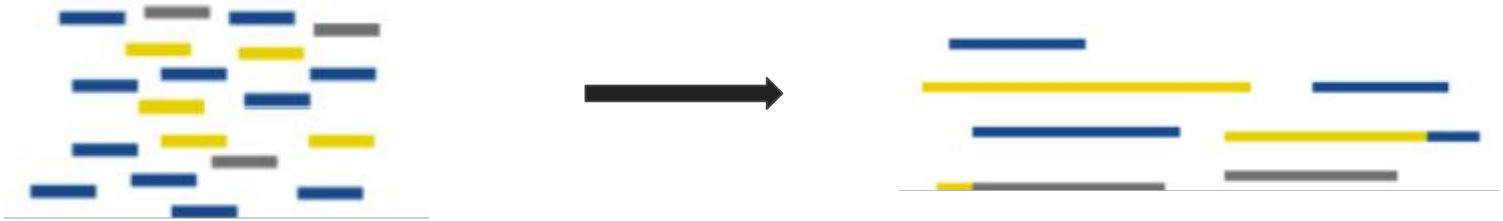
Tutorial 7: Metagenome De Novo Assembly and Binning - KBase

Gavin Hearne, Hyunwoo Yoo

Background

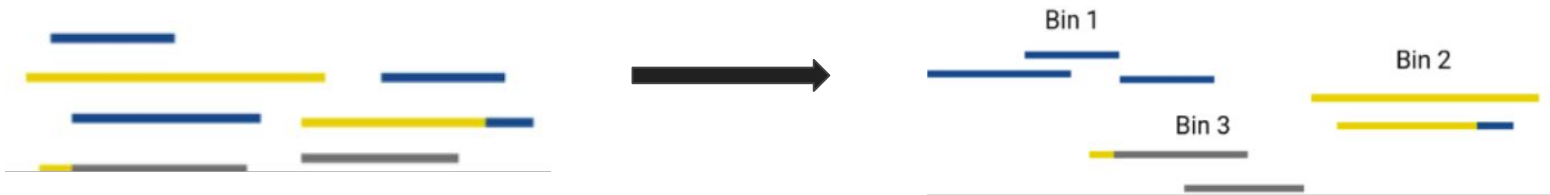
What is Metagenome De Novo Assembly?

- Uses small DNA pieces called reads from environmental samples to rebuild the original genome sequence
- Assembles the genome using only the sequencing data, without a reference genome
- Outcome of this process is called 'contigs,' which are long pieces of DNA sequences
- Contigs are the reconstructed segments of the original genome



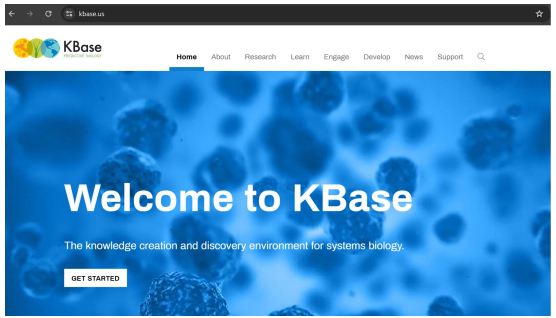
What is Binning?

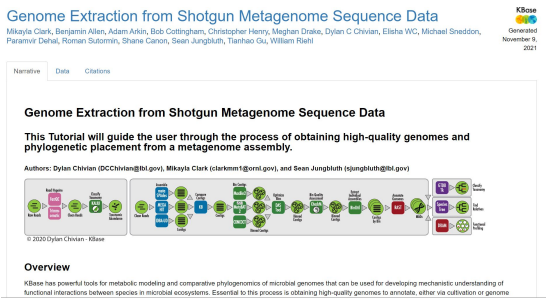
- The process of grouping contigs into bins
- Bin represents a potential genome from different organisms in the sample.
- This is done by analyzing patterns in the DNA sequences and their abundance
- Helps in organizing the complex data from environmental samples, making it easier to study the microbial diversity and the genetic makeup of individual species within the sample.



What is KBase?

- Web platform for performing De Novo Assembly and other bioinformatics tasks
- Provides tools and workflows to analyze environmental DNA samples
- Users can rebuild genomes from sequencing data using KBase's resources
- Makes it easier to study microbial communities and their functions





Workflows

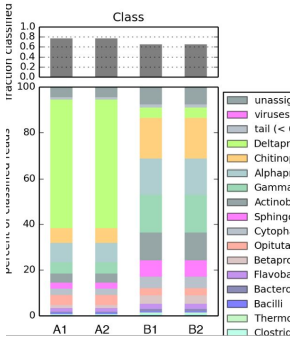
- Read Hygiene
- Classify Taxonomy
- Assemble
- Compare Contigs
- Bin Contigs
- Optimize Binned Contigs by Consensus
- Bin Quality Assessment
- Extract Individual Assemblies
- Annotate Genomes
- Taxonomic Classification of MAGs

Read Hygiene

- Paired-end reads in FASTQ format are imported
  - Paired-end reads?
    - Refer to sequences where both ends of a DNA or RNA fragment are sequenced.
- Improves the quality using FastQC and Trimmomatic
  - FastQC?
    - Provides a comprehensive analysis and visualization of the quality of data
  - Trimmomatic?
    - Removes adapter sequences, trimming low-quality sequence regions

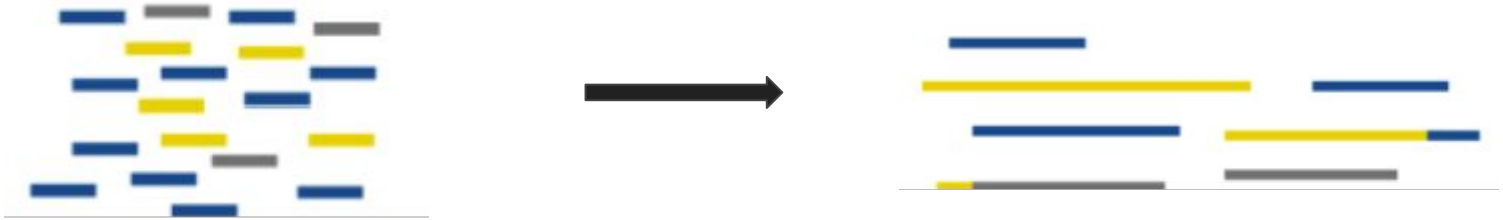
Classify Taxonomy

- Kaiju is used to predict microbial composition based on protein similarity, and from this, a species tree is generated
  - Kaiju?
    - Tool designed for analyzing the taxonomic composition of microbial communities
    - Identifies prokaryotes, viruses, and eukaryotes quickly and accurately



Assemble

- Reads are assembled to create scaffolding for the entire genome
- Multiple assembly apps are run to compare their results
  - using metaSPAdes, MEGAHIT, IDBA-UD
    - All uses De Bruijn graph approach



