Metagenomics with only three commands

Silas Kieser

Content

- Assembly based metagenomics
- Metagenome-atlas
- Compositional-data analysis

To assemble or not to assemble

Assembly based metagenomics

Why doing metagenomics?

Who is there?

What are they doing?

Composition

Functional potential

Problem:

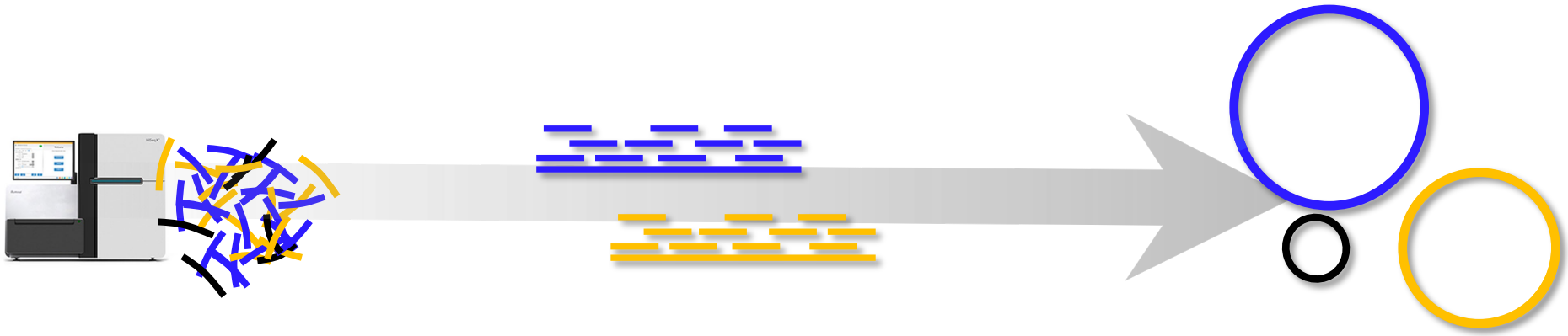
Insufficient references

de novo assembly

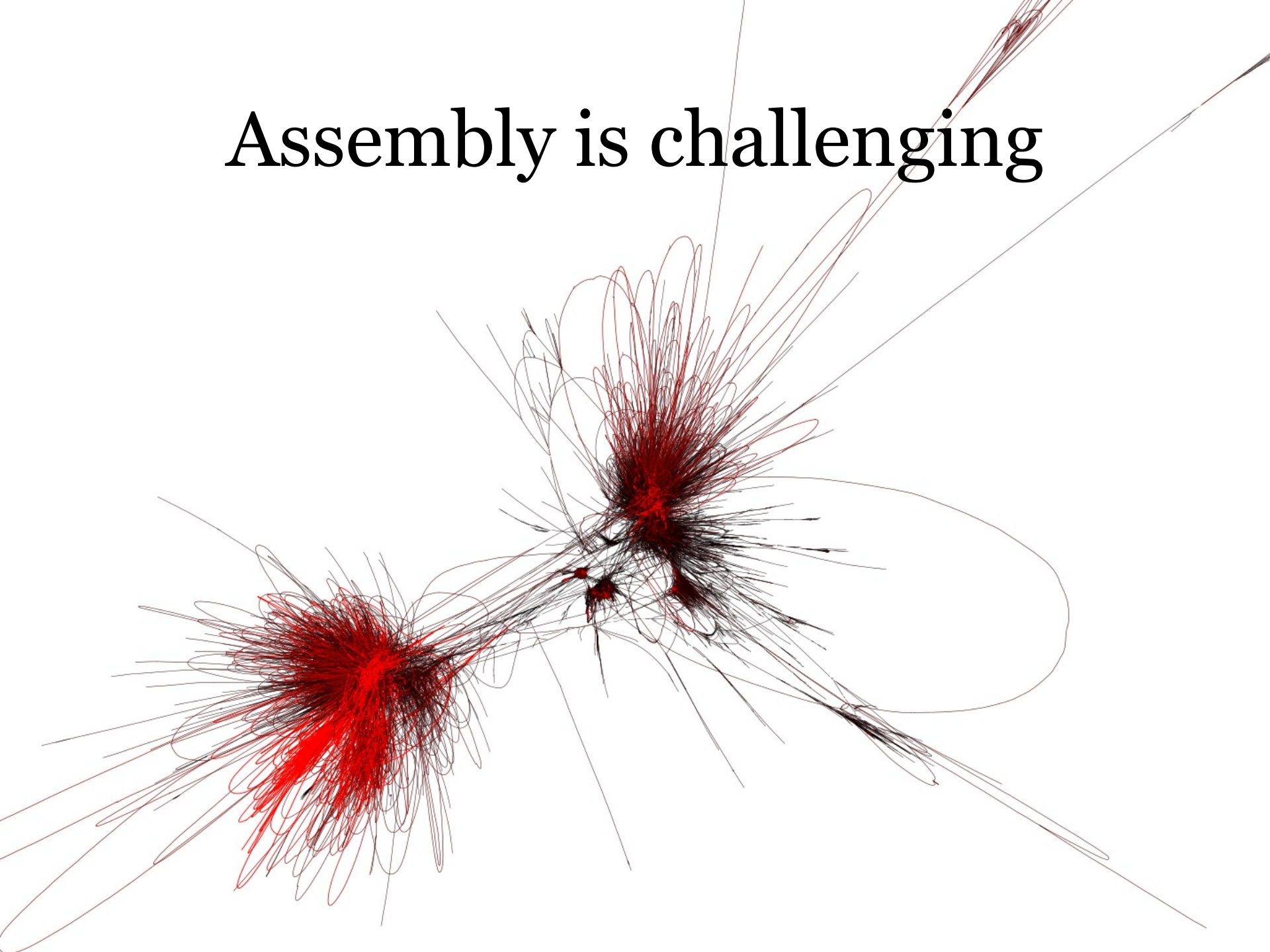
Reads

Assemble

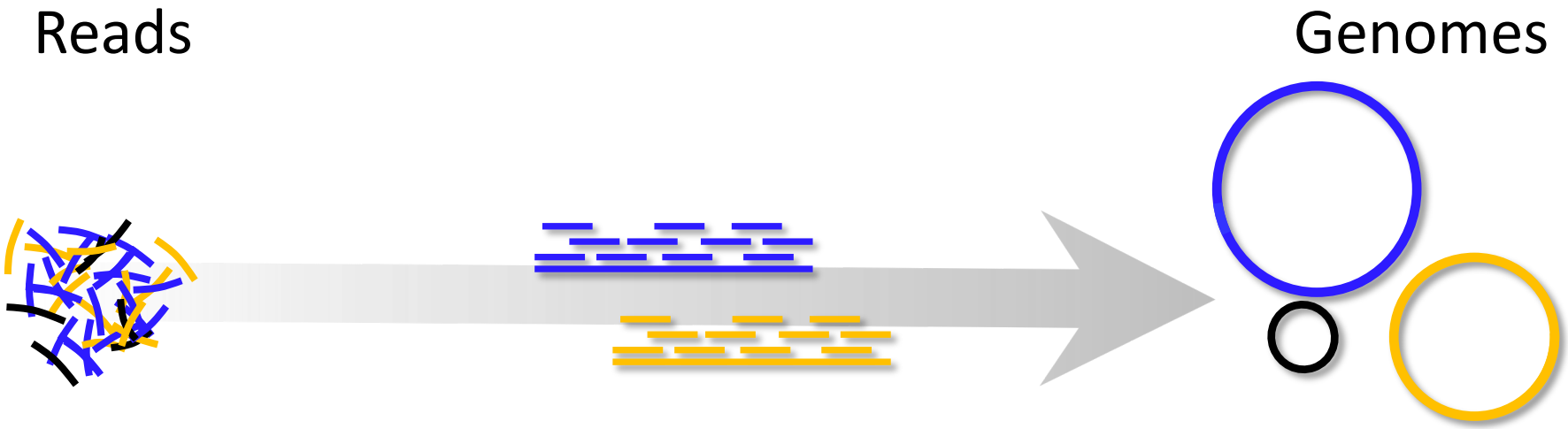
Genomes



Assembly is challenging



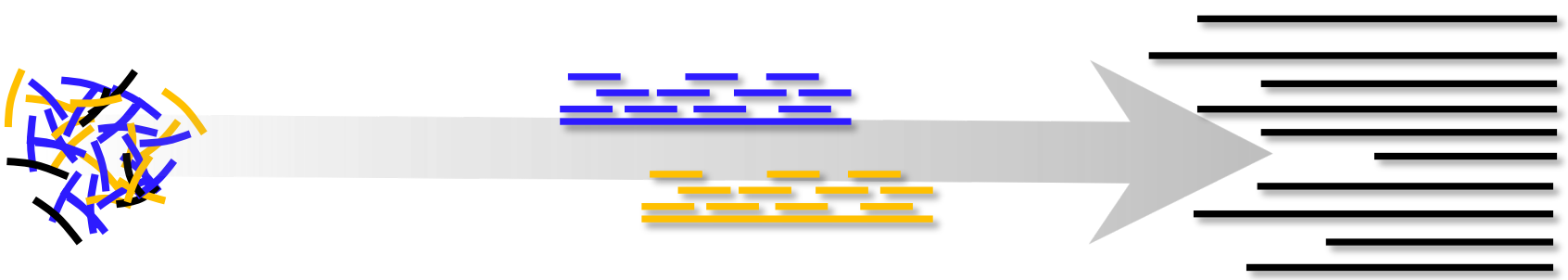
Goal: Assemble genomes from metagenomes



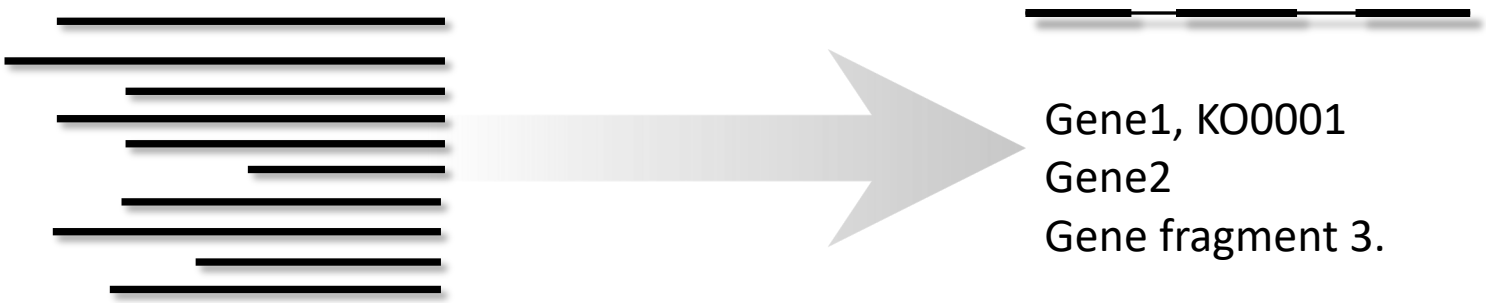
Goal: Assemble genomes from metagenomes

Reads

Contigs



Annotate genes on contigs

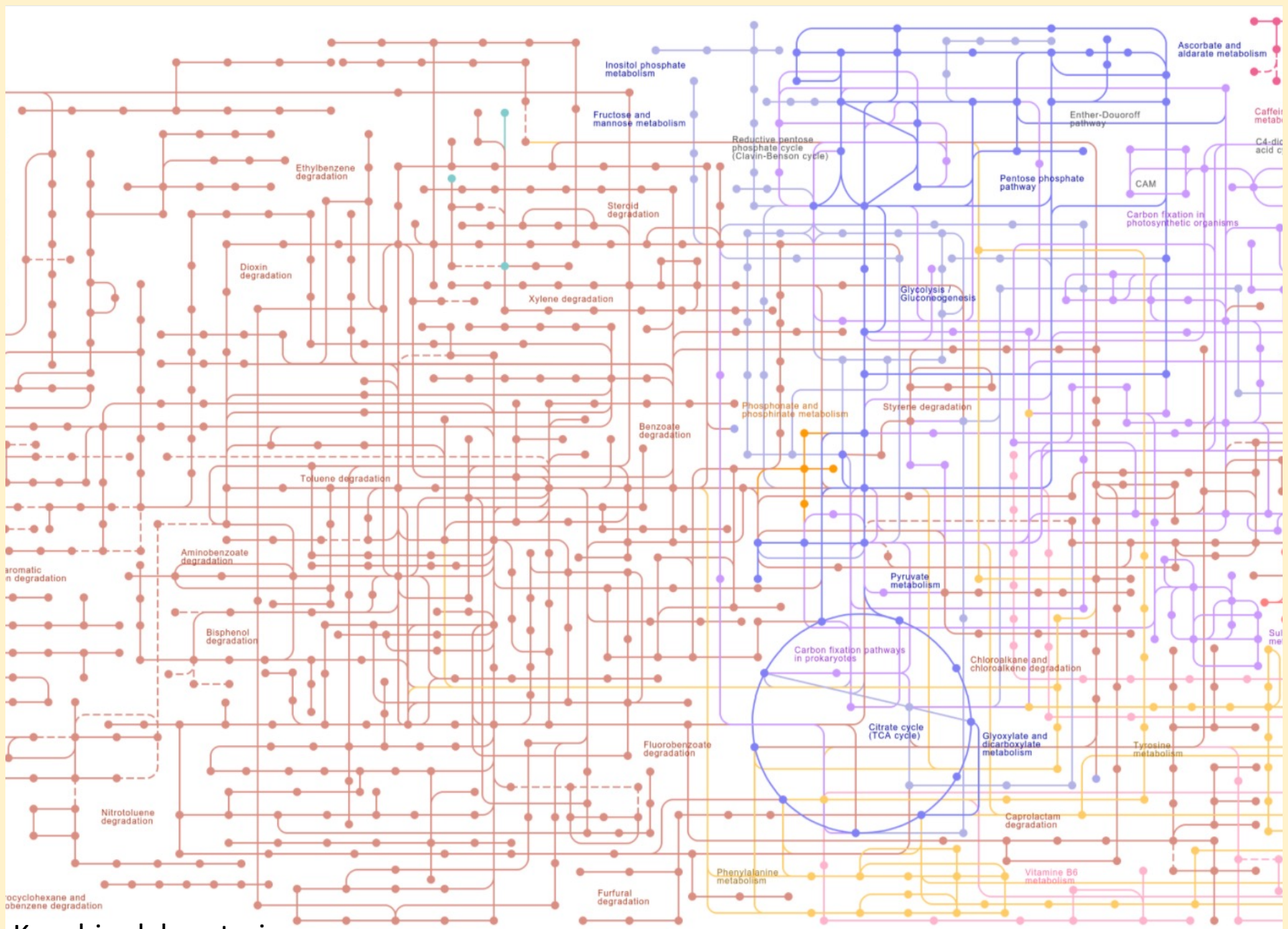


Era of gene catalogs

Functional potential

Bacteroidetes sp

Taxonomy



Kanehisa laboratories

Analyze a metagenome

Who is there?

What are they doing?

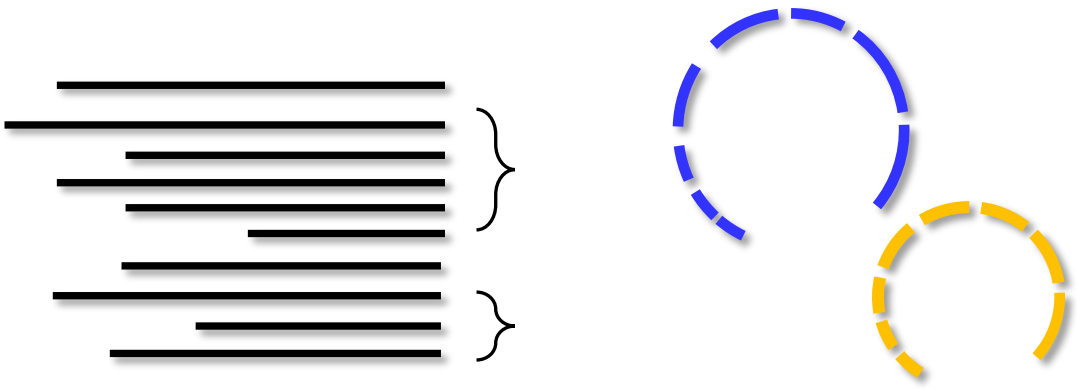
Analyze a metagenome

Who is doing
what?

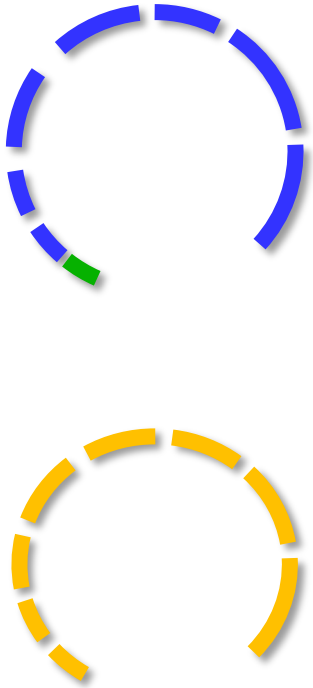
“Genes are expressed within cells, not in a homogenized cytoplasmic soup.”

