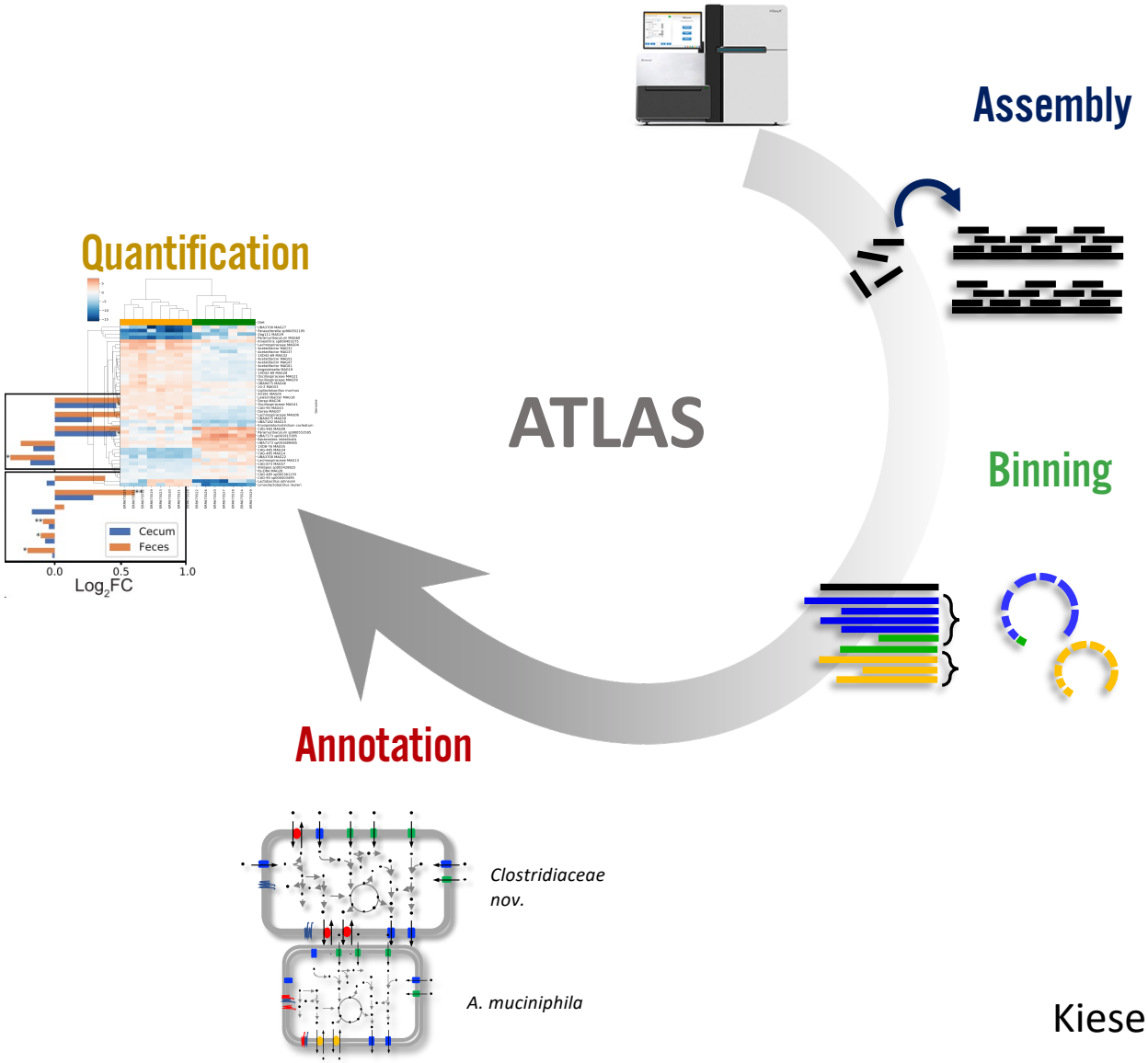
Metagenome atlas

And the bioinformatics behind it



Kieser *et al.* 2020

Others on Metagenome-Atlas

Aria Hahn, Co-founder Koonkie inc.
Thanks for the great tool! I've been using it in my research and telling everyone about it!

Taylor Reiter Graduate from UC Davis.
Learners were excited about all of the functionality that **just worked** without them having to type out all of the steps.

Josh Neufeld, Professor at University of Waterloo.
Very useful package for my lab.

Start in three commands!

```
conda install metagenome-atlas
atlas init path/to/fastq
atlas run all
```

1 Dependency


ANACONDA®

Snakemake

Why do I nead a pipeline?

- Install of dependencies
- Parallelization
- Multiple samples
- Log and control of completion
- Cluster submission on different systems

Snakemake

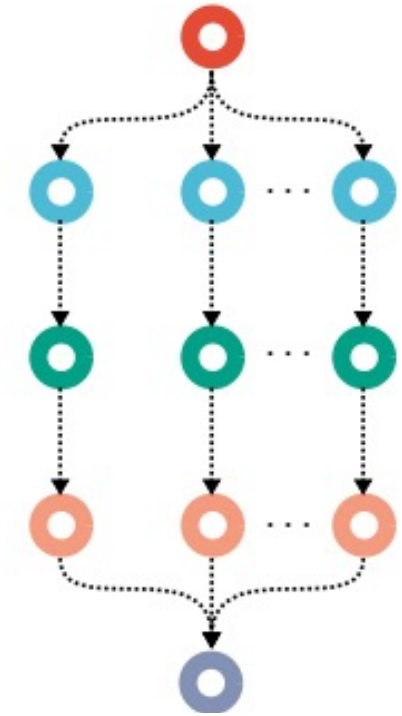
Create rules

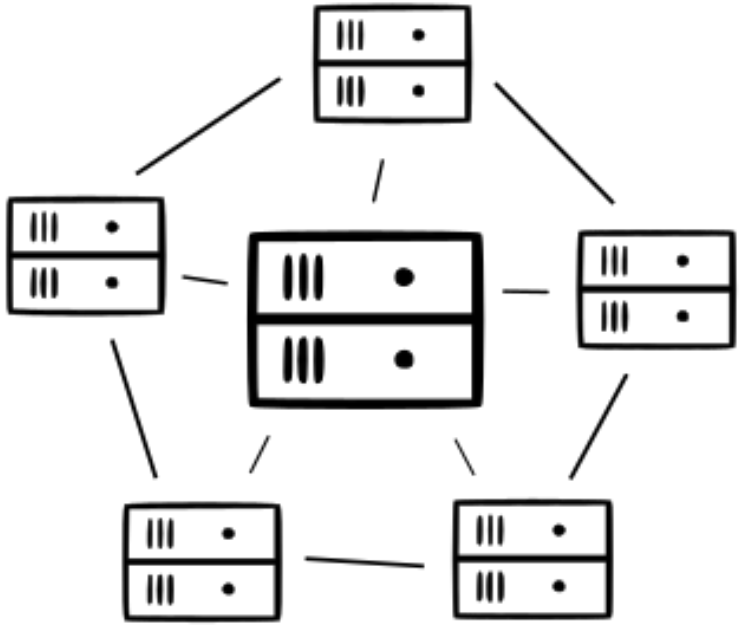
```
rule plot:
    input:
        "raw/{dataset}.csv"
    output:
        "plots/{dataset}.pdf"
    shell:
        "somecommand {input} {output}"
```

Install dependencies automatically

```
channels:
  - bioconda
  - r
dependencies:
  - python=2.7
  - checkm-genome=1.0.7
  - prodigal >=2.6.1
```

Snakemake





Cluster submission

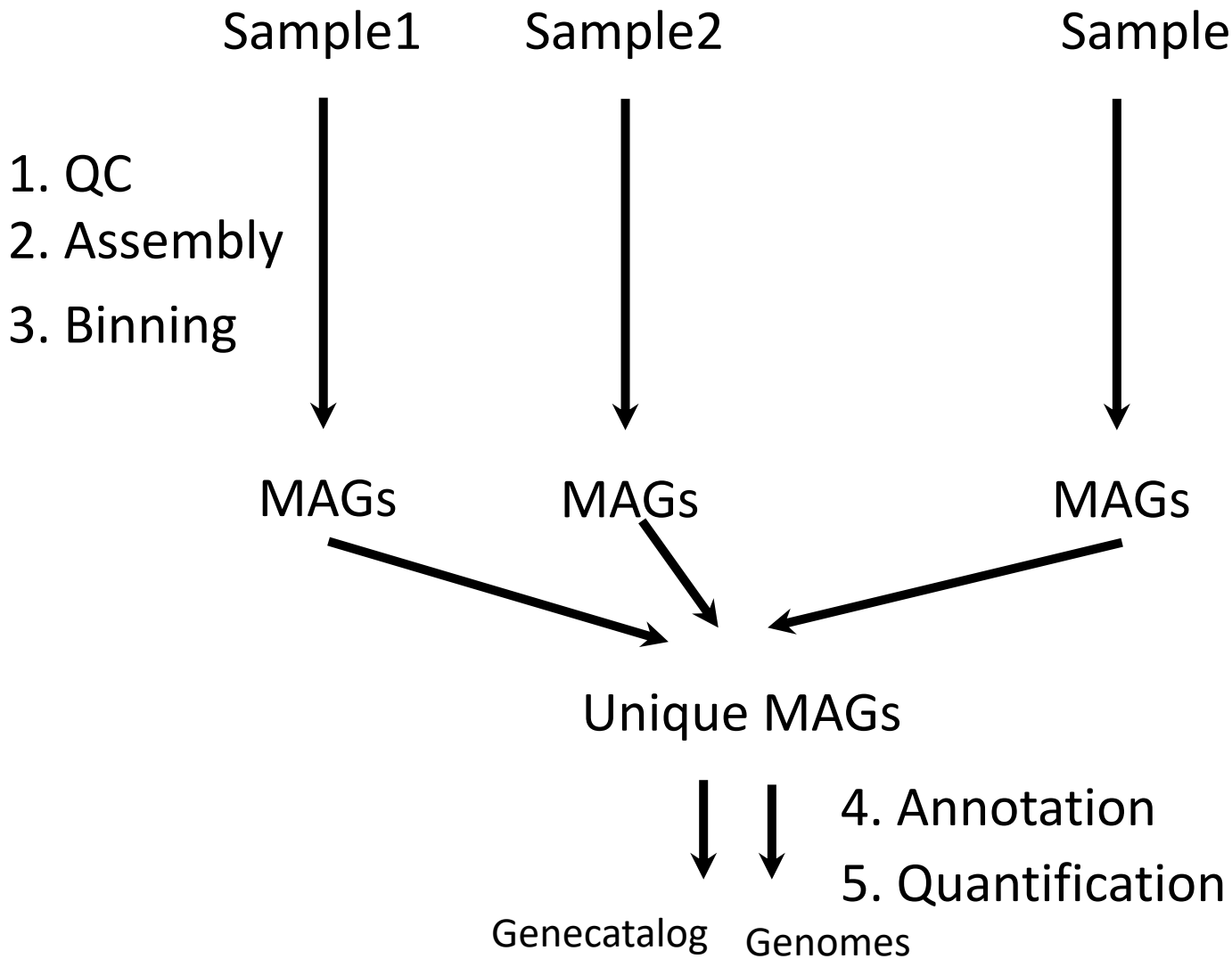
- Different cluster systems
- Different resource-limits
- Different queues
- Error handling