De Bruijn graph approach

1. **k-mer Splitting**

Breaking down DNA sequences (reads) into all possible subsequences of length k, known as k-mers

2. **Graph Construction**

Connecting nodes in the graph based on the continuity of k-mers

3. **Path Tracing and Assembly**

Finding paths from the start node to the end node in the De Bruijn graph and following these paths to assemble the sequence

Assemble

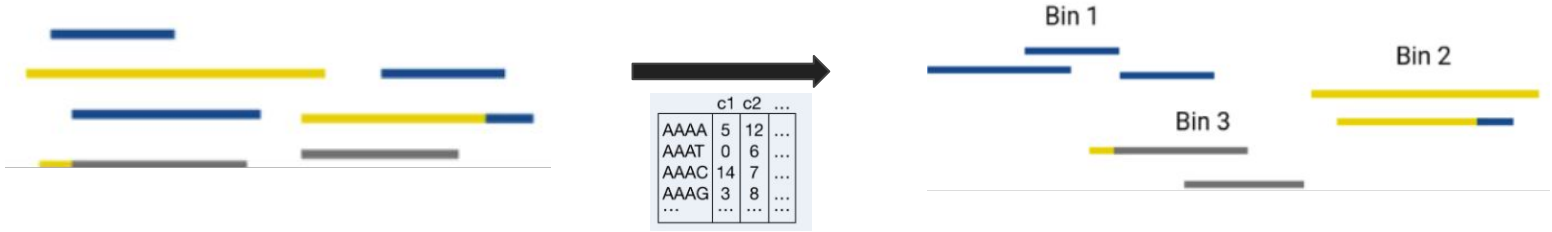
- Reads are assembled to create scaffolding for the entire genome
- Multiple assembly apps are run to compare their results
  - metaSPAdes vs. MEGAHIT vs. IDBA-UD
    - MEGAHIT is optimized for memory efficiency
    - IDBA-UD employs iterative k-mer size increases
    - Generally, MEGAHIT is faster and sound for large data but slightly inaccurate
    - metaSPAdes and IDBA-UD are more accurate but computationally expensive

Compare Contigs

- Sets of contigs generated using assemblers are compared with the quality metrics N50, L50
    - N50 refers to the minimum length of contigs, required to cover more than half of the total genome length
    - A high N50 value is desirable because it implies that the assembly has successfully captured longer regions of the genome without breaking them into smaller pieces
    - L50 signifies the minimum number of contigs required to cover half of the total genome length
    - A lower L50 value is preferred because it indicates that a significant portion of the genome can be represented with fewer, longer contigs

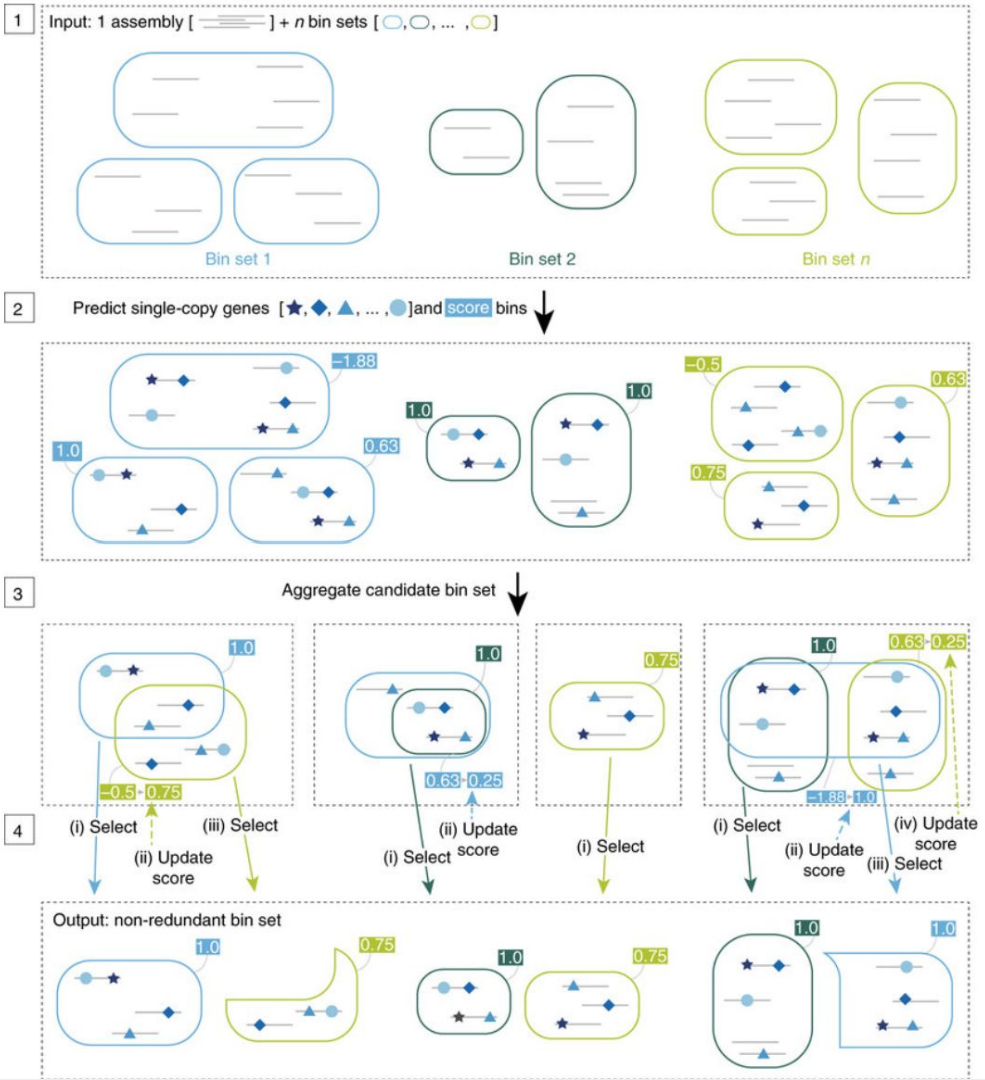
Bin Contigs

- Assembled contigs are clustered into bins, representing a hypothetical genome
  - MaxBin2
    - Employes k-mer frequencies and differential coverage
  - MetaBAT2
    - Utilizes tetranucleotide frequency and contig coverage patterns to group contigs
  - CONCOAT
    - Combines both the coverage and the nucleotide composition to cluster contigs



Optimize Binned Contigs by Consensus

- The quality of binned contigs is improved through consensus assignments using DAS-Tool, employing multiple methods
  - DAS-Tool?
    - Differential Abundance Score Tool is a comprehensive and advanced tool designed for integrating and refining the results of microbial genome reconstructions from metagenomic data


1
Input: 1 assembly [ ] + n bin sets [, , ... ,]
Bin set 1
Bin set 2
Bin set n
2
Predict single-copy genes [★, ◆, ▲, ... , ●] and score bins
−1.88
1.0
0.63
1.0
1.0
−0.5
0.63
0.75
3
Aggregate candidate bin set
1.0
−0.5
0.75
1.0
0.63
0.25
0.75
1.0
0.63
0.25
−1.88
1.0
4
(i) Select
(ii) Update score
(iii) Select
(i) Select
(ii) Update score
(i) Select
(i) Select
(ii) Update score
(iii) Select
(iv) Update score
Output: non-redundant bin set
1.0
0.75
1.0
0.75
1.0
1.0

Bin Quality Assessment

- The quality of bins is evaluated using CheckM, and high-quality bins are filtered out
  - CheckM?
    - CheckM is a software tool designed for assessing the quality of microbial genomes recovered from isolates, single-cell sequencing, or metagenomic assemblies.