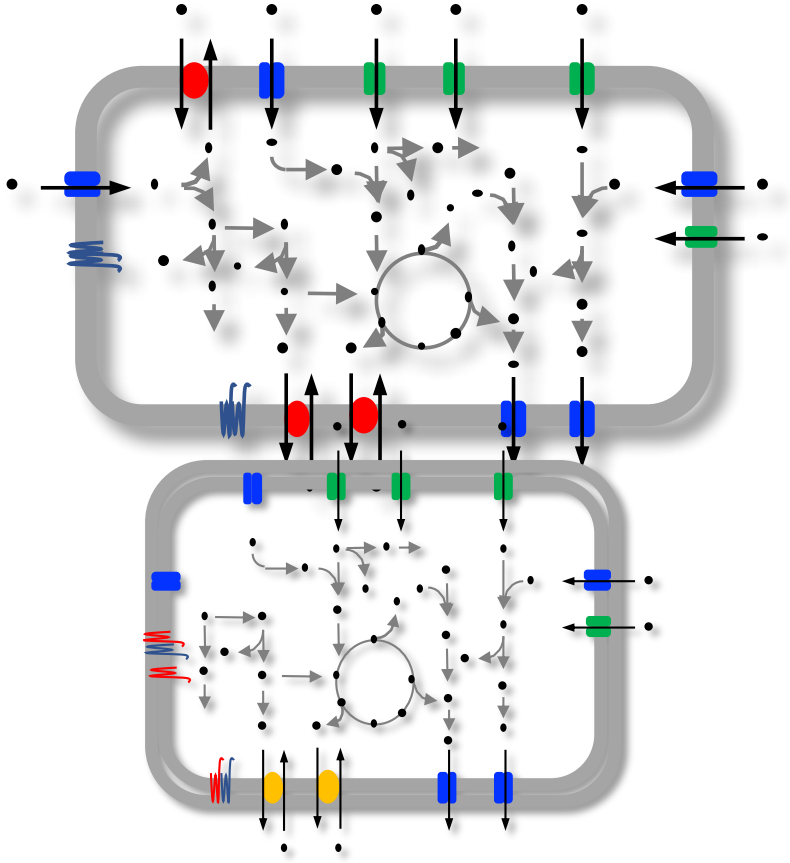
Katherine McMahon

Binning



Functional annotation

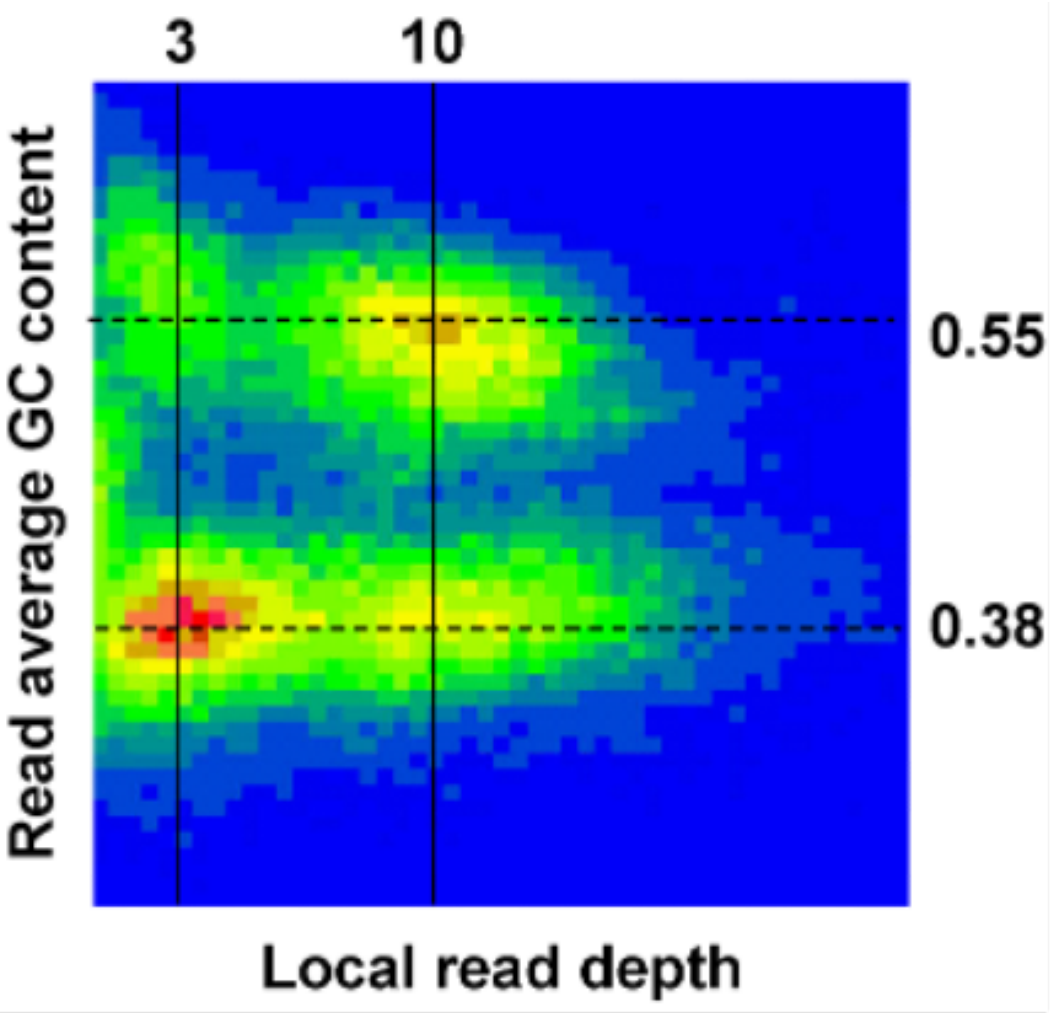




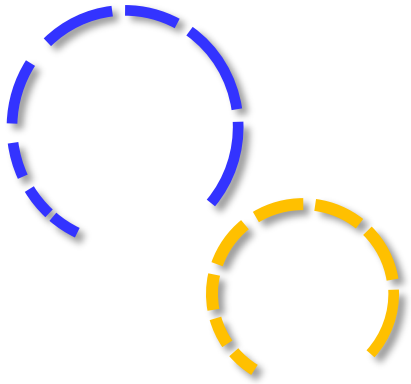
Why is is called “binning”?

The beginning of binning

Tyson et al. 2004


3
10
Read average GC content
0.55
0.38
Local read depth

Metagenome-assembled genomes (MAGs)




Trefl
PUZZLE 500

1. Genecatalogs

2. MAGs

Metagenome-Atlas

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!

```
mamba install metagenome-atlas
atlas init path/to/fastq
atlas run genomes
```

1 Dependency


```
conda install <the tool I need>
```

Snakemake

Why you need **Snake**make?

Snakemake

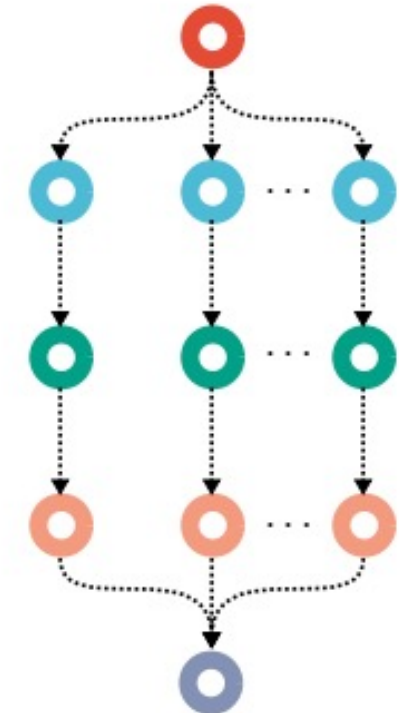
Create rules

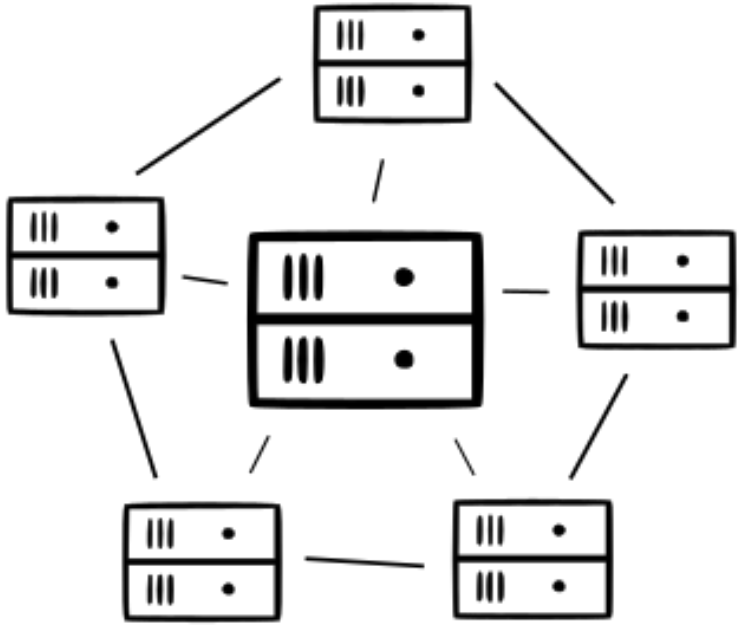
```
rule plot:
    input:
        "raw/{dataset}.csv"
    output:
        "plots/{dataset}.pdf"
    shell:
        "somecommand {input} {output}"
```

Install dependencies automatically

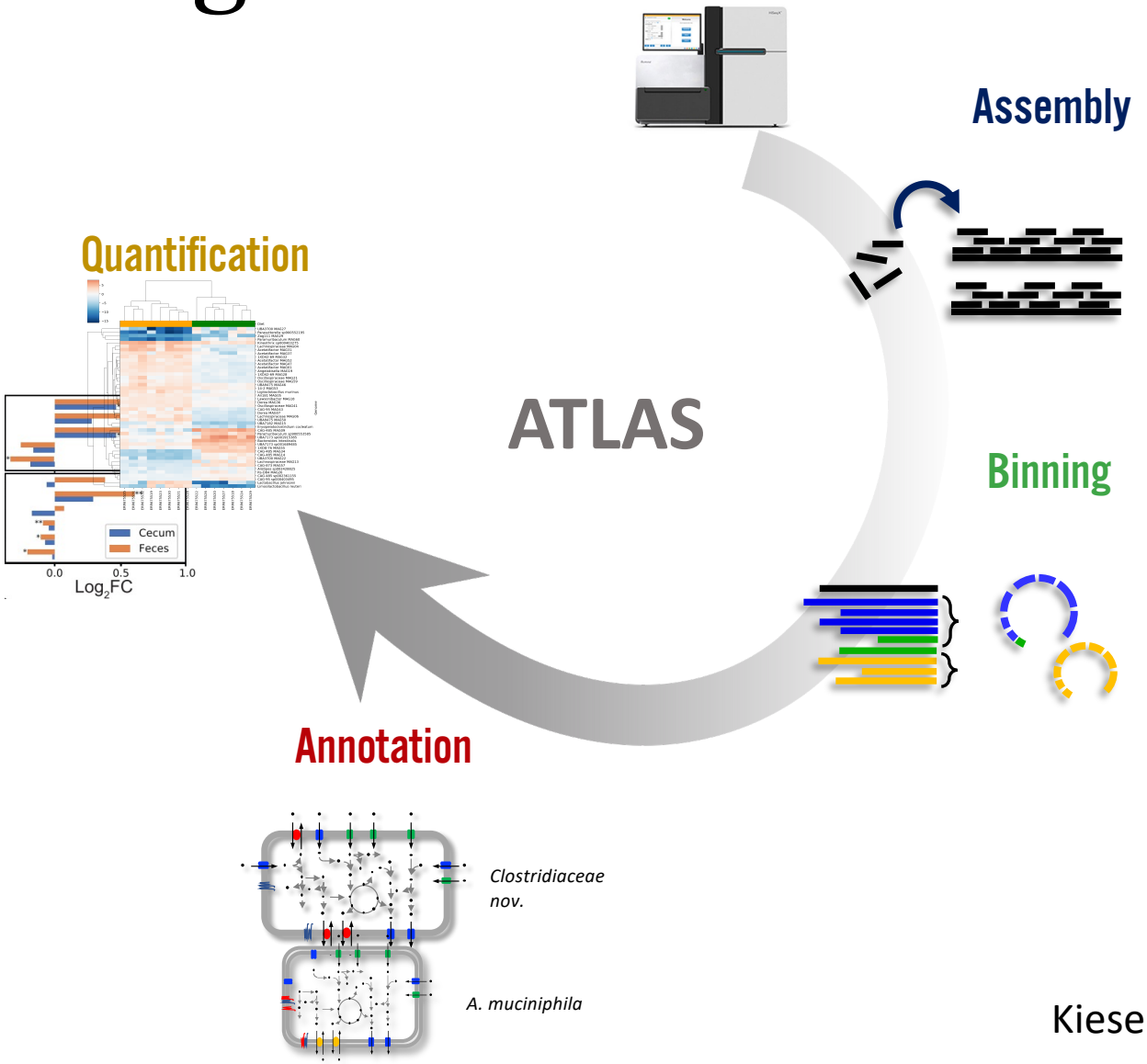
```
channels:
  - bioconda
  - r
dependencies:
  - python=2.7
  - checkm-genome=1.0.7
  - prodigal >=2.6.1
```

Snakemake





Metagenome-Atlas in detail



Kieser *et al.* 2020

Start a project

```
mamba install metagenome-atlas
atlas init path/to/fastq
```

Samples.tsv

	Reads_R1	Reads_R2
sample1		
sample2		

Config.yaml

```yaml
#######################
# Binning
#######################

final_binner: DASTool             # [SemiBin, DASTool, vamb, maxbin]

binner:                           # If DASTool is used as final_binner,
  - metabat
  - maxbin
#  - vamb


metabat:
  sensitivity: sensitive
  min_contig_length: 1500 # metabat needs >1500

maxbin:
  max_iteration: 50
  prob_threshold: 0.9
  min_contig_length: 1000
```

Start a project

```
mamba install metagenome-atlas
atlas init-public SRA_ID
```

Run atlas

```
atlas run genecatalog
```

```
atlas run genomes
```

1. Quality control

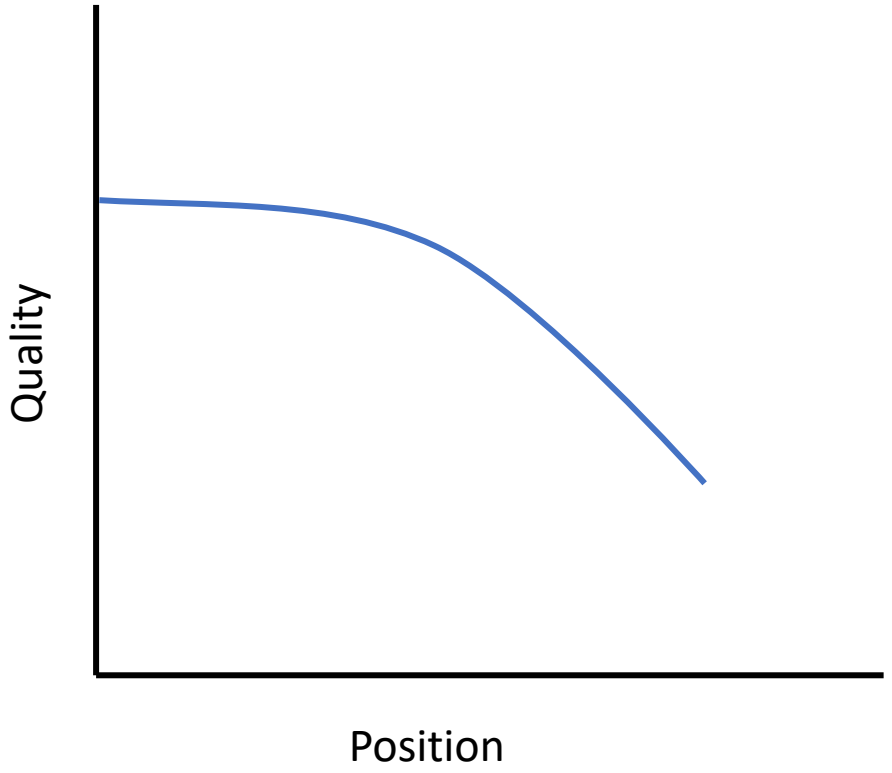


Image: Illumina Inc

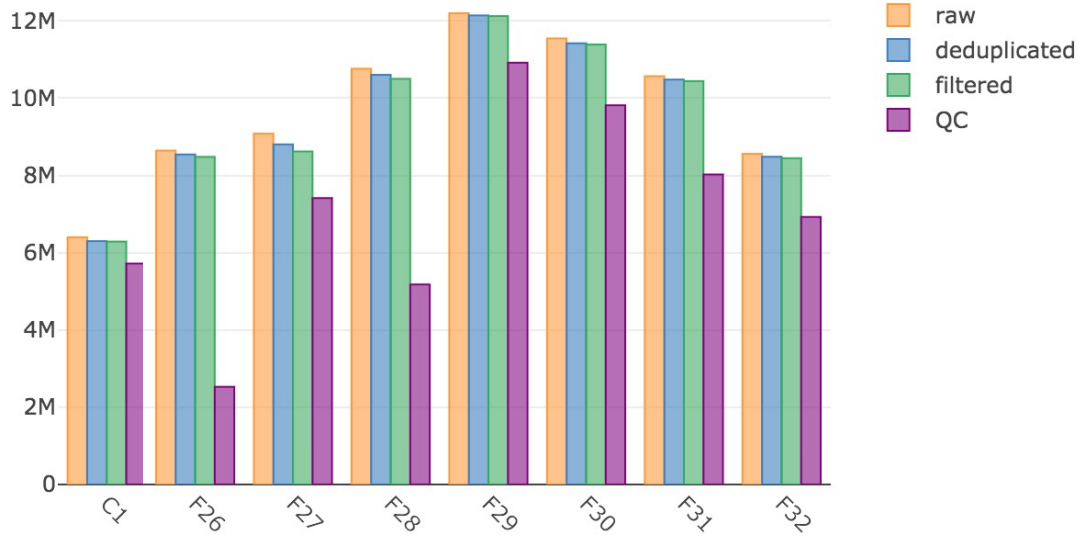
1. Quality control

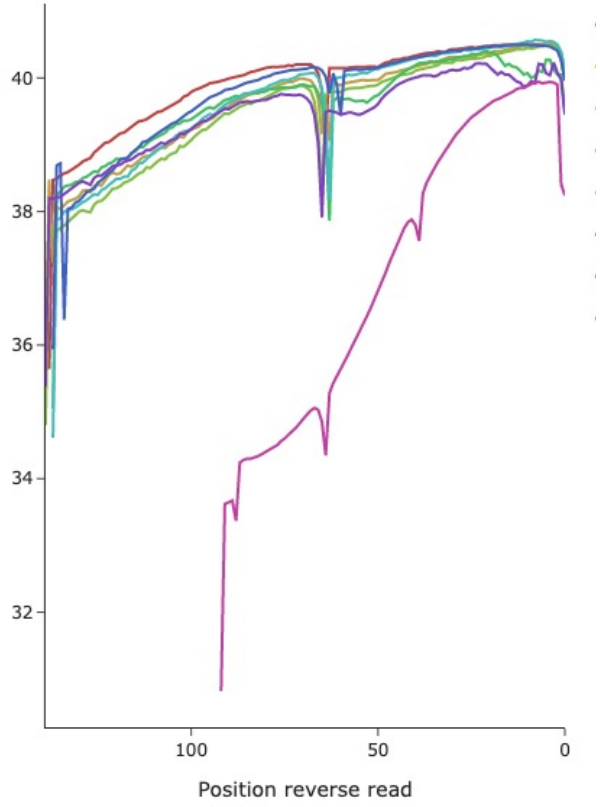
- Host removal

➢ **Good-quality reads**

Quality report

Total reads per sample





2. Assembly

2. Assembly

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

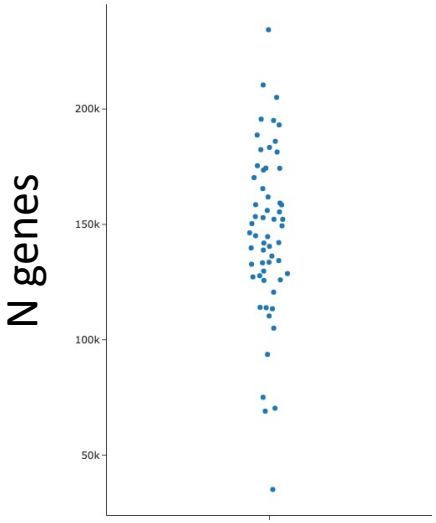
2. Assembly

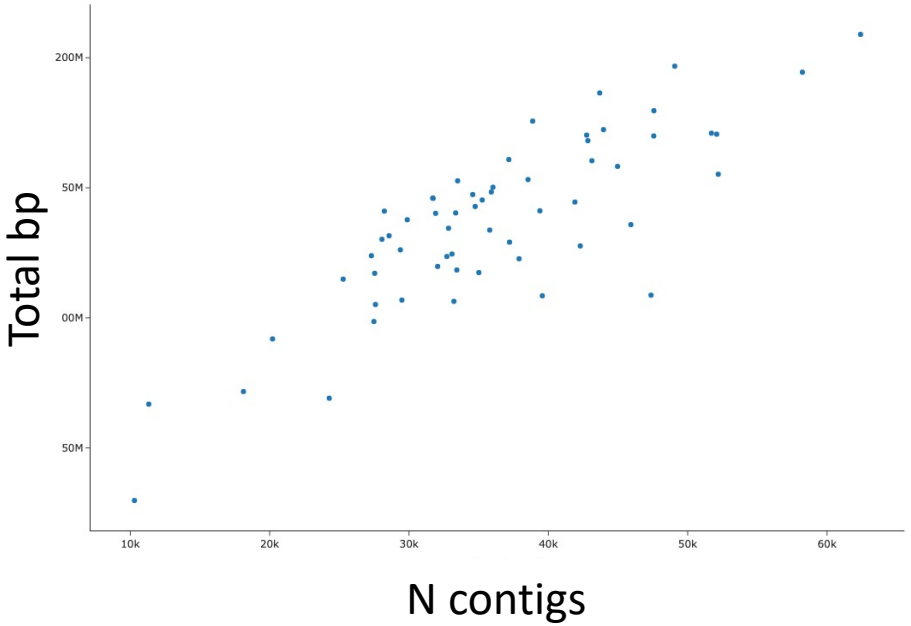
- Uses metaSpades or megahit
- Pre-processing
  - Error correction
  - Paired-end merging (pre-assembly)

2. Assembly

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

Assembly report





3. Binning

3. Binning

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

How do we bin contigs into genomes?