→ Atlas cluster wrapper

Metagenome-Atlas in detail

```
atlas run genomes
```

Atlas workflow



1. Quality control

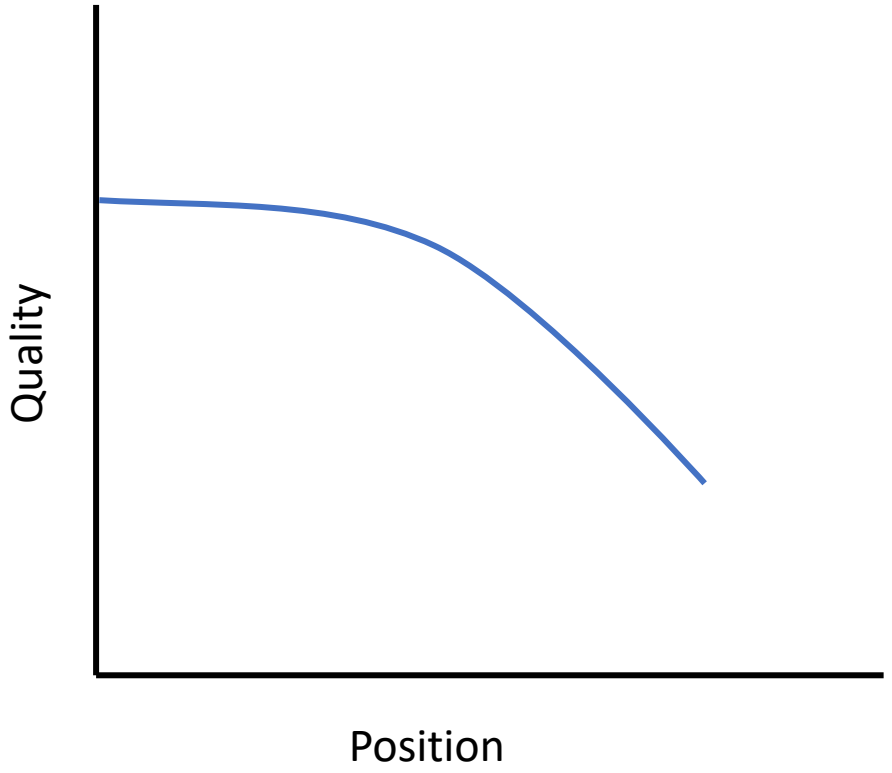


Image: Illumina Inc

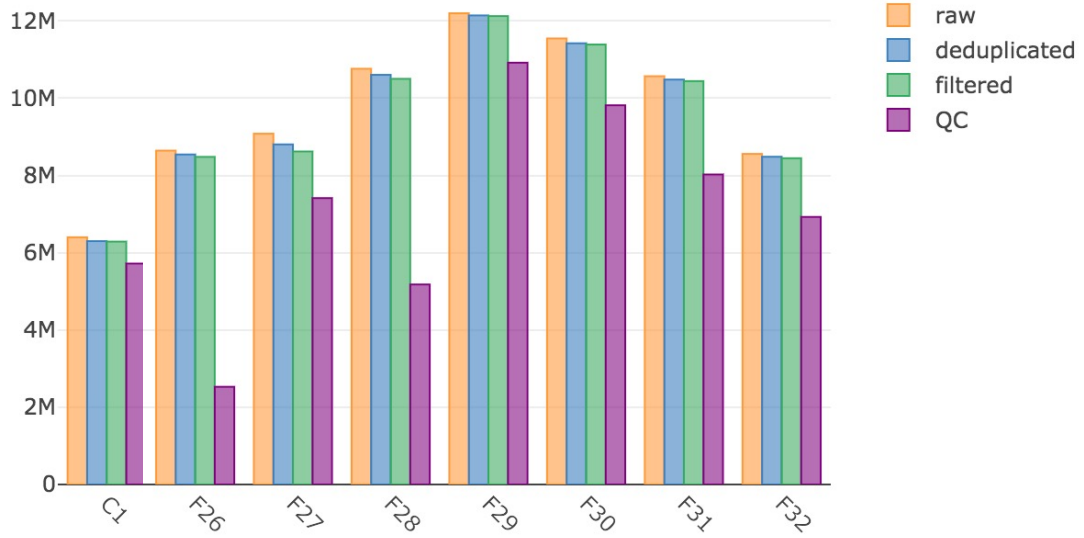
1. Quality control

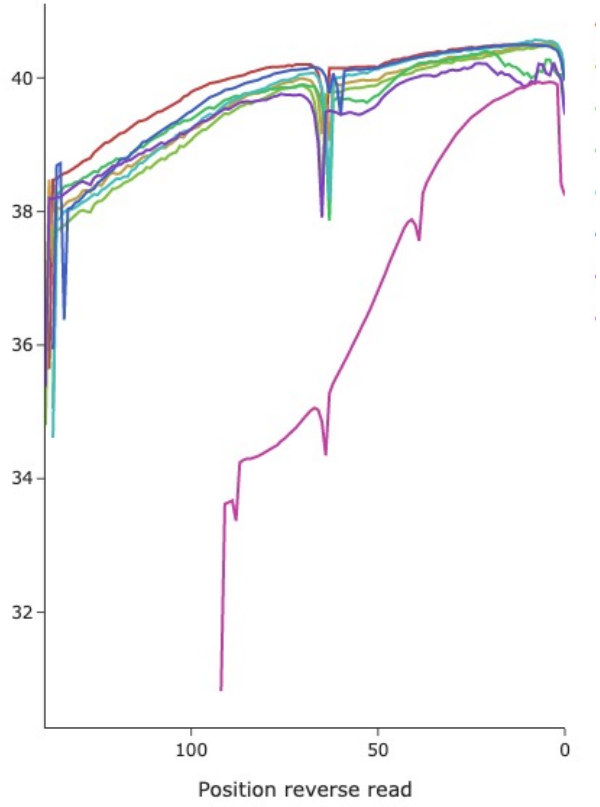
- Using bbmap-tools
- Remove low quality bases
- Contaminant removal
- Host removal

➢ **Good-quality reads**

Quality report

Total reads per sample





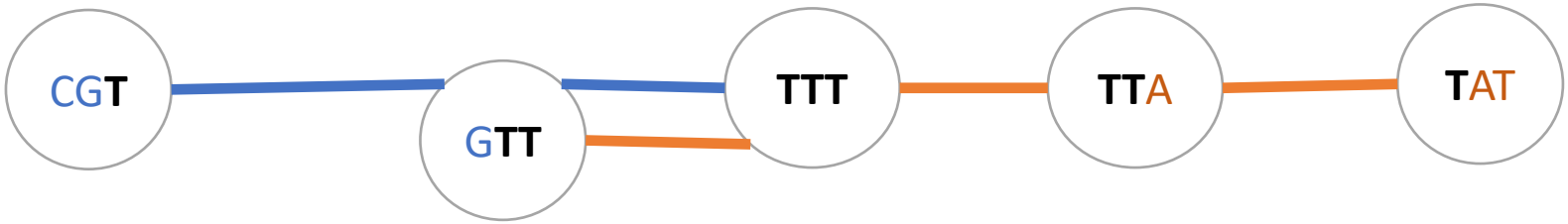
2. Assembly

Building assembly graphs

K=3

ATCGTCACG**TTT**

G**TTT**ATCGTCTG



Building assembly graphs

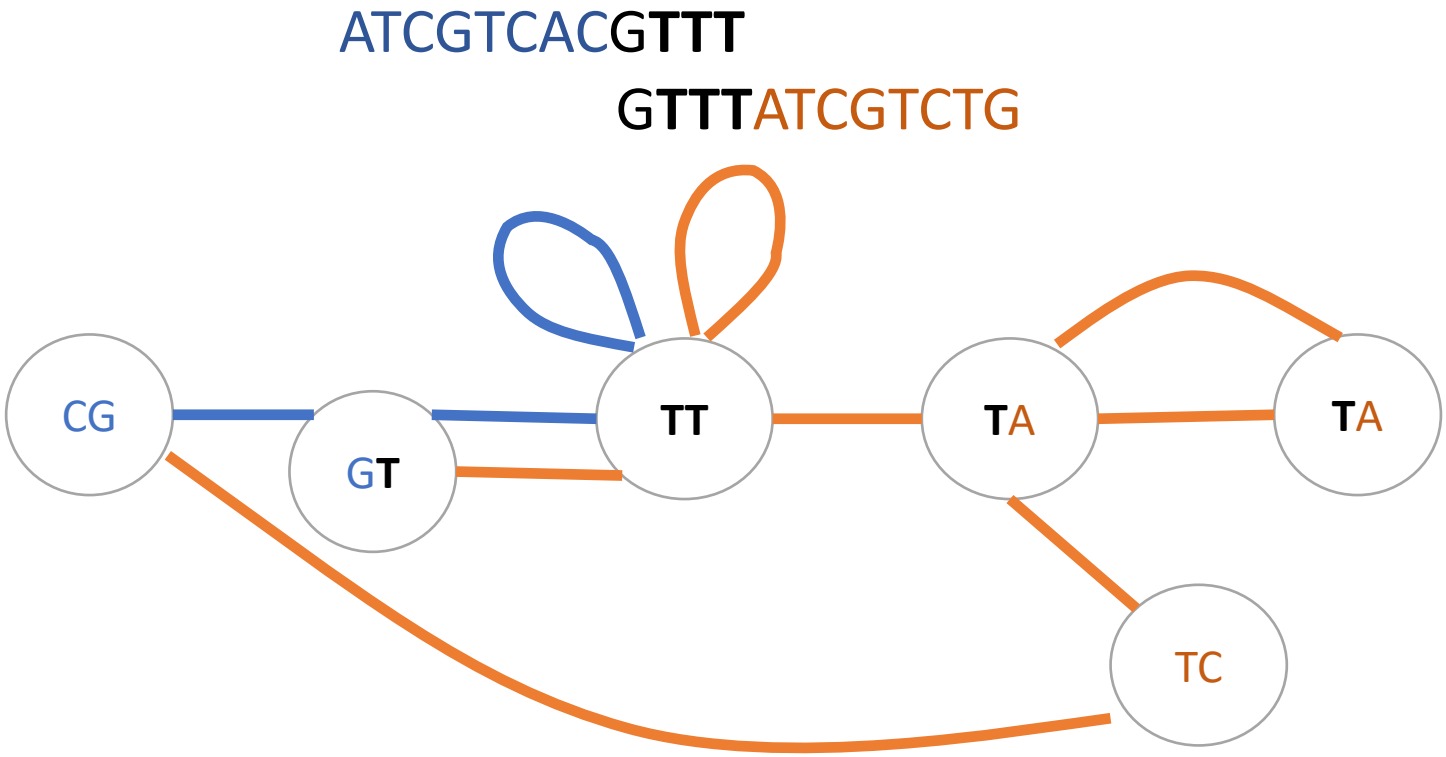
K=5

ATCGTCAC**GTTT**

GTTTATCGTCTG

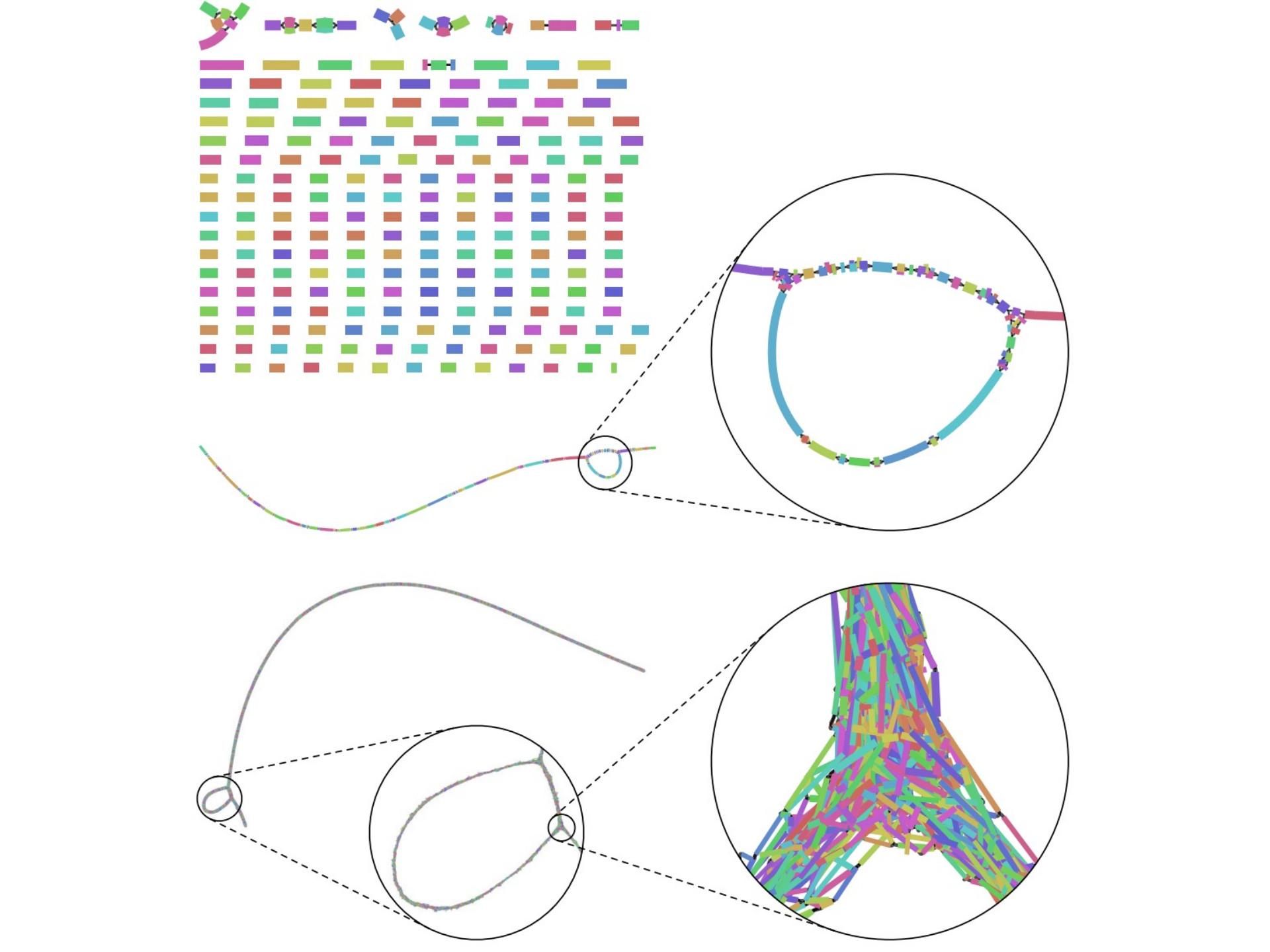
Building assembly graphs

K=2



2. Assembly

- Assembly graph with multiple k-mers
- Sophisticated graph simplification
- Error correction



