DTU Health Technology

Bioinformatics

Metagenomics & Binning

Trine Zachariasen, PhD

Post doc at Copenhagen University

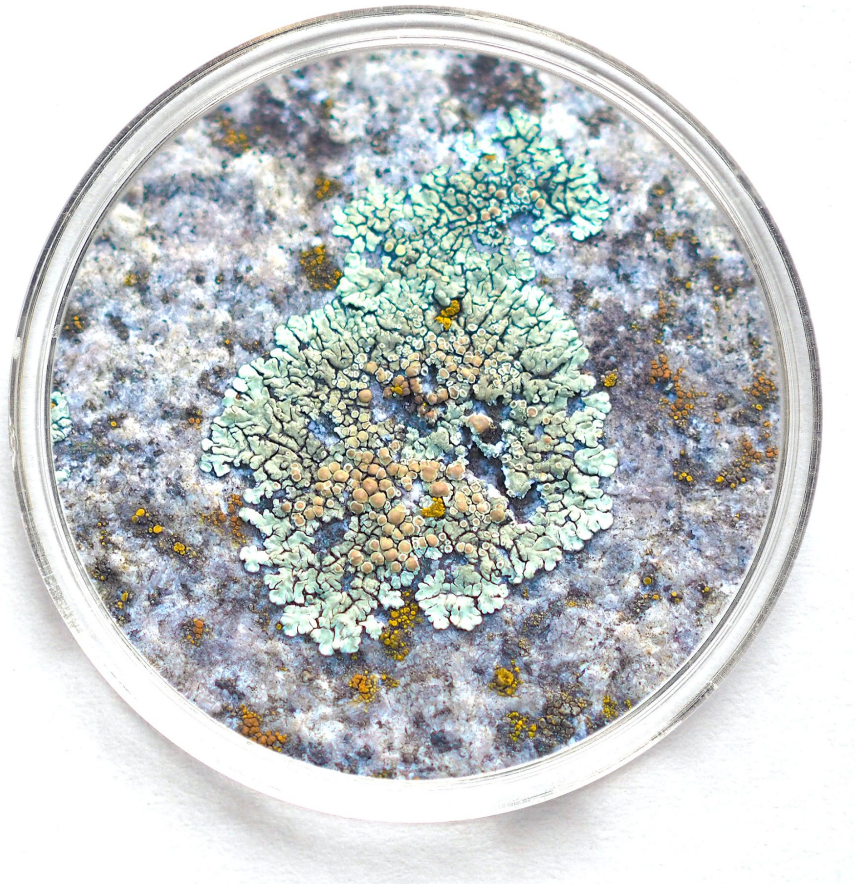
Menu

<u>Introduction to metagenomics</u>

- What is metagenomics
- Methods in metagenomics
- Challenges in metagenomics
- Uses of metagenomics

<u>Metagenomics binning</u>

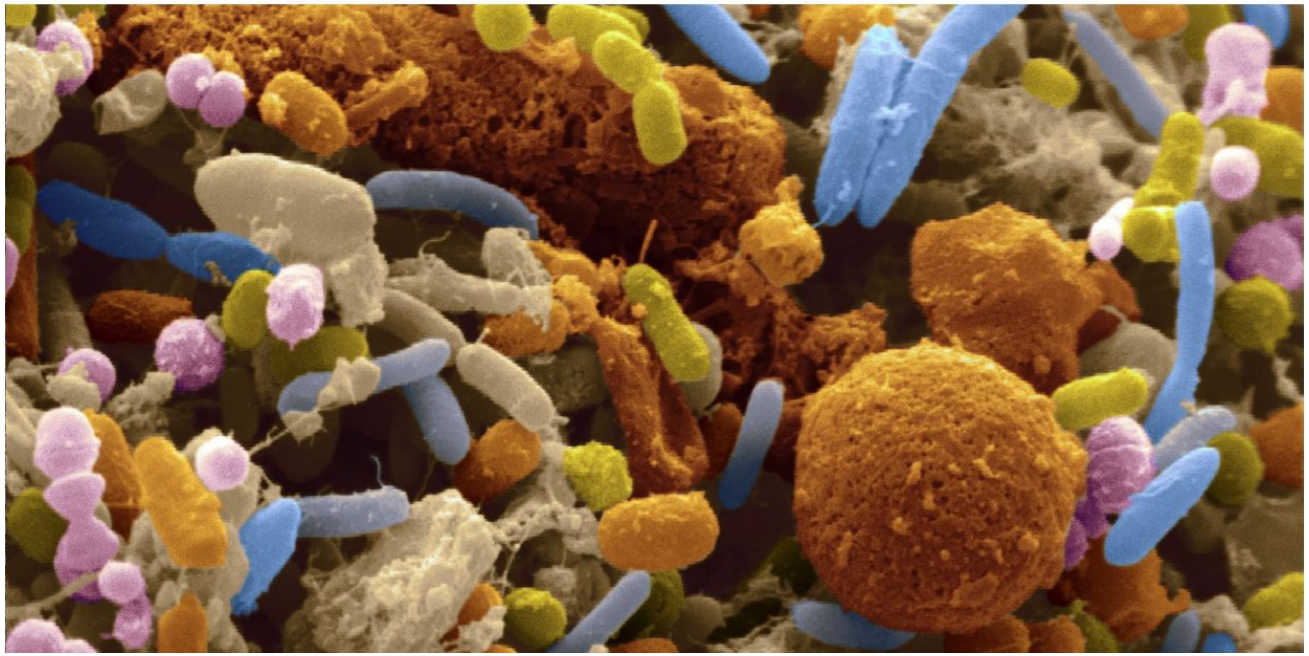
- What is binning?
- Types of metagenomic binners
- Assesing bin quality
- Binned genes for profiling



What is metagenomics?

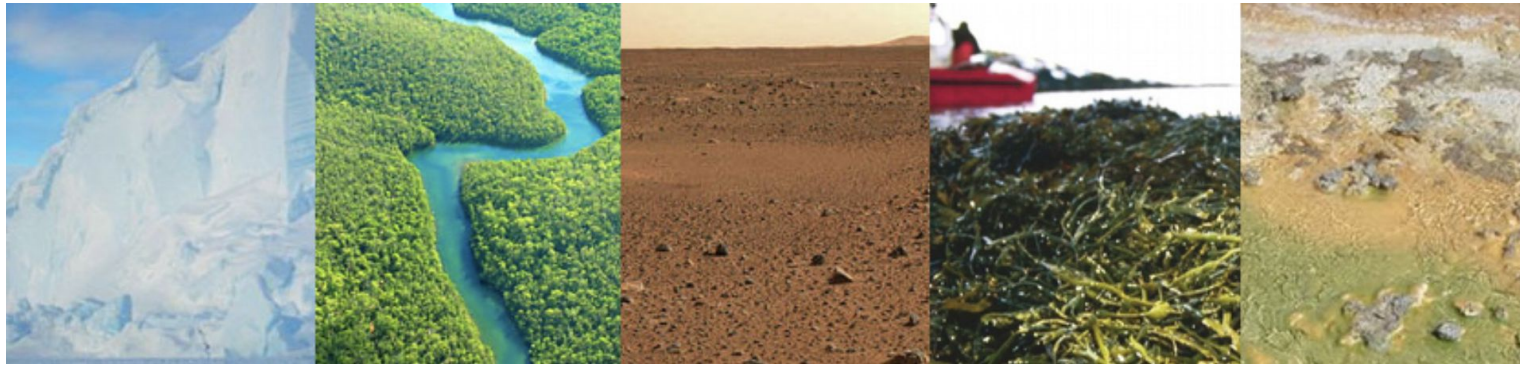
Metagenomics is the study of genetic material recovered directly from the environment

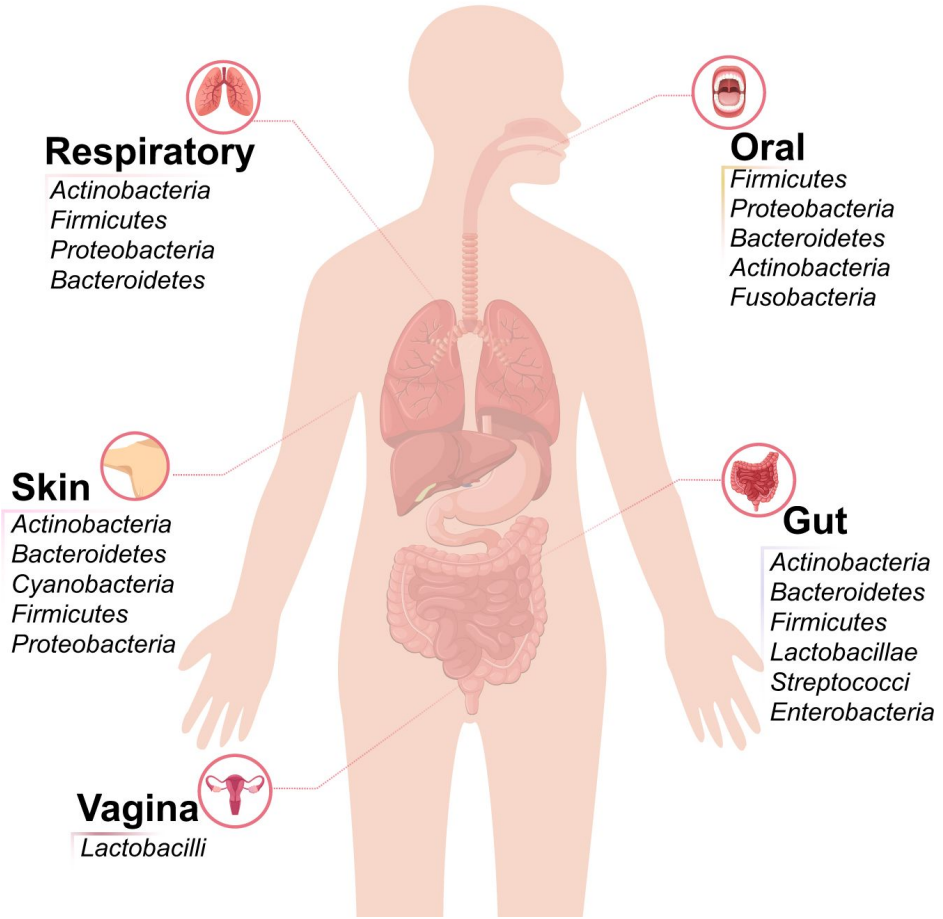
- No need for culturing
- Collection of organisms



Microbes exist everywhere and metagenomics allow us to study them

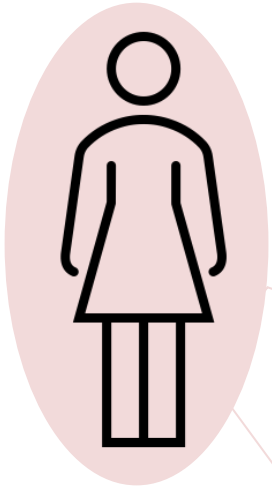
What is the difference between a microbiome and a metagenome?





Hou, K. *et al.* Microbiota in health and diseases. *Sig Transduct Target Ther* **7**, 135 (2022).