Sequence Features

ATCGTCACGTAA

Abundance Features

How do we bin contigs into genomes?

Sequence Features

ATCGTCACGTAA

Abundance Features

Annotations

Co-Abundance Features

3 Binning

Single-sample / Cross mapping:

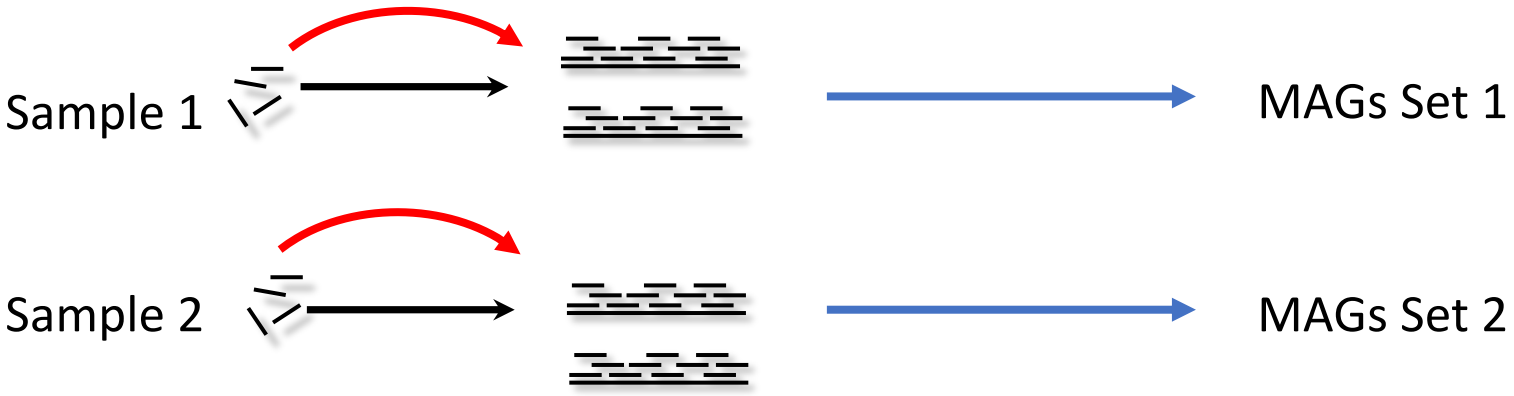
- Metabat2
- Maxbin2

Co-Binning

- Vamb
- SemiBin

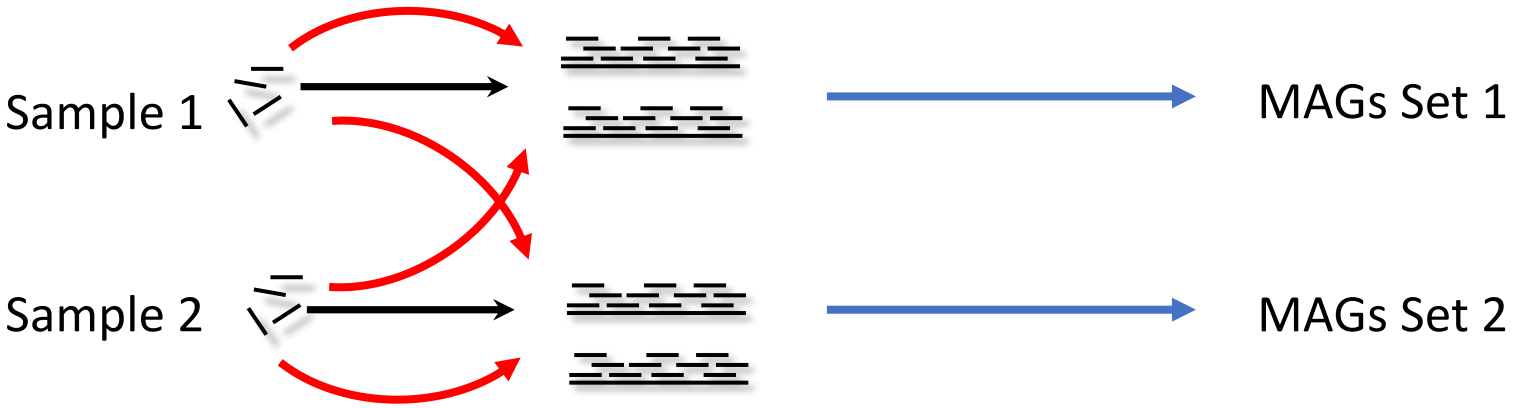
Co-abundance

Option 1: Single-sample assembly/Binnig

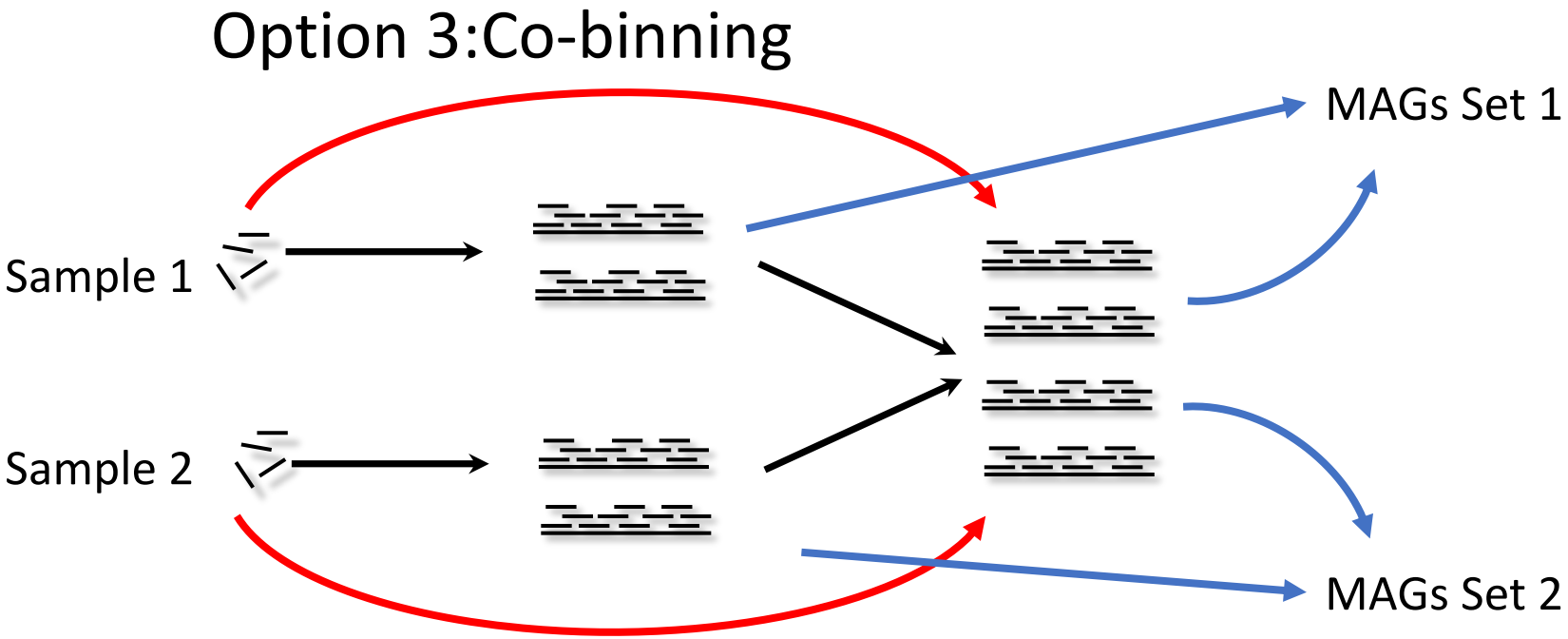


Co-abundance

Option 2: Cross mapping



Co-abundance



3 Binning

Single-sample / Cross mapping:

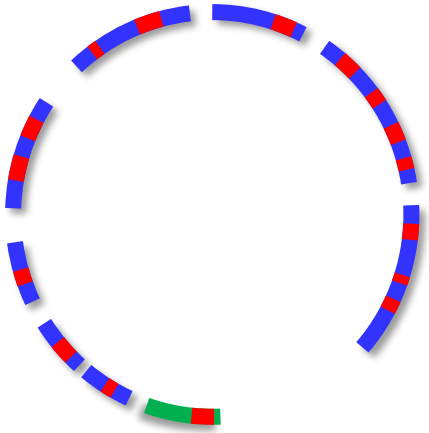
- Metabat2
- Maxbin2

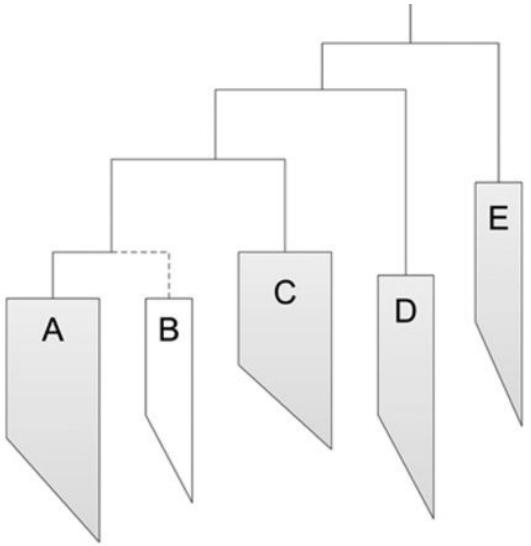
Co-Binning

- Vamb
- SemiBin

Quality estimation

(Essential) single-copy genes

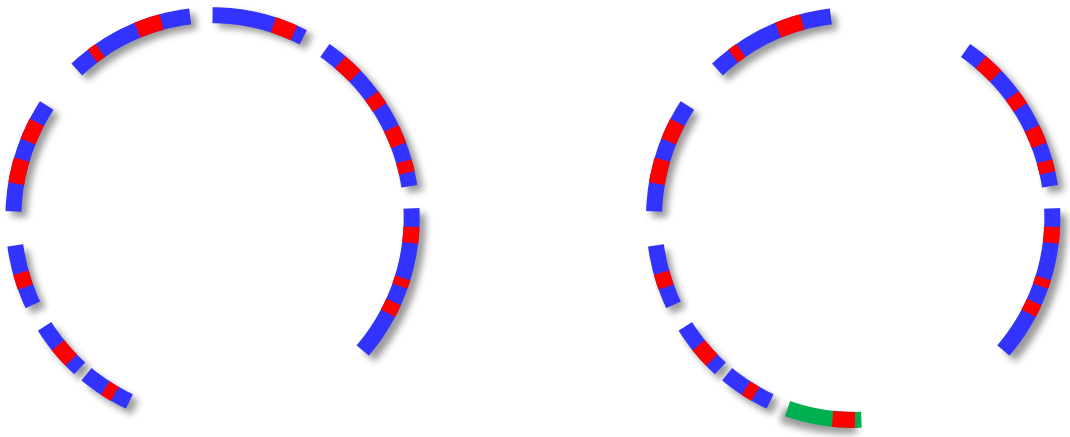




CheckM

Bin Refinement

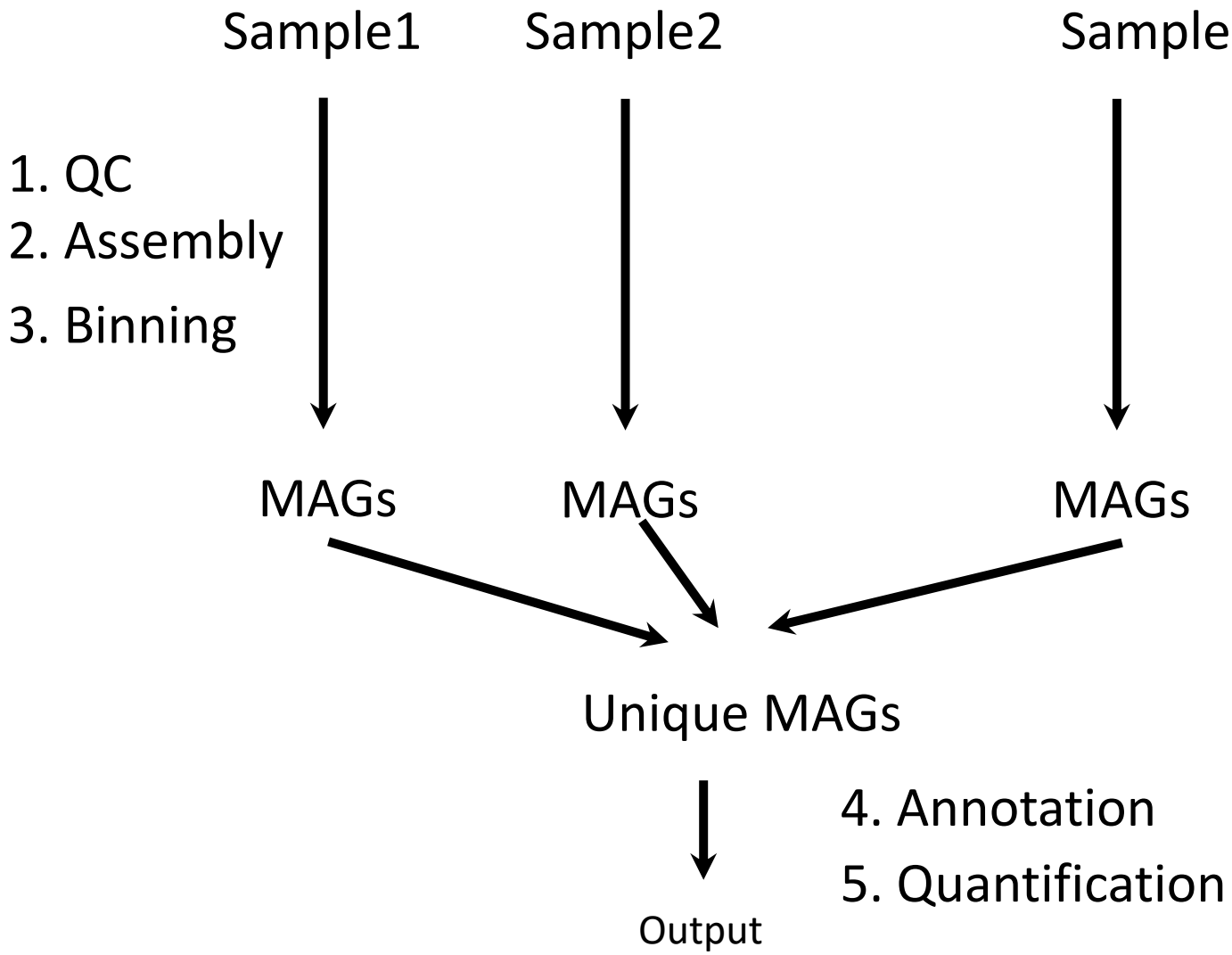
DAS Tool: Choose best Bin



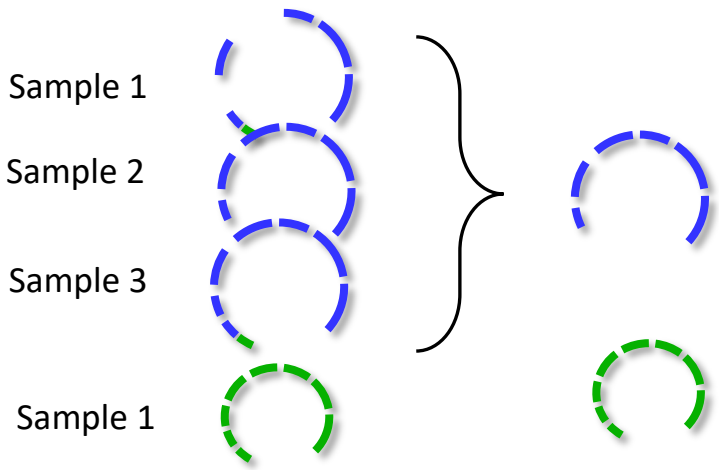
Atlas uses the same tools as large-scale studies on the Human microbiome