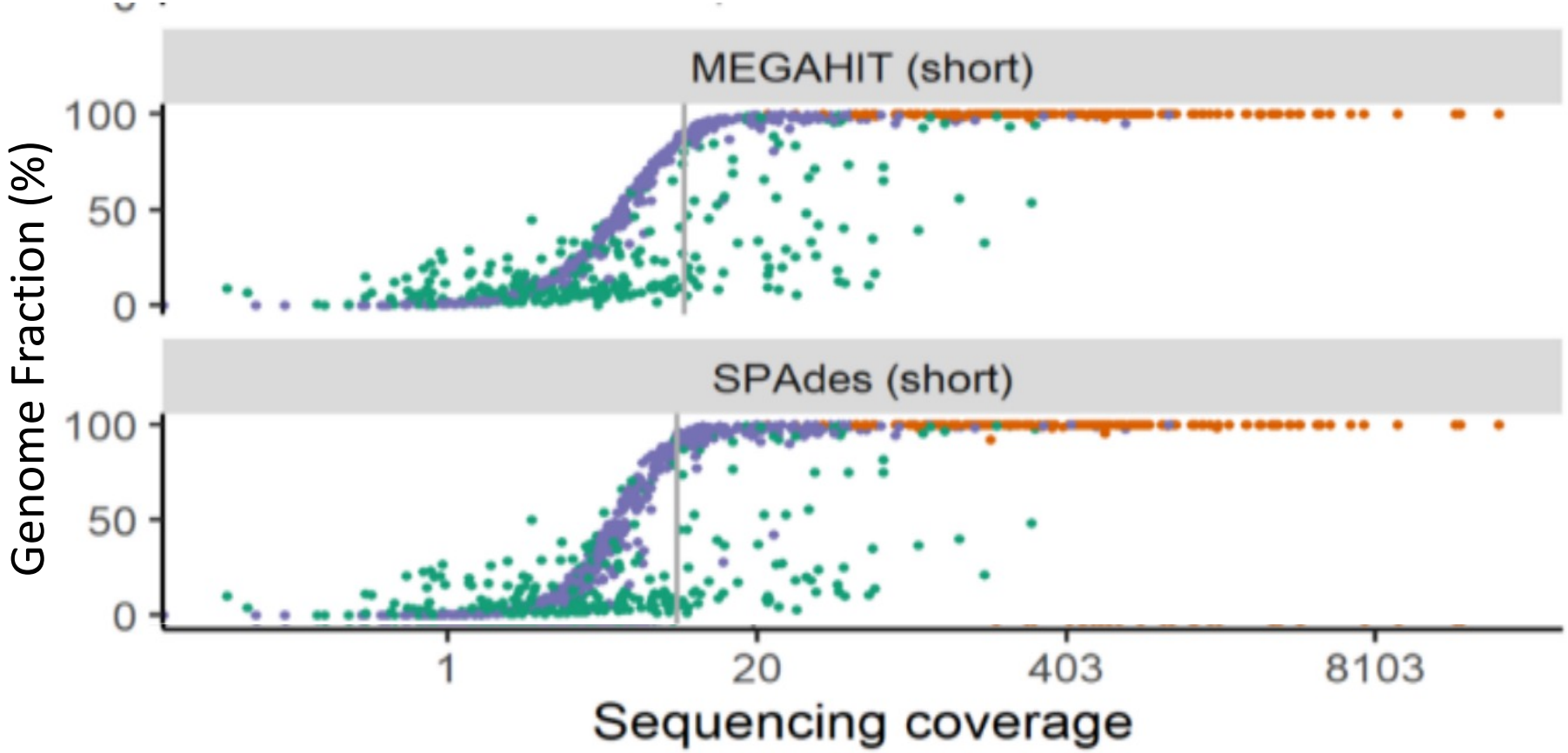
2. Assembly

- Uses metaSpades or megahit
- Pre-processing
  - Error correction
  - Paired-end merging (pre-assembly)

2. Assembly

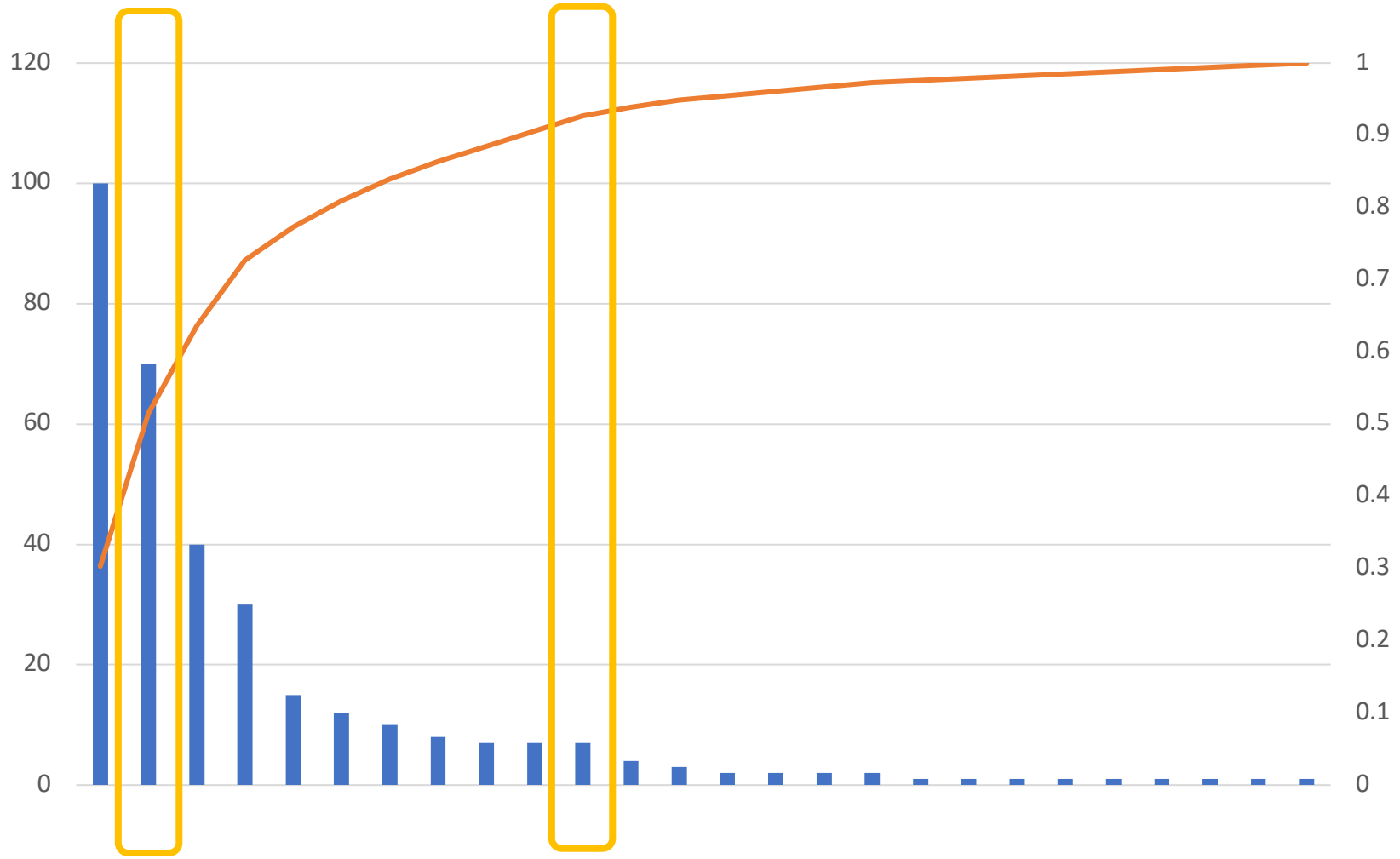
- Uses metaSpades or megahit
- Pre-processing
  - Error correction
  - Paired-end merging (pre-assembly)
- Post-processing
  - Filtering based on length and coverage
- Hybrid assembly supported

A minimum coverage is needed for good assembly



doi: 10.1101/2021.07.12.451567

N50/N90

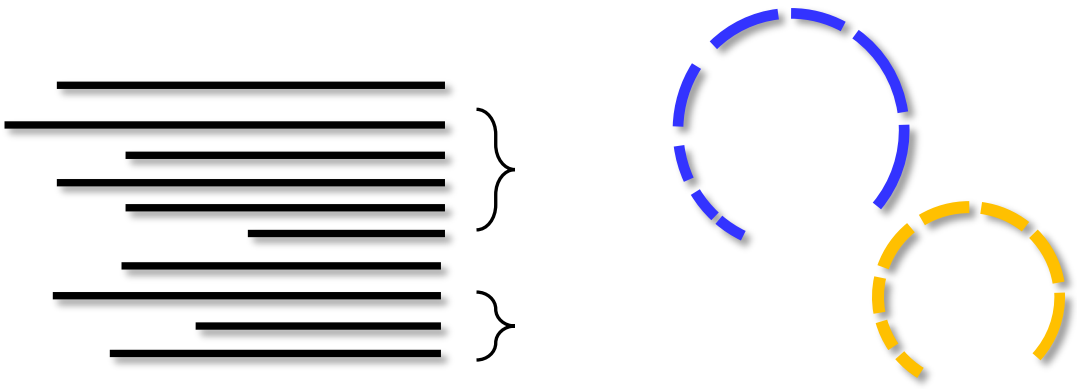


3. Binning

3. Binning

a) Binning
b) Quality estimation & Bin refinement
c) Dereplication

Binning: Clustering of Contigs



How do we bin contigs into genomes?

