against DB to see which bin has cyanobacteria seqs
Bin 8!

Binning With Metabat2 – Just Bin 8



Bin 8 did not bin any further!

Quast on Bin 8

```
Assembly                    bin.8
# contigs (>= 0 bp)         281
# contigs (>= 1000 bp)      281
# contigs (>= 5000 bp)      242
# contigs (>= 10000 bp)     161
# contigs (>= 25000 bp)     58
# contigs (>= 50000 bp)     16
Total length (>= 0 bp)      4940314
Total length (>= 1000 bp)   4940314
Total length (>= 5000 bp)   4791211
Total length (>= 10000 bp)  4207022
Total length (>= 25000 bp)  2522908
Total length (>= 50000 bp)  1130416
# contigs                   281
Largest contig              101391
Total length                4940314
GC (%)                      46.34
N50                         25426
N75                         14369
L50                         56
L75                         120
# N's per 100 kbp           0.00
```

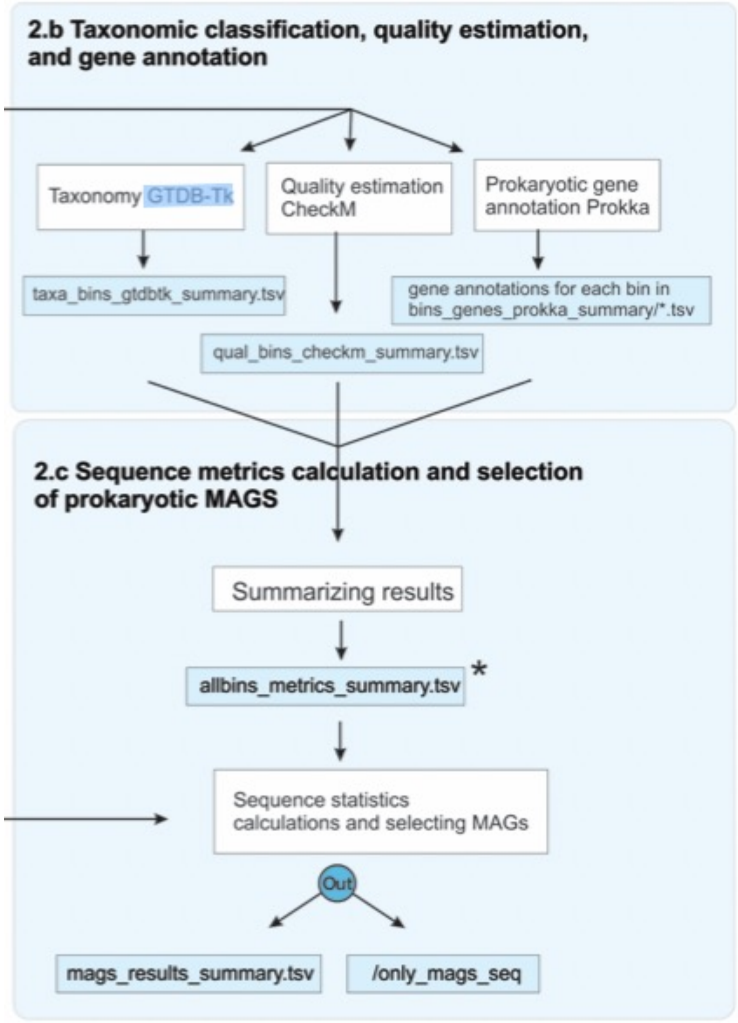
Different GC content!

SILVA to categorize Phylogenetic assignments

```
71 Bacteria;Cyanobacteria;Cyanobacteriia;Cyanobacteriales;Oscillatoriaceae;Oscillatoria
62 Bacteria;Actinobacteriota;Actinobacteria;Micrococcales;Microbacteriaceae;Microbacterium;Microbacterium
59 Bacteria;Cyanobacteria;Cyanobacteriia;Cyanobacteriales;Nostocaceae;Nostoc
53 Bacteria;Actinobacteriota;Actinobacteria;Micrococcales;Micrococcaceae;Arthrobacter;Arthrobacter
43 Bacteria;Cyanobacteria;Cyanobacteriia;Synechococcales;Cyanobiaceae;Cyanobium
42 Bacteria;Cyanobacteria;Cyanobacteriia;Cyanobacteriales;Phormidiaceae;Tychonema
35 Bacteria;Cyanobacteria;Cyanobacteriia;Pseudanabaenales;Pseudanabaenaceae;Pseudanabaena
25 Bacteria;Cyanobacteria;Cyanobacteriia;Cyanobacteriales;Nostocaceae;Rivularia
25 Bacteria;Actinobacteriota;Actinobacteria;Micrococcales;Micrococcaceae;Pseudarthrobacter;Arthrobacter
```