	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			

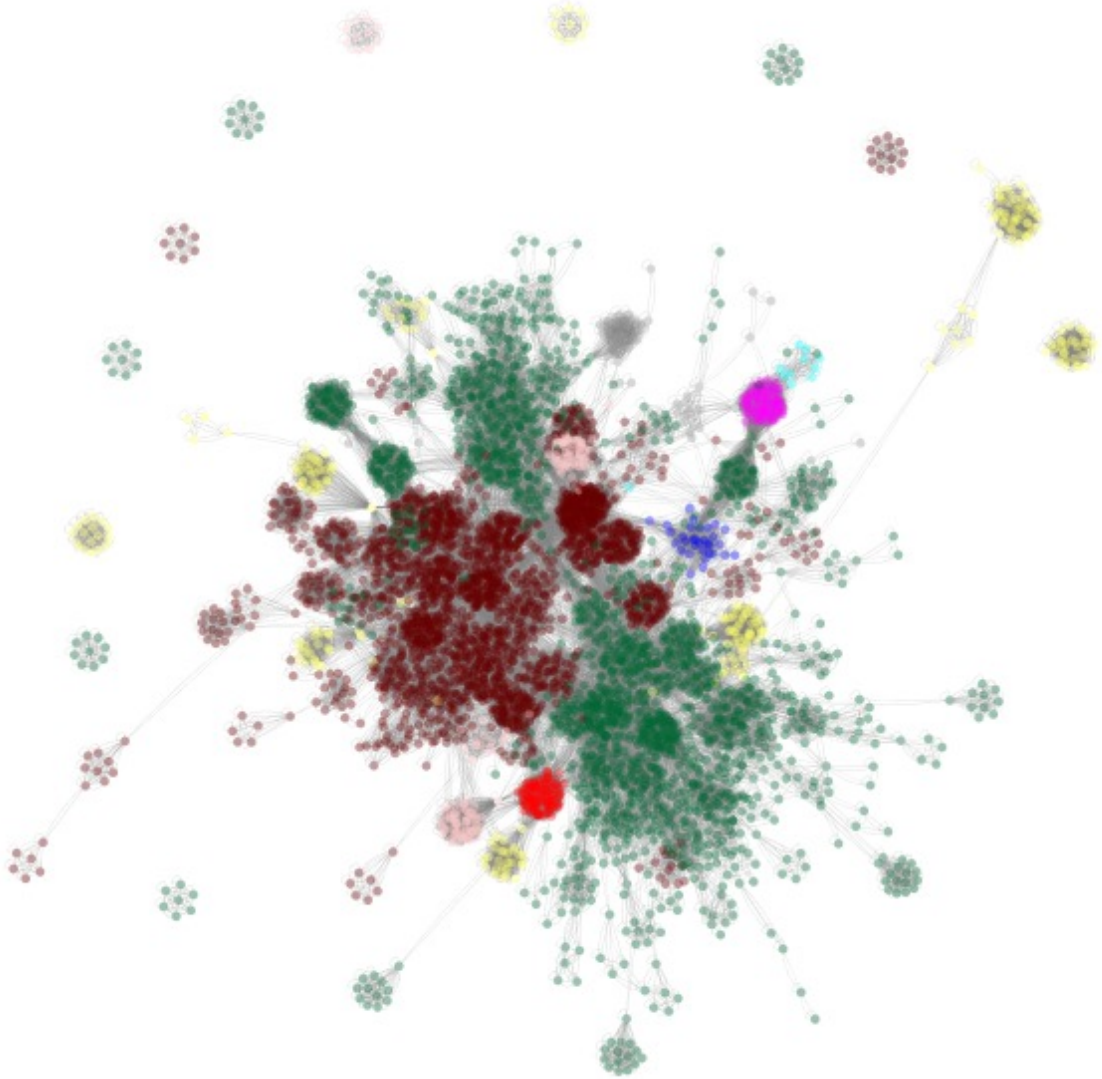
Atlas workflow



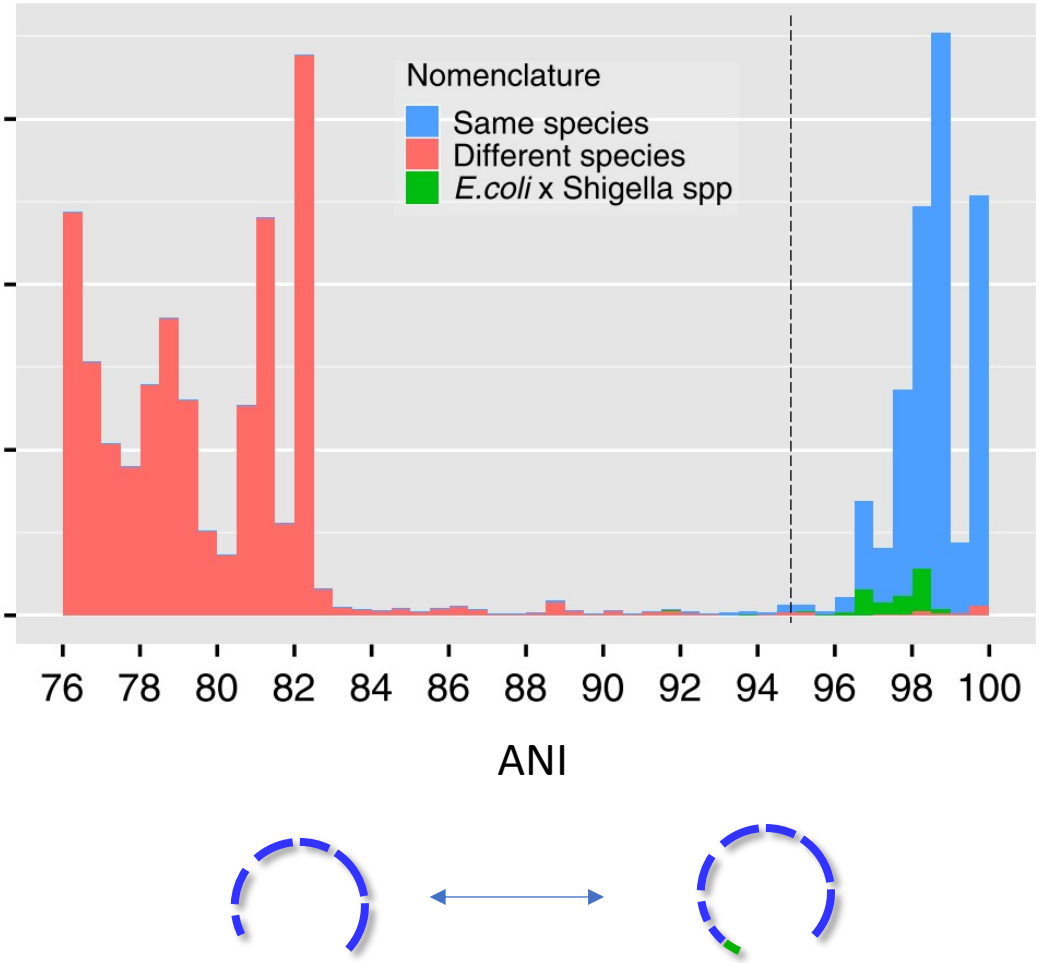
De-replication



Average nucleotide Identity (ANI)

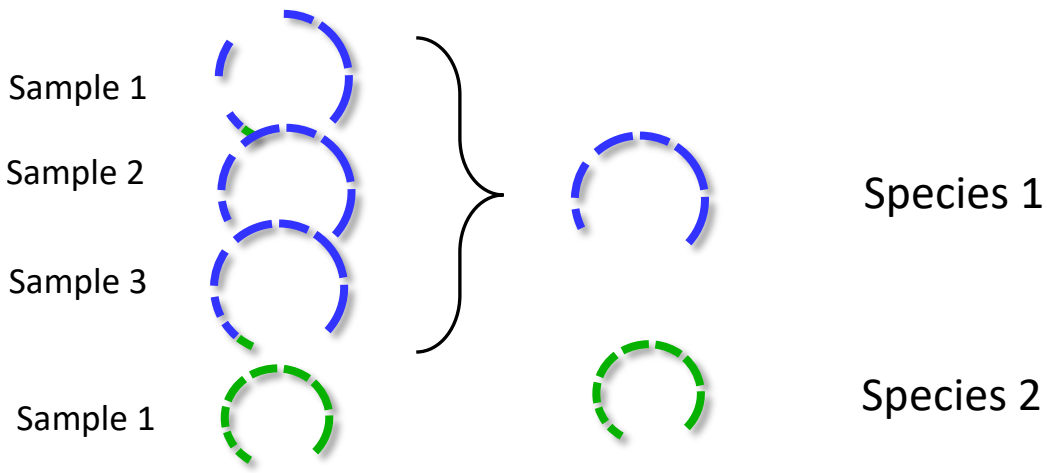


95% ANI used as species threshold



Jain et al. 2018

De-replication



4. Annotation

4. Annotation

What does it all mean?

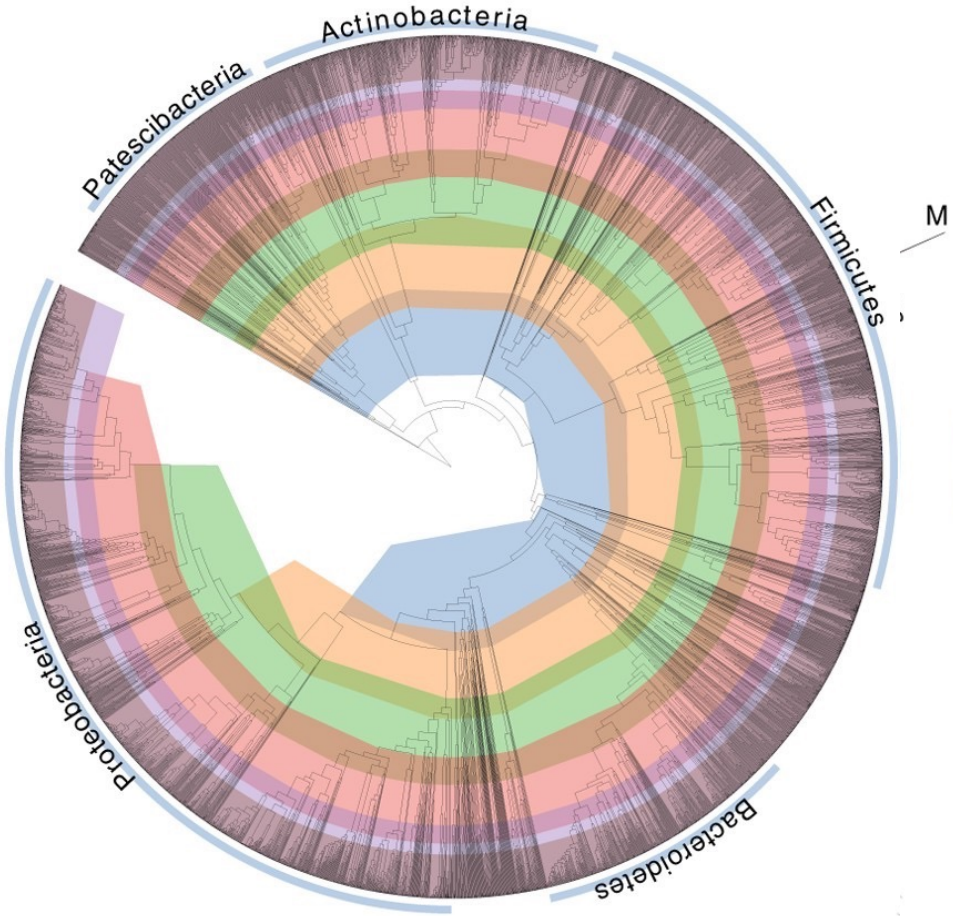
4. Annotation

a) Functions
b) Taxonomy

Taxonomic annotation

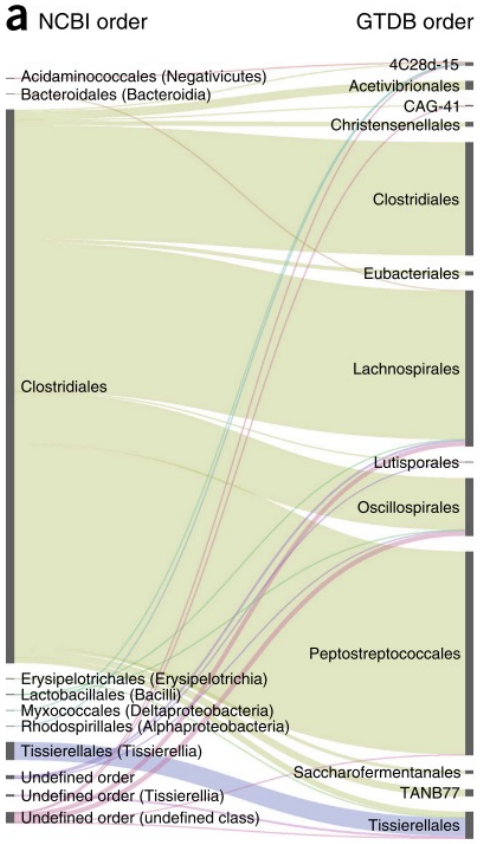
Genome Taxonomy database (GTDB)

Genome Taxonomy database



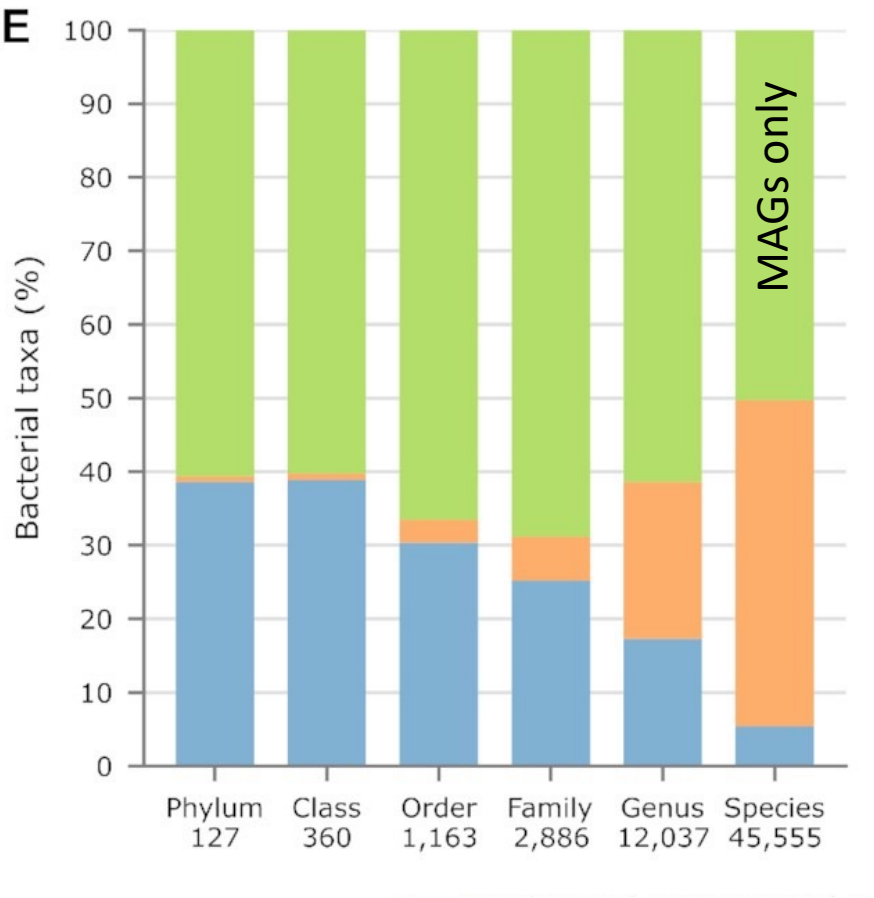
Parks et al. 10.1038/nbt.4229.

Proposed rearrangements



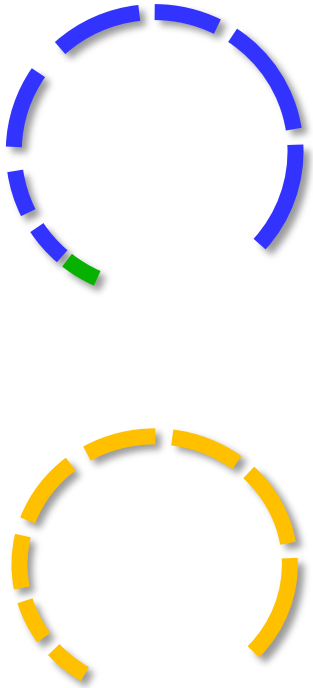
Parks et al. 10.1038/nbt.4229.