Extract Individual Assemblies

- High-quality bins are extracted as Assembly objects to be utilized in downstream applications.
  - BinUtil?
    - BinUtil is a tool designed for extracting high-quality bins as independent Assembly objects from metagenomic datasets
    - It facilitates the reconstruction and analysis of individual microbial genomes within complex communities
    - It plays a crucial role in downstream bioinformatics analyses, enabling detailed study of microbial diversity and function

Annotate Genomes

- High-quality bins are converted into annotated genomes using RASTtk
  - RASTtk?
    - RASTtk (Rapid Annotation using Subsystem Technology toolkit) is a software tool used for annotating bacterial and archaeal genomes
    - It uses a combination of similarity-based and model-based methods to predict genes and their functions

Taxonomic Classification of MAGs

- GTDB-Tk Classify is used to provide a phylogenetic classification for MAGs
  - GTDB-Tk Classify?
    - GTDB-Tk (Genome Taxonomy Database Toolkit) classifies microbial genomes based on the Genome Taxonomy Database
    - GTDB is a updated database that provides a standardized taxonomy for bacteria and archaea, which is based on genome phylogeny
  - MAG?
    - A Metagenome-Assembled Genome (MAG) is a collection of contigs derived from metagenomic datasets that are assembled and binned together
    - It represents the genome of a single microbial species or strain present within a complex microbial community

Kbase Usage

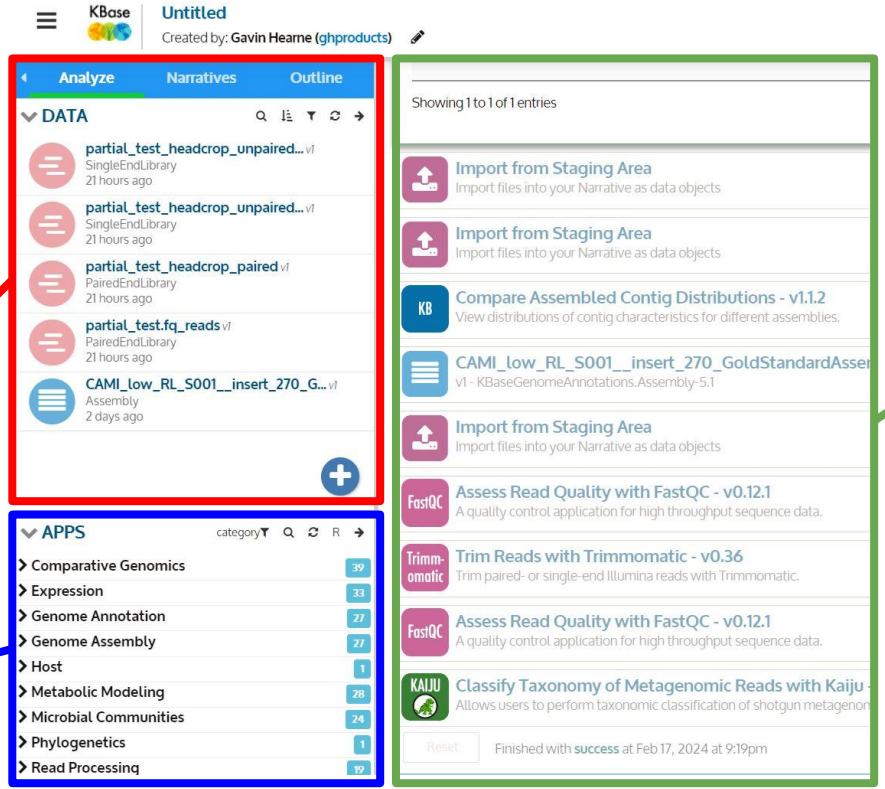
Kbase Narratives

The narrative is the workspace for kbase.

Here you can import data, utilize all the functions of kbase through apps, and view outputs.

Import and select from available data

Select apps and functions



Narrative interface: here you can see the apps you have selected and their outputs

Data Pre-processing

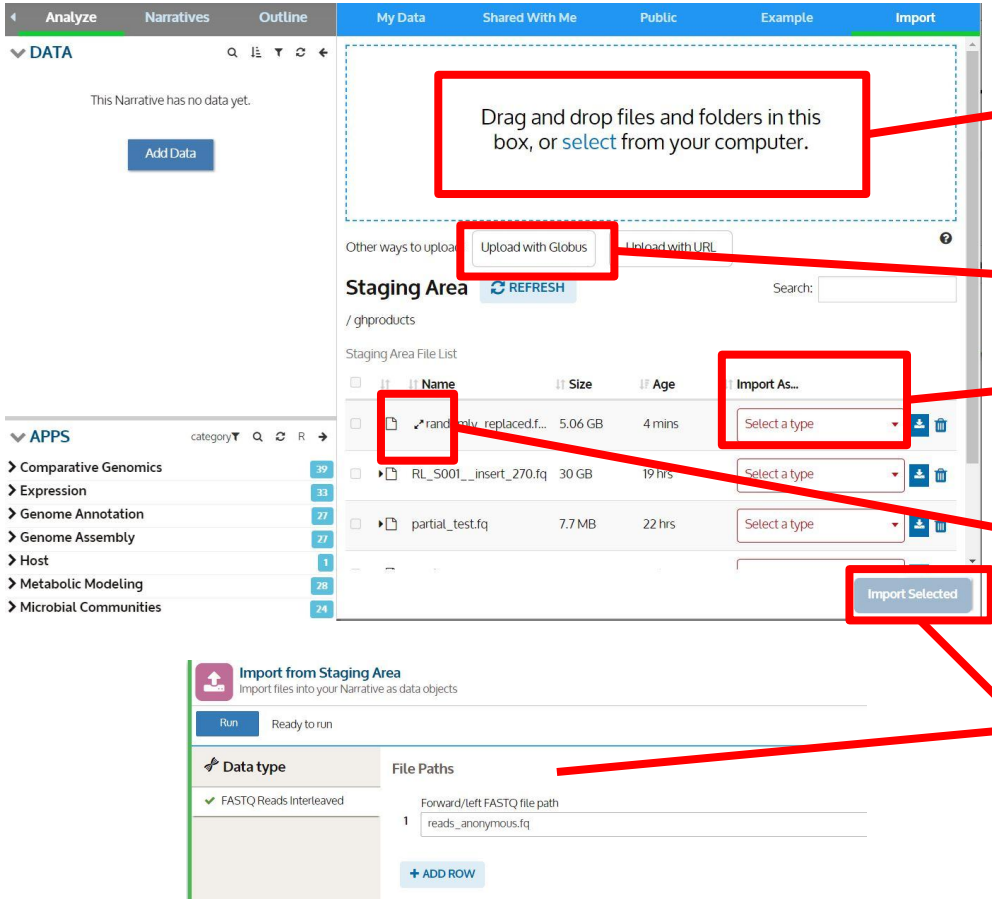
Due to invalid characters preventing the upload of the fastq metagenome sample, we needed a data preprocessing tool.

```
81   ERROR on Line 93258: Invalid character ('S') in base sequence.
82   ERROR on Line 94738: Invalid character ('S') in base sequence.
83   ERROR on Line 105246: Invalid character ('W') in base sequence.
84   ERROR on Line 134718: Invalid character ('Y') in base sequence.
85   ERROR on Line 144210: Invalid character ('K') in base sequence.
86   ERROR on Line 168330: Invalid character ('R') in base sequence.
87   ERROR on Line 168330: Invalid character ('R') in base sequence.
```