The 16S data does not show Gloeocapsa as the top hit

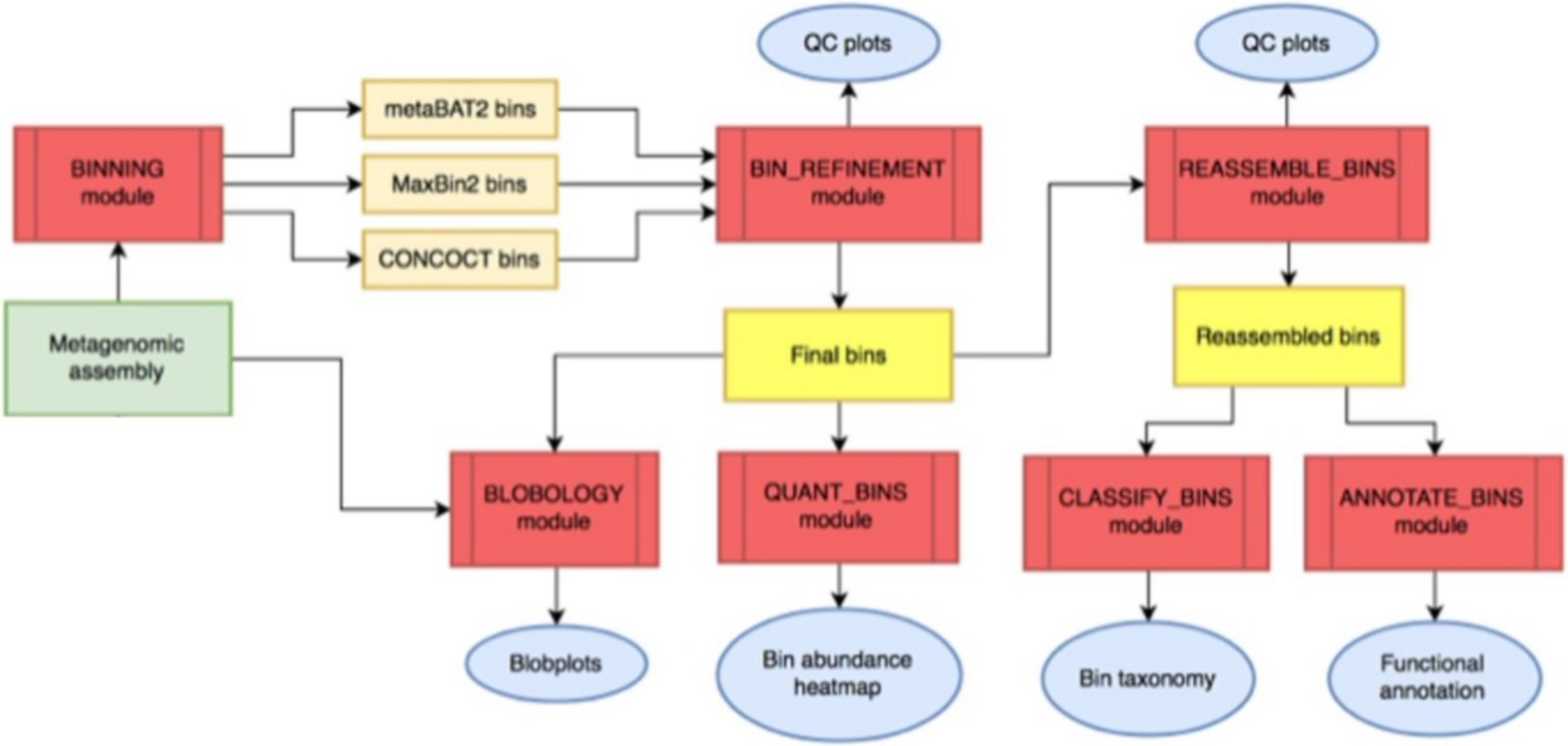
Next Steps Metawrap


2.b Taxonomic classification, quality estimation, and gene annotation
Taxonomy GTDB-Tk
Quality estimation CheckM
Prokaryotic gene annotation Prokka
taxa_bins_gtdbtk_summary.tsv
gene annotations for each bin in bins_genes_prokka_summary/*.tsv
qual_bins_checkm_summary.tsv
2.c Sequence metrics calculation and selection of prokaryotic MAGS
Summarizing results
allbins_metrics_summary.tsv
*
Sequence statistics calculations and selecting MAGs
Out
mags_results_summary.tsv
/only_mags_seq