Sequence Features

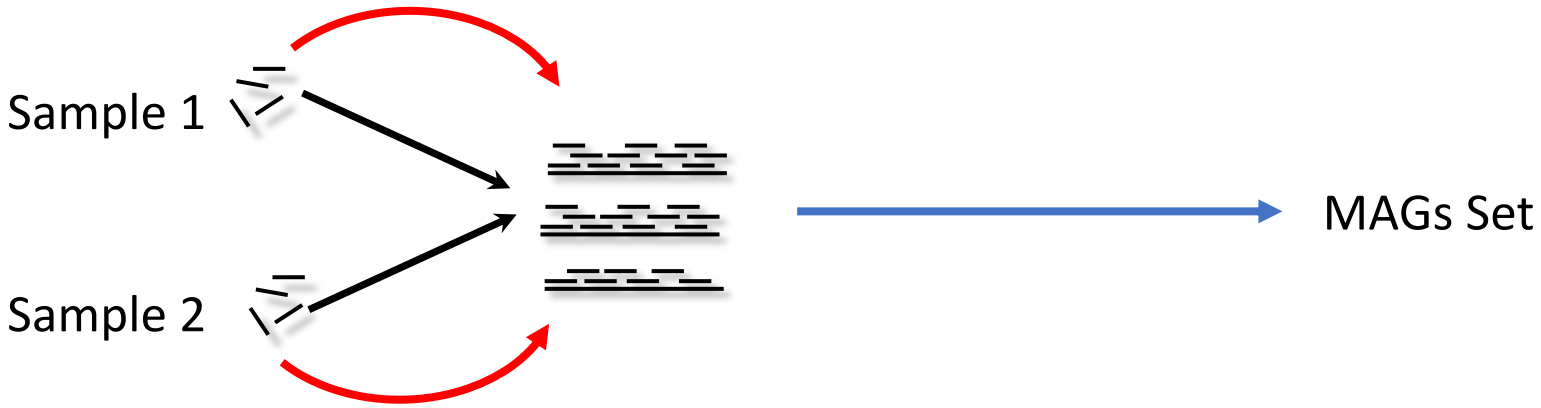
ATCGTCACGTAA

Abundance Features

Co-Abundance Features

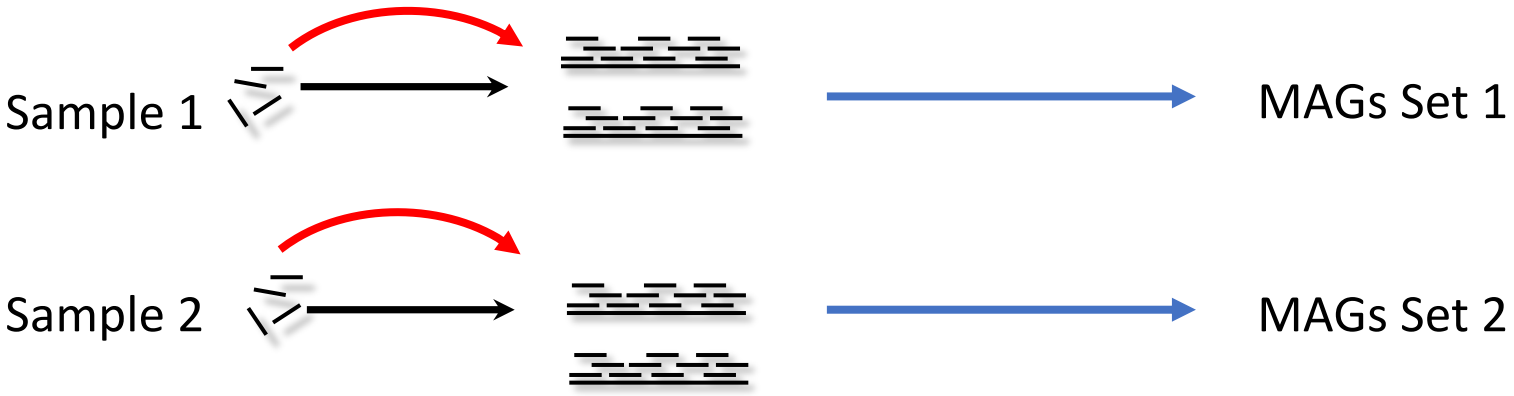
Co-abundance

Option 1: Co-assembly



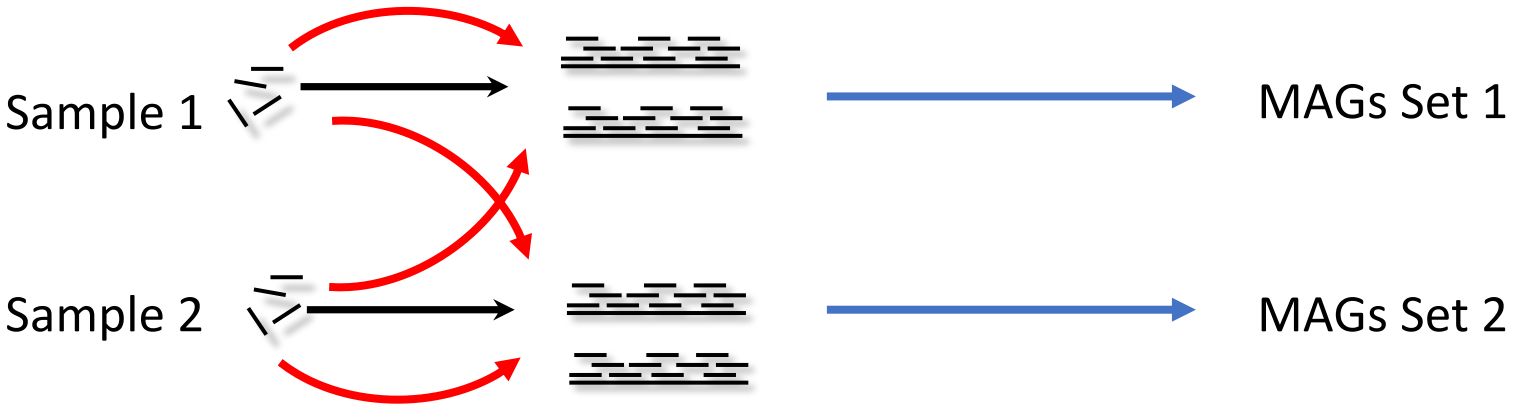
Co-abundance

Option 2: Single-sample assembly/Binnig

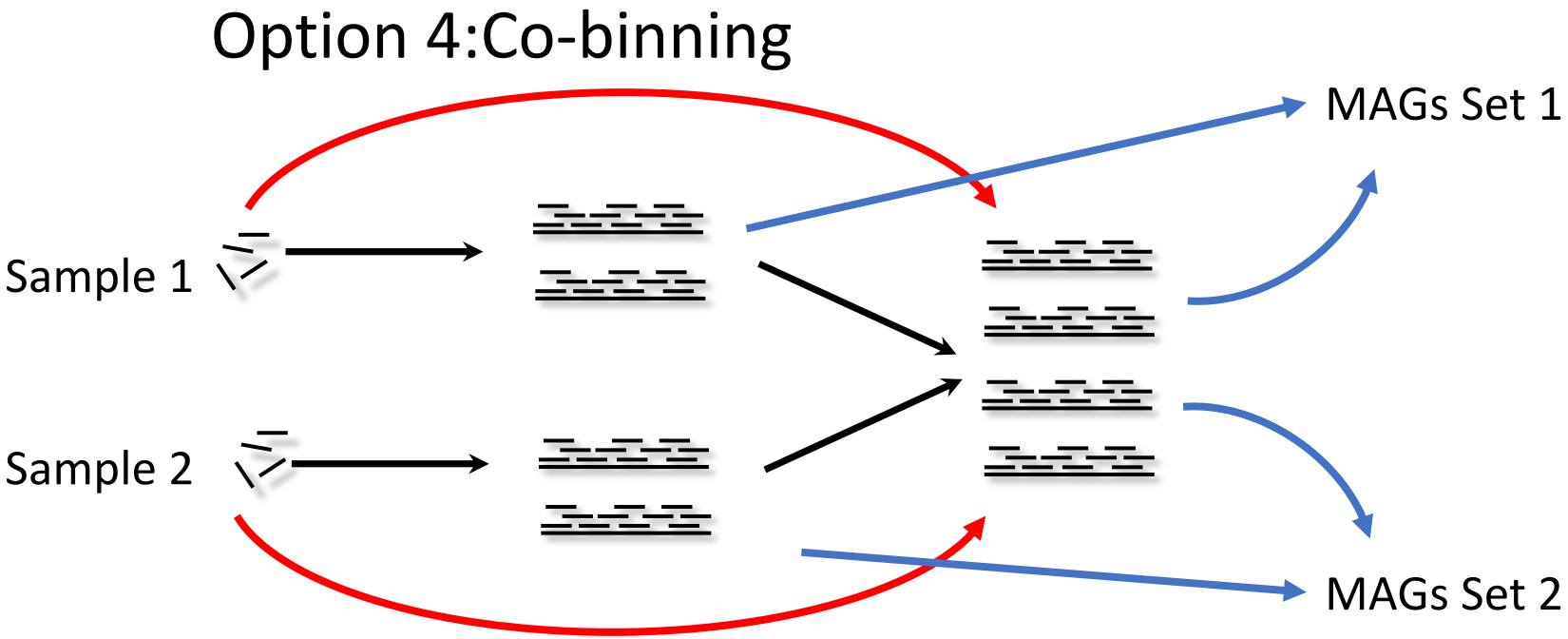


Co-abundance

Option 3: Cross mapping



Co-abundance



3 Binning

Single-sample / Cross mapping:

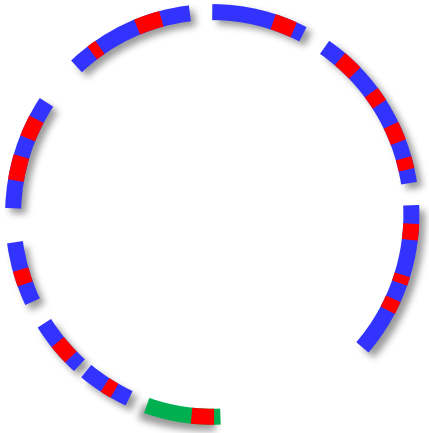
- Metabat2
- Maxbin2

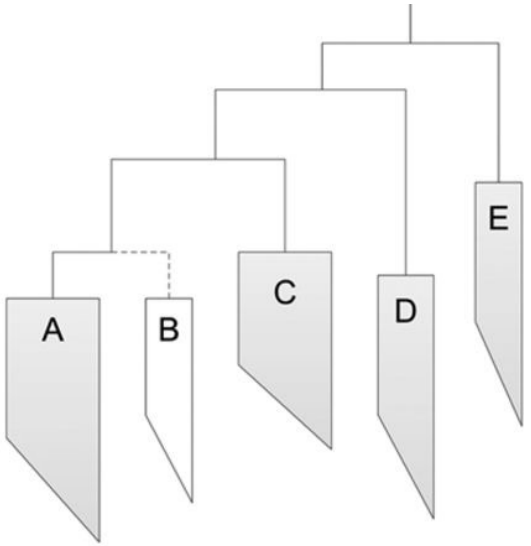
Co-Binning

- Vamb
- SemiBin

Quality estimation

(Essential) single-copy genes



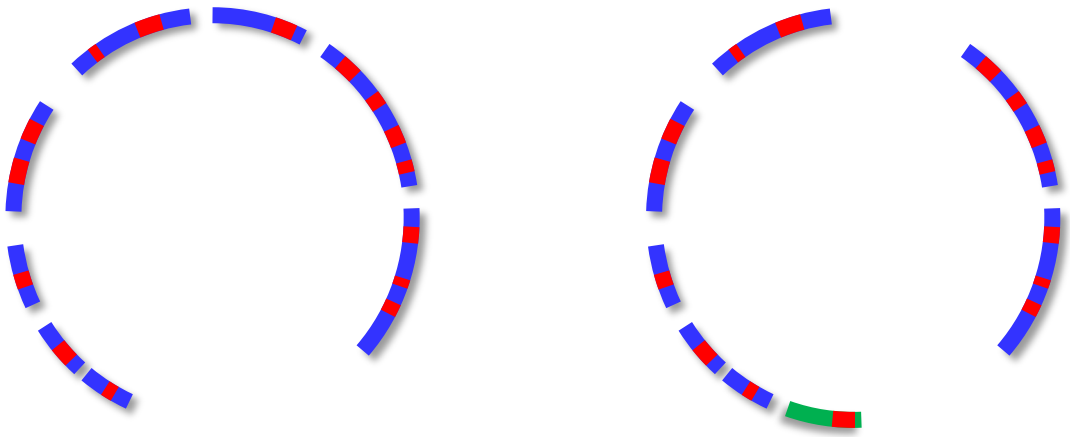


CheckM

BUSCO

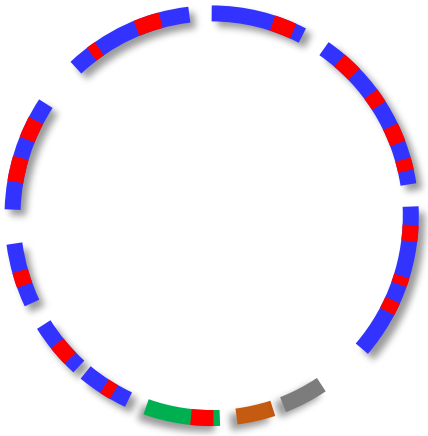
Bin Refinement

DAS Tool: Choose best Bin



Bin Refinement

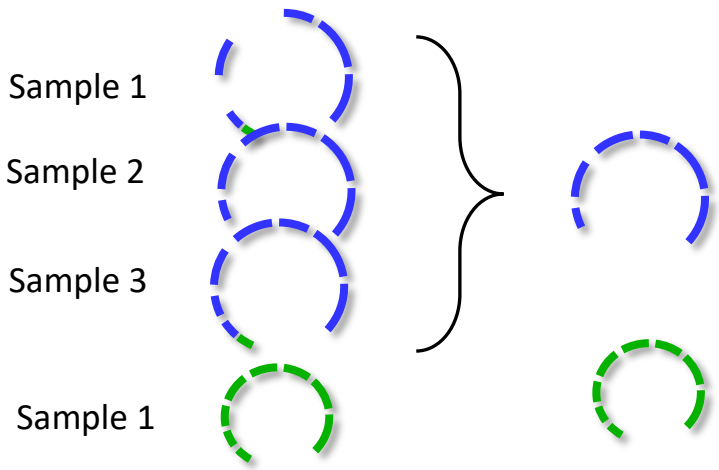
GUNC: Filtering based on Taxonomy



Atlas uses the same tools as large-scale studies on the Human microbiome

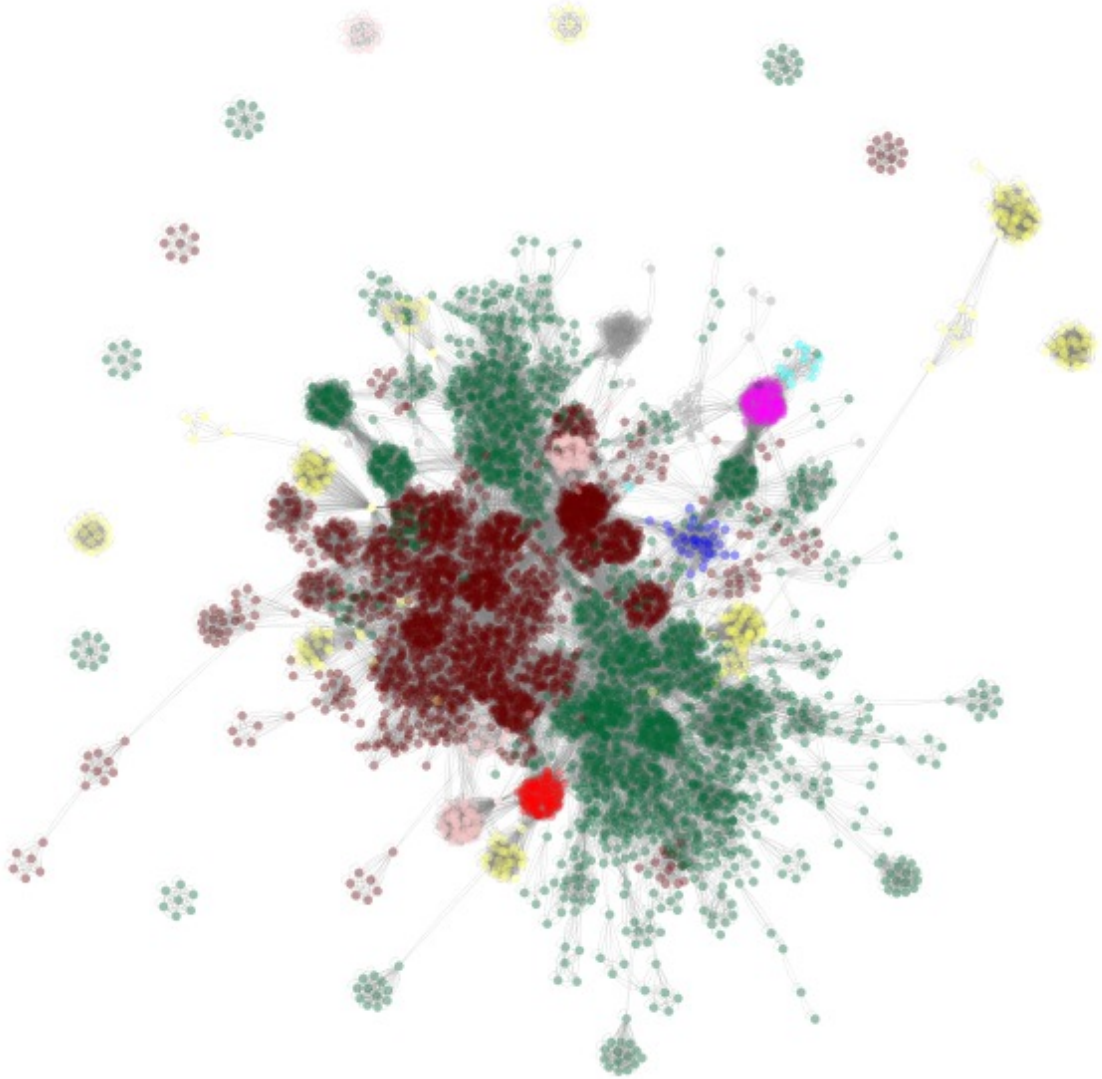
	CIBO	EBI	JGI	ATLAS
	Pasolli et al. 2019	Almeida et al. 2019	Nayfach et al. 2019	Kieser et al. 2020
Assembly	metaSpades Megahit			
Binning	Metabat	Metabat	Metabat Maxbin Concoct DASTool	Metabat Maxbin DASTool VAMB SemiBin
Quality estimation	CheckM			