As a quick fix, we replace the incorrect nucleotides with a random ATCG, and reduce the corresponding quality scores to zero

```python
def remove_incorrect(fasta_dir):
    '''
    replace with random but with a quality of zero
    '''
    new_records = []
    bases = ['A', 'T', 'C', 'G']
    for record in SeqIO.parse(fasta_dir, "fastq"):
        #remove incorrect bases
        sequence = str(record.seq)
        index = re.search('[^ATCG]', sequence)

        if index is None:
            new_records.append(record)
            continue

        new_sequence = re.sub('[^ATCG]', random.choice(bases), sequence)
        record.seq = Seq(new_sequence)

        #remove annotations
        letter_annotations = record.letter_annotations
        record.letter_annotations = {}
        letter_annotations['phred_quality'][index.start()] = 0
        new_letter_annotations = {'phred_quality': letter_annotations['phred_quality']}
        record.letter_annotations = new_letter_annotations

        new_records.append(record)

    with open("incorrect_removed_" + fasta_dir, 'w') as output_handle:
        SeqIO.write(new_records, output_handle, "fastq")
    print(len(new_records))
```

Importing data



For small files (>5gb), it is possible to upload data directly through the kbase interface.

Otherwise, it is necessary to use Globus

Our filetype is fastq interleaved

Compressed files can be uploaded, but they need to be uncompressed before importing to the narrative

Importing the selected data will automatically add this app which verifies and imports the file as a data object

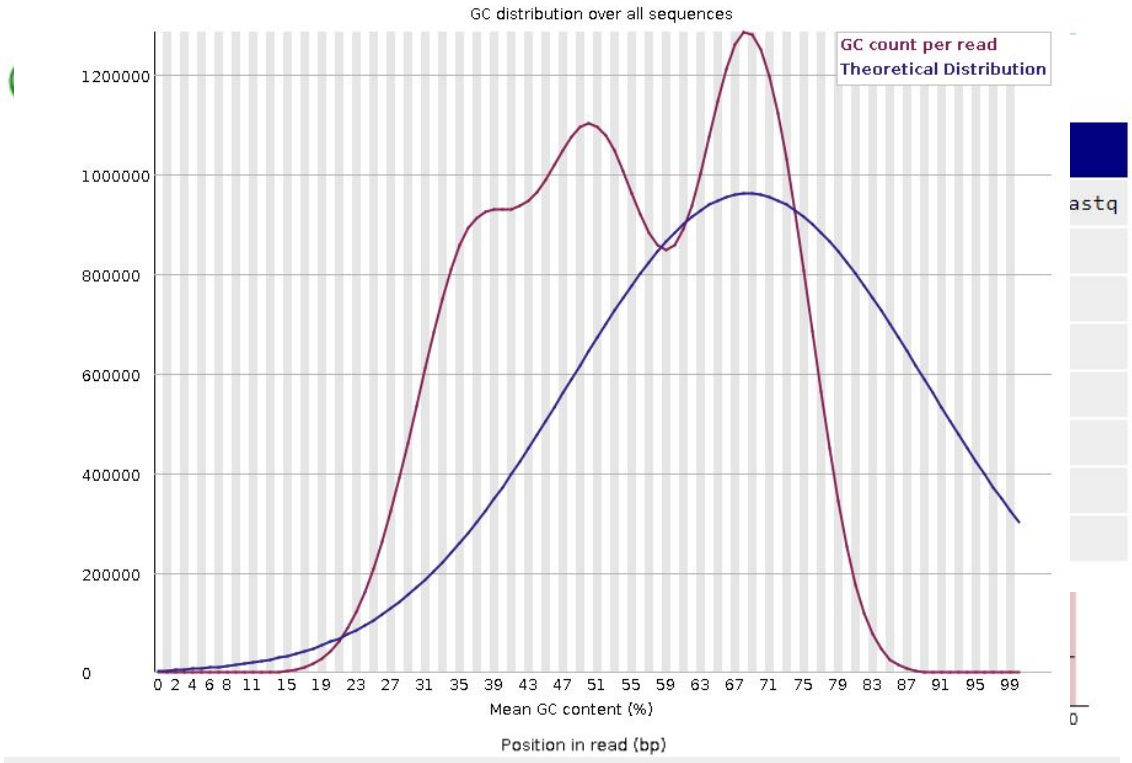
Assess Read Quality

FastQC Assess Read Quality with FastQC - v0.12.1
A quality control application for high throughput sequence data.