More bacteria in you than the human population

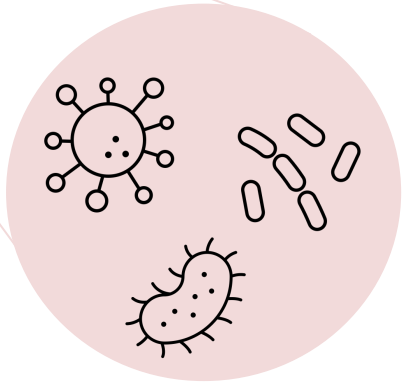


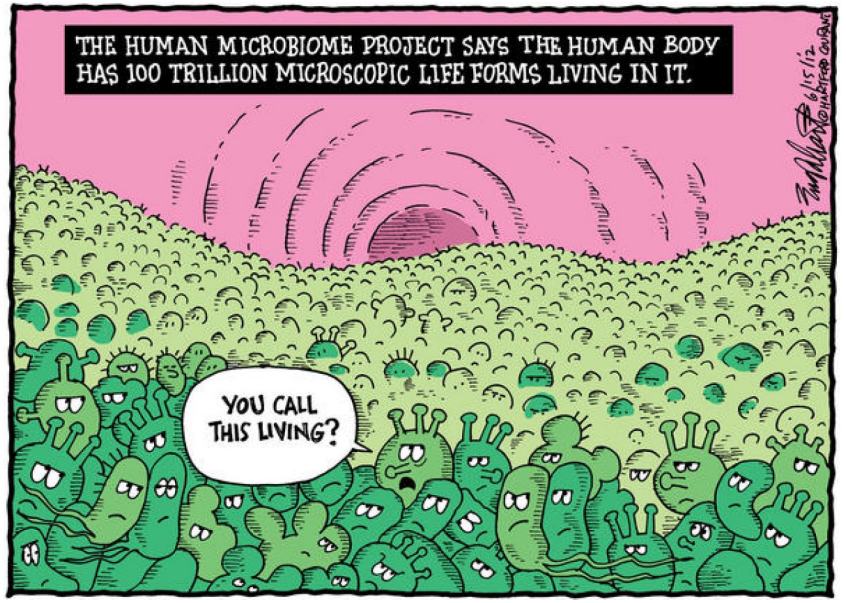
Human genome

30 trillion cells[1]

22,000 genes[2]

99.9% similar between individuals[2]





Human microbiome

10-100 trillion microbes[1]

3.3 million genes[2]

As little as 10-20% similar between body sites[2]

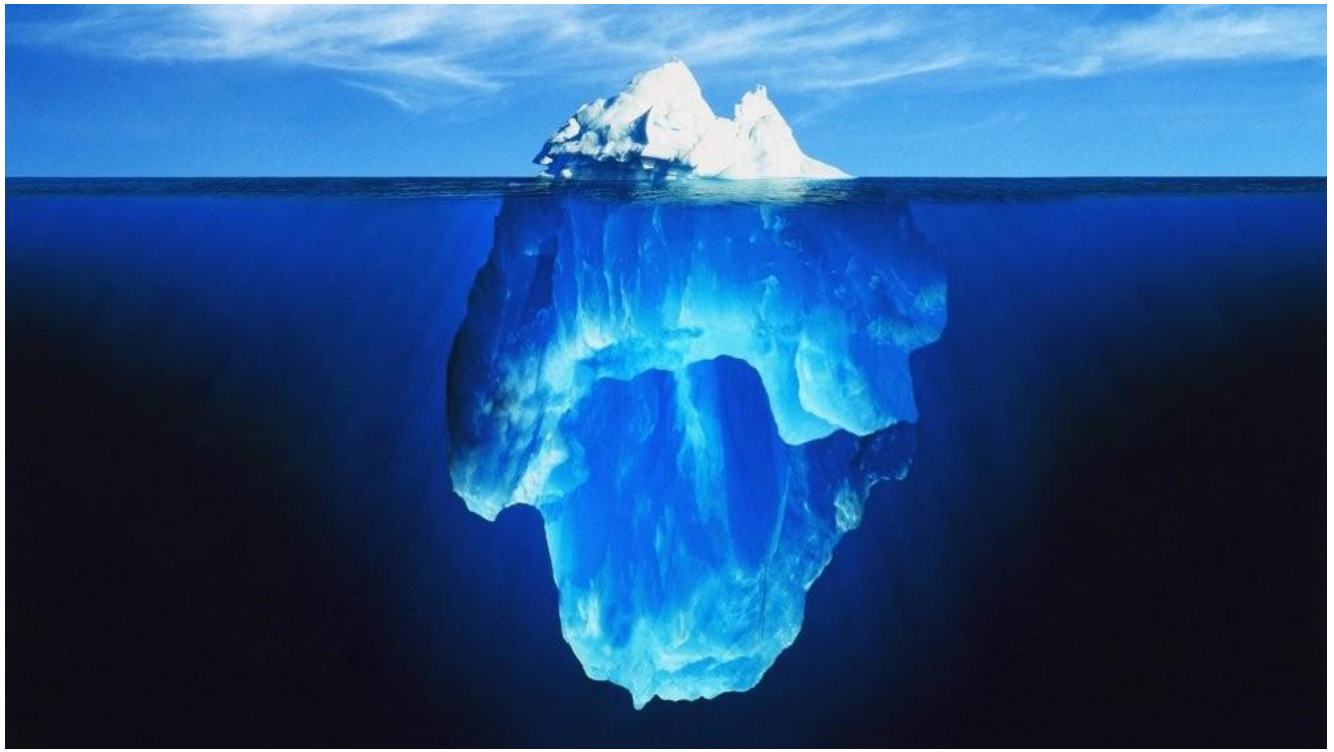
1. L. K. Ursell et al., Nutrition Reviews, 2012
2. P. J. Turnbaugh et al., Nature, 2007

Microbiome research in pre-sequencing days

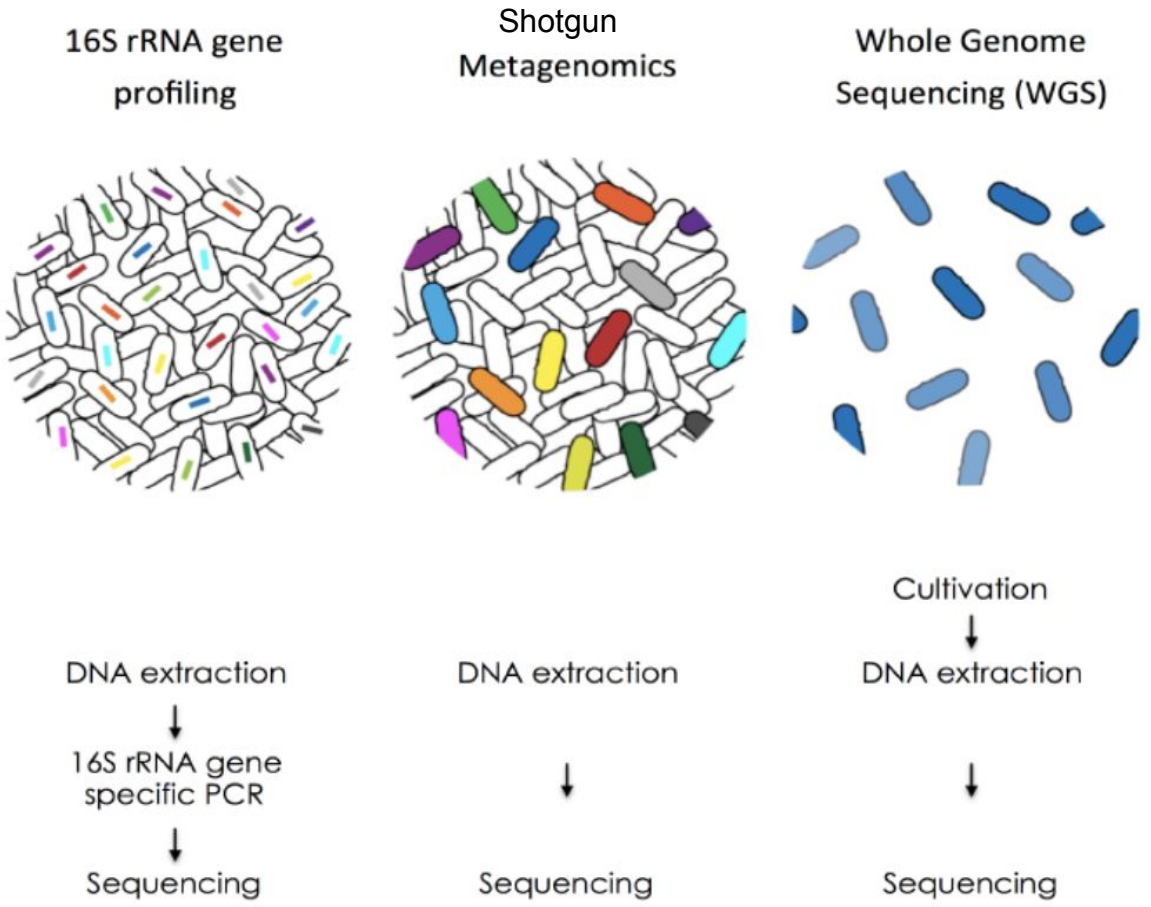
- Culturable organism chosen as models
- Might not be representative even for close relatives

From culturing to sequencing

- Microbiome research previously limited to culturable organisms
- ~95% of bacteria in the environment cannot be cultured



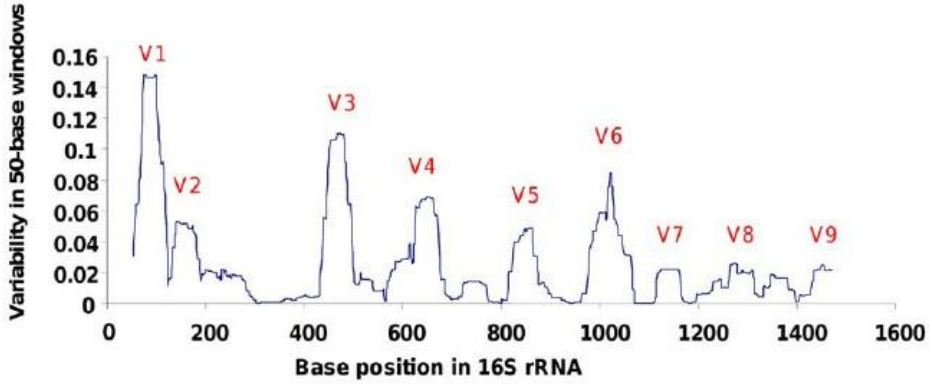
Methods

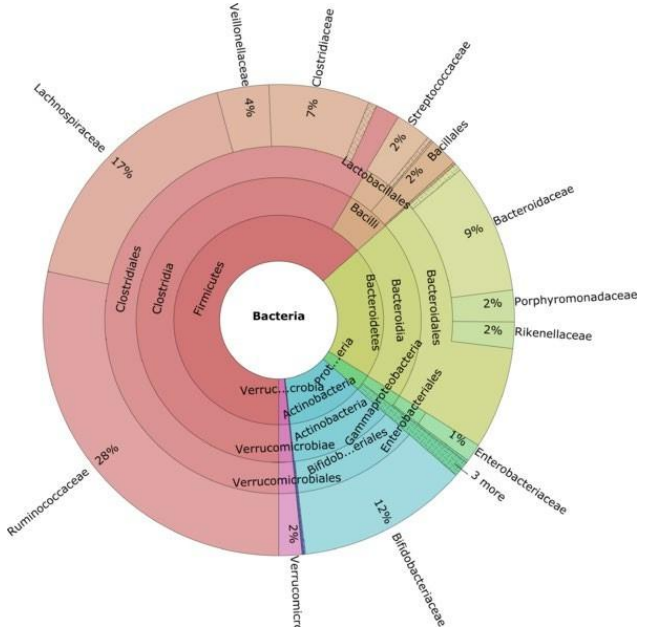

16S rRNA gene profiling
Shotgun Metagenomics
Whole Genome Sequencing (WGS)
DNA extraction
16S rRNA gene specific PCR
Sequencing
DNA extraction
Sequencing
Cultivation
DNA extraction
Sequencing

16S rRNA – who is there?