De-replication

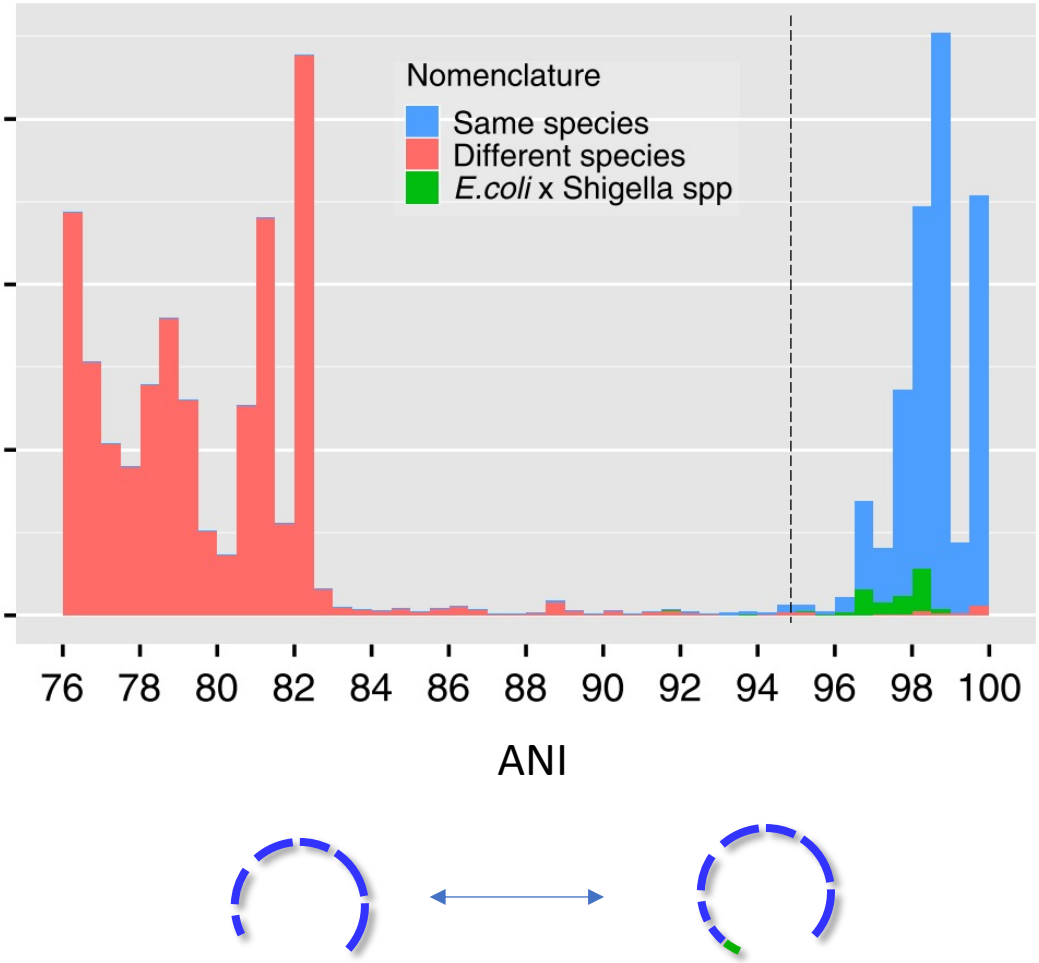


Average nucleotide Identity (ANI)

Mash

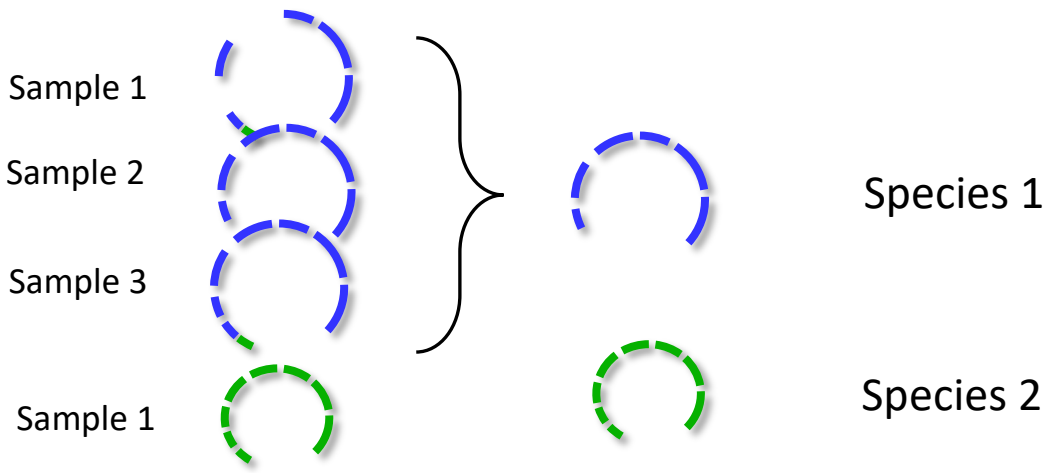


95% ANI used as species threshold



Jain et al. 2018

De-replication



4. Annotation

4. Annotations

What does it all mean?

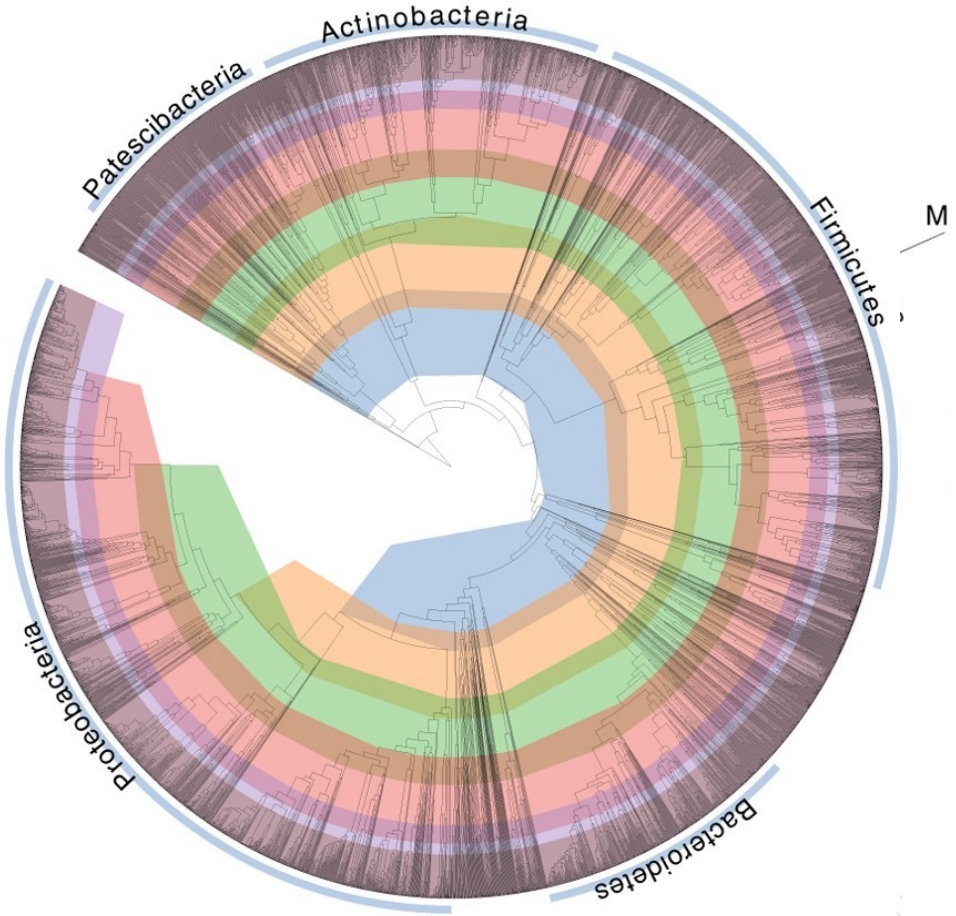
4. Annotations

a) Functions
b) Taxonomy

Taxonomic annotation

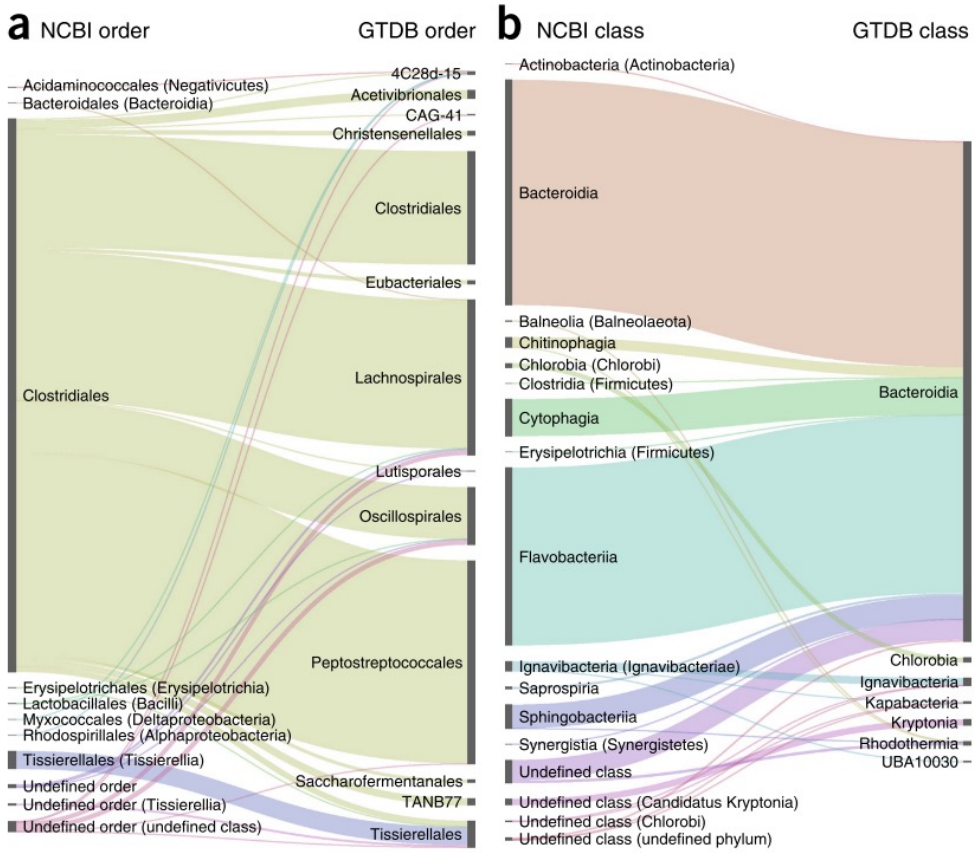
Genome Taxonomy database (GTDB)

Genome Taxonomy database



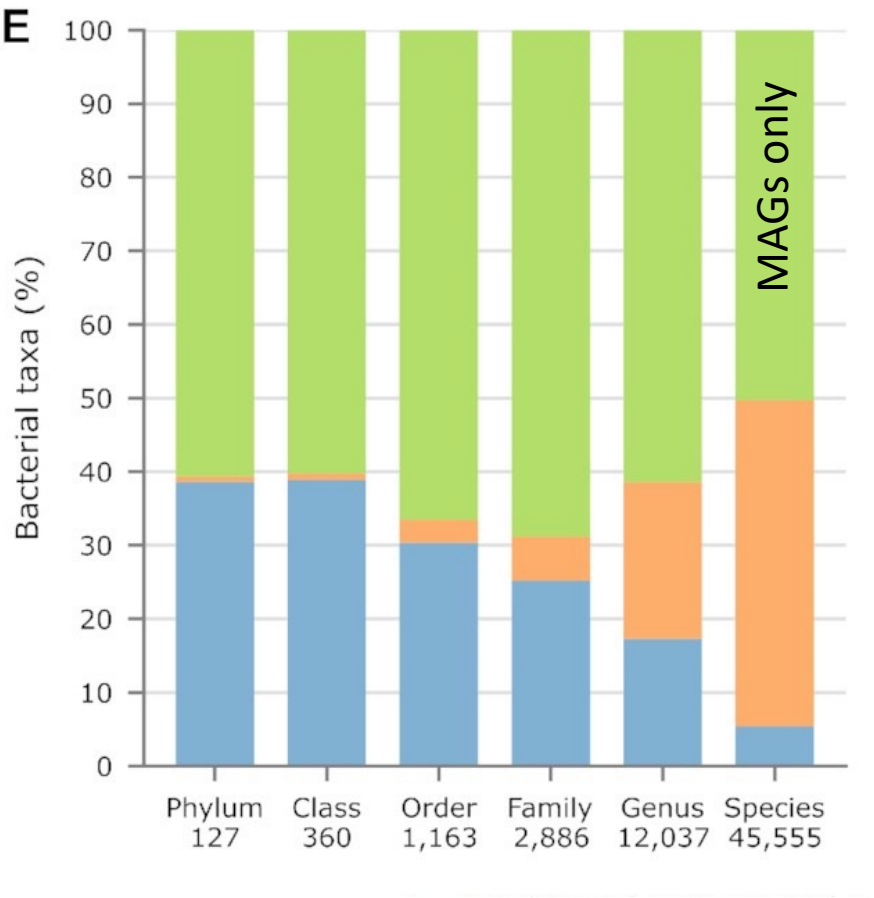
Parks et al. 10.1038/nbt.4229.