Results of Metawrap Analysis

Adelaide Rhodes and Jason Laird

Metawrap



https://github.com/bxlab/metaWRAP/blob/master/Usage_tutorial.md

Metawrap

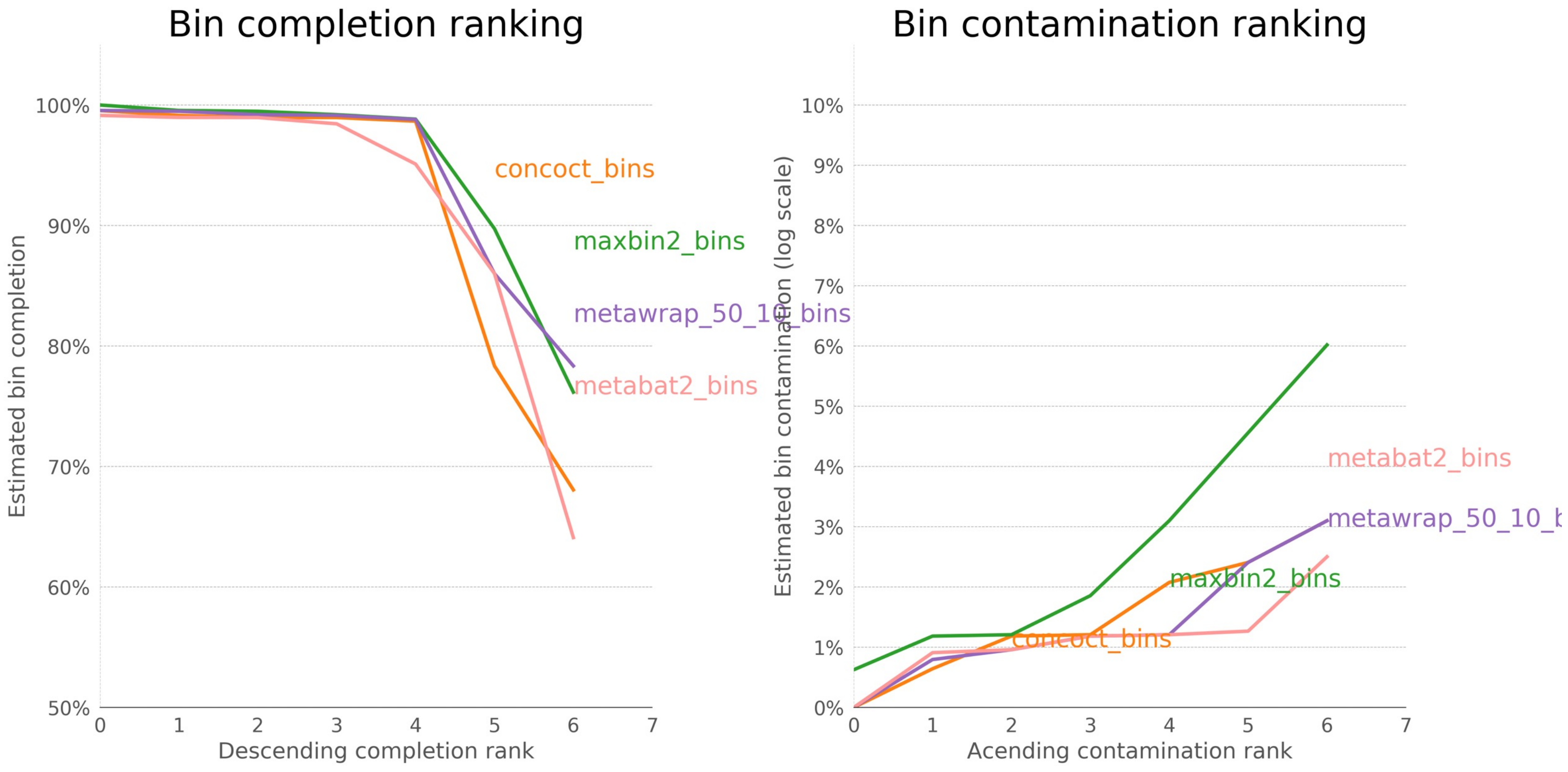
Metawrap is an open source bioinformatics program. It is actually a wrapper script that runs several different programs.
Version: metaWRAP v=1.3.2