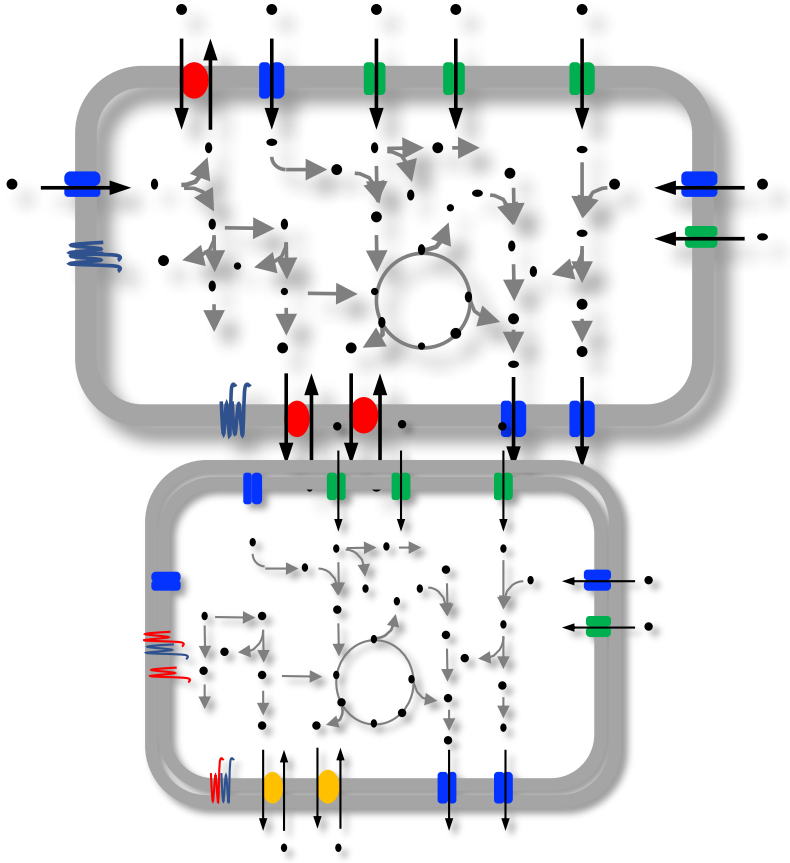
Genome Taxonomy database



Doi: 10.1093/nar/gkab776

Functional annotation

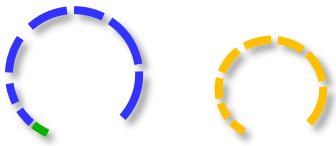




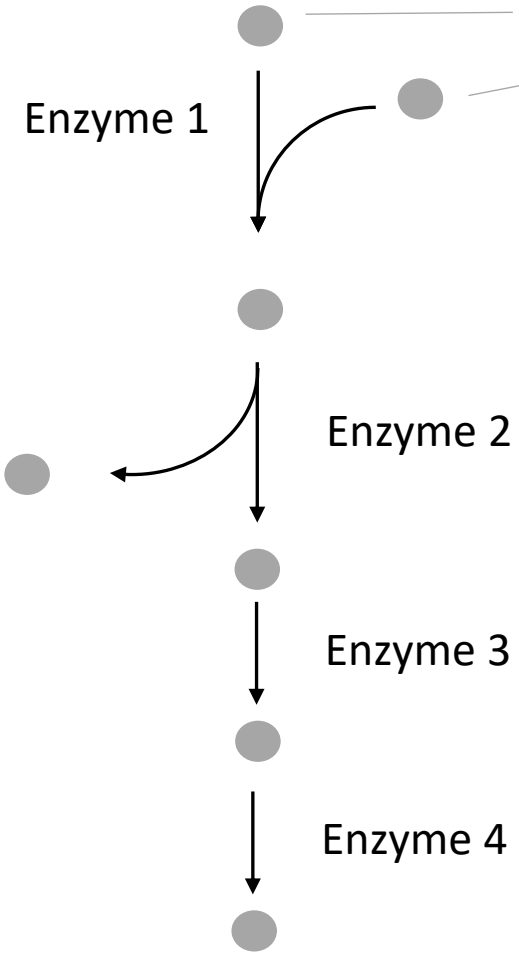
Functional annotation

Pathway inference

MAG 001 **MAG002**



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

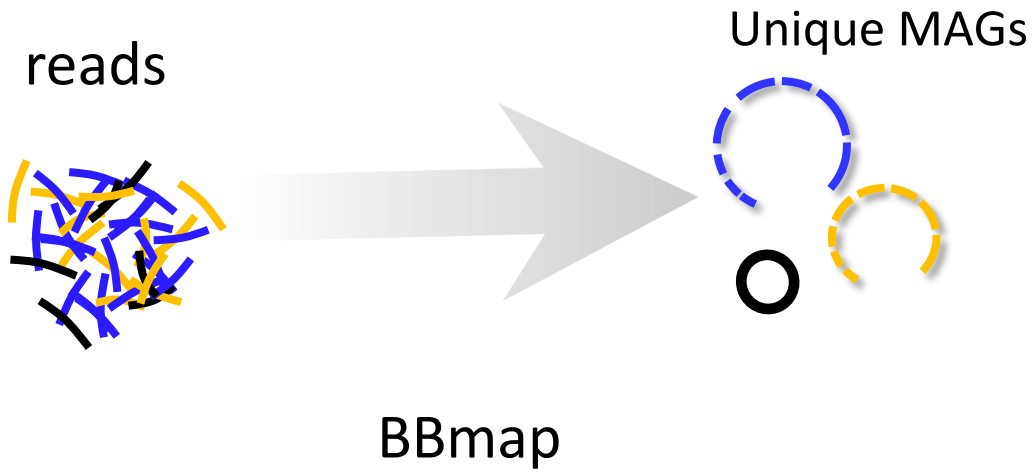


KEGG Kyoto Encyclopedia of Genes and Genomes

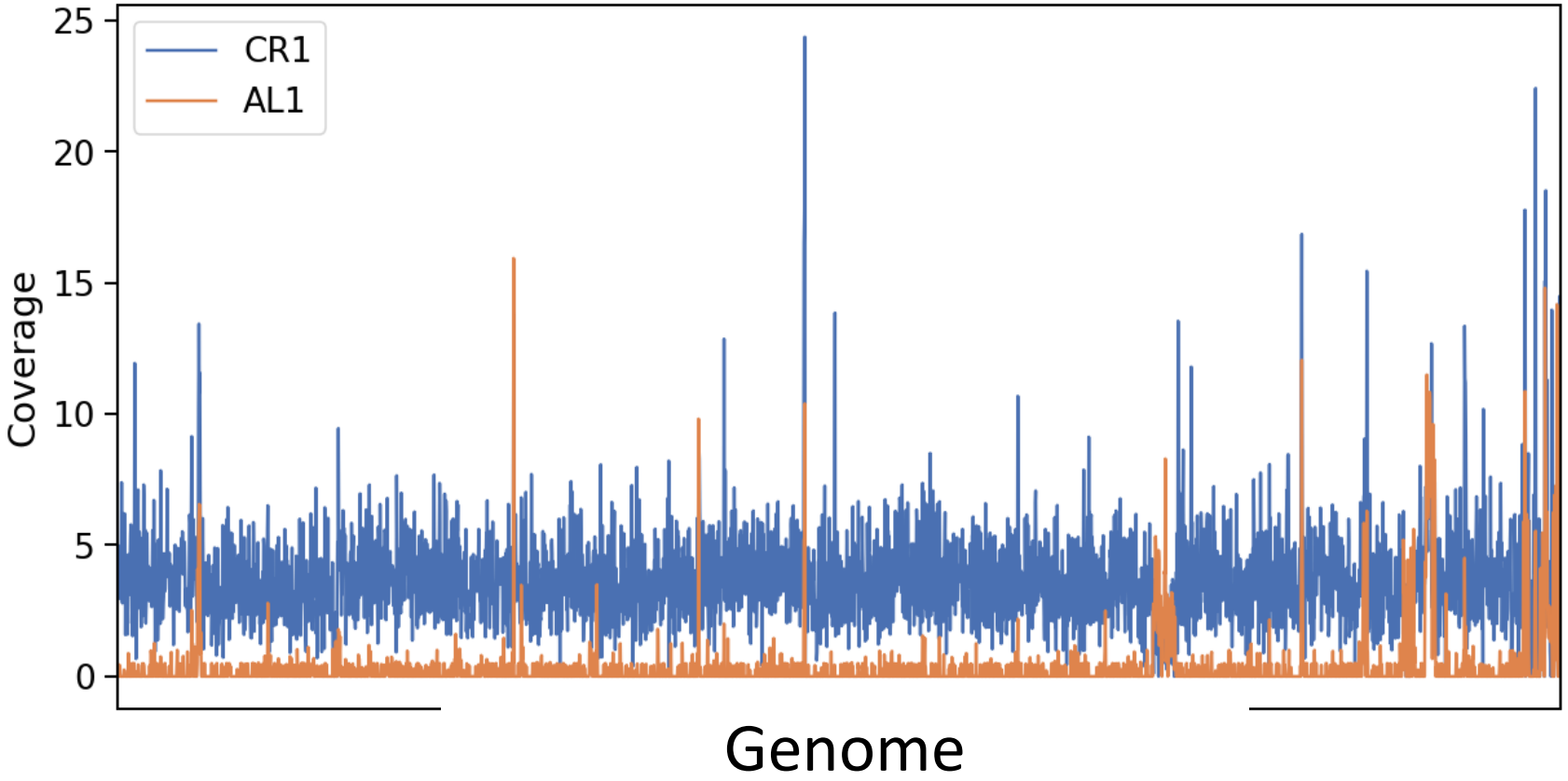
DRAM

Quantification

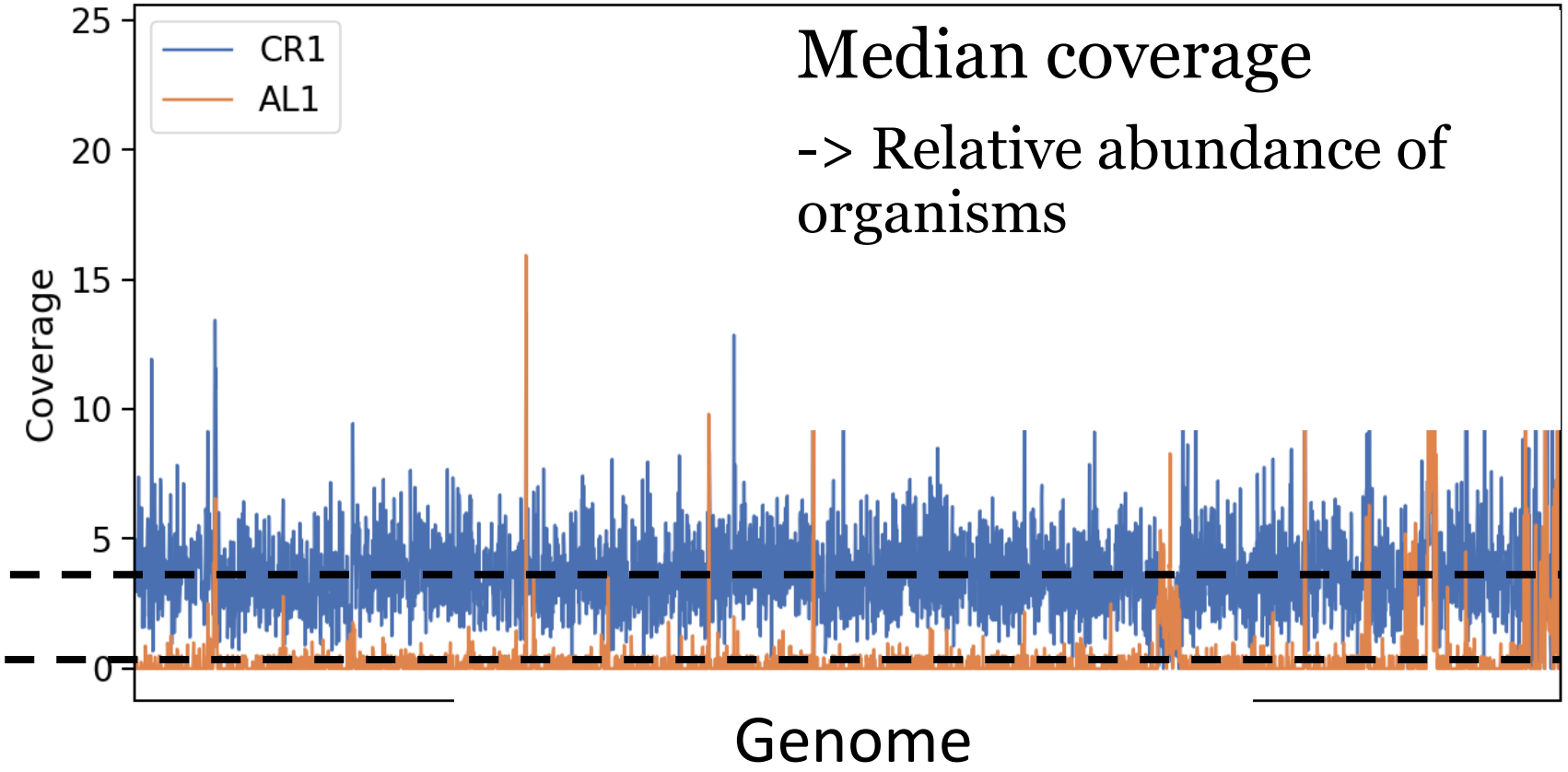
Quantification



What is the abundance of a genome?



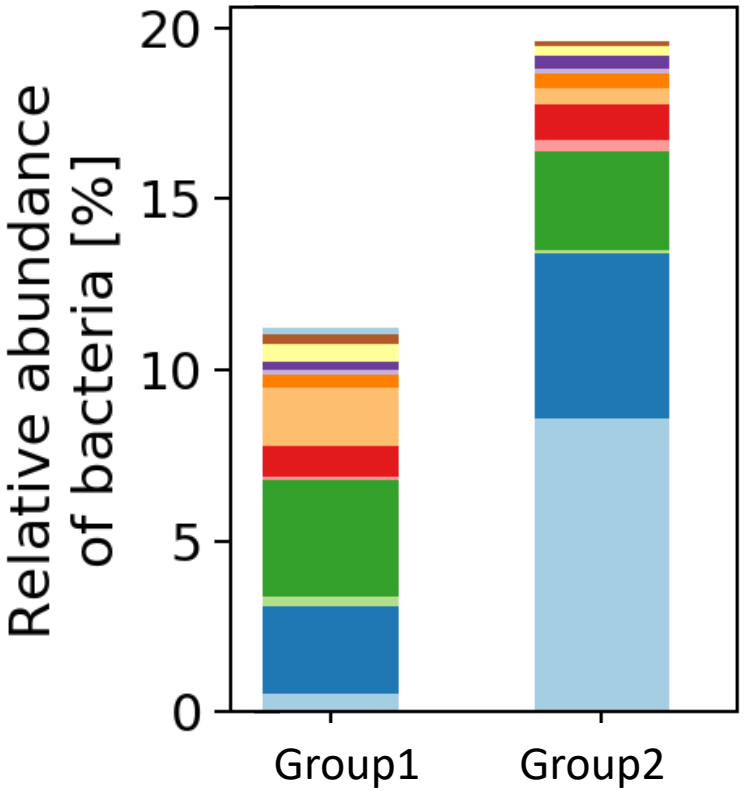
What is the abundance of a genome?



Abundance of pathways

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

Abundance of pathways

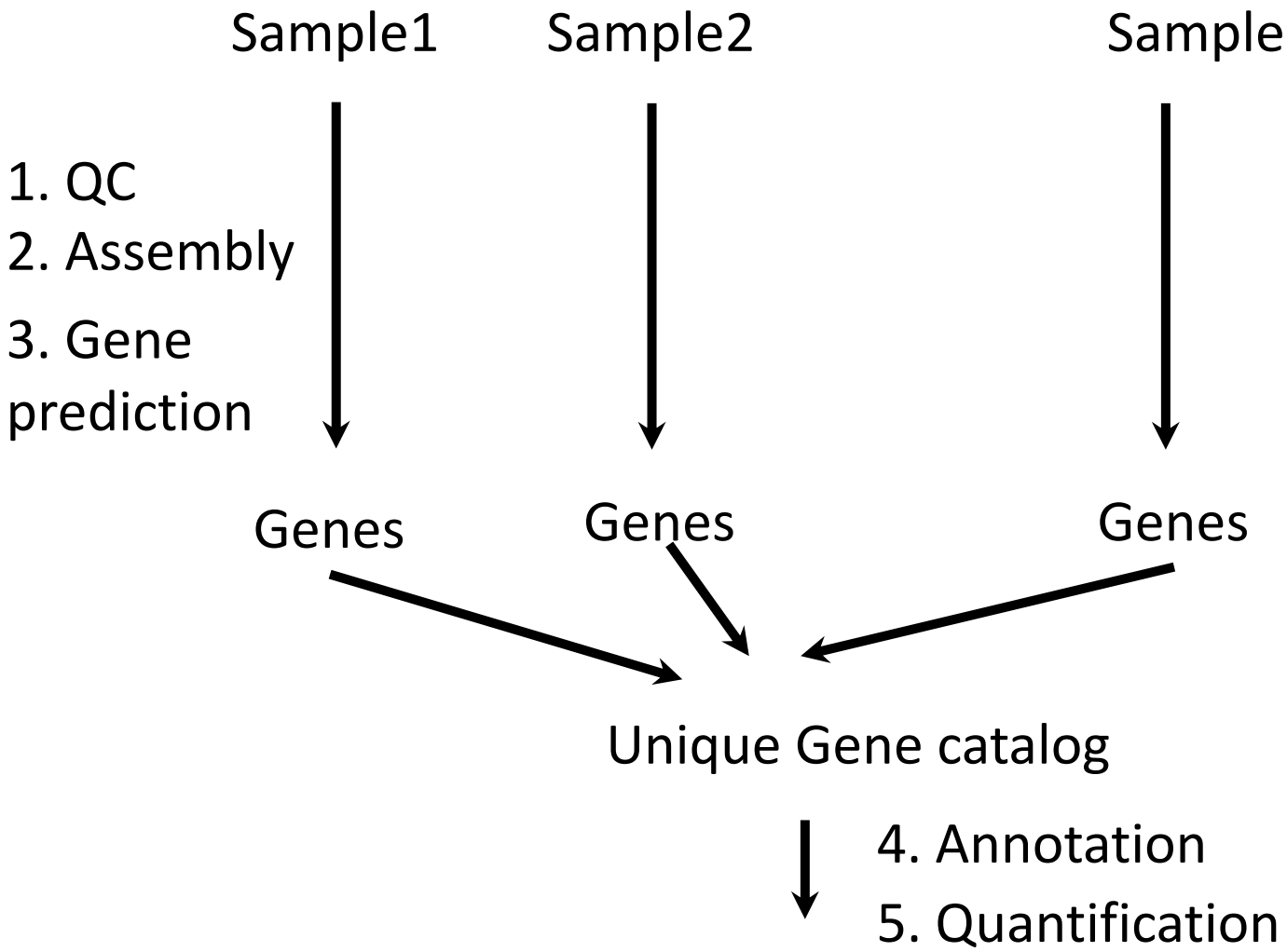


Gene catalog

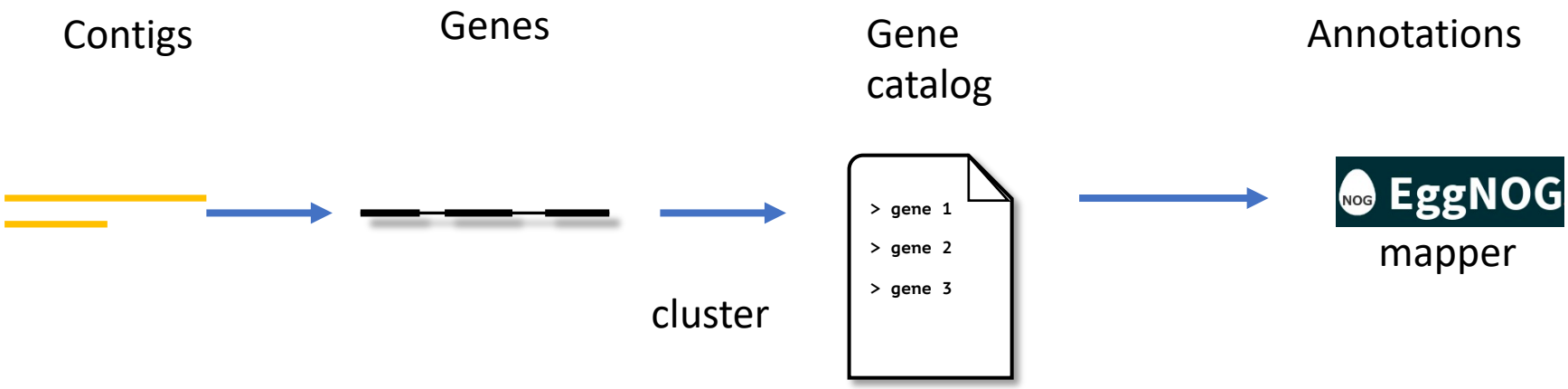
```
atlas run genecatalog
```

```
atlas run genomes
```

Atlas workflow



Annotation



- CAZy
- COGs
- (KEGG)