Proposed rearrangements



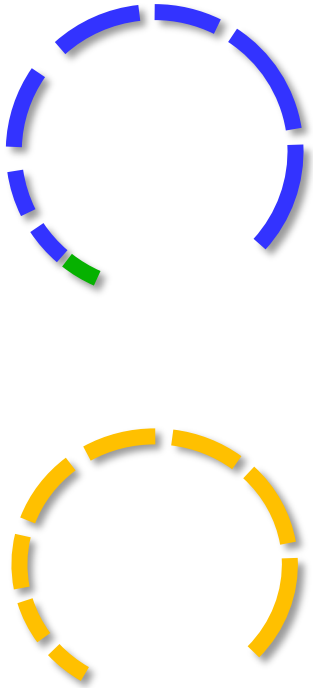
Parks et al. 10.1038/nbt.4229.

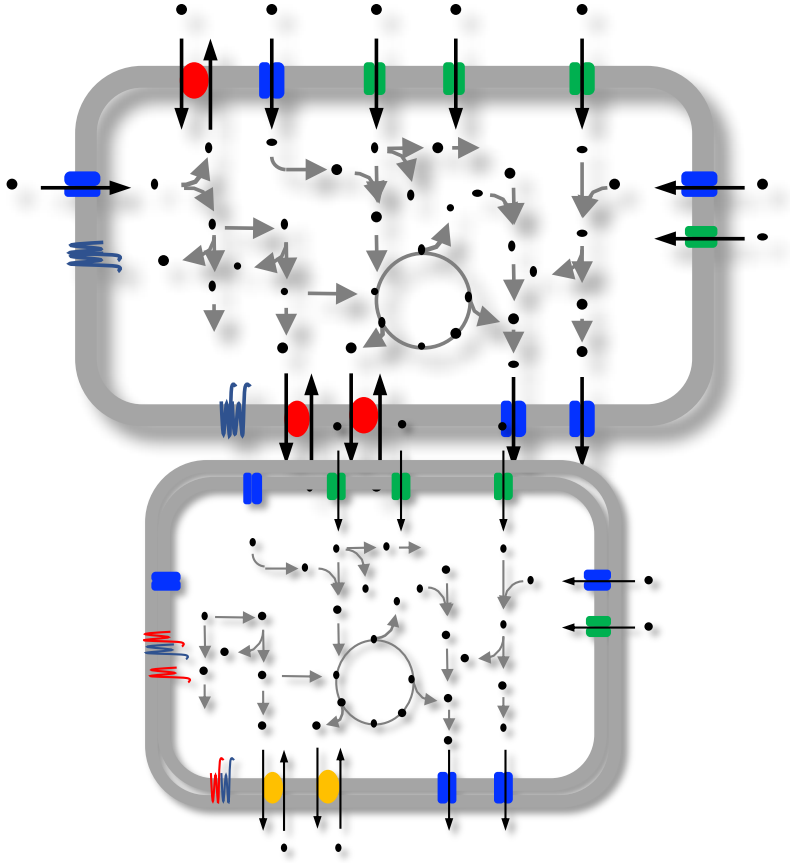
Genome Taxonomy database



Doi: 10.1093/nar/gkab776

Functional annotation

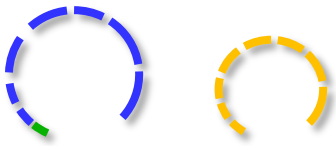




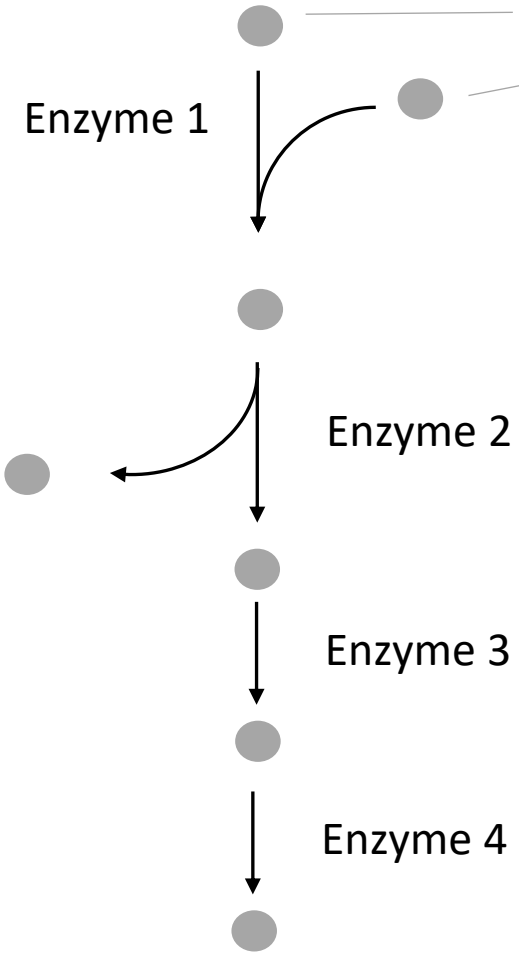
Functional annotation

Pathway inference

MAG 001 **MAG002**



Enzyme 1 Enzyme 2	Enzyme 1 Enzyme 2 Enzyme 4
X	✓

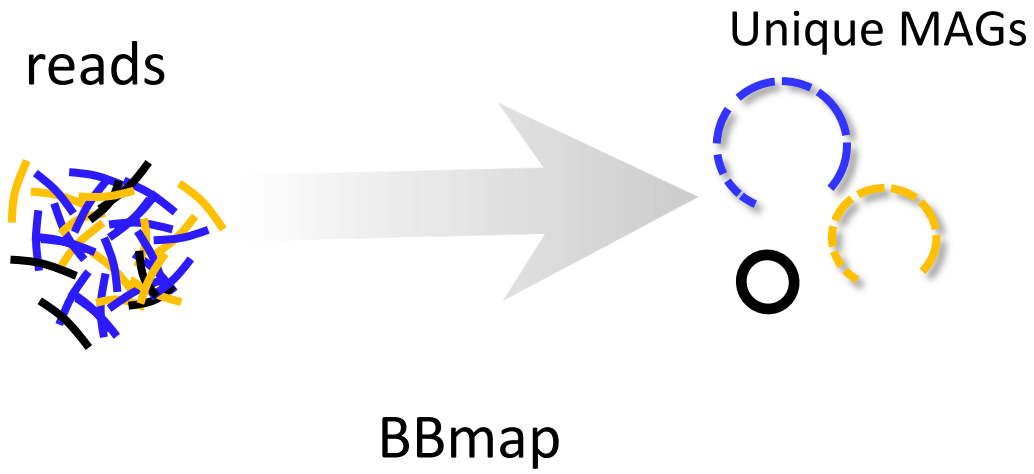


KEGG Kyoto Encyclopedia of Genes and Genomes

DRAM

Quantification

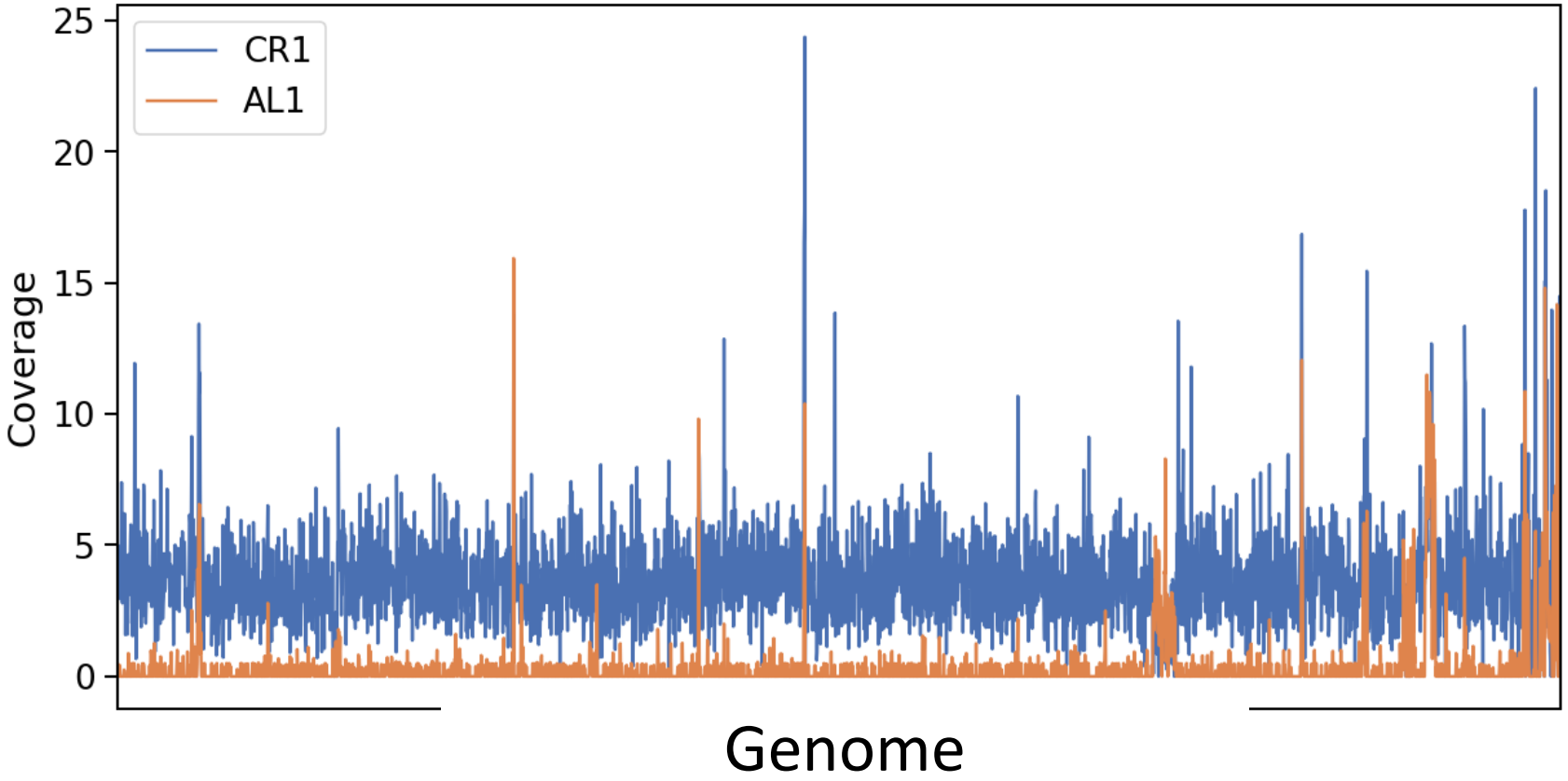
Quantification



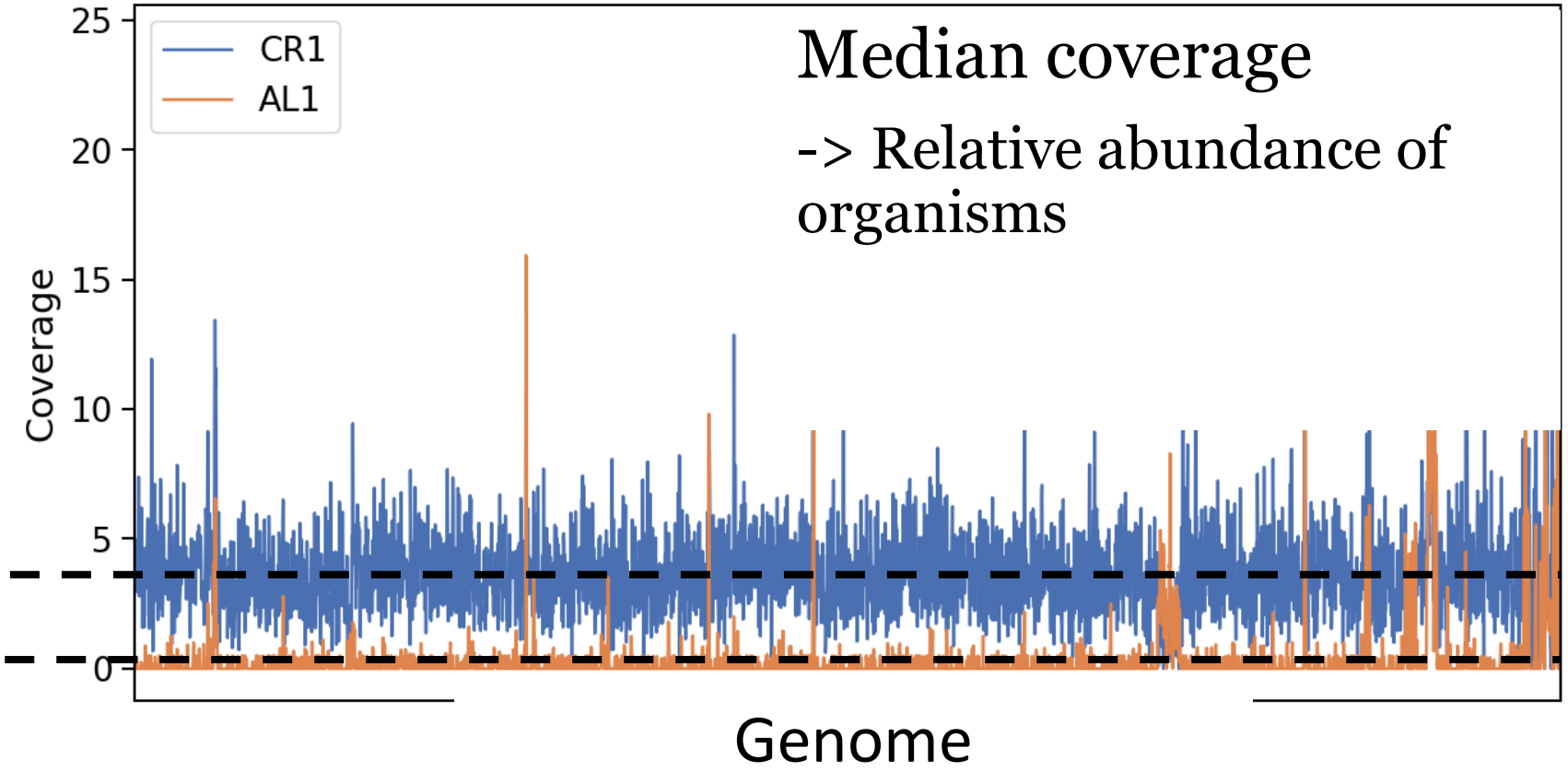
Quantifying genomes is not straight forward

- Unmapped reads
- Ambiguous mapped reads
- Variability in coverage
- **Compositional nature of microbiome data**

What is the abundance of a genome?