- Ribosomal RNA gene conserved in Bacteria and Archaea
- Conserved areas interspaced with hypervariable regions
- Allows the determination of the phylogenetic composition

- Strains often share same 16s sequence – impossible to distinguish with this method

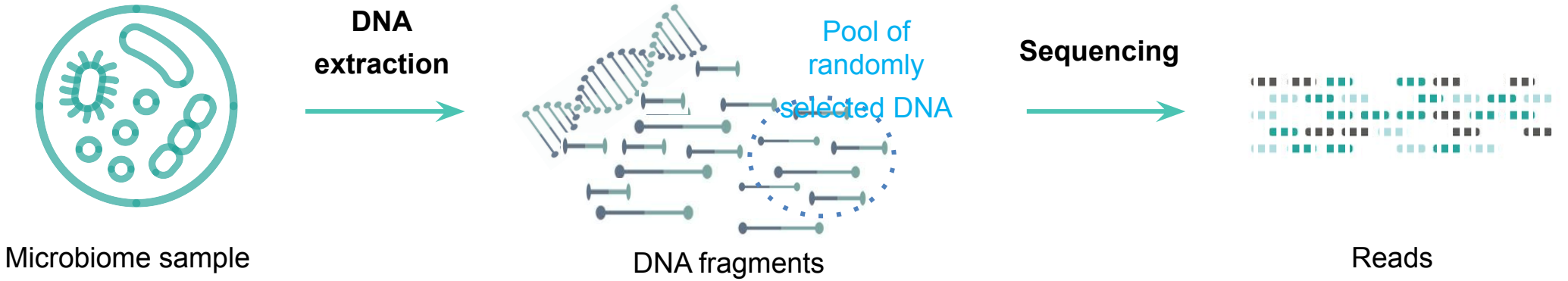


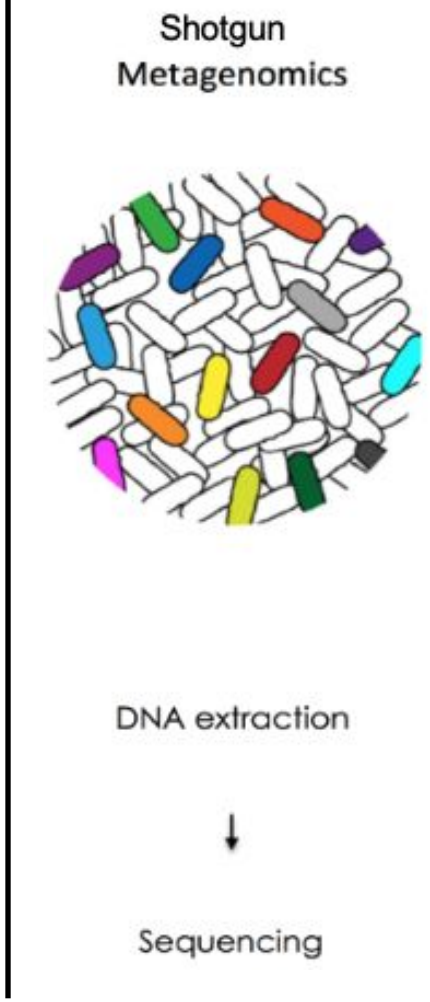


DNA extraction
↓
16S rRNA gene specific PCR
↓
Sequencing

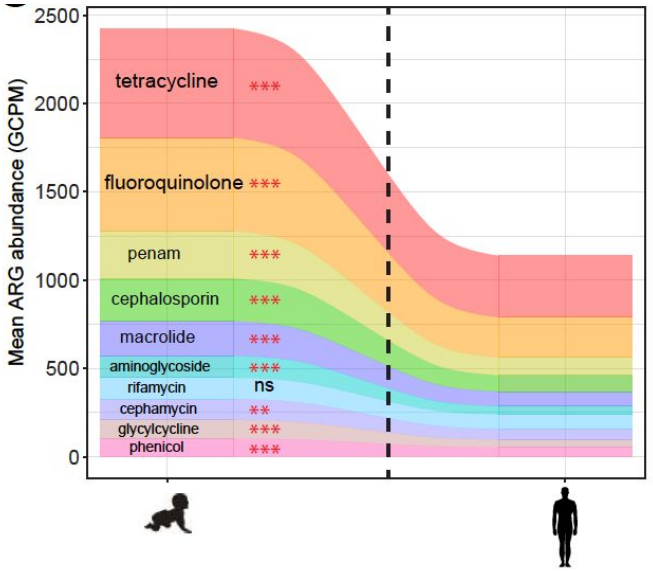
Shotgun metagenomics – what do they do?

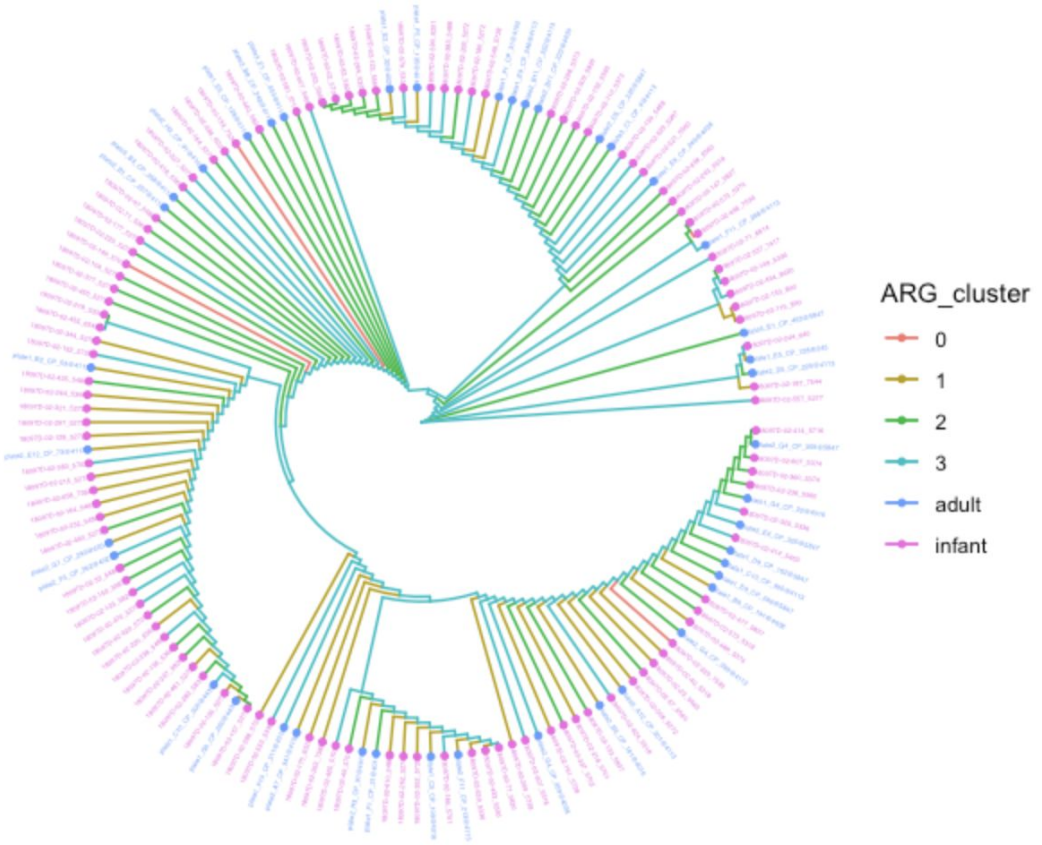
- Capture all the diversity IF sequencing depth is high enough
- Requires enough biomass for DNA extraction
- Functional analysis possible
- Metagenome diversity analysis is possible

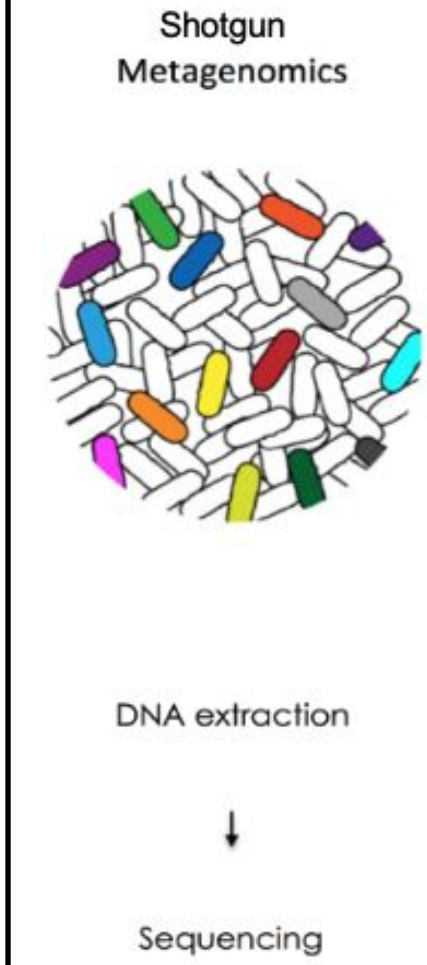




Shotgun metagenomics – what do they do?