Please cite: MetaWRAP - a flexible pipeline for genome-resolved metagenomic data analysis.
These are the other programs used by the wrapper that were integral and shoud be cited:

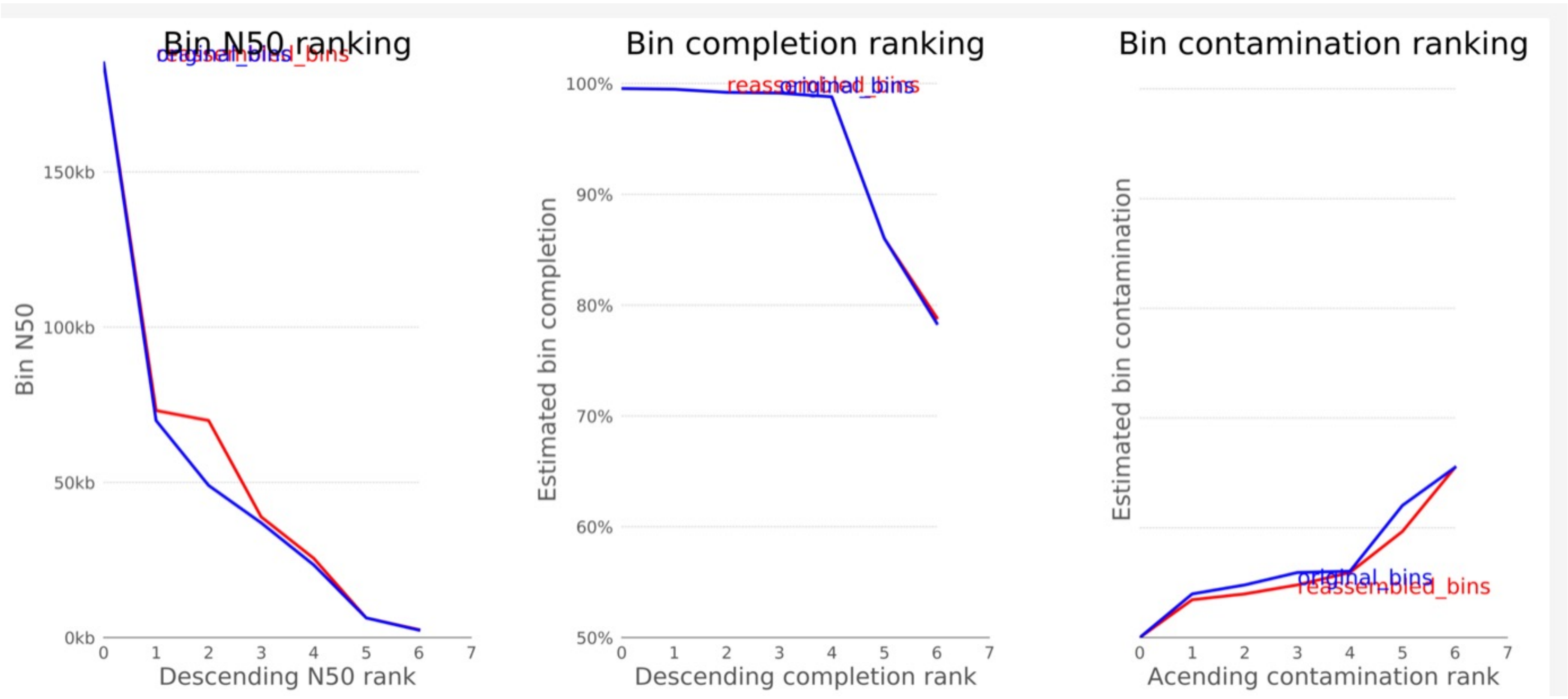
prokka v1.13
Blast-plus v. 2.13.0
QUAST v5.0.2
CheckM v1.0.12
FastQC v0.11.8
Trimmomatic v0.36
metaspades –
  SPAdes version: 3.15.4
  Python version: 3.8.13
  OS: Linux-3.10.0-1127.el7.x86_64-x86_64-with-glibc2.10
MetaBAT: Metagenome Binning based on Abundance and Tetranucleotide frequency (version 2:2.15 (Bioconda); 2020-01-04T21:10:40)
by Don Kang (ddkang@lbl.gov), Feng Li, Jeff Froula, Rob Egan, and Zhong Wang (zhongwang@lbl.gov)
Bowtie2 – v2.2.3
Samtools - v0.1.18
Blobology – which version? https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3843372/
Concoct – v 1.0.0
MaxBins – v2.2.6
Minimap v2.24
Alvis v1.2
TaxatorTK v1.3.3e