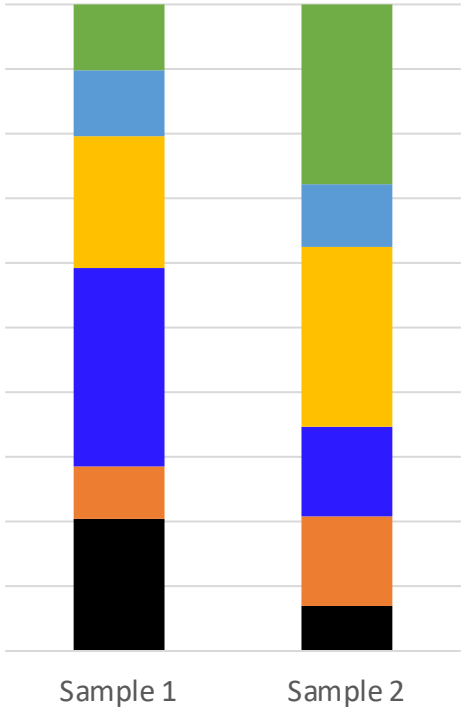


What is the abundance of a genome?



Statistical analysis

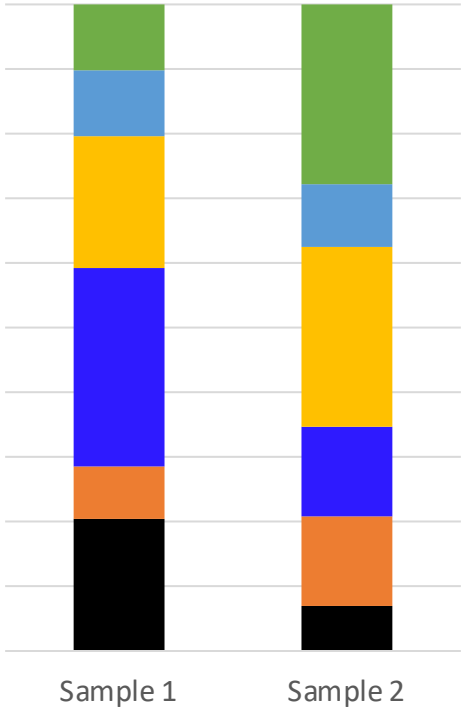
Relative abundance



What to do with the unmapped reads?

Interpret microbial abundances as ratios

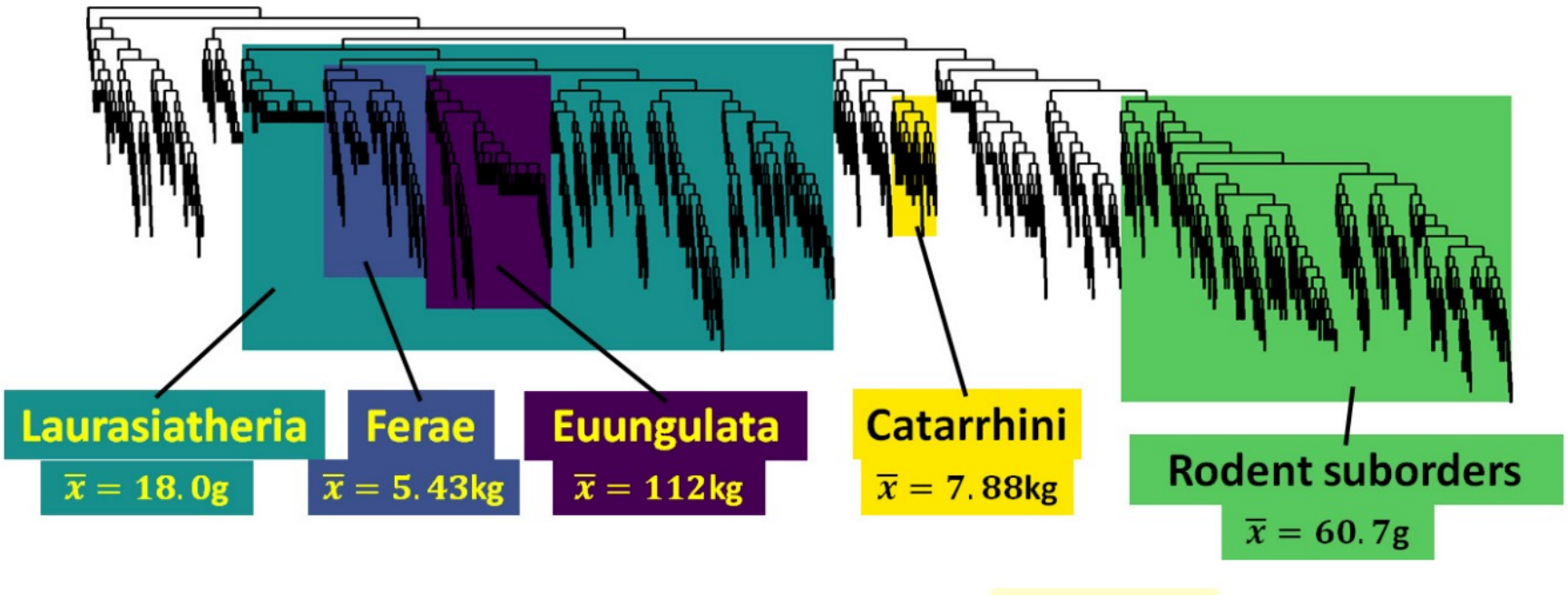
Ratios



Calculate ratios

A) Based on phylogeny
B) Centered log-ratios (CLR)
C) Machine-learning based on ratios