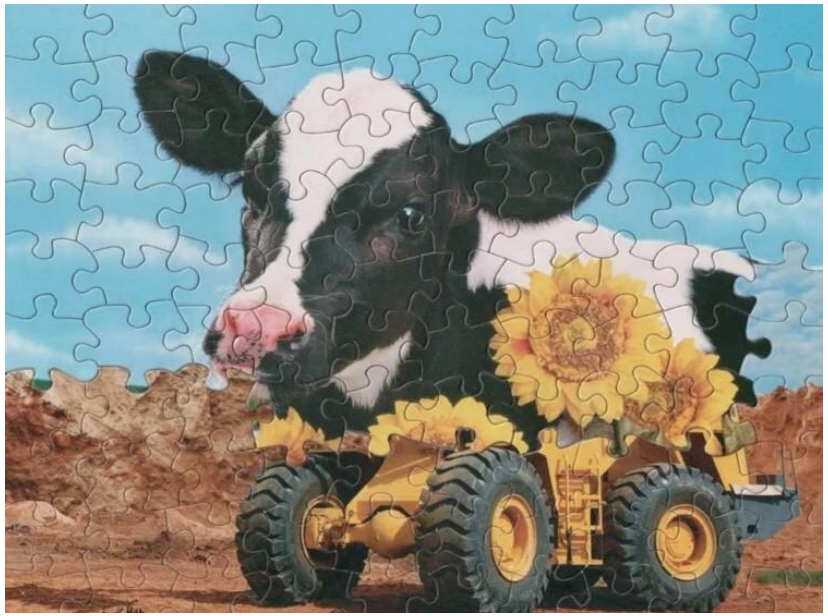


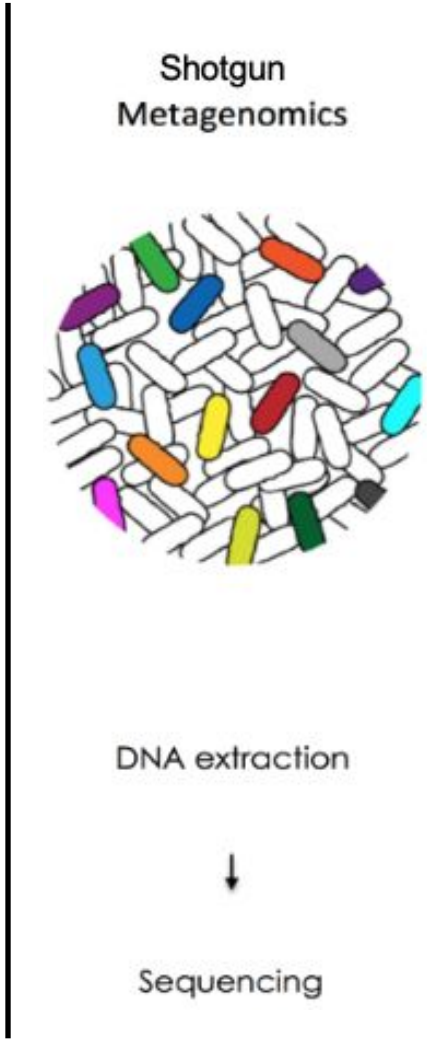




Shotgun metagenomics – the downside

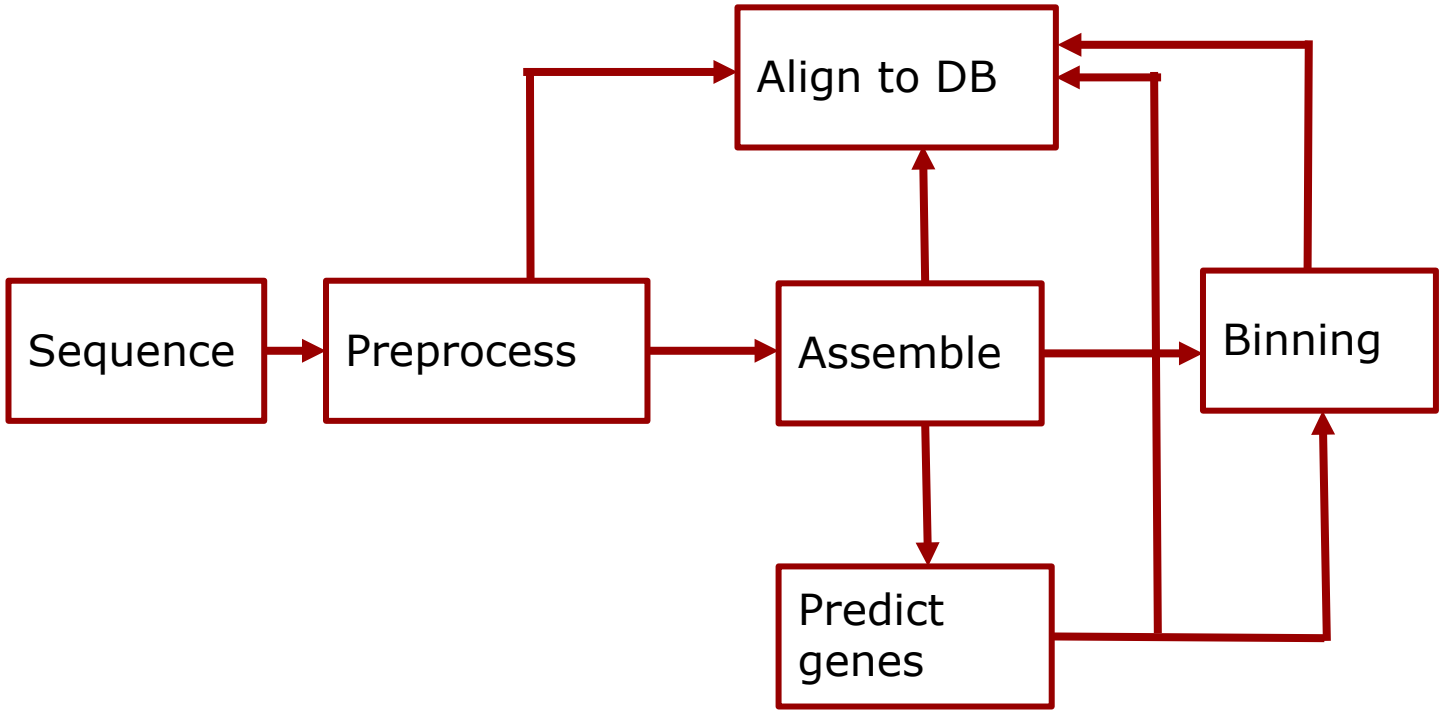
- Enormous datasets
- Varying abundance, detection problems due to low depth or bias
- Expensive to sequence, analyse and store
- Lack of references to genes and organisms
- Shared and/or similar regions hinders assembly





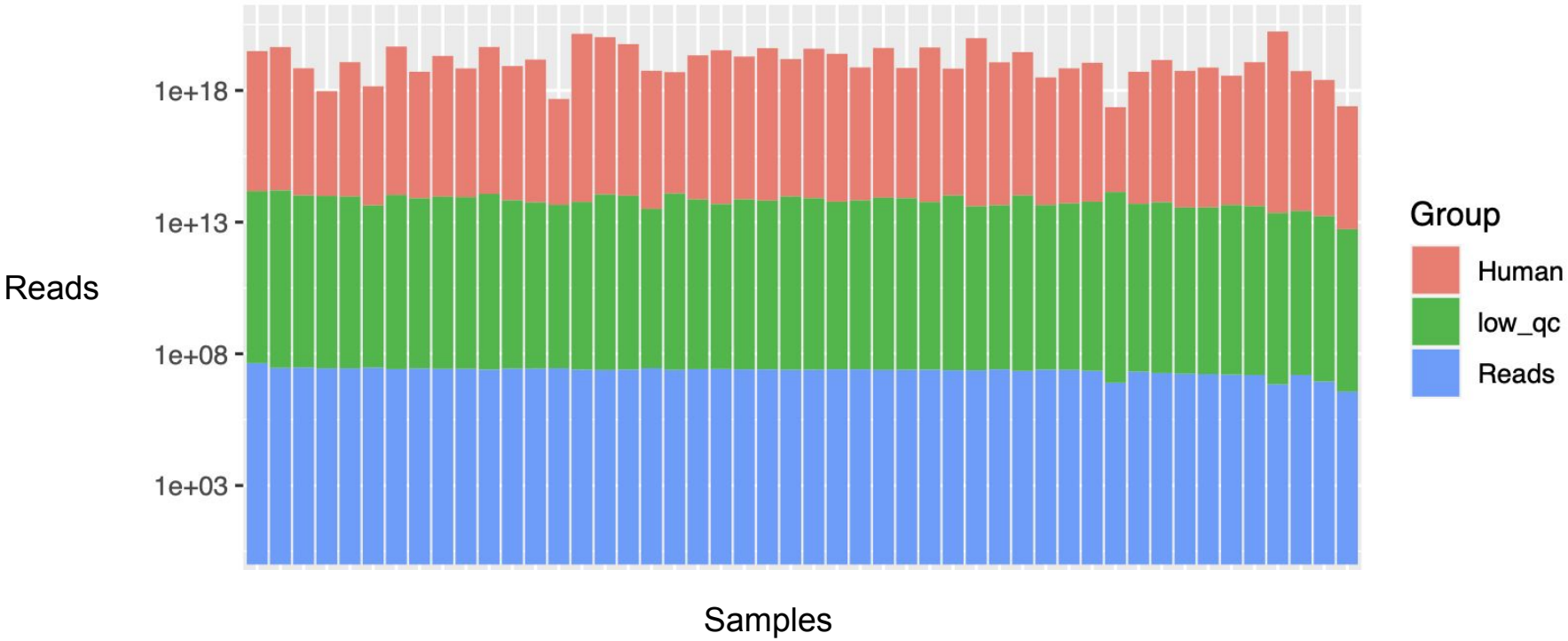
So much data – so many options

So what do we do with metagenomic data?



- We **always** preprocess reads
- We *may* assemble the reads
- All seqs (reads, contigs, genes) can be aligned to database.
- With assembled contigs, we can predict genes and investigate their activity, *in silico* or *in vivo*
- We can bin contigs or genes

Preprocessing



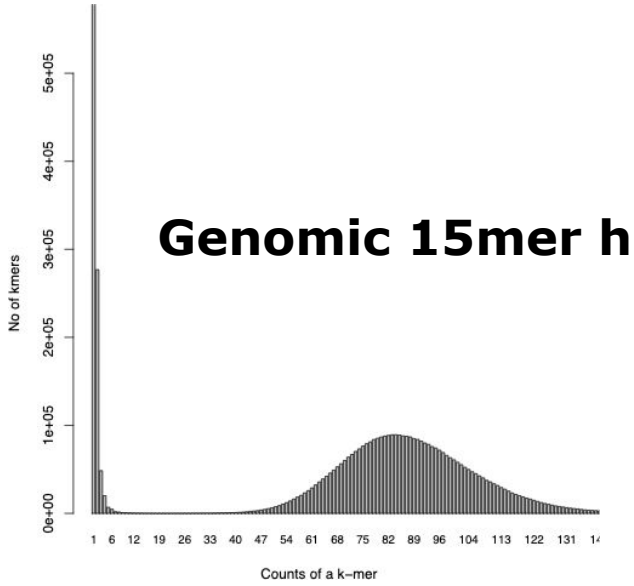
Preprocessing problems

Preprocess

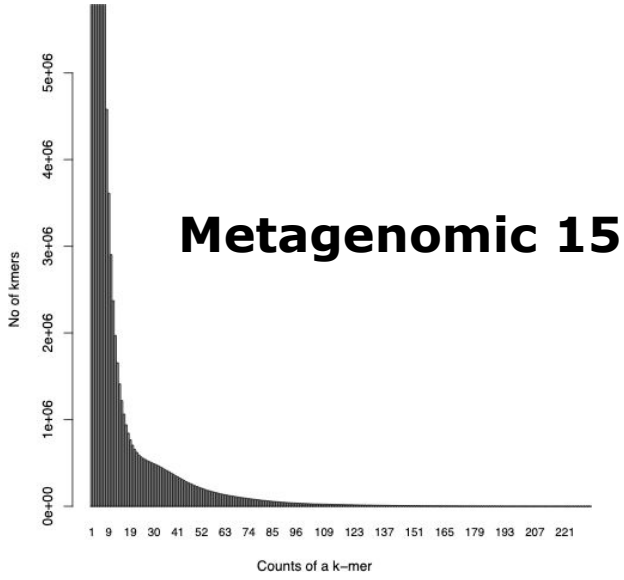
We can still trim reads for quality… however:

Kmer correction is close to impossible.

Genomic 15mer histogram



Metagenomic 15mer histogram



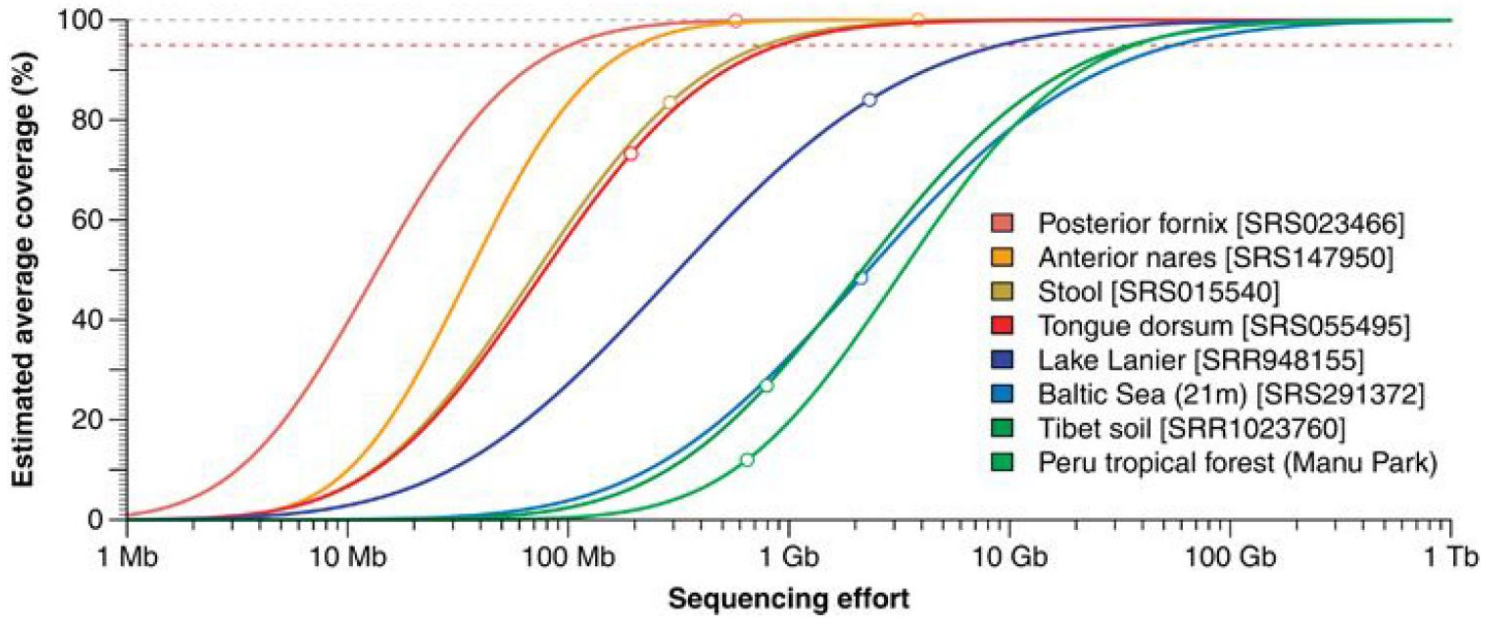
Sequence needed to describe a microbiome

Depth

- No reference database like 16s, therefore we cannot use rarefaction
- No K-mer count like with single genomes

Nonpareil: How often do I find the same read in a dataset?