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

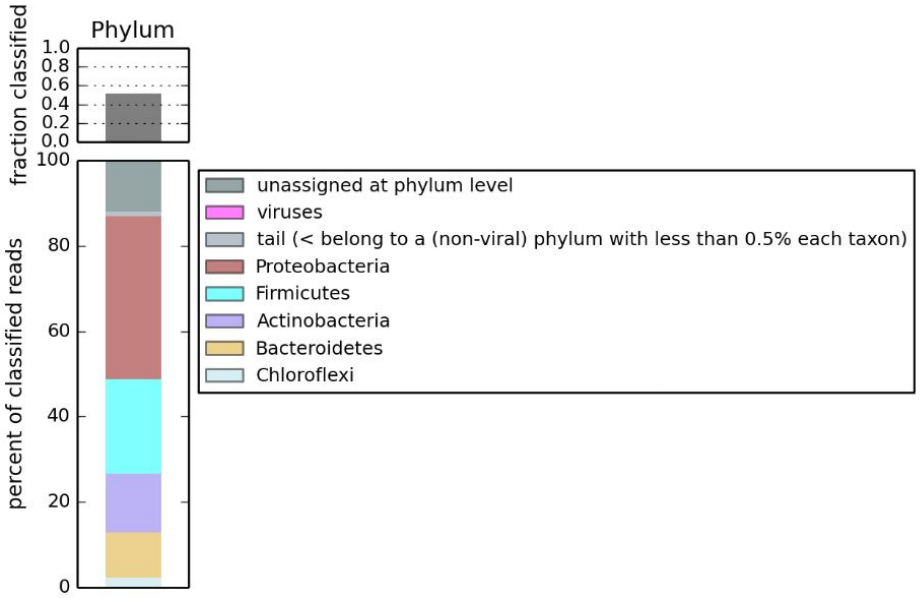
Per sequence GC content

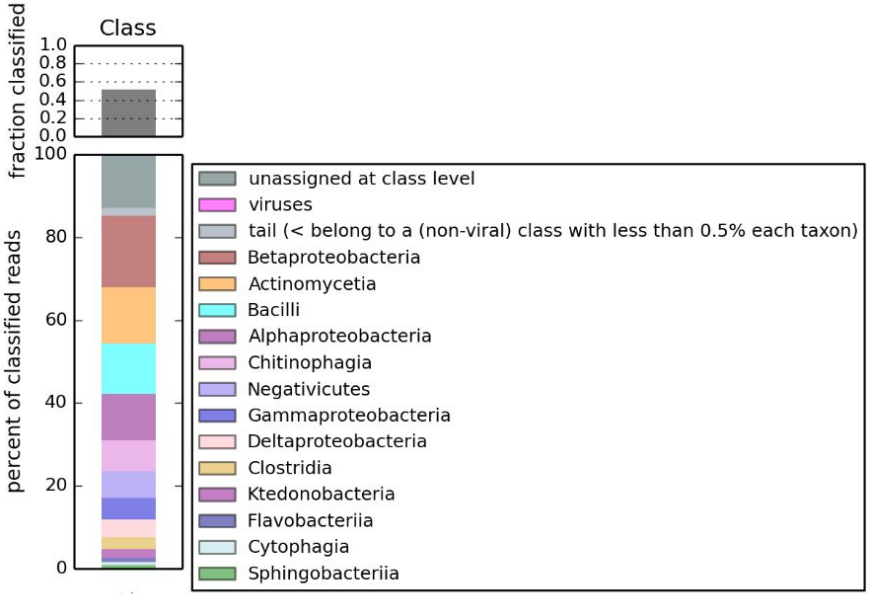


Classify Taxonomy - Kaiju

Classify Taxonomy of Metagenomic Reads with Kaiju - v1.9.0

Allows users to perform taxonomic classification of shotgun metagenomic read data with Kaiju.





Assembly results

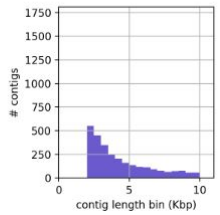
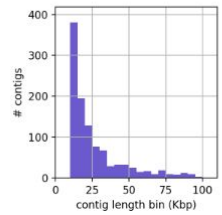
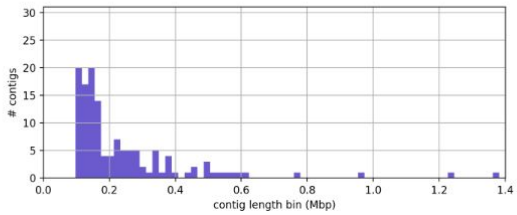
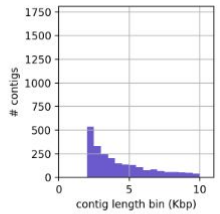
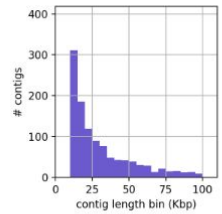
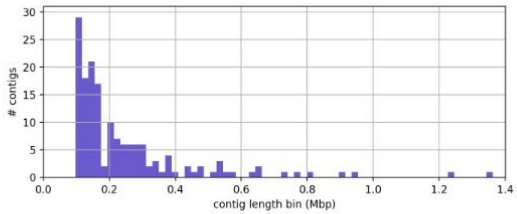
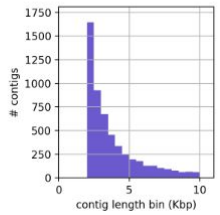
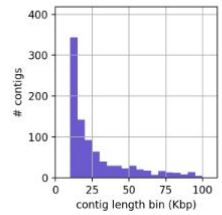
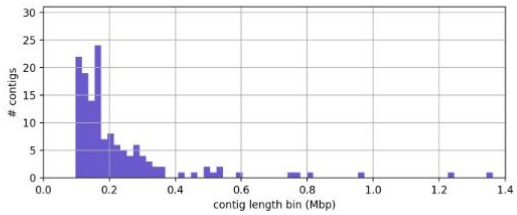
Assemble Reads with metaSPAdes - v3.15.3
Assemble metagenomic reads using the SPAdes assembler.

Assemble Reads with MEGAHIT v1.2.9
Assemble metagenomic reads using the MEGAHIT assembler.

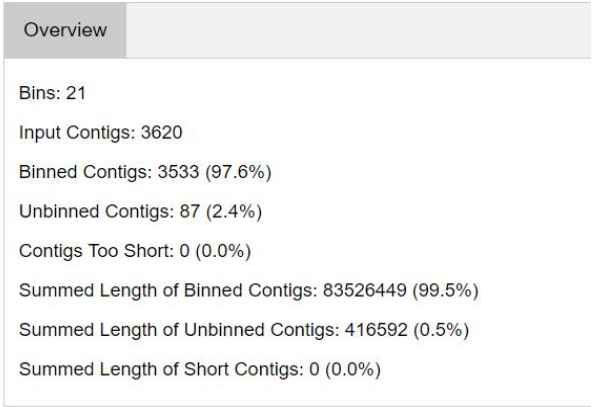
Compare Assembled Contig Distributions - v1.1.2
View distributions of contig characteristics for different assemblies.