Phylofactor



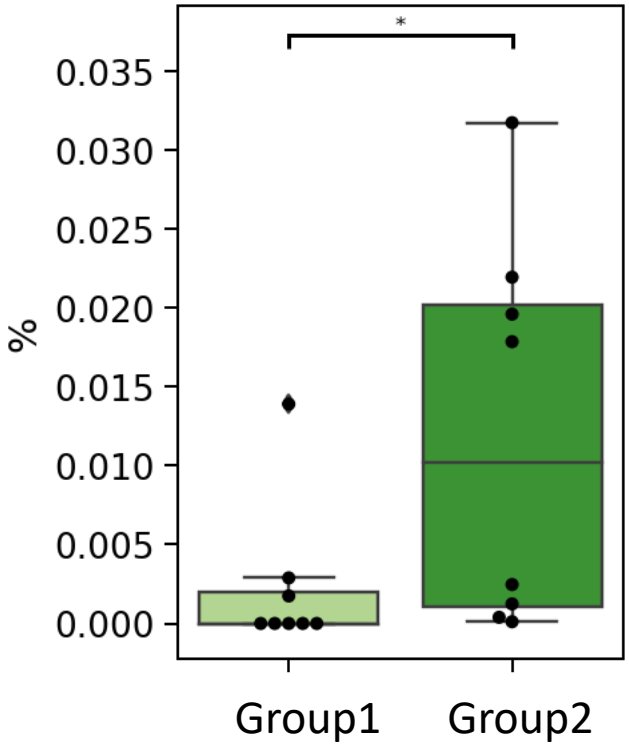
Centered log ratios

Impute zeros

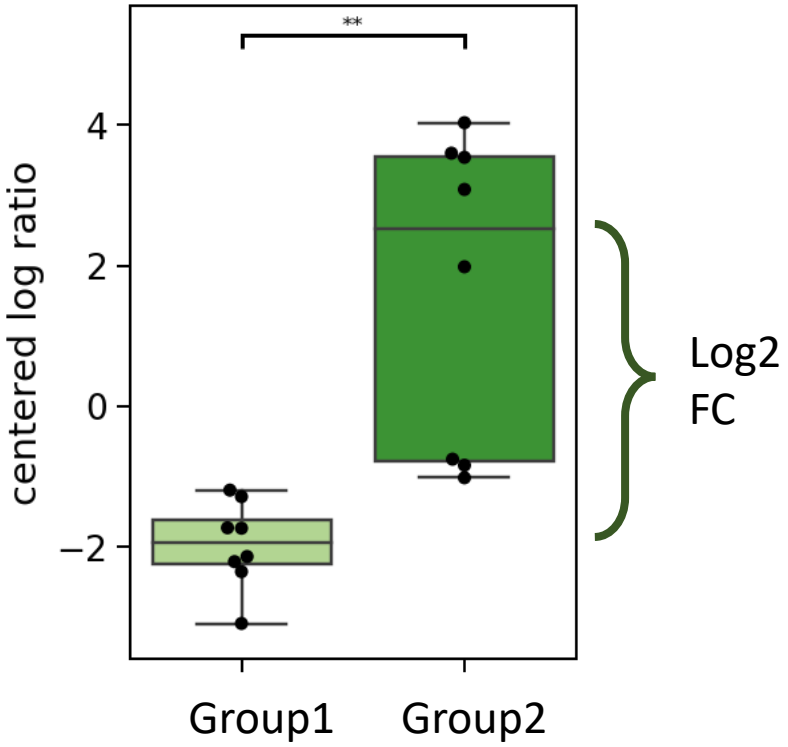
1. Take log
2. Subtract sample-mean

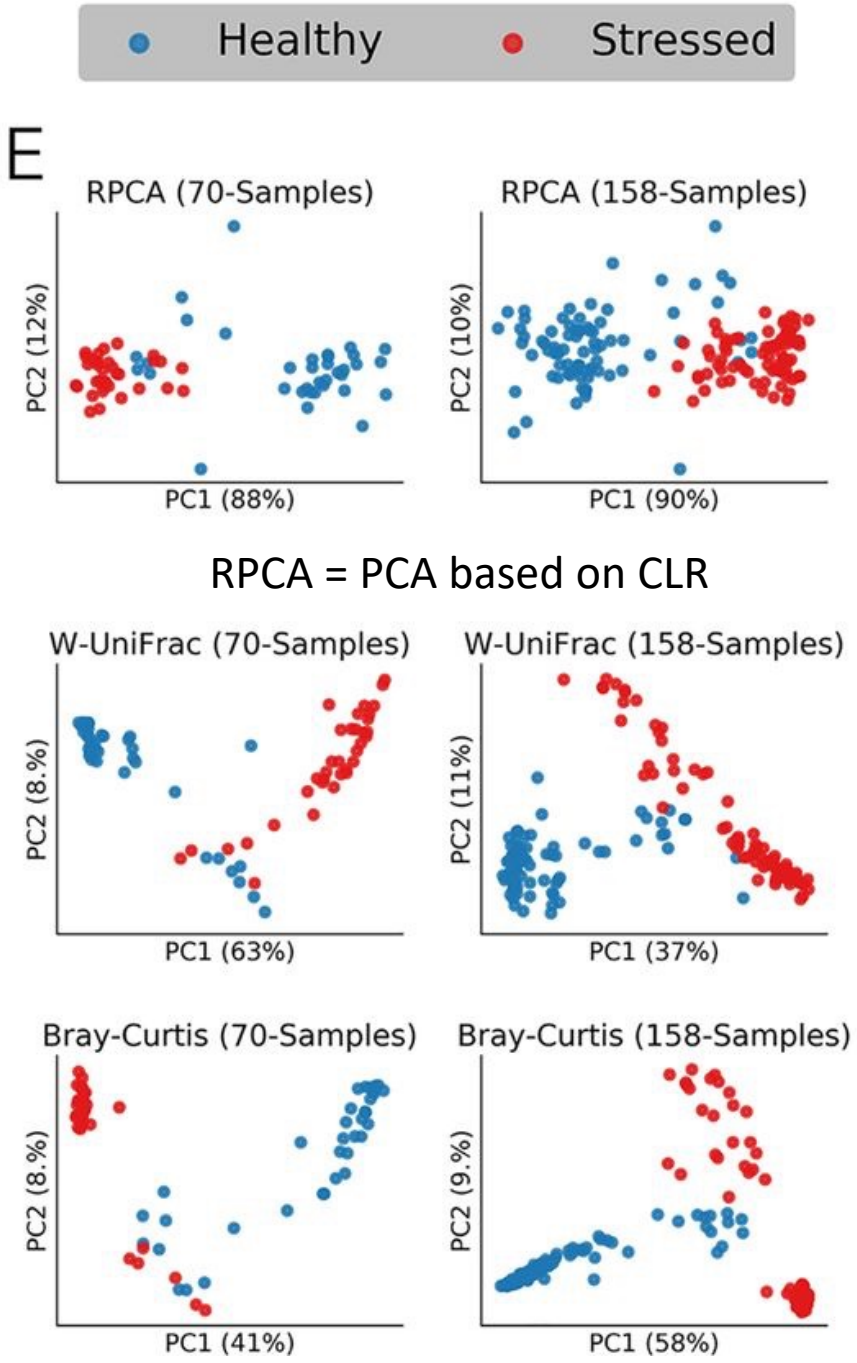
Centered log ratios

Relative abundance



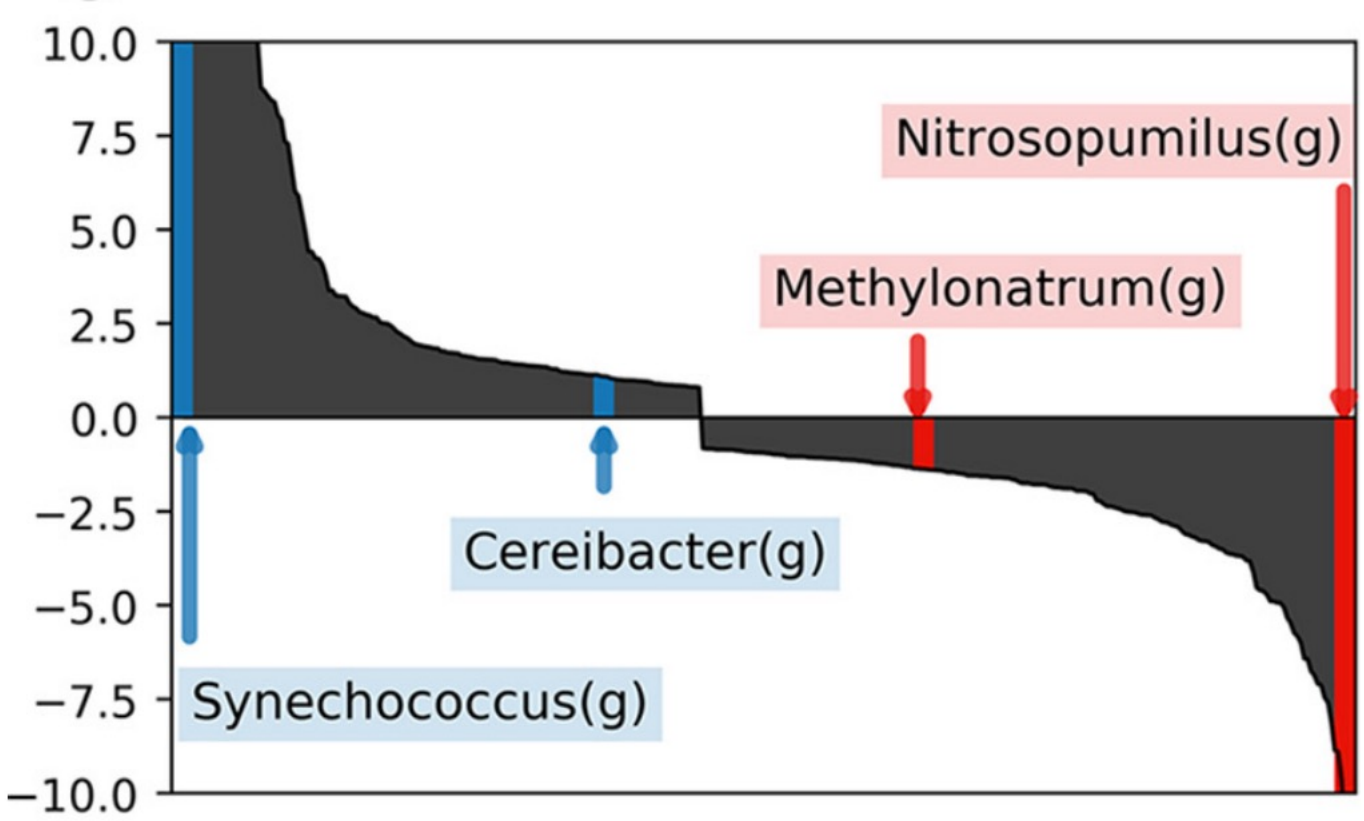
Centered log ratios





RPCA = PCA based on CLR

Martino *et al.* 2019



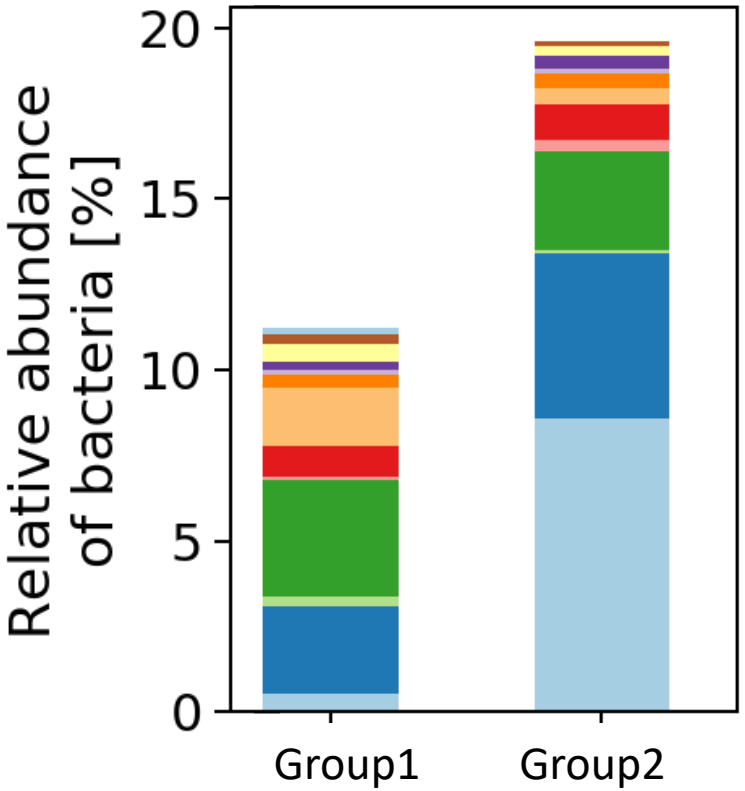
Martino *et al.* 2019

silask.github.io
Chapter 5 of my thesis

Abundance of pathways

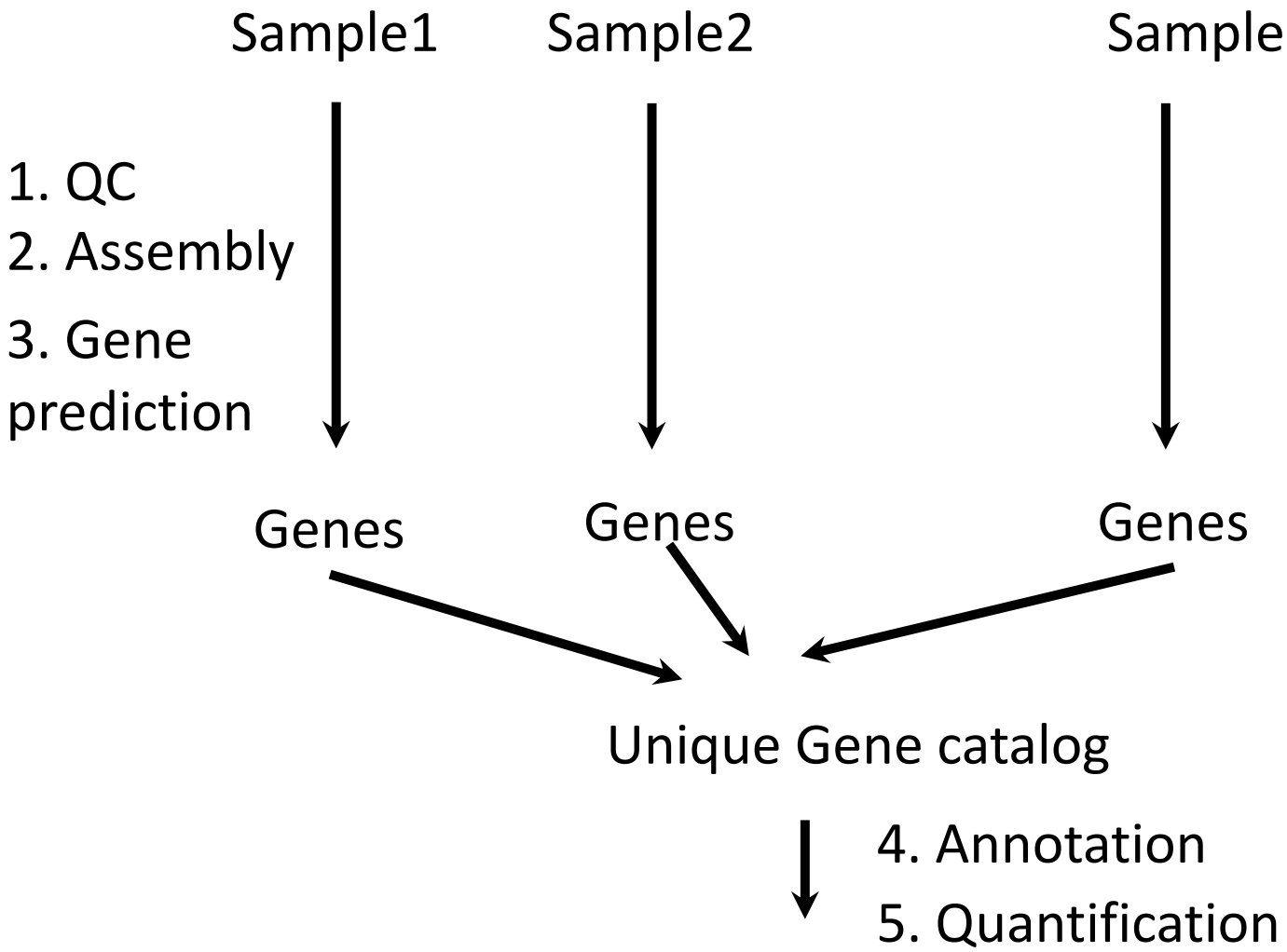
Sum of the species-abundance for all species where the pathway is present

Abundance of pathways



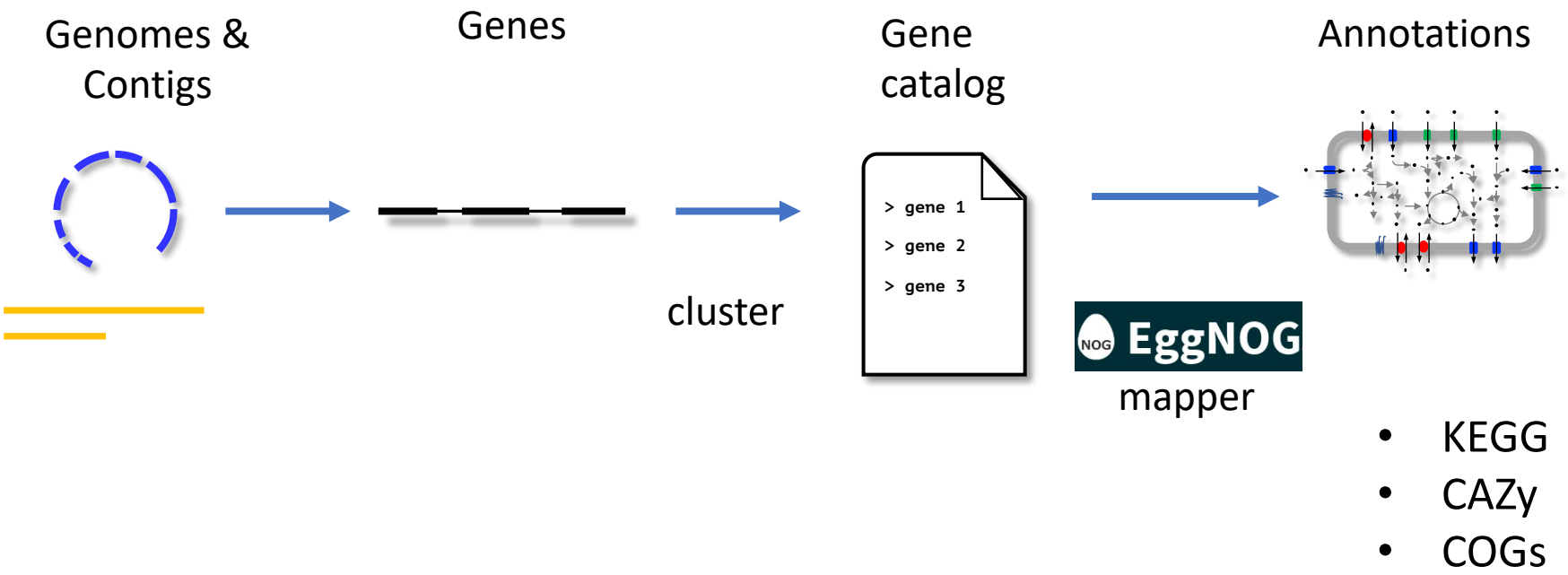
Gene catalog

Atlas workflow



```
atlas run genecatalog
```

Annotation


Genomes &
Contigs
Genes
cluster
Gene
catalog
> gene 1
> gene 2
> gene 3
Annotations
EggNOG
mapper
• KEGG
• CAZy
• COGs