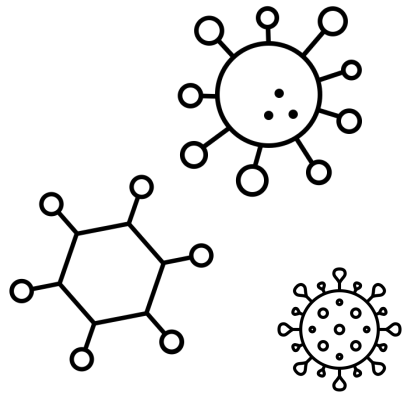
Use redundancy of reads to estimate average coverage and predict the amount of reads needed to achieve "nearly complete coverage"



Alignment problems

Align to DB

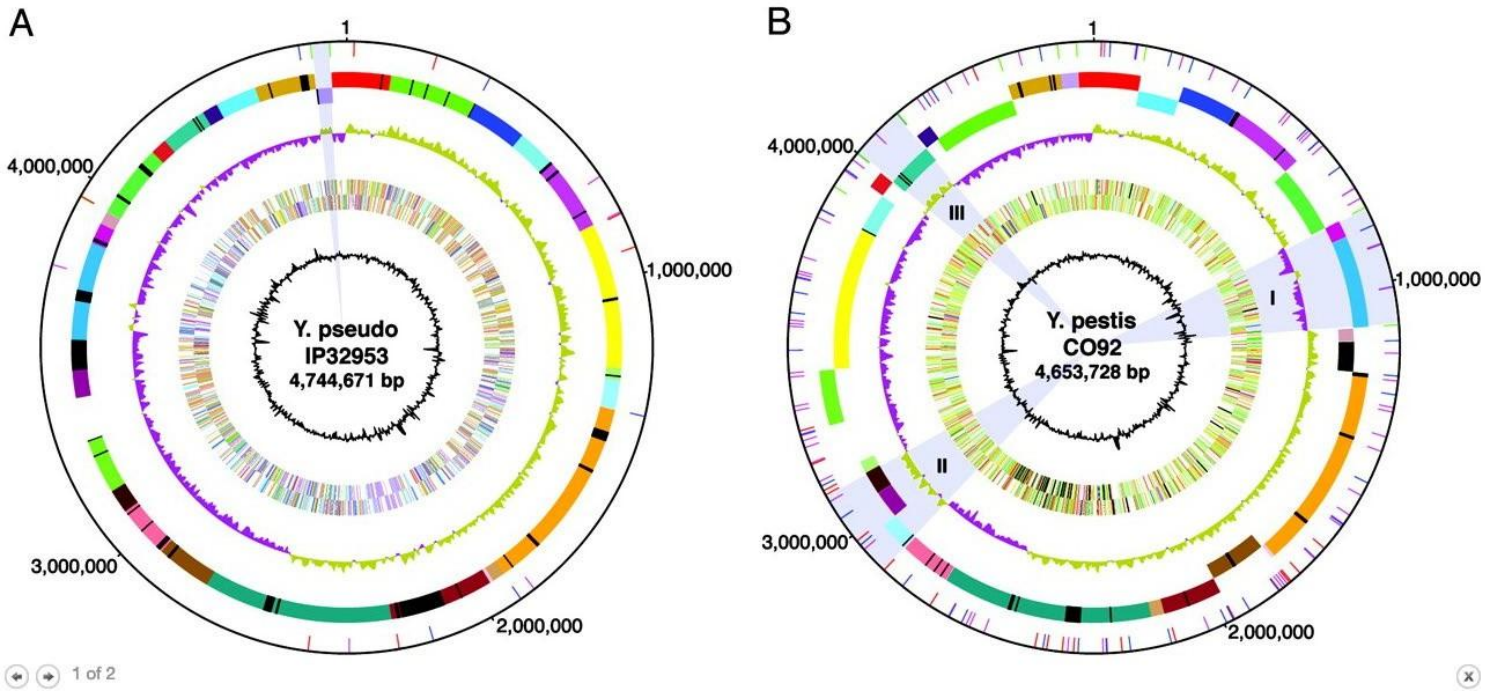
- Alignments are slow enough already. If we want to align against a database of ALL known species, this will take ages (and take up HUGE amounts of RAM + disk space)
- What if your species is not in your database? Then you will not find it. Most aren't.
- Suppose your species IS in your database, and you can align to it with high nucleotide identity (say 96%). Can you think of a situation where this does not tell you all you need to know about the bacteria?



Alignment problems

Align to DB

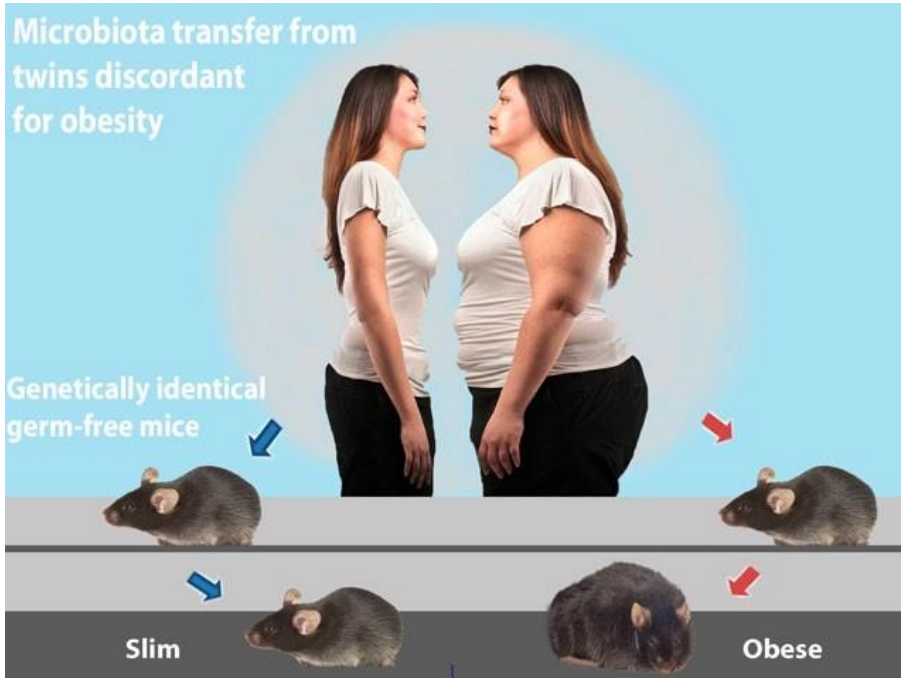
Yersinia pseudotuberculosis vs *Y. pestis.*
Have more than 97% nuc. identity between them! Yet quite different niches these bacteria occupy.
If one didn't cause plague, we would consider these the same species.
Reads from one bacteria aligns to the other bacteria, no problem!



Uses for metagenomics

- Find new enzymes for industrial use
- Find new antibiotics
- Many examples of transitions from fundamental research to application such as CRISPR-Cas systems
- Examine spread of COVID19

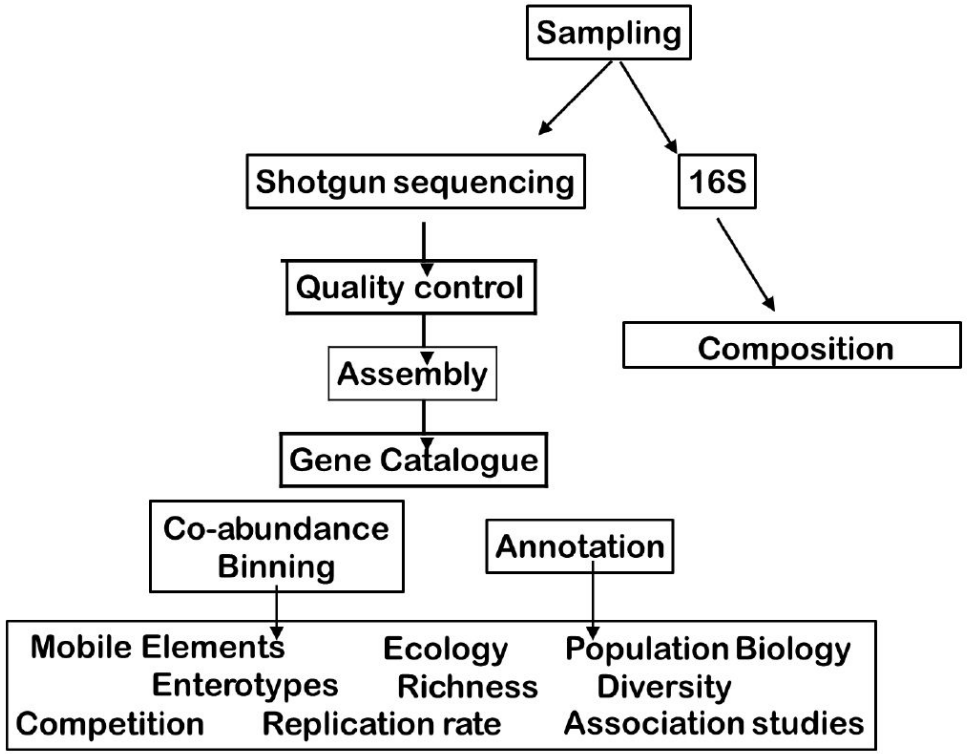
- Identify microorganisms affecting human health
  - Obesity
  - Asthma
  - Allergies



Gut microbiome from a fat and a skinny human transplanted into genetically identical mice (Turnbaugh 2006 Nature)

Horses for courses

- Question dictates the appropriate tool
- 16s rRNA sequencing is easy & cheap
- Shotgun metagenomics allows more questions
  – Discovery of novel proteins, antibiotics etc.
- Same microbiome composition does not mean same metagenome!

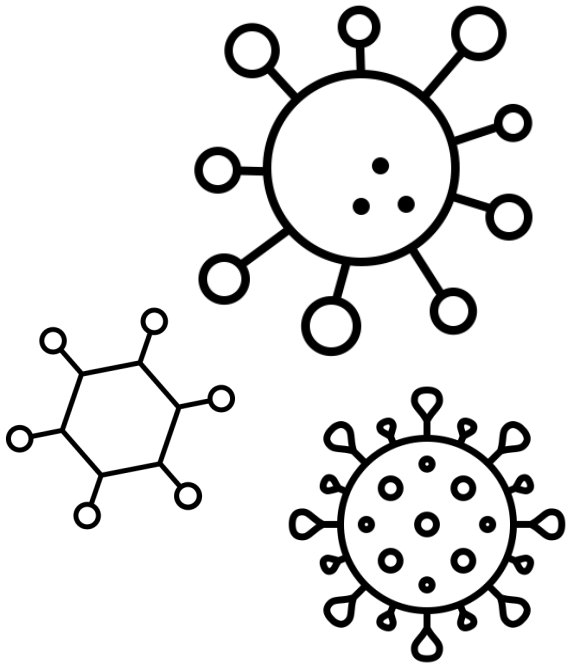


Next up…

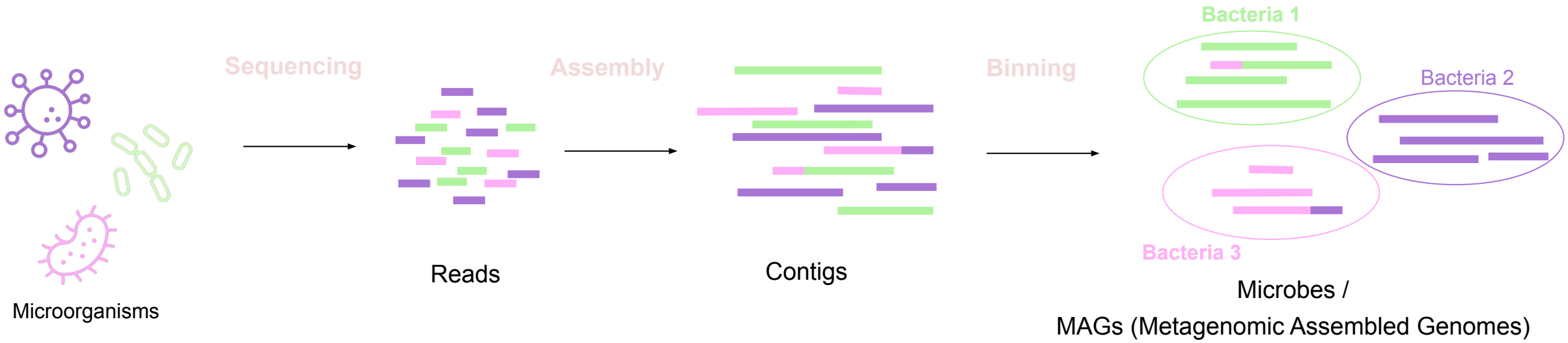
Metagenomic binning

Menu

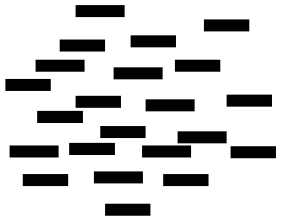
- What is binning?
- Types of metagenomic binners
- Assessing bin quality
- Binned genes for profiling



What is binning?