Bin completion - Bin Refinement

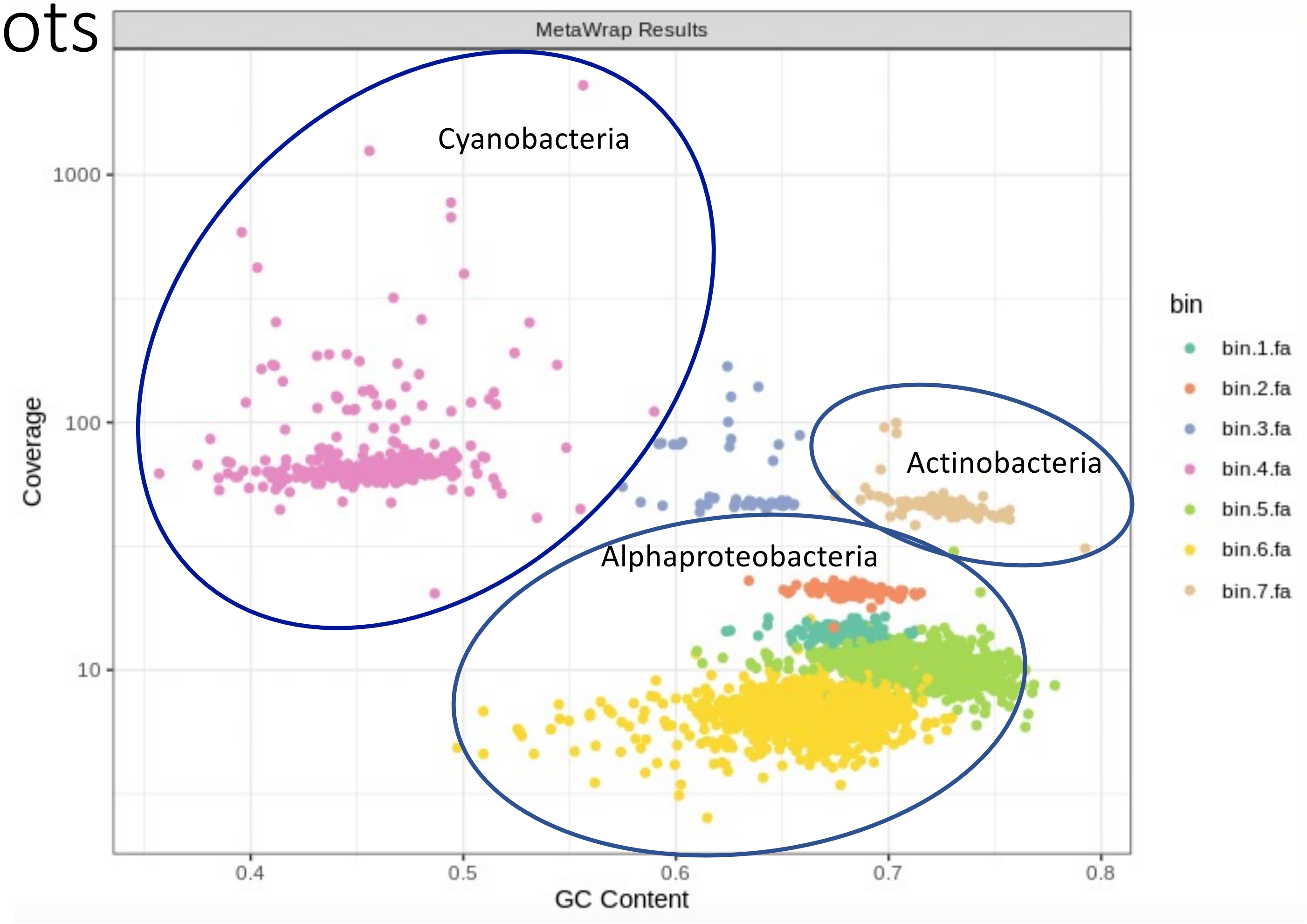
– use best results after three different binning programs – 7 bins identified