Assemble Reads with IDBA-UD - v1.1.3
Assemble paired-end reads from single-cell or metagenomic se

BEST WORST

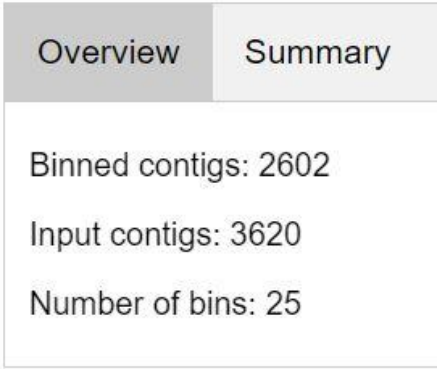
ASSEMBLY	LONGEST CONTIG (bp)	Nx (Lx)		LENGTH (bp)	NUM CONTIGS	SUM LENGTH (bp)	Contig Length Histogram (1bp <= len < 10Kbp)	Contig Length Histogram (10Kbp <= len < 100Kbp)	Contig Length Histogram (len >= 100Kbp)
IDBA.contigs	1362924	N50:	73077	>= 10^6	2	2591304			
		L50:	(183)	>= 10^5	131	32044837			
		N75:	16441	>= 10^4	1205	60309663			
		L75:	(749)	>= 10^3	4050	72691075			
		N90:	5767	>= 500	4050	72691075			
		L90:	(1878)	>= 1	4050	72691075			
metaSPAdes.Assembly	1362255	N50:	89948	>= 10^6	2	2590385			
		L50:	(184)	>= 10^5	160	39747091			
		N75:	27588	>= 10^4	1276	73840281			
		L75:	(615)	>= 10^3	3620	83943041			
		N90:	8142	>= 500	3620	83943041			
		L90:	(1467)	>= 1	3620	83943041			
MEGAHIT.assembly	1362697	N50:	63929	>= 10^6	2	2591041			
		L50:	(222)	>= 10^5	140	32595372			
		N75:	9410	>= 10^4	1052	58036617			
		L75:	(1125)	>= 10^3	6441	78320115			
		N90:	3360	>= 500	6441	78320115			
		L90:	(3352)	>= 1	6441	78320115			

Bin Contigs



MaxBin

CONCOT



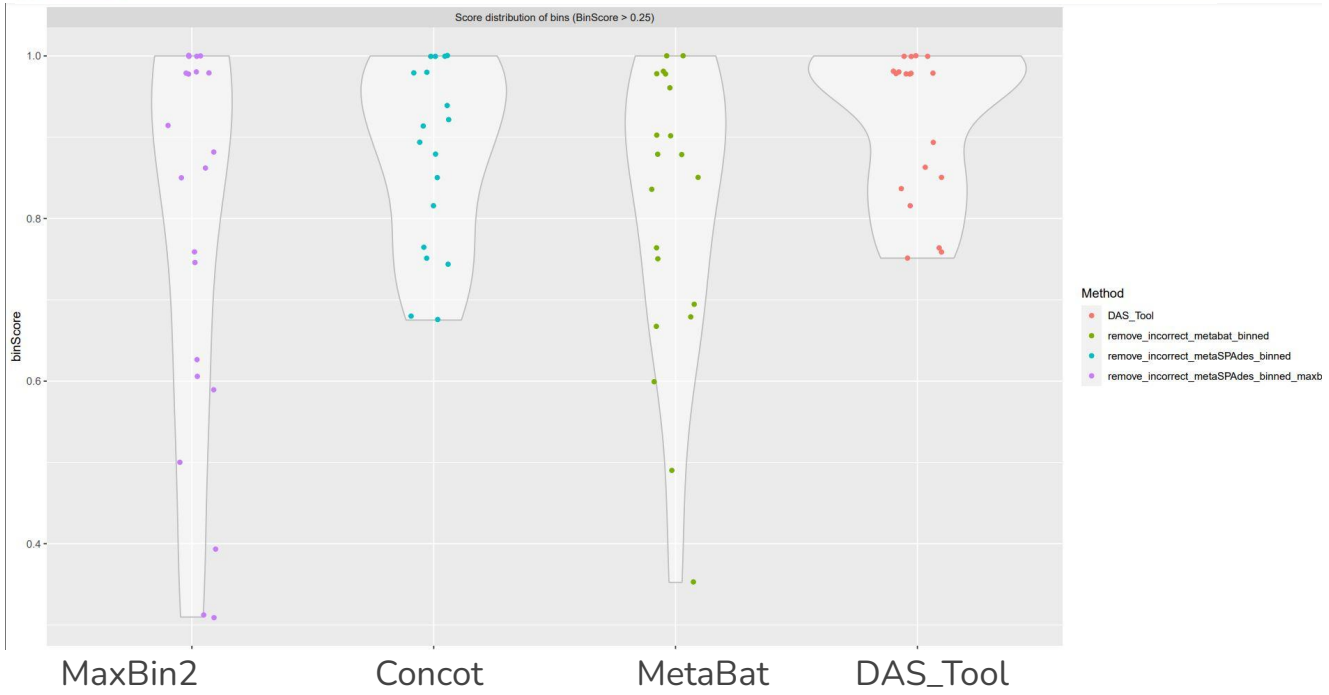
MetaBAT2

Refine Bins by Consensus

DAS-Tool

Optimize Bacterial or Archaeal Binned Contigs using DAS Tool - v1.1.2

Optimize bacterial or archaeal genome bins using a dereplication, aggregation and scoring strategy



Bin Quality Assessment

Assess Genome Quality with CheckM - v1.0.18
Runs the CheckM lineage workflow to assess the genome quality of isolates

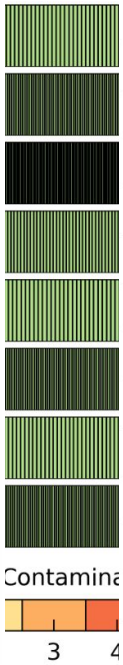
Filter Bins by Quality with CheckM - v1.0.18
Runs the CheckM lineage workflow to assess the genome quality of isolates,

Extract Bins as Assemblies from BinnedContigs - v1.0.2
Extract a bin as an Assembly from a BinnedContig dataset

From an initial 19 bins, we assess the quality and extract 12 for downstream use

The removed bins are marked in red. They do not satisfy the ≥95% completeness and <2% contamination threshold

Bin Name	Marker Lineage	# Genomes	# Markers	# Marker Sets	0	1	2	3	4	5+	Completeness	Contamination
bin.001	f__Rhodobacteraceae	84	568	330	2	562	4	0	0	0	99.6	0.5
bin.002	f__Rhodobacteraceae	84	568	330	6	560	2	0	0	0	98.56	0.3
bin.003	c__Betaproteobacteria	323	387	234	2	383	2	0	0	0	99.15	0.43
bin.004	o__Actinomycetales	488	309	185	0	308	1	0	0	0	100.0	0.18
bin.005	o__Burkholderiales	193	427	214	82	337	8	0	0	0	82.63	2.76
bin.006	o__Clostridiales	304	250	143	7	243	0	0	0	0	96.5	0.0
bin.007	p__Bacteroidetes	364	303	203	71	229	3	0	0	0	73.48	0.54
bin.008	c__Gammaproteobacteria	67	481	276	7	468	6	0	0	0	97.96	1.14
bin.009	c__Deltaproteobacteria	83	247	155	3	244	0	0	0	0	98.06	0.0
bin.010	o__Actinomycetales	148	572	276	6	559	7	0	0	0	98.99	0.71
bin.011	c__Bacilli	750	273	152	7	266	0	0	0	0	96.85	0.0
bin.012	k__Bacteria	924	151	101	2	130	19	0	0	0	98.68	10.89
bin.013	p__Bacteroidetes	364	302	203	14	285	3	0	0	0	98.81	0.99
bin.014	c__Alphaproteobacteria	468	388	250	39	345	4	0	0	0	91.86	1.13
bin.015	p__Firmicutes	100	295	158	1	289	5	0	0	0	99.37	2.22
bin.016	k__Bacteria	3167	126	75	30	96	0	0	0	0	81.25	0.0
bin.017	o__Pseudomonadales	185	813	308	5	805	3	0	0	0	99.02	0.32
bin.018	f__Xanthomonadaceae	55	659	290	152	493	14	0	0	0	79.54	2.72
bin.019	o__Burkholderiales	107	574	251	0	566	8	0	0	0	100.0	1.49