Remember - all sequences look the same to us!



Why do we care?

If we didn’t bin, all pieces of DNA would have *no context*.

Sometimes, this is okay:

- We look for a specific bacteria, where we already have the reference
- When scanning for interesting genes with a certain signature

Sometimes, we really want context:

- We find resistance genes. *Which bacteria* are resistant?
- This gene looks interesting. *Which operon* is it (likely) part of?
- If we want to find new bacterial species
- We find virulence genes. How worried should we be?
- Why is this bacteria sometimes a bad guy?

Microbiome Composition – why do we care?

Microbiome composition – identification and stratification of the microbes, (which is crucial for understanding influence on human health)

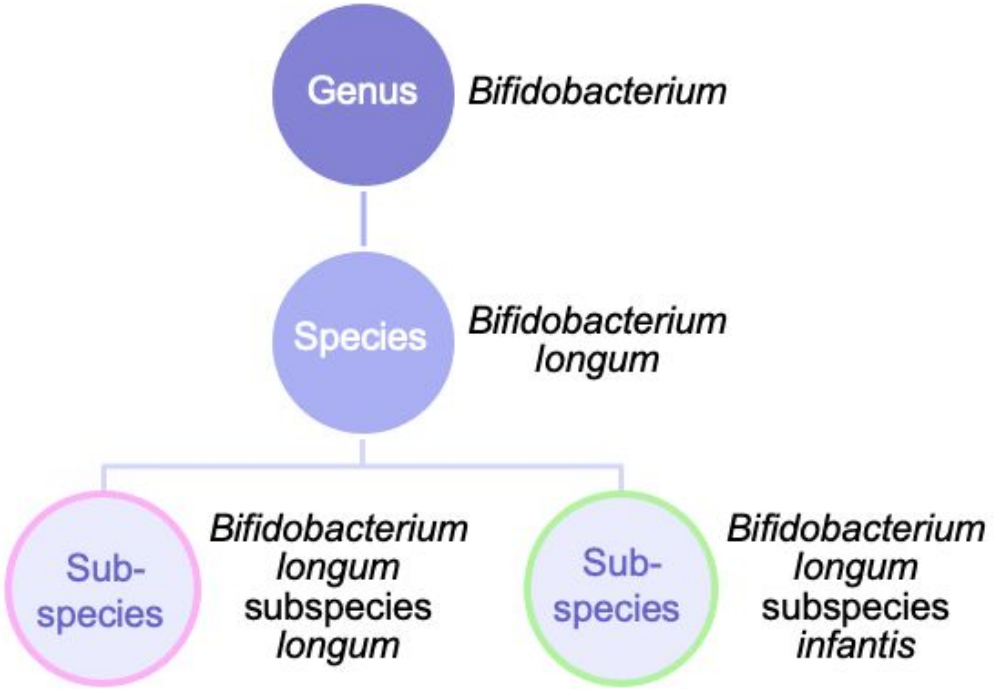
Bifidobacterium longum subspecies *infantis*

↓

Metabolise specific types of human milk oligosaccharides

↓

Healthier babies



Main methods for binning

- Composition-based
- Abundance and co-abundance-based
- Why not both!

Composition-based binning

It turns out that related organisms have similar small-scale patterns in their DNA e.g. a similar frequency of 1-mers, 2-mers, 3-mers and 4-mers (Tetra Nucleotide Frequency - TNF).

We can use statistics of those patterns to bin our DNA. Fast and easy, BUT:

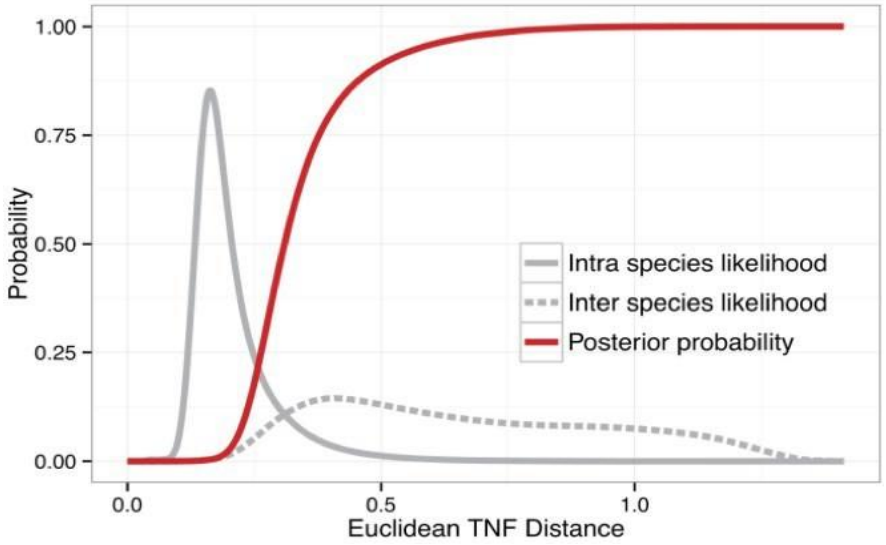
- Several species might have the same signal
- No guarantee that same species have same signal across the genome
- You need long pieces of DNA for statistics
- Composition deviation does not necessarily track anything you care about

Composition-based binning: Example (MetaBAT)*

*MetaBAT combines compositional and co-abundance binning.

They sampled 1 billion contig pairs from known genomes. Calculated 4-mer/TNF frequency for between- and within-species pairs.

Applies Bayes' Theorem for determining probability of being different species.
This probability can then be converted to a distance.



Binning algorithm:

1) Pick a seed contig (say, most coverage)
2) Get all contigs with distance less than D
3) Find the middlemost member of that set
4) Set that member as the seed
5) Repeat 2-4 until seed doesn't change
6) These are a bin. Remove them and repeat until no contigs are unbinned

Why does it not work with small contigs?

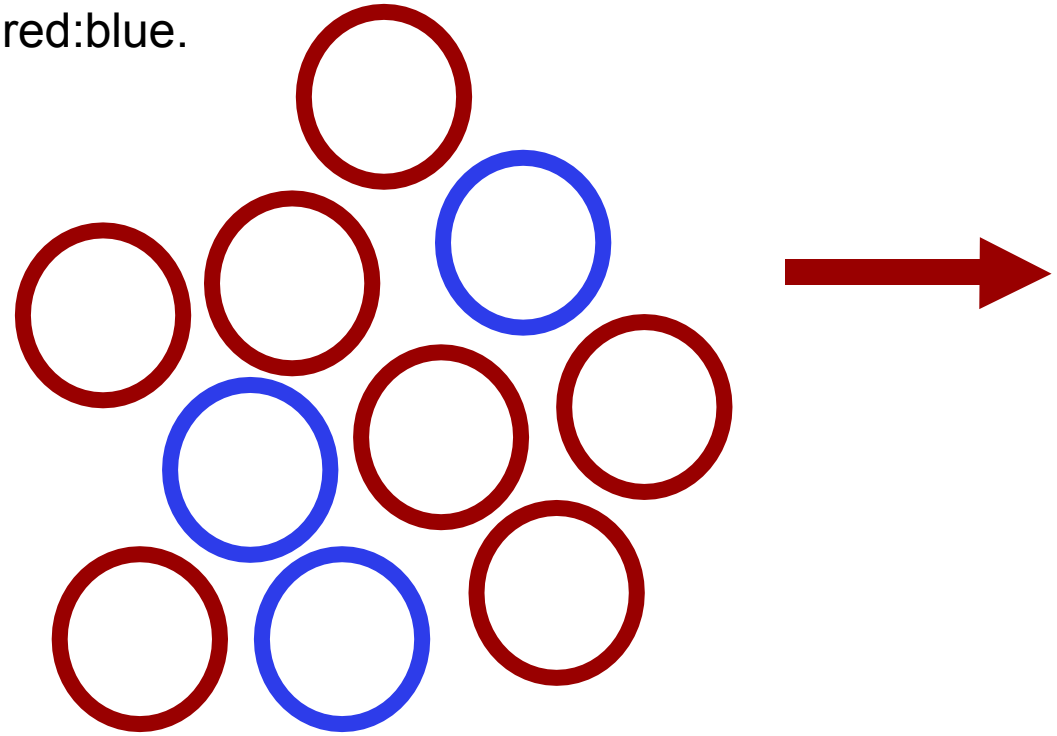
It works by comparing frequencies of kmers between contigs.
In short sequences, there are few kmers, frequencies are inaccurate.
It’s like trying to compare two parliament elections asking only 100 people!

Experiments show that 500 bp is enough to gain *some* information, 3,000 bp is enough to do rough binning, and the accuracy still increases up to about 25,000 bp!

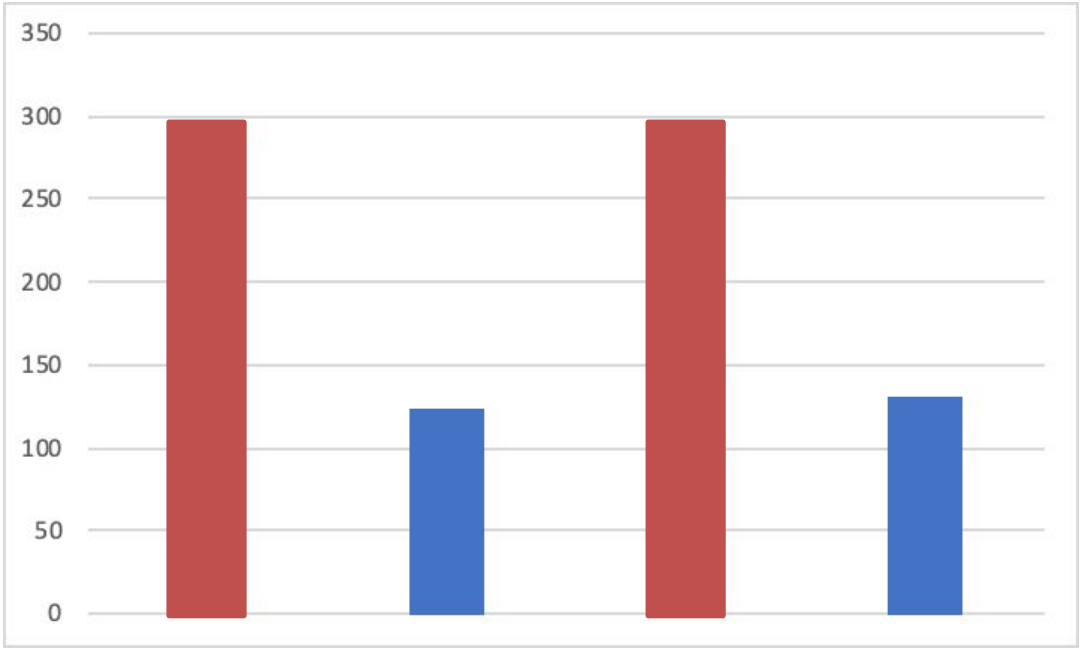
Abundance-based binning

Principle: If read/contig A and B both come from the same genome/plasmid, then they should exist in approximately equal amount in all samples. Therefore, they should have similar depth.

Ratio of microorganisms in the environment is 7:3 red:blue.