Reassembly Resulting Bins



Blob Plots

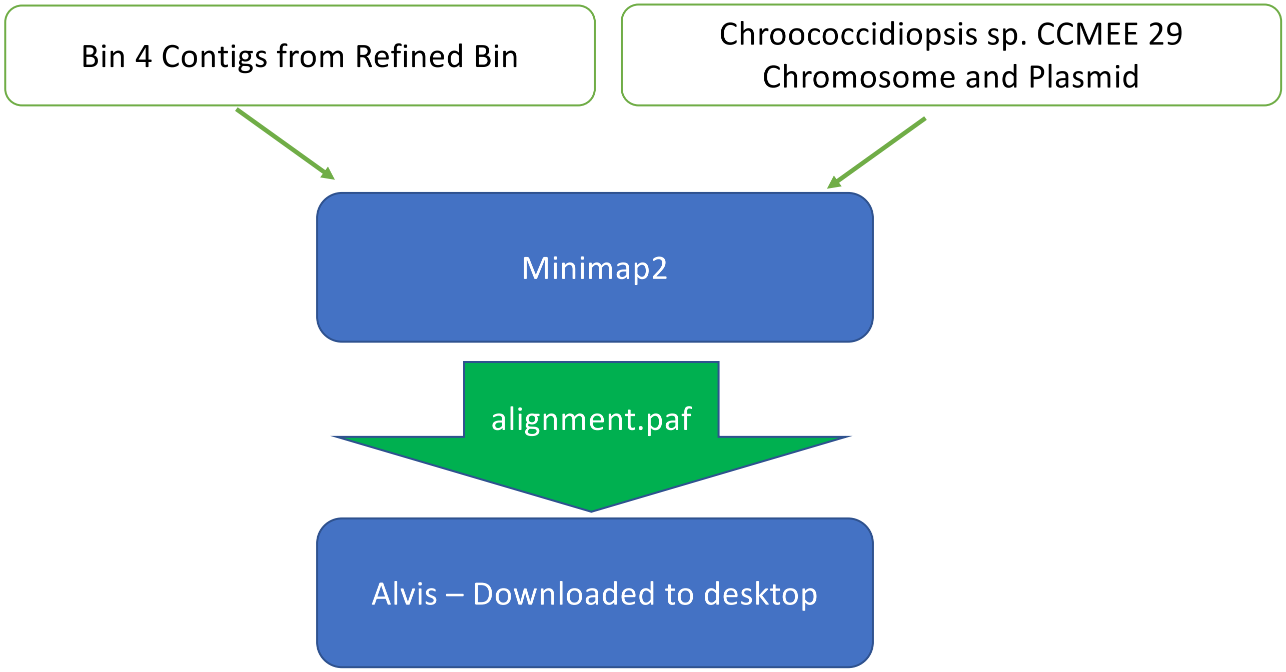

MetaWrap Results
Cyanobacteria
Actinobacteria
Alphaproteobacteria
1000
100
10
Coverage
0.4
0.5
0.6
0.7
0.8
GC Content
bin
bin.1.fa
bin.2.fa
bin.3.fa
bin.4.fa
bin.5.fa
bin.6.fa
bin.7.fa

Classification of bins

- In the next table we present the classification combining a few results from CheckM (reassembled_bin.stats) and TaxatorTK (bin_taxnonomy.tab)
- CheckM: While a rough set of markers is used to identify higher taxonomic orders (e.g. Sphingomonadales) is used to define "completeness", it is not a measure of completeness of the genome. It just measures the presence and absence of expected genes for that taxa.
- https://github.com/Ecogenomics/CheckM/wiki/Genome-Quality-Commands#qa
- We used the entire ncbi_nt database, megablast and Taxator TK to make more specific calls.

