

GORG-Dark QC

Input:
13,113 SAGs
8,220,159,968 bp

Method details:

/mnt/storage/data/daily-data/day2/sag_qc

Assembly size
cut-off



Removal of :
< 10% completeness
Unclassified by GTDB-tk

9,220 SAGs
7,322,312,463 bp

Removal of
reagent
contaminants



Removal of :
'Bradyrhizobium'
'Ralstonia'
...

9,218 SAGs
7,320,830,119 bp

Removal of
multiple-genomes
assemblies

Examined by TFA,
GUNC, and
MDMcleaner

Removal of :
assemblies or
contigs



Manual
examination

Output:
9,182 SAGs
7,273,645,506 bp

- *Assembly size cut-off* (CheckM1-estimated completion $\geq 10\%$ + GTDB-tk r202 classified)
- *Removal of reagent contaminants* (e.g., Bradyrhizobium, BLASTN—nt)
- *Assemblies containing sequences from multiple marine prokaryoplankton genomes*
 - *removal of the entire assembly*
 - tetramer PCA, GUNC, MDMcleaner, CheckM + BLASTN—nr
 - *removal of individual contigs*
 - tetramer PCA outliers + MDMcleaner + BLASTN—nr +
 - keep virus-like: DeepVirFinder2, VIBRANT, Virsorter2 + BLASTN—nr