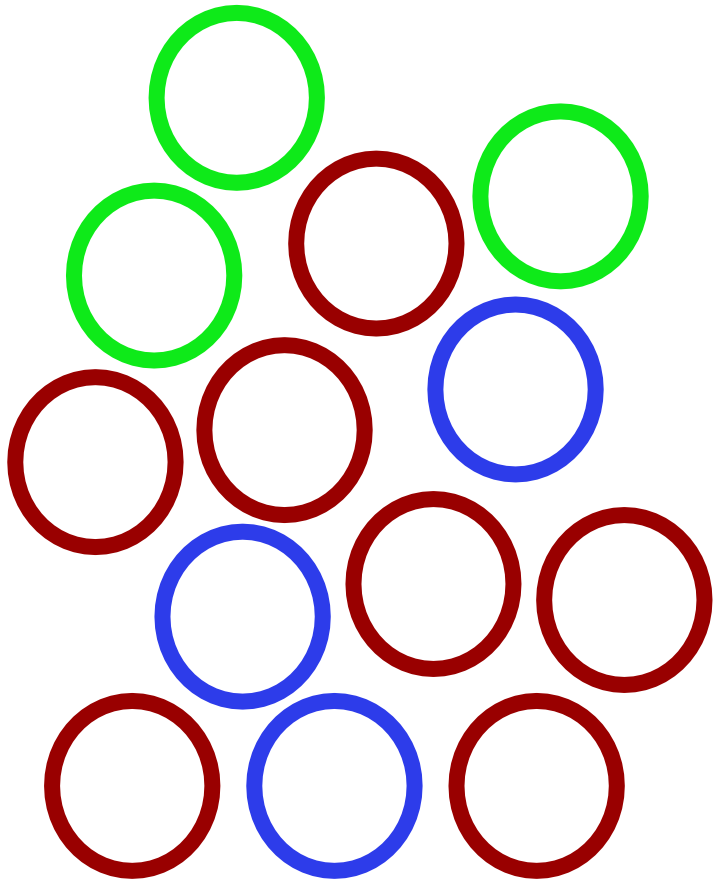


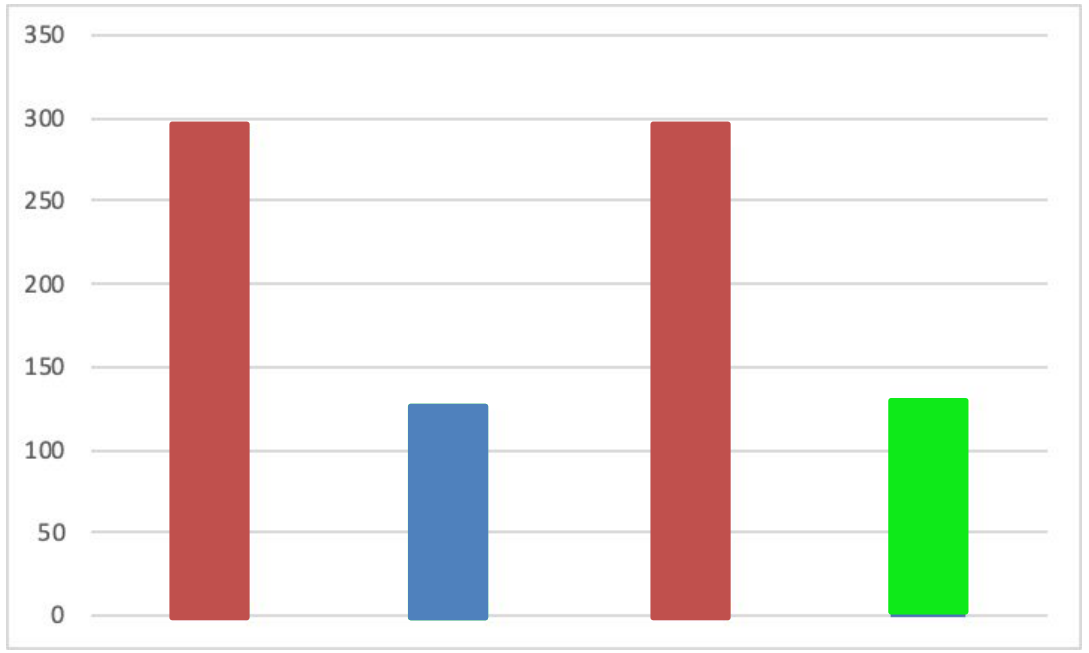
Depth of 4 random contigs. Which are from the blue microorganisms, and which are from the red?



Abundance-based binning

If the organisms are present in same abundance – impossible to separate the contigs

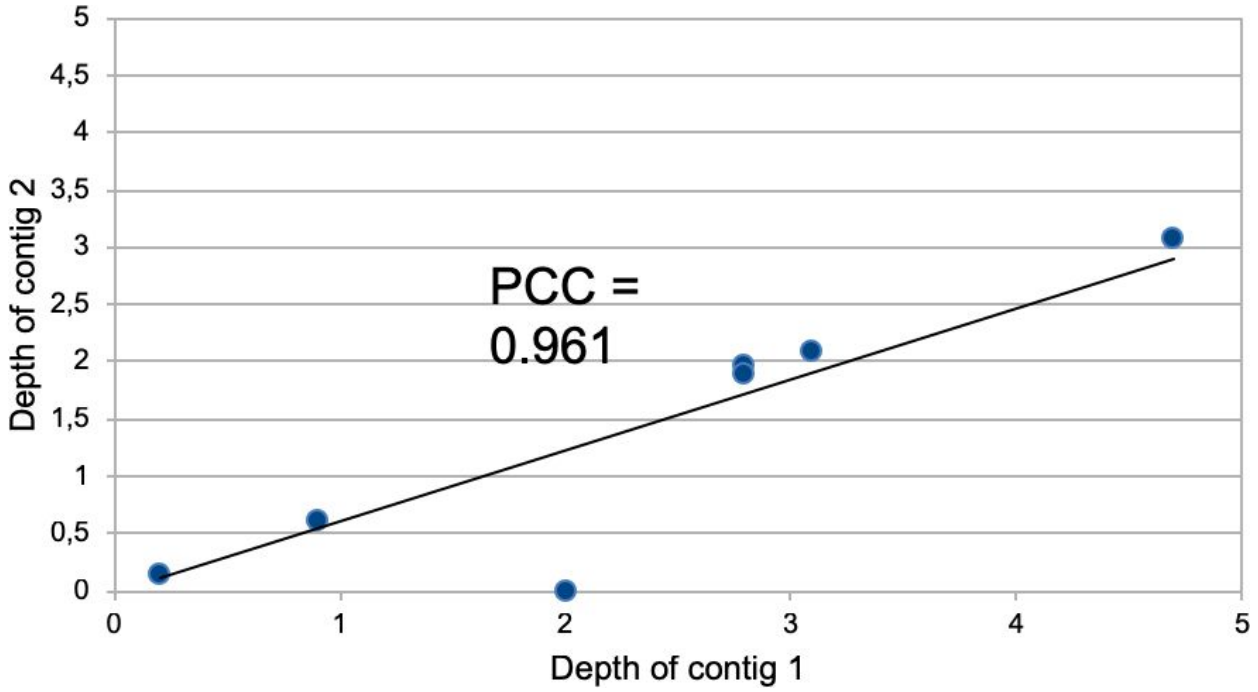




Co-abundance-based binning

- To get the depth, we map reads to the contigs and see how many reads map.

- If we have multiple samples, we can check the *correlation* of the depths between two contigs across all the samples. I.e. two contigs from different microorganisms may have the same depth by chance, but highly unlikely they have a similar depth in 10 independent samples...



- There's typically LOTS of noise, so itis only reliable with many samples!
- For low abundant organisms the contigs will not be present in most samples

Co-abundance-based binning

It does not rely on a database and we understand why it works, BUT:

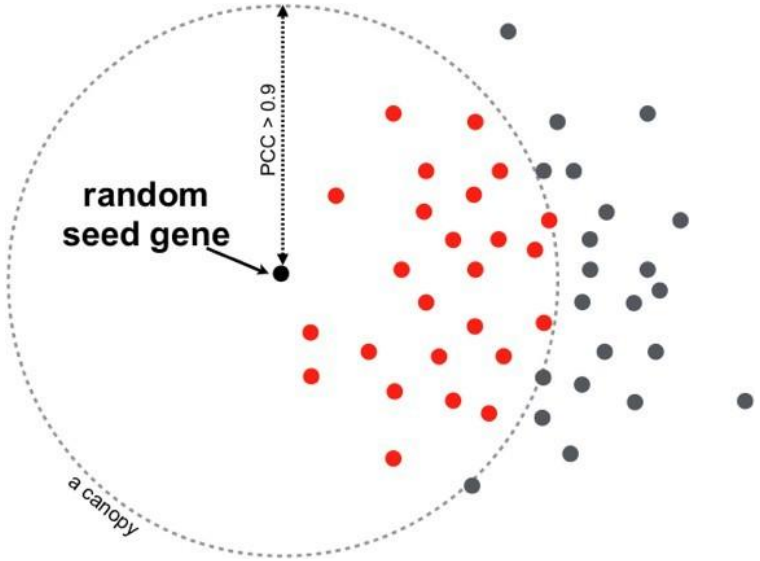
- It takes a long time to do it (LOTS of correlations to calculate).
- It's better with many samples with different abundances.
- You need to have a minimum level of depth.

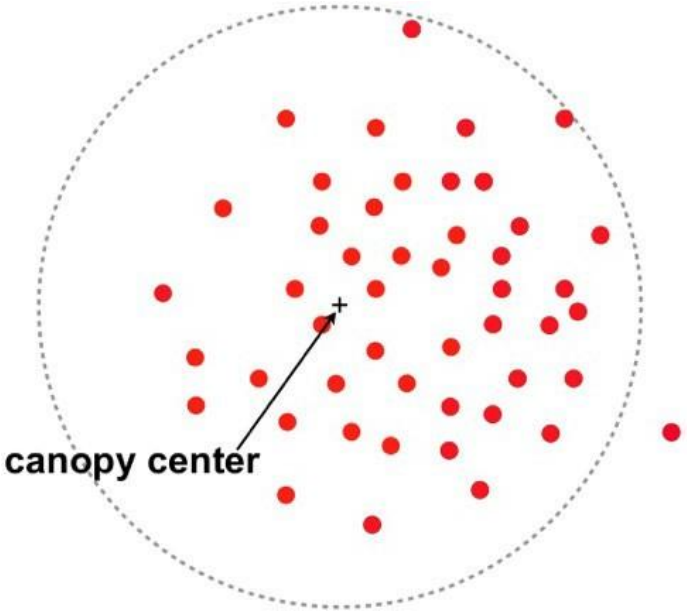
  Can you think of what happens to the correlation if most contigs have 5-10 reads mapping to them?
- You assume that each genome is present in many of your samples.
- Sensitive to having too many contigs to map against (random hits, reads attracted to the contigs they are part of)