Bin	Completeness %	Contamination %	GC %	Lowest Taxonomic ID	N50	Size
bin.1.orig	98.8	0.796	0.678	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales; **Sphingomonadaceae**	38,889	3,050,608
bin.2.orig	99.14	0	0.68	Bacteria;Proteobacteria;Alphaproteobacteria;Hyphomicrobiales;Boseaceae;Bosea; **Bosea sp. RAC05**	69,930	4,833,064
bin.3.orig	99.2	3.1	0.639	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales; **Sphingomonadaceae**	185,638	4,538,885
bin.4.permissive	**99.55**	**1.185**	**0.464**	**Bacteria;Cyanobacteria;Chroococcidiopsidales;Chroococcidiopsidaceae;Chroococcidiopsis; Chroococcidiopsis sp. CCMEE 29**	**25,544**	**5,554,811**
bin.5.orig	86.01	0.959	0.709	Bacteria;Proteobacteria;Alphaproteobacteria; **Hyphomicrobiales**	6,331	4,787,219
bin.6.permissive	78.86	1.934	0.662	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae; **Brevundimonas**	2,570	2,523,321
bin.7.permissive	99.48	0.69	0.722	Bacteria;Actinobacteria;Actinomycetia;Propionibacteriales; **Nocardioidaceae**	73,144	3,770,891

Aligning to Reference and Visualizing



Bin 4 Contigs from Refined Bin

Chroococcidiopsis sp. CCMEE 29 Chromosome and Plasmid

Minimap2

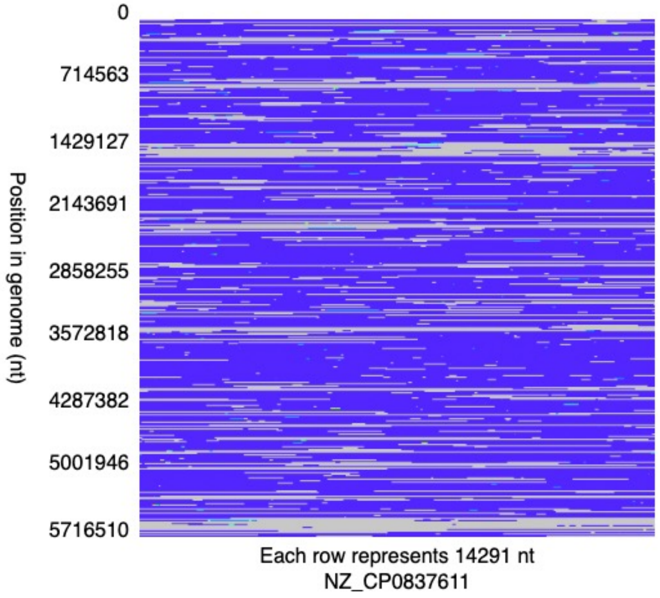
alignment.paf

Alvis – Downloaded to desktop

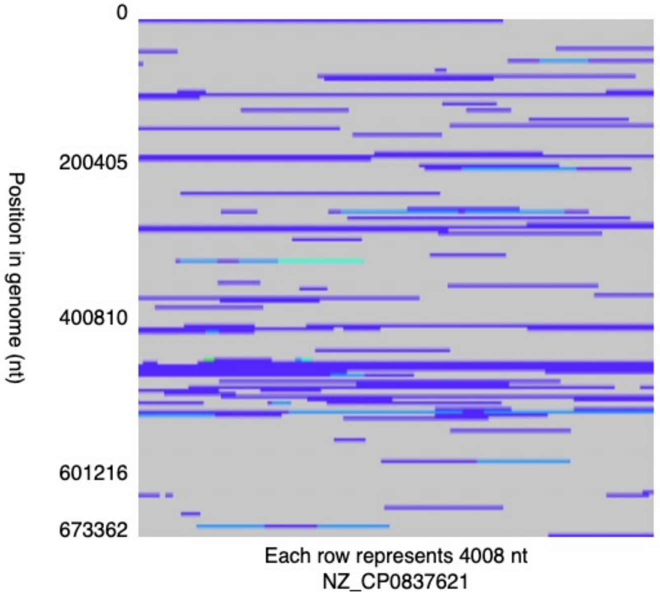
Aligned bin.4 to Reference from NCBI

Reference: Chroococcidiopsis sp. CCMEE 29

Chromosome



Plasmid



0
Coverage
6

The coverage is low because we aligned assembled contigs to the genome

The plasmid should be different because we are very far apart and have different symbionts

Proposed Next Steps to Publish or Get a Complete Genome

TO PUBLISH METAGENOME:

https://ena-docs.readthedocs.io/en/latest/submit/assembly/metagenome/primary.html