Created Object Name
Bin.001.fasta_assembly
Bin.002.fasta_assembly
Bin.003.fasta_assembly
Bin.004.fasta_assembly
Bin.006.fasta_assembly
Bin.008.fasta_assembly
Bin.009.fasta_assembly
Bin.010.fasta_assembly
Bin.011.fasta_assembly
Bin.013.fasta_assembly
Bin.017.fasta_assembly
Bin.019.fasta_assembly
extracted_bins.AssemblySet

Annotate Metagenome

Annotate Multiple Microbial Assemblies with RASTtk - v1.073

Annotate bacterial or archaeal assemblies and/or assembly sets using RASTtk (Rapid Annotations using Subsystems Technology toolkit).

Created Object Name
annotated_metagenome
Bin.001.fasta_assembly.RAST
Bin.002.fasta_assembly.RAST
Bin.003.fasta_assembly.RAST
Bin.004.fasta_assembly.RAST
Bin.006.fasta_assembly.RAST
Bin.008.fasta_assembly.RAST
Bin.009.fasta_assembly.RAST
Bin.010.fasta_assembly.RAST
Bin.011.fasta_assembly.RAST
Bin.013.fasta_assembly.RAST
Bin.017.fasta_assembly.RAST
Bin.019.fasta_assembly.RAST

```
The RAST algorithm was applied to annotating a genome sequence comprised of 141 contigs containing 4517581 nucleotides.
No initial gene calls were provided.
Standard features were called using: glimmer3; prodigal.
A scan was conducted for the following additional feature types: rRNA; tRNA; selenoproteins; pyrrolysoproteins; repeat r
egions; crispr.
The genome features were functionally annotated using the following algorithm(s): Kmers V2; Kmers V1; protein similarit
y.
In addition to the remaining original 0 coding features and 0 non-coding features, 4592 new features were called, of whi
ch 163 are non-coding.
Output genome has the following feature types:
        Coding gene                    4429
        Non-coding crispr_array           1
        Non-coding crispr_repeat         20
        Non-coding crispr_spacer         19
        Non-coding repeat                79
        Non-coding rna                   44
Overall, the genes have 0 distinct functions.
The genes include 0 genes with a SEED annotation ontology across 0 distinct SEED functions.
The number of distinct functions can exceed the number of genes because some genes have multiple functions.
Bin.003.fasta_assembly succeeded!
```

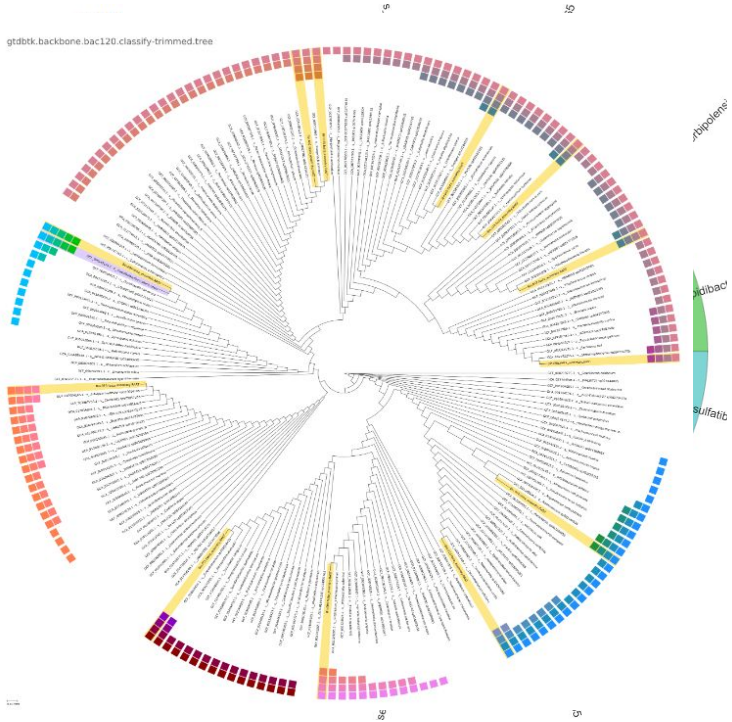
Taxonomic Classification

Classify Microbes with GTDB-Tk - v2.3.2

Obtain objective taxonomic assignments for bacterial and archaeal genomes based on the Genome Taxonomy Database (GTDB)

GTDB Trees | Krona Chart | Bacteria | Archaea | Bacteria Marker Summary | Archaea Marker Summary

User Genome	Classification	FastA Refere	Closest	ssification Method	Note	Other Related References	MSA AA Percent
Bin.001.fasta_assembly	d__Bacteria; p__Pseudomonadota; c__Alphaproteobacteria; o__Rhodobacterales; f__Rhodobacteraceae; g__Albidovulum; s__Albidovulum denitrificans	GCF_0029		screen	classification based on ANI only	GCA_020161655.1, s__Albidovulum sp0201...	97.46
Bin.002.fasta_assembly	d__Bacteria; p__Pseudomonadota; c__Alphaproteobacteria; o__Rhodobacterales; f__Rhodobacteraceae; g__Paracoccus; s__Paracoccus denitrificans	GCF_9001				001447385.1, acoccus 14...	97.74
Bin.003.fasta_assembly	d__Bacteria; p__Pseudomonadota; c__Gammaproteobacteria; o__Burkholderiales; f__Chitinibacteraceae; g__Andreprevotia; s__Andreprevotia lacus	GCF_9001				009848685.1, dreprevotia	97.76
Bin.004.fasta_assembly	d__Bacteria; p__Actinomycetota; c__Actinomycetia; o__Actinomycetales; f__Dermatophilaceae; g__Phycicoccus; s__Phycicoccus sp001428065	GCA_0014		screen	classification based on ANI only	GCF_002846495.1, s__Phycicoccus duodec...	94.92