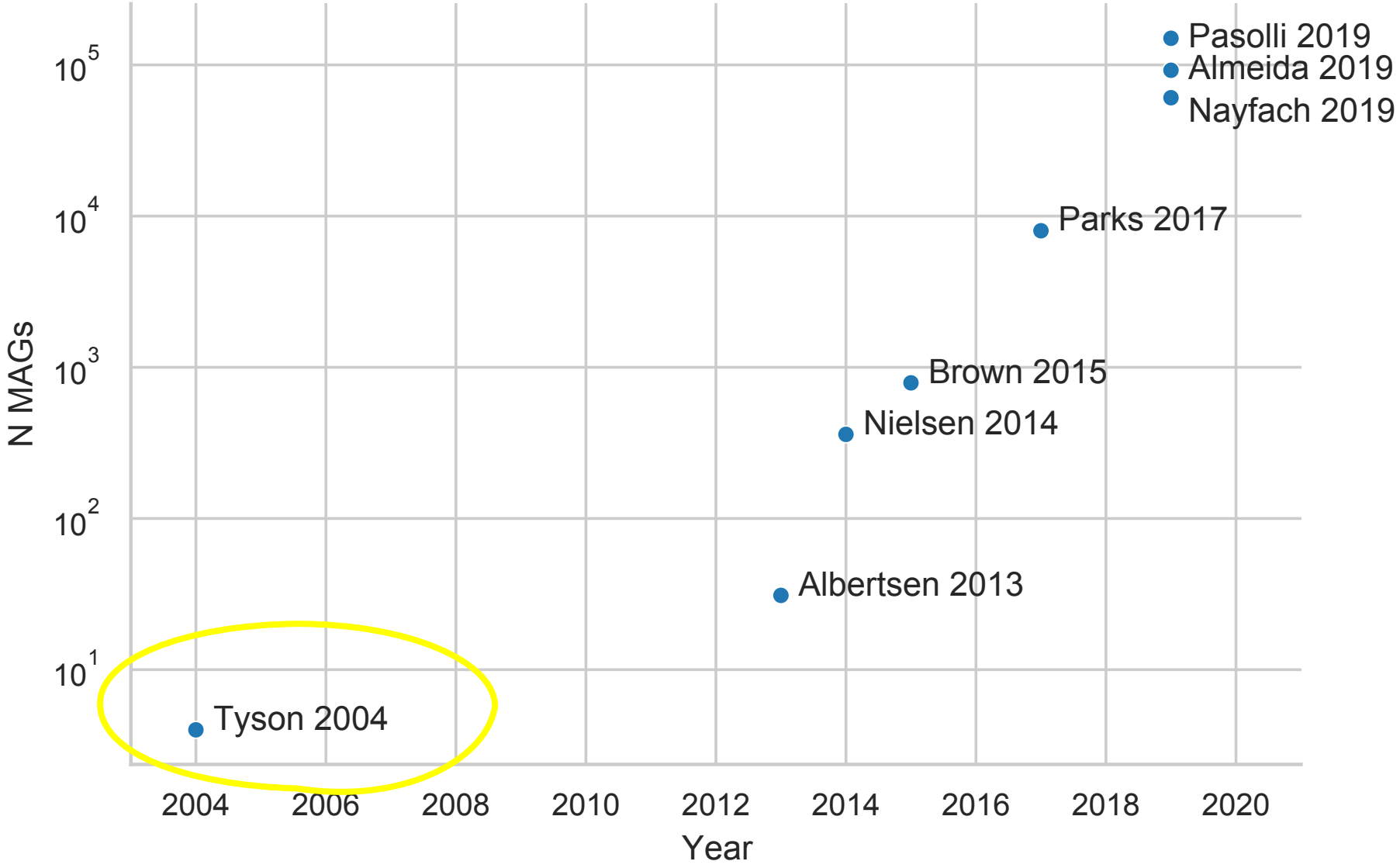
New features in V2.9

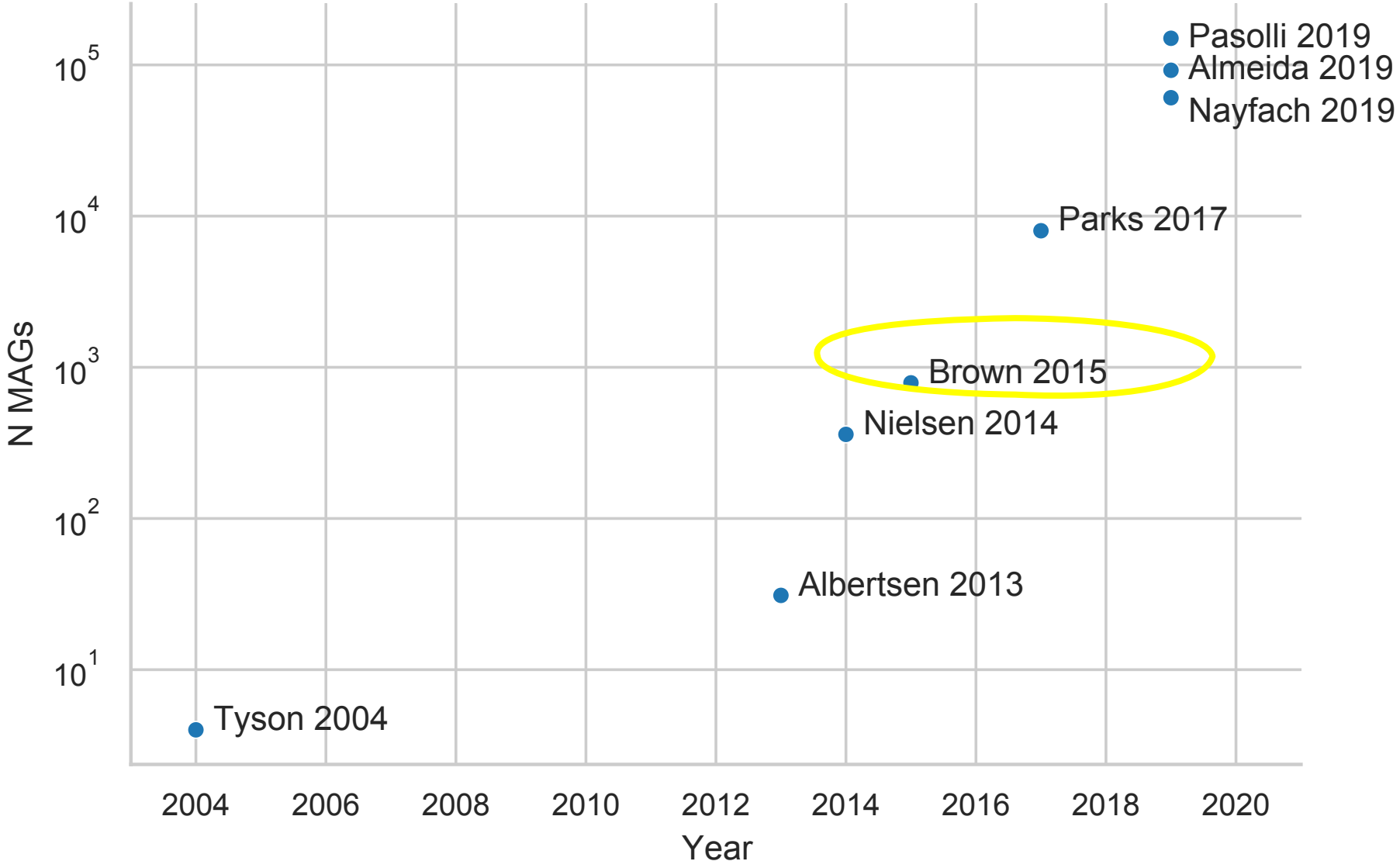
- Direct download from NCBI
- Strains

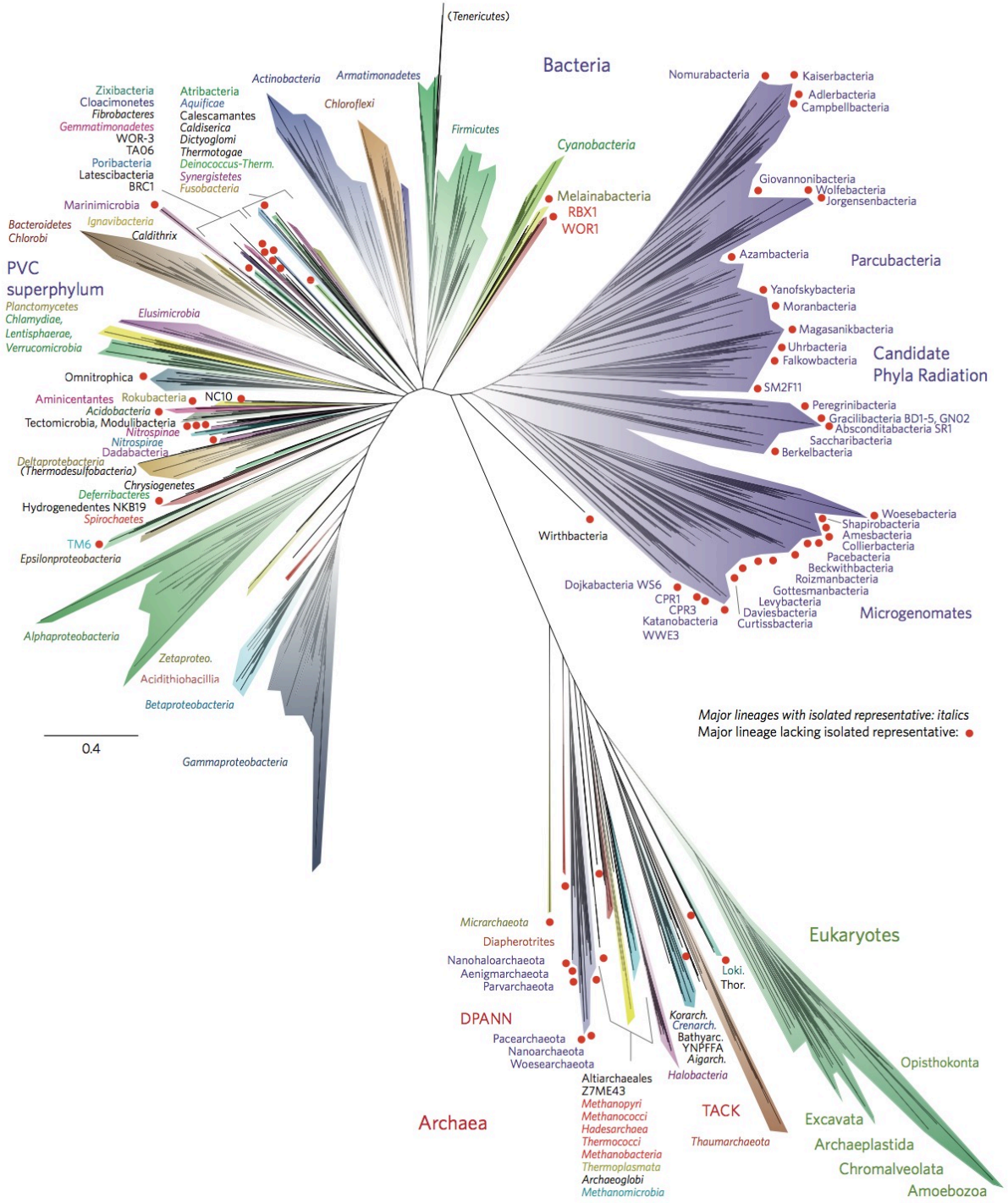
```
atlas run genomes strains
```

Questions

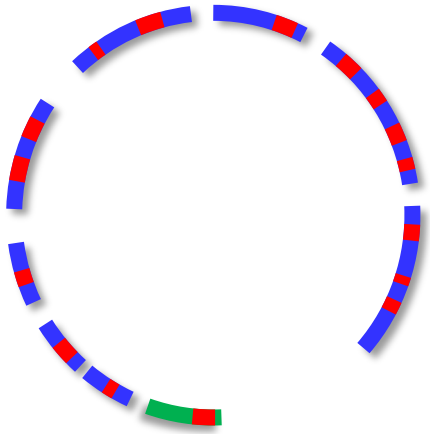
The future of assembly-based metagenomics

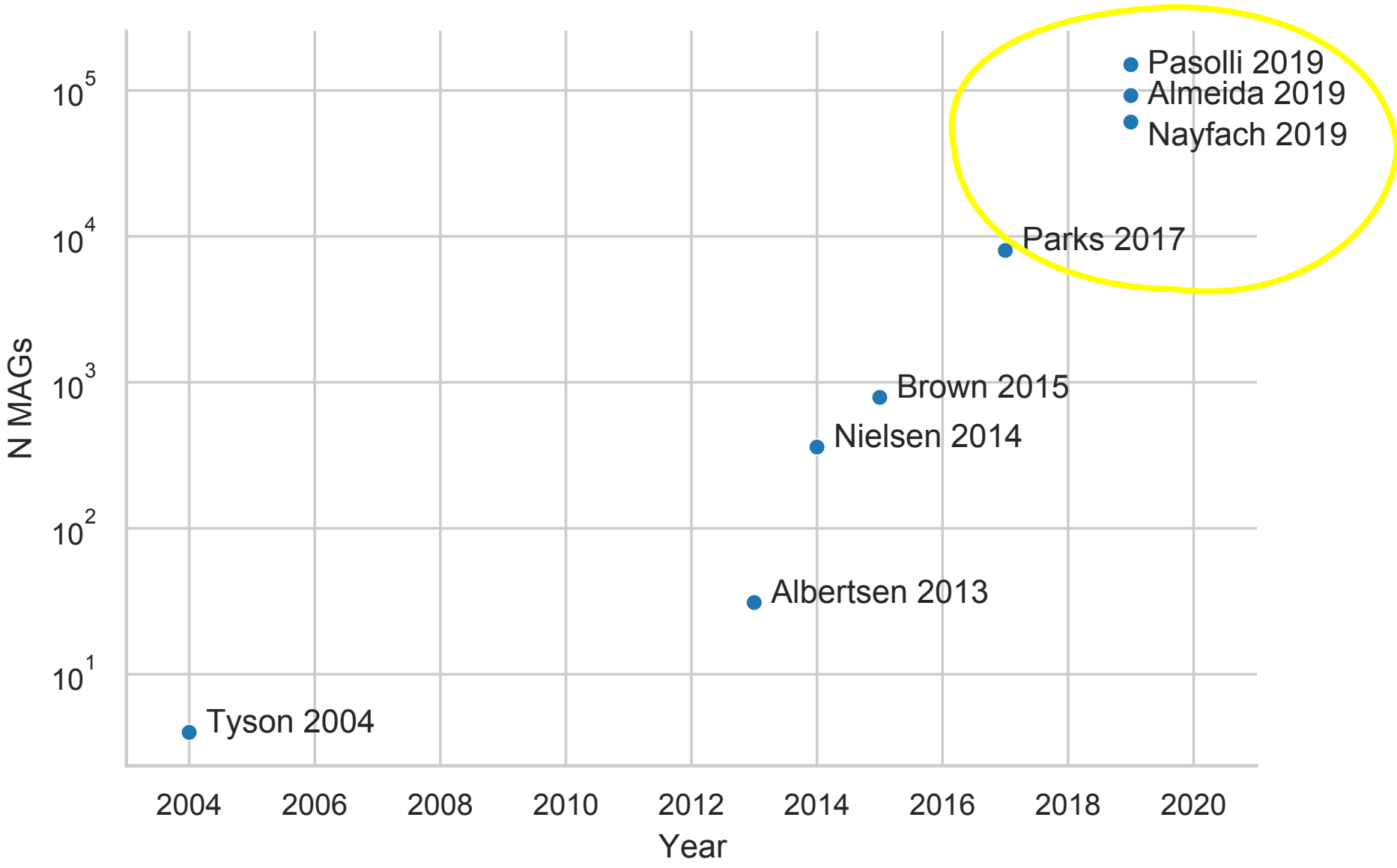

Pasolli 2019
Almeida 2019
Nayfach 2019
Parks 2017
Brown 2015
Nielsen 2014
Albertsen 2013
Tyson 2004
N MAGs
10^5
10^4
10^3
10^2
10^1
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year


Pasolli 2019
Almeida 2019
Nayfach 2019
Parks 2017
Brown 2015
Nielsen 2014
Albertsen 2013
Tyson 2004
N MAGs
10^5
10^4
10^3
10^2
10^1
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year


(Tenericutes)
Bacteria
Actinobacteria
Armatimonadetes
Chloroflexi
Firmicutes
Cyanobacteria
Melainabacteria
RBX1
WOR1
Zixibacteria
Cloacimonetes
Fibrobacteres
Gemmatimonadetes
WOR-3
TA06
Poribacteria
Latescibacteria
BRC1
Atribacteria
Aquificae
Calescamantes
Caldiserica
Dictyoglomi
Thermotogae
Deinococcus-Therm.
Synergistetes
Fusobacteria
Marinimicrobia
Bacteroidetes
Chlorobi
Ignavibacteria
Caldithrix
PVC
superphylum
Planctomycetes
Chlamydiae,
Lentisphaerae,
Verrucomicrobia
Elusimicrobia
Omnitrophica
Aminicentantes
Rokubacteria
NC10
Acidobacteria
Tectomicrobia, Modulibacteria
Nitrospinae
Nitrospirae
Dadabacteria
Deltaprotebacteria
(Thermodesulfobacteria)
Chrysiogenetes
Deferribacteres
Hydrogenedentes NKB19
Spirochaetes
TM6
Epsilonproteobacteria
Alphaproteobacteria
Zetaproteo.
Acidithiobacillia
Betaproteobacteria
Gammaproteobacteria
0.4
Nomurabacteria
Kaiserbacteria
Adlerbacteria
Campbellbacteria
Giovannonibacteria
Wolfebacteria
Jorgensenbacteria
Azambacteria
Parcubacteria
Yanofskybacteria
Moranbacteria
Magasanikbacteria
Uhrbacteria
Falkowbacteria
Candidate
Phyla Radiation
SM2F11
Peregrinibacteria
Gracilibacteria BD1-5, GN02
Absconditabacteria SR1
Saccharibacteria
Berkelbacteria
Wirthbacteria
Woesebacteria
Shapirobacteria
Amesbacteria
Collierbacteria
Pacebacteria
Beckwithbacteria
Roizmanbacteria
Gottesmanbacteria
Levybacteria
Daviesbacteria
Curtissbacteria
Microgenomates
Dojkabacteria WS6
CPR1
CPR3
Katanobacteria
WWE3
Major lineages with isolated representative: italics
Major lineage lacking isolated representative:
Micrarchaeota
Diapherotrites
Nanohaloarchaeota
Aenigmarchaeota
Parvarchaeota
DPANN
Pacearchaeota
Nanoarchaeota
Woesearchaeota
Loki.
Thor.
Korarch.
Crenarch.
Bathyarc.
YNPFFA
Aigarch.
Halobacteria
Altiarchaeales
Z7ME43
Methanopyri
Methanococci
Hadesarchaea
Thermococci
Methanobacteria
Thermoplasmata
Archaeoglobi
Methanomicrobia
Archaea
TACK
Thaumarchaeota
Eukaryotes
Opisthokonta
Excavata
Archaeplastida
Chromalveolata
Amoebozoa

Patescibacteria don't have all universal marker genes!