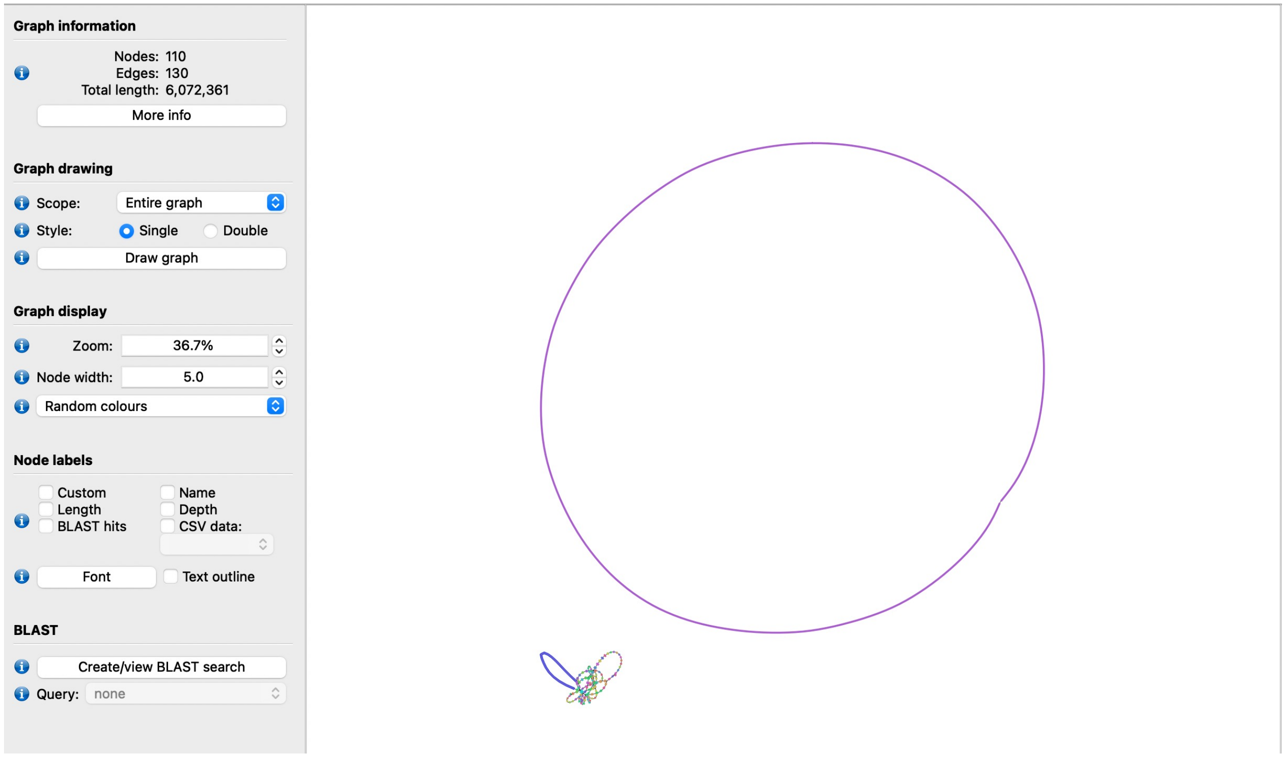
- 4316384 paired end reads (4316384 forward + 4316384 reverse)
- Illumina? (which sequencer) (/assembly.meta.quast/report.txt)
- Assemble with metaSPAdes from SPAdes v3.15.4

Assembly	scaffolds
# contigs	9076
Largest contig	580104
Total length	34675725
GC (%)	64.68
N50	22935
N75	3570
L50	262
L75	1328
# N's per 100 kbp	25.98

TO CONTINUE:

- Get a long read assembly if possible
  - Oxford Nanopore – relatively inexpensive for multiple runs
- Separate plasmids using PCR and sequence independently
- https://www.sciencedirect.com/science/article/pii/S1319562X19302529
- Some of the other bins have almost complete genomes – what do you want to do with those?

Current Status of the Assembly After Long Read Data



Graph information

Nodes: 110
Edges: 130
Total length: 6,072,361

Graph drawing

Scope: Entire graph
Style: ● Single ○ Double

Graph display

Zoom: 36.7%
Node width: 5.0
Random colours

Node labels

□ Custom □ Name
□ Length □ Depth
□ BLAST hits □ CSV data:

□ Text outline

BLAST

Query: none