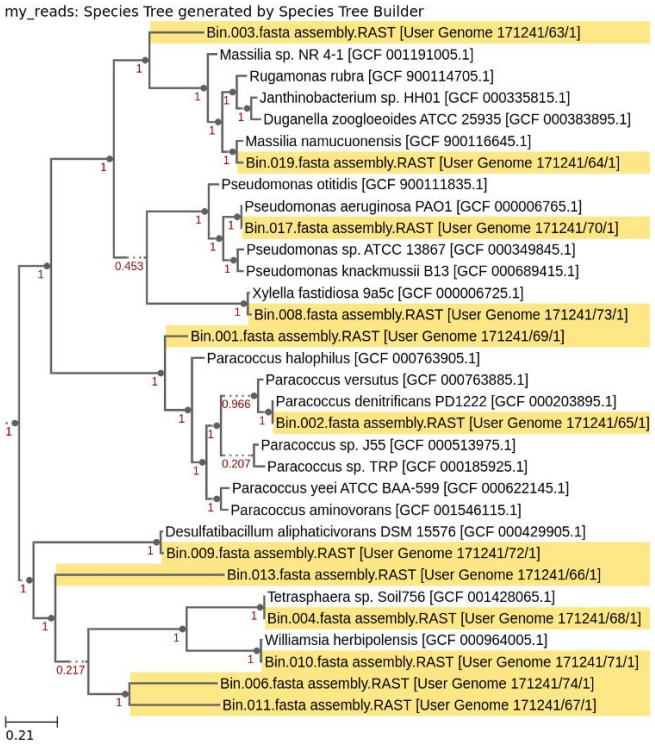
p__Actinomycetota
c__Actinomycetia
o__Actinomycetales
f__Dermatophilaceae
o__Mycobacteriales
f__Mycobacteriaceae
p__Bacillota
c__Bacilli
o__Exiguobacterales
f__Exiguobacteraceae
p__Bacillota_A
c__Clostridia
o__Peptostreptococcales
f__Peptostreptococcaceae
p__Bacteroidota
c__Bacteroidia
o__Chitinophagales
f__Chitinophagaceae
p__Desulfobacterota
c__Desulfobacteria
o__Desulfobacterales
f__Desulfatibacillaceae
g__Desulfatibacillum

p__Pseudomonadota
c__Alphaproteobacteria
o__Rhodobacterales
f__Rhodobacteraceae
c__Gammaproteobacteria
o__Burkholderiales
f__Burkholderiaceae
f__Chitinibacteraceae
o__Pseudomonadales
f__Pseudomonadaceae
o__Xanthomonadales
f__Xanthomonadaceae

Find Relatives with Species Tree

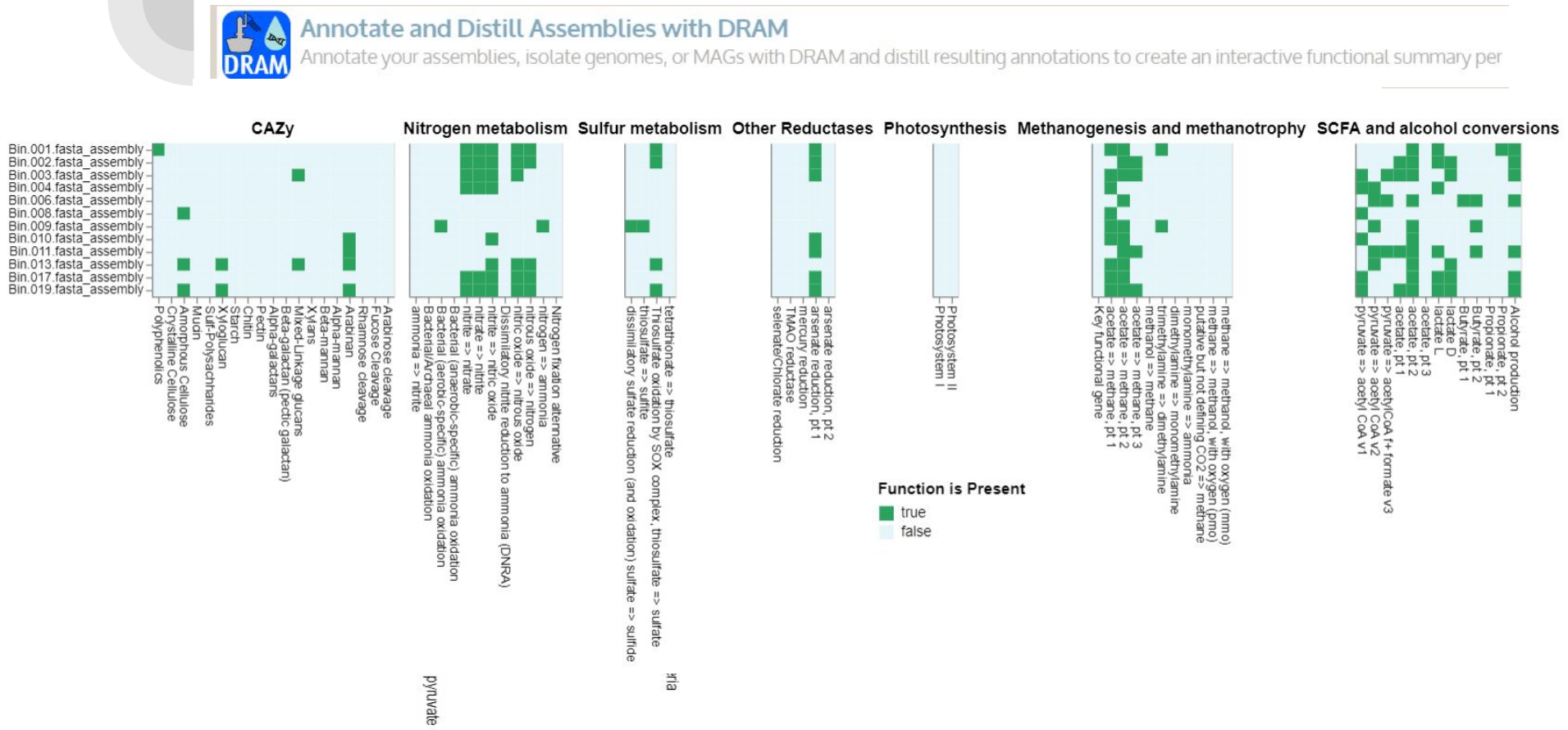
Insert Set of Genomes Into SpeciesTree - v2.2.0

Add a user-provided GenomeSet to a KBase SpeciesTree.



The highlighted genomes are the ones we have

Functional Classification of MAGs



Reference

- Clark, M., Allen, B., Arkin, A., Cottingham, B., Henry, C., Drake, M., Chivian, D. C., Elisha, W. C., Sneddon, M., Dehal, P., Sutormin, R., Canon, S., Jungbluth, S., Gu, T., & Riehl, W. (2021, November 9). Genome Extraction from Shotgun Metagenome Sequence Data. KBase. https://kbase.us/n/33233/606/
- Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology, 3(7). https://doi.org/10.1038/s41564-018-0171-1