Pasolli 2019
Almeida 2019
Nayfach 2019
Parks 2017
Brown 2015
Nielsen 2014
Albertsen 2013
Tyson 2004
N MAGs
10^5
10^4
10^3
10^2
10^1
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year

Large-scale re-assembly era

nature microbiology

ARTICLES

DOI: 10.1038/s41564-017-0012-7

OPEN

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz* and Gene W. Tyson*

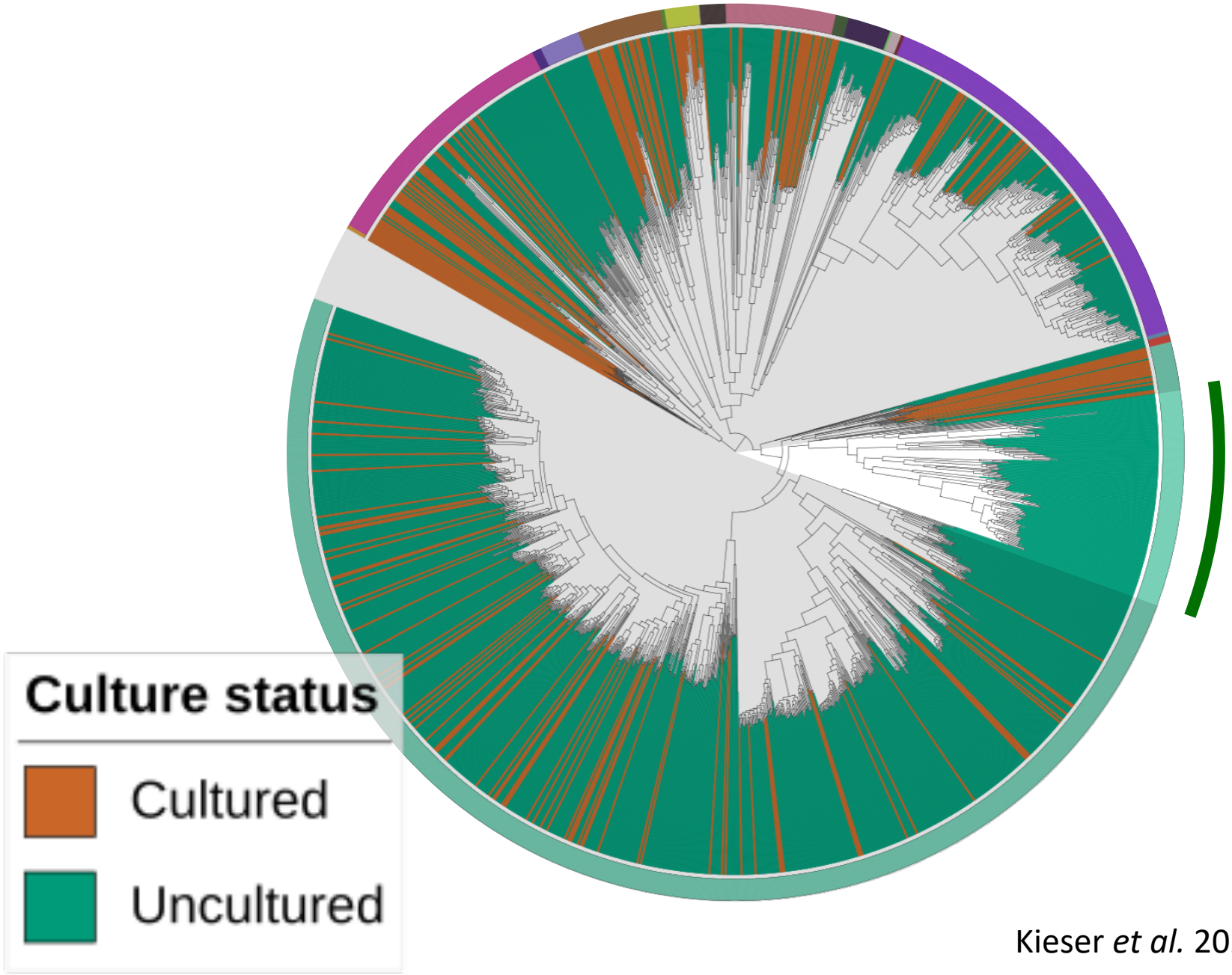
nature biotechnology

RESOURCE

https://doi.org/10.1038/s41587-020-0603-3

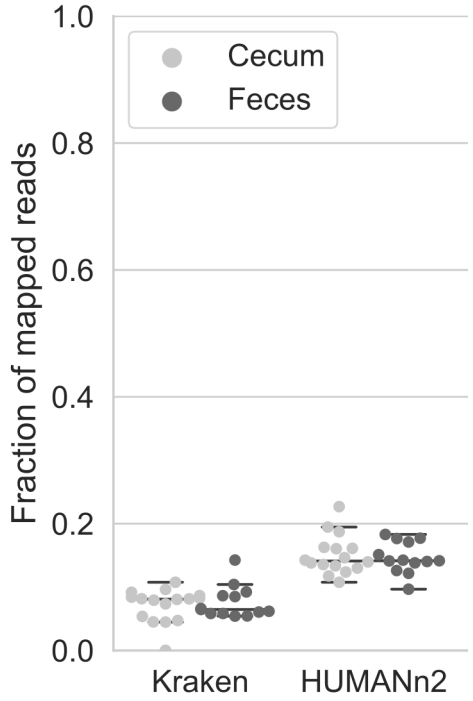
OPEN

A unified catalog of 204,938 reference genomes from the human gut microbiome



Kieser *et al.* 2022

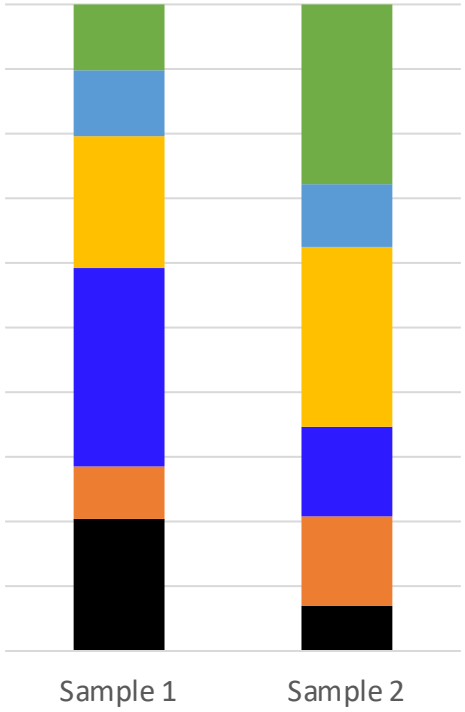
Improved mapping rate



Post-assembly era

Compositional data analysis

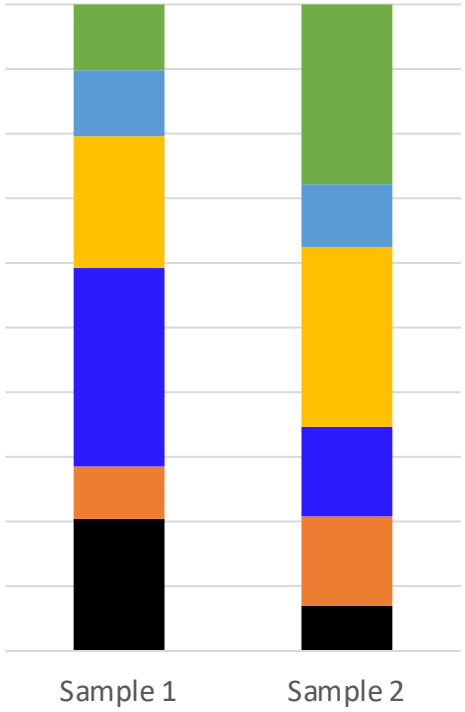
Relative abundance



What to do with the unmapped reads?

Interpret microbial abundances as ratios

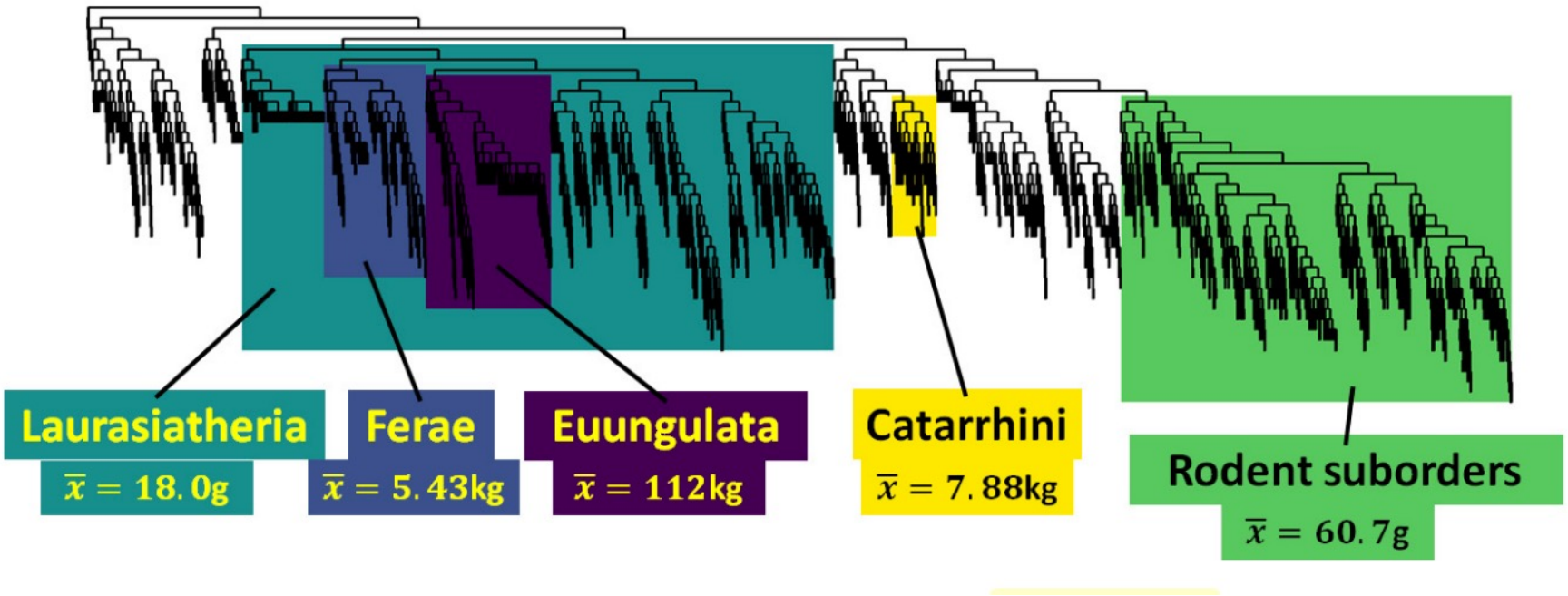
Ratios



Calculate ratios

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

Phylofactor



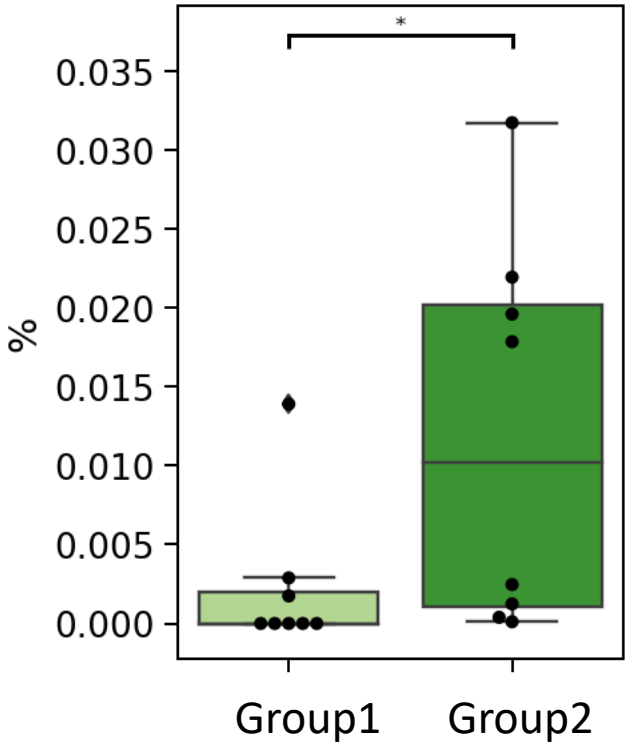
Centered log ratios

Impute zeros

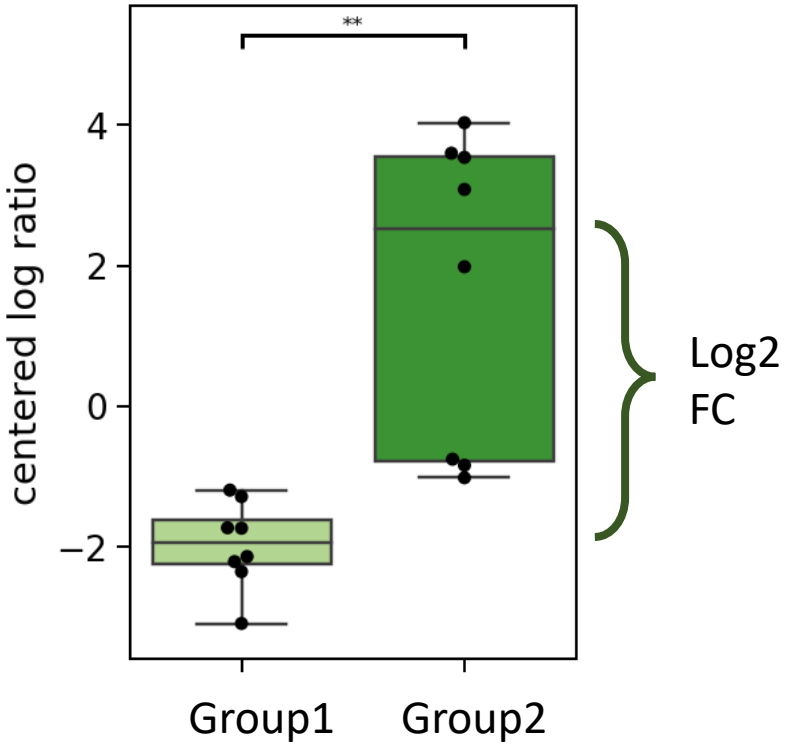
1. Take log
2. Subtract sample-mean

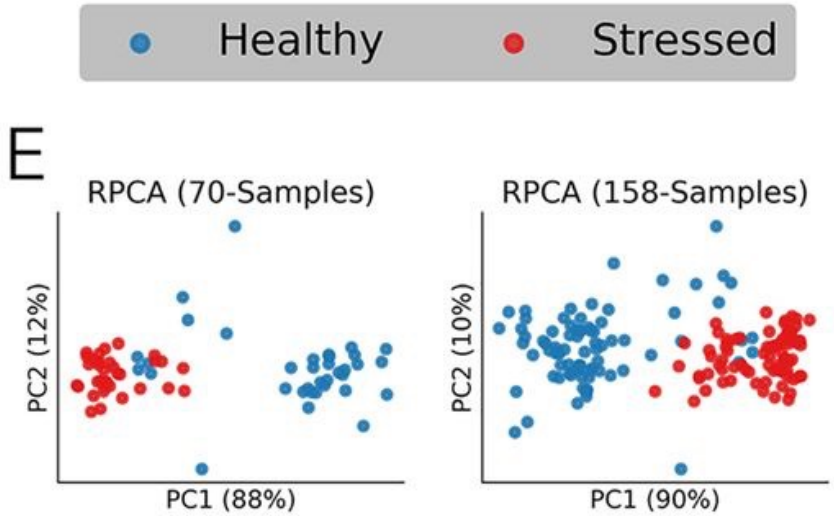
Centered log ratios

Relative abundance



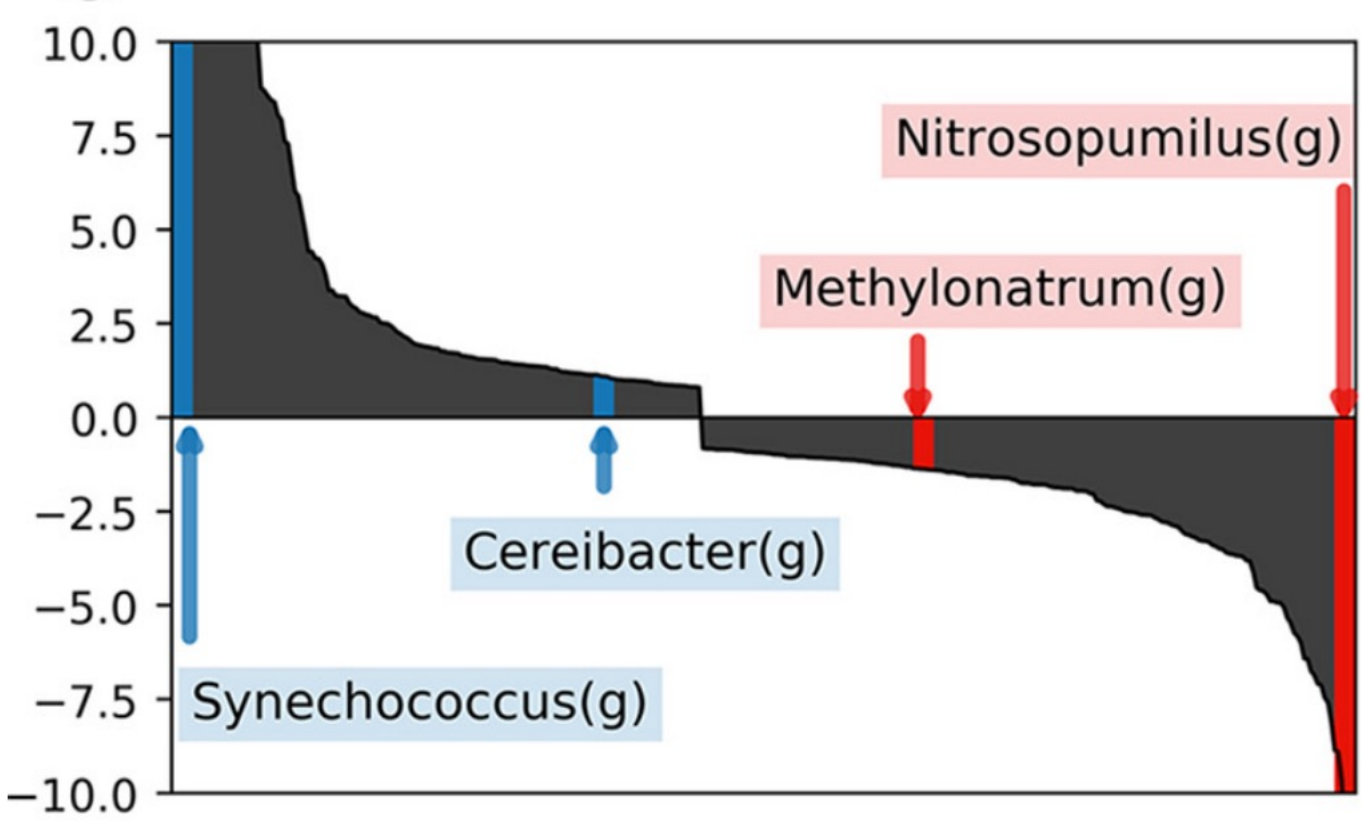
Centered log ratios





RPCA = PCA based on CLR

Martino *et al.* 2019



Martino *et al.* 2019

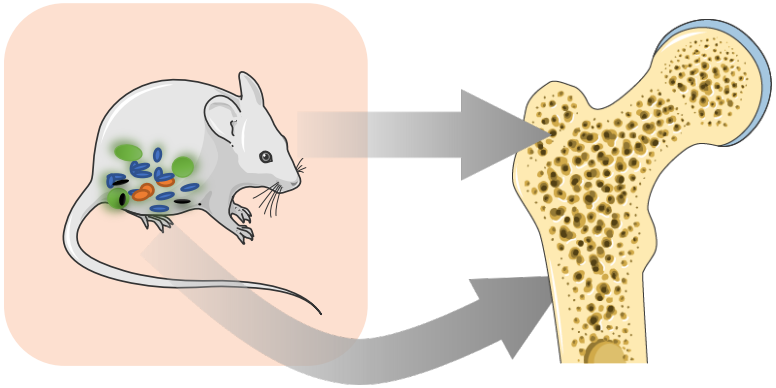
silask.github.io
Chapter 5 of my thesis


Thank you for your attention.
Questions?