Example of co-abundance-binning: Canopy

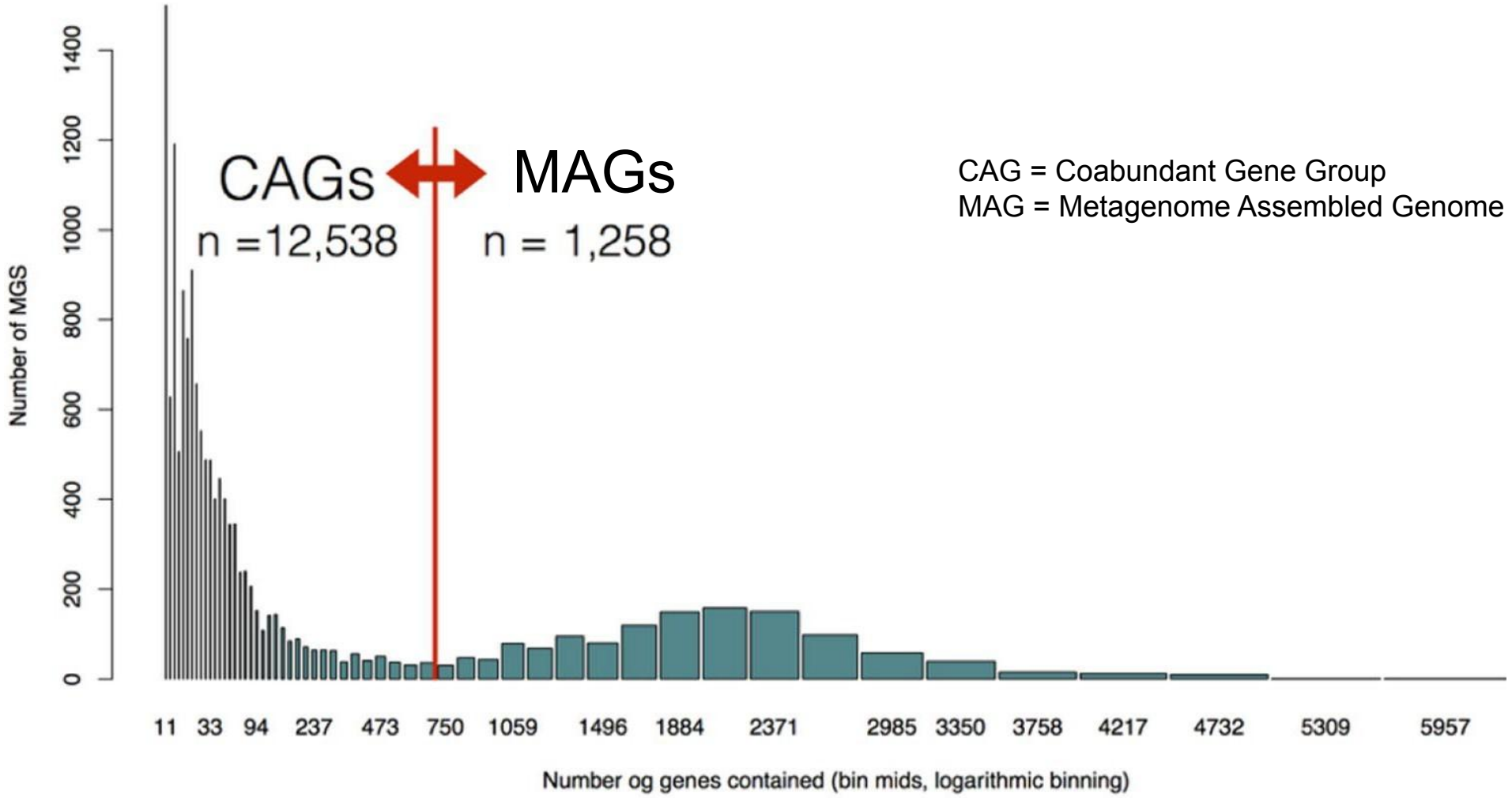
Algorithm:

1) Pick random seed contig
2) Pick all contigs with Pearson correlation > 0.9
3) Select centre of cluster
4) Repeat 2 and 3 until centre is stable
5) Continue until all contigs have been assigned to a cluster





Not just microbial genomes gets binned….



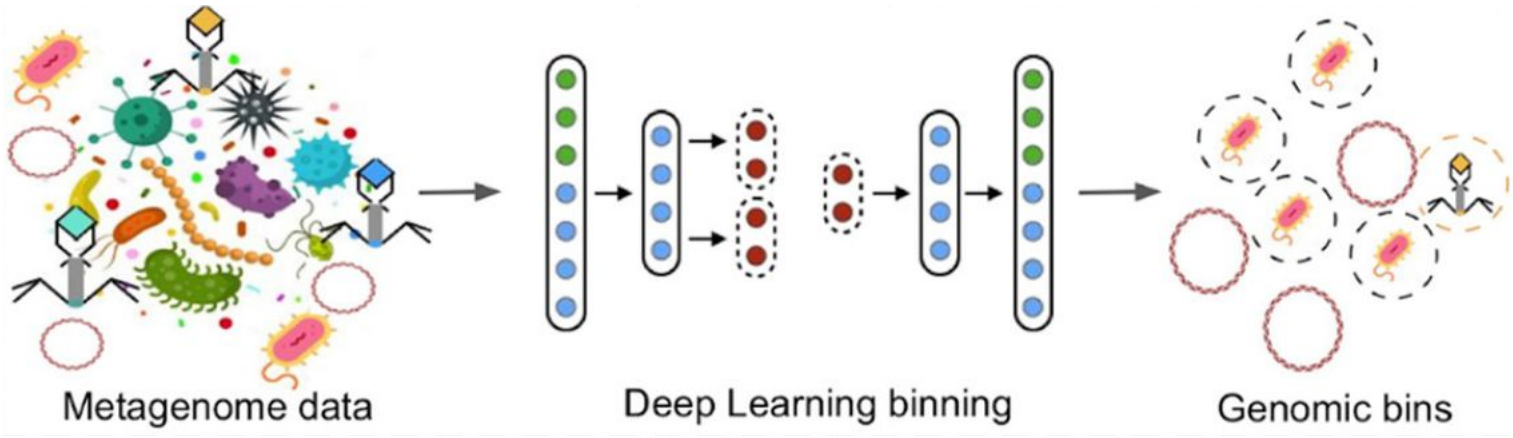
CAG = Coabundant Gene Group
MAG = Metagenome Assembled Genome

VAMB – Variational Autoencoders for Metagenomic Binning

- Uses both coabundance and sequence composition
- Can also assemble smaller biological entities, such as plasmids

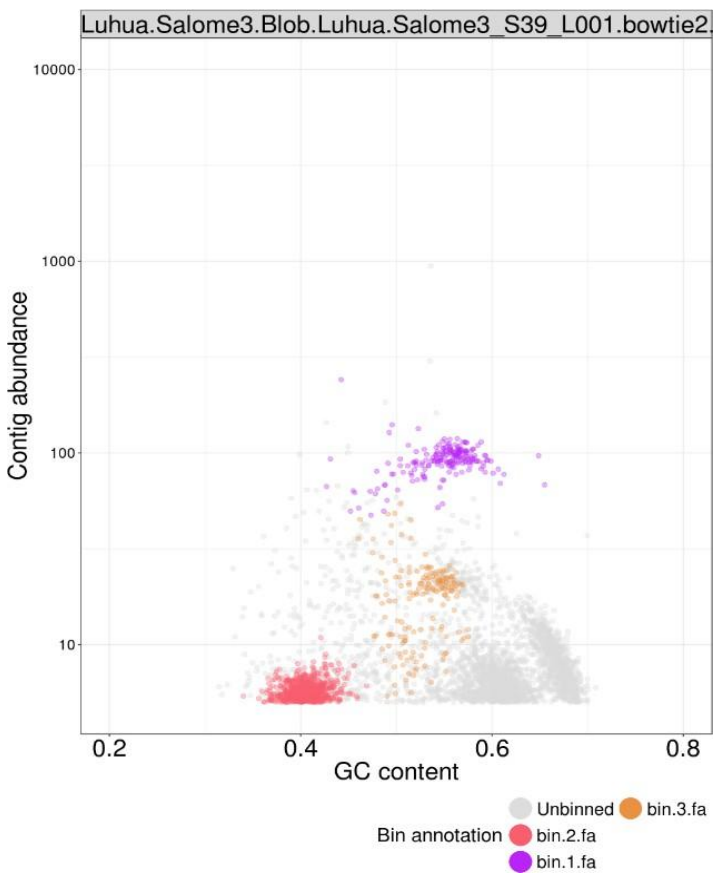
The Vamb pipeline consist of a series of tasks each which have a dedicated module:

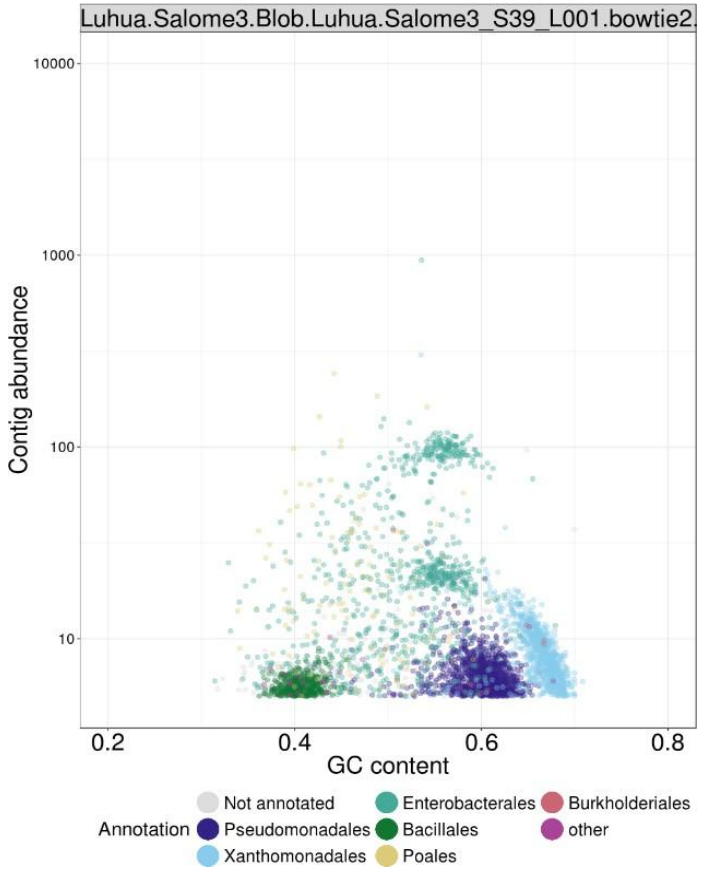
1. Parse fasta file and get TNF of each sequence, as well as sequence length and names (module `parsecontigs`)
2. Parse the BAM files and get abundance estimate for each sequence in the fasta file (module `parsebam`)
3. Train a VAE with the abundance and TNF matrices, and encode it to a latent representation (module `encode`)
4. Cluster the encoded inputs to metagenomic bins (module `cluster`)



Assesing bin quality - Blobology

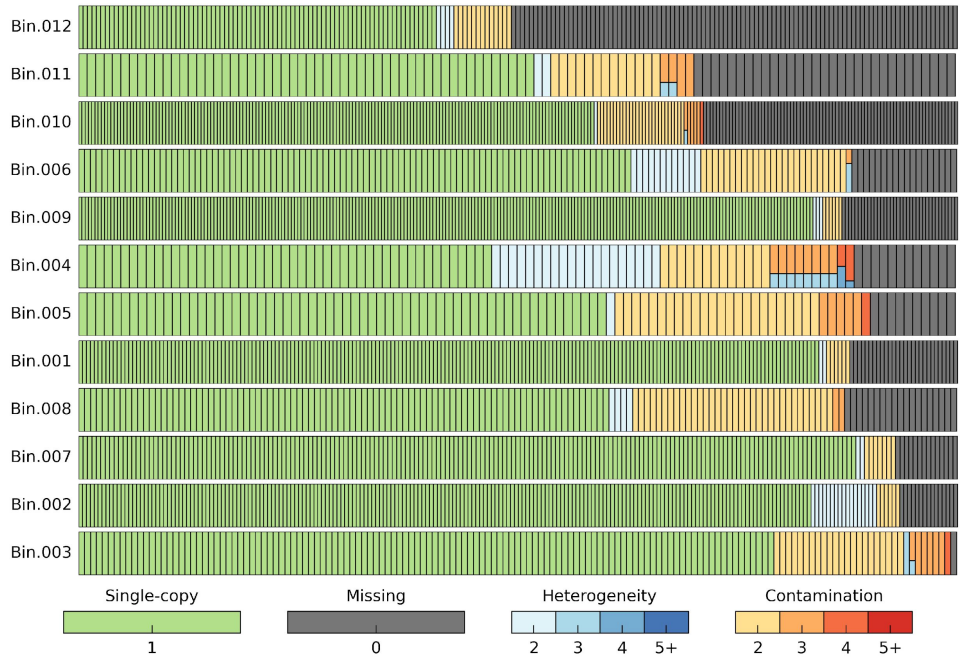
- Blobology plot shows contig abundance vs taxonomy or bin





Assesing bin quality - CheckM