Article

Cell Metabolism

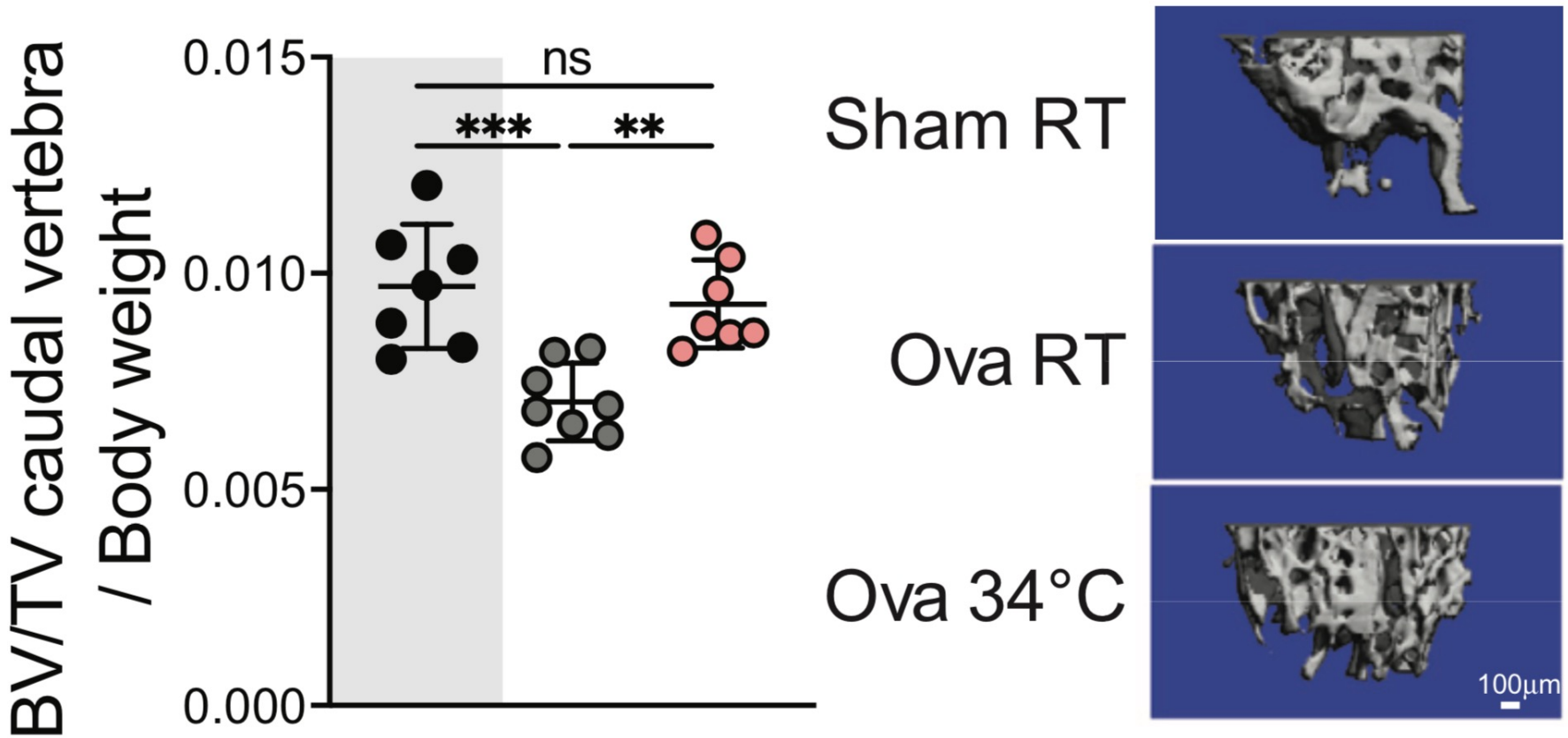
Warmth Prevents Bone Loss Through the Gut Microbiota



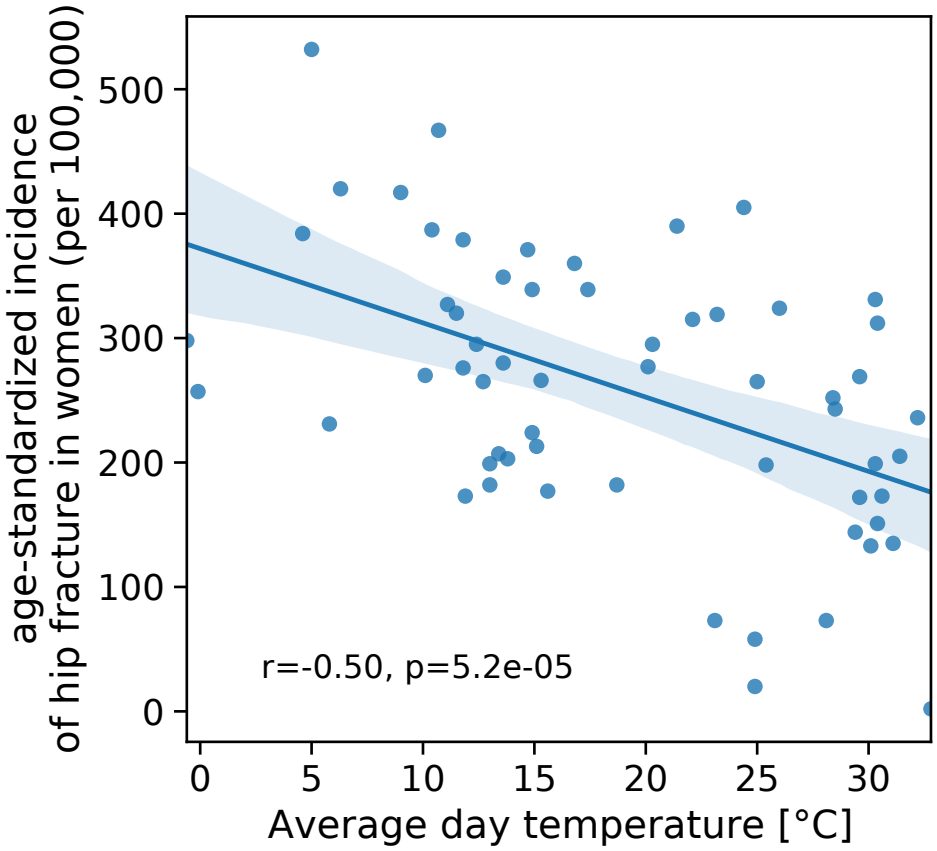
Authors

Claire Chevalier, Silas Kieser,
Melis Çolakoğlu, ...,
Andrew Macpherson, Nicolas Bonnet,
Mirko Trajkovski

L

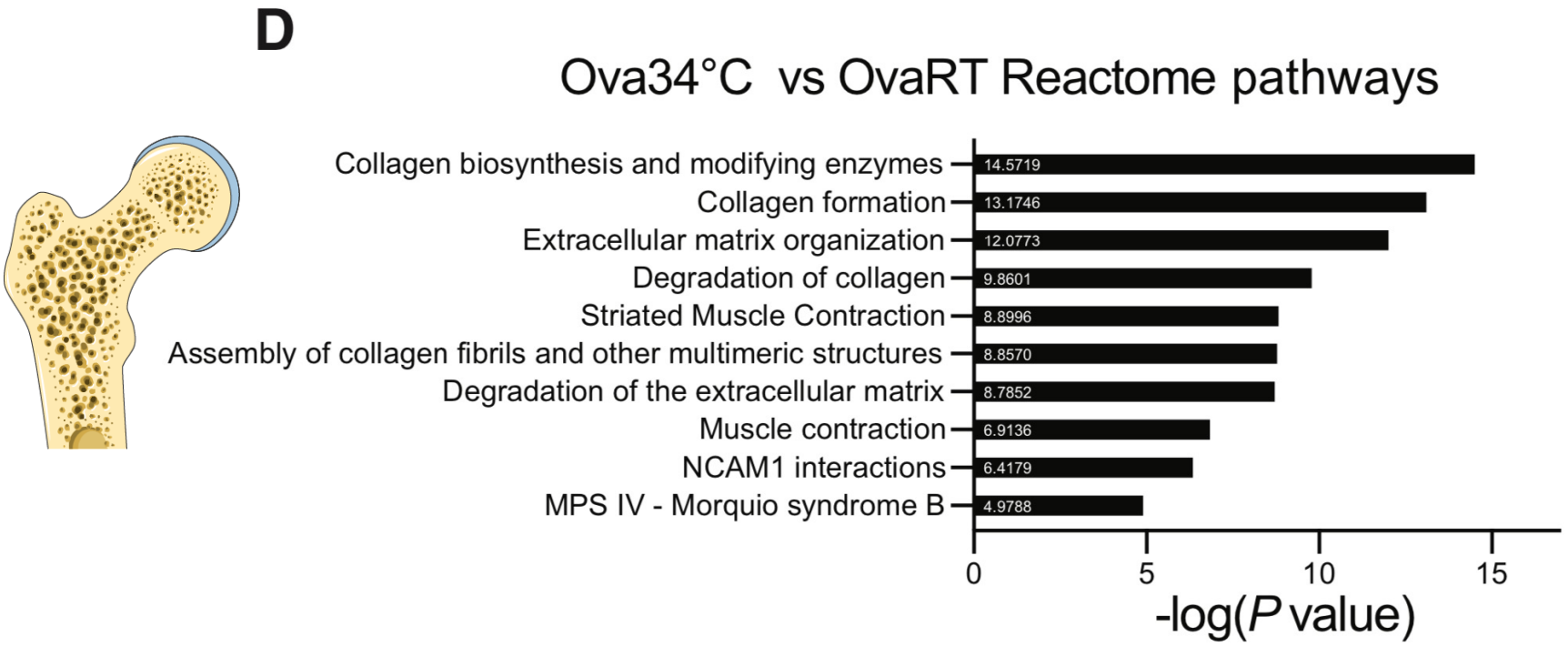

BV/TV caudal vertebra
/ Body weight
0.015
0.010
0.005
0.000
ns
***
**
Sham RT
Ova RT
Ova 34°C
100μm

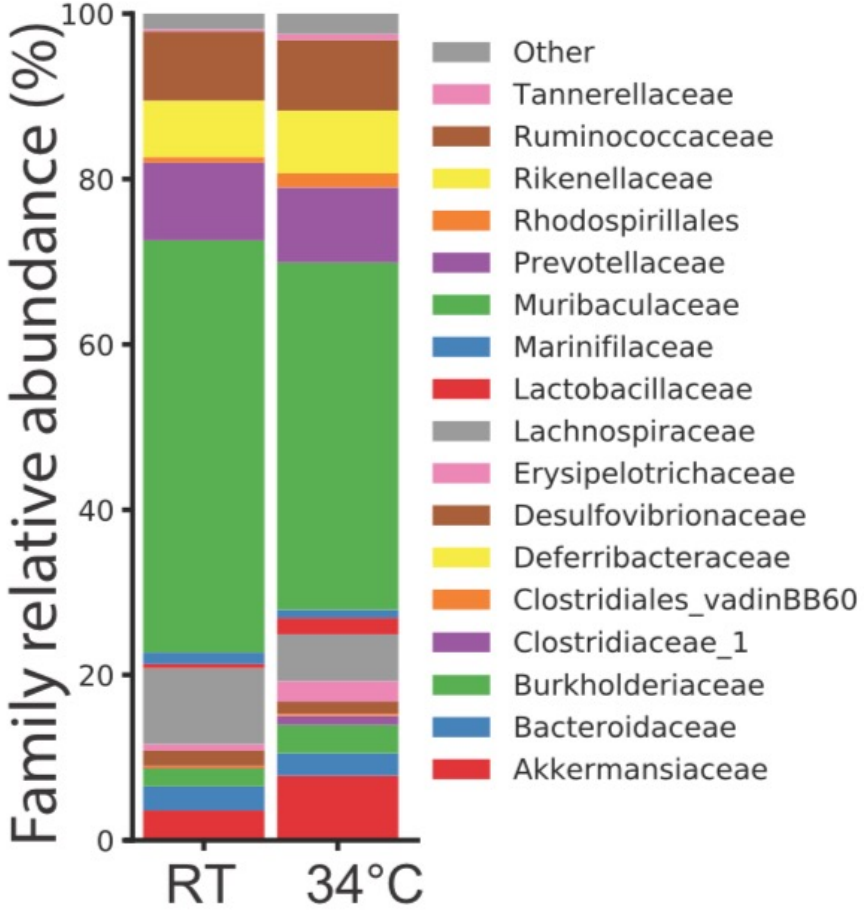
Osteoporosis incidence correlates with lower temperature



How does it work?

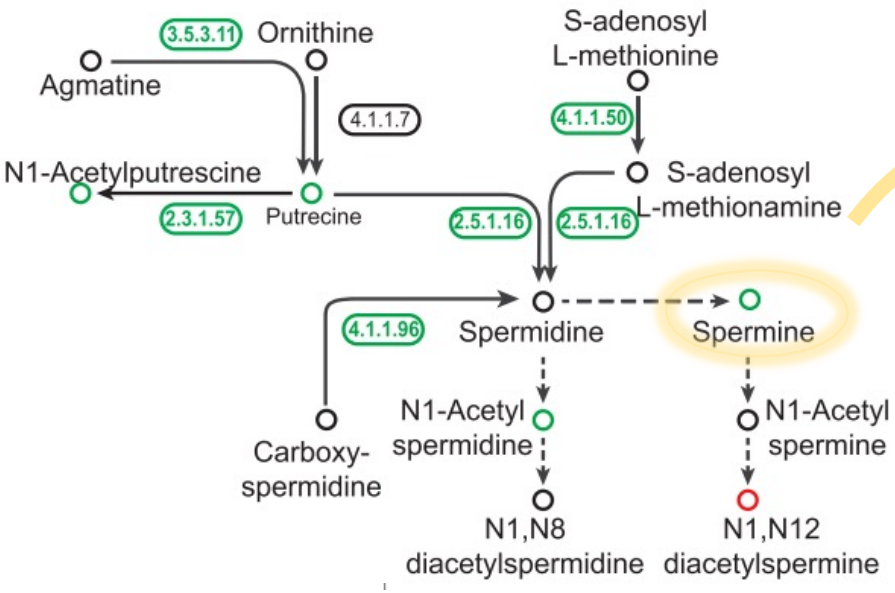
Gene expression indicates significant bone remodeling upon Warm exposure

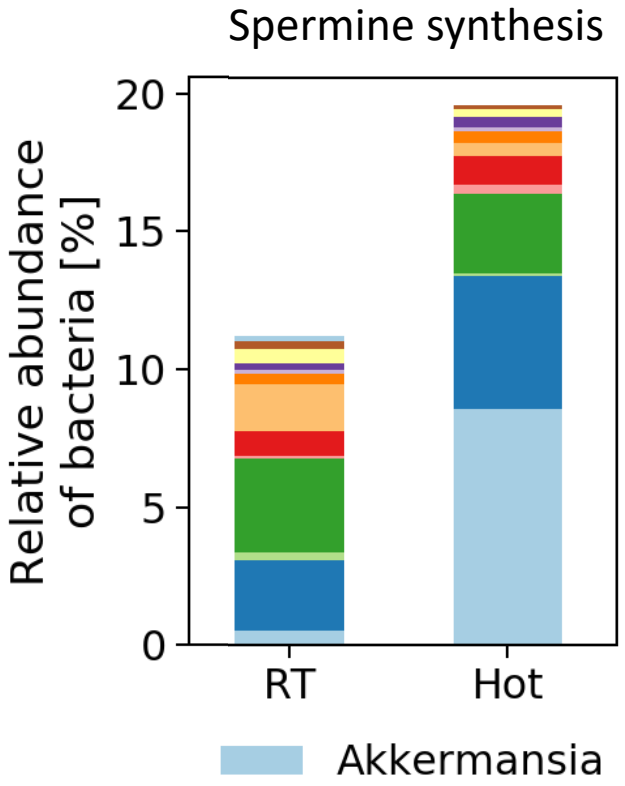



Family relative abundance (%)
100
80
60
40
20
0
RT
34°C
Other
Tannerellaceae
Ruminococcaceae
Rikenellaceae
Rhodospirillales
Prevotellaceae
Muribaculaceae
Marinifilaceae
Lactobacillaceae
Lachnospiraceae
Erysipelotrichaceae
Desulfovibrionaceae
Deferribacteraceae
Clostridiales_vadinBB60
Clostridiaceae_1
Burkholderiaceae
Bacteroidaceae
Akkermansiaceae

Cell Metabolism

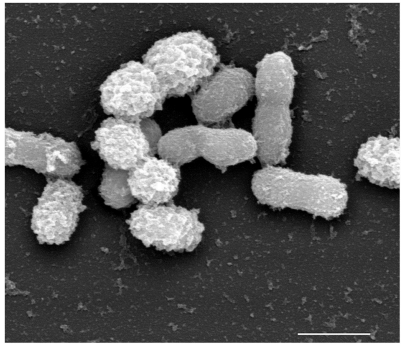
Warmth Prevents Bone Loss Through the Gut Microbiota





Pathway ⟷ Species

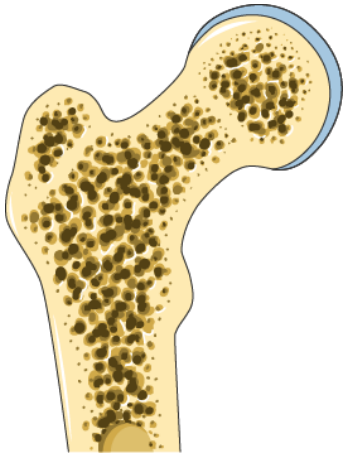
How does it work?



Akkermansia

Spermidine

Osteoclasts



Images: Derrien et al. 2016, Servier Art

Run atlas

Install metagenome-atlas

```
conda install metagenome-atlas
```

Install metagenome-atlas

```
mamba install metagenome-atlas
```

Start a project

```
mamba install metagenome-atlas
atlas init path/to/fastq
```

Samples.tsv

	Reads_R1	Reads_R2
sample1		
sample2		

Config.yaml

```yaml
#######################
# Binning
#######################

final_binner: DASTool             # [SemiBin, DASTool, vamb, maxbin]

binner:                           # If DASTool is used as final_binner,
  - metabat
  - maxbin
#  - vamb


metabat:
  sensitivity: sensitive
  min_contig_length: 1500 # metabat needs >1500

maxbin:
  max_iteration: 50
  prob_threshold: 0.9
  min_contig_length: 1000
```

Run atlas

```
atlas run genecatalog
```

```
atlas run genomes
```

Run atlas

```
atlas run genecatalog
```

```
atlas run genomes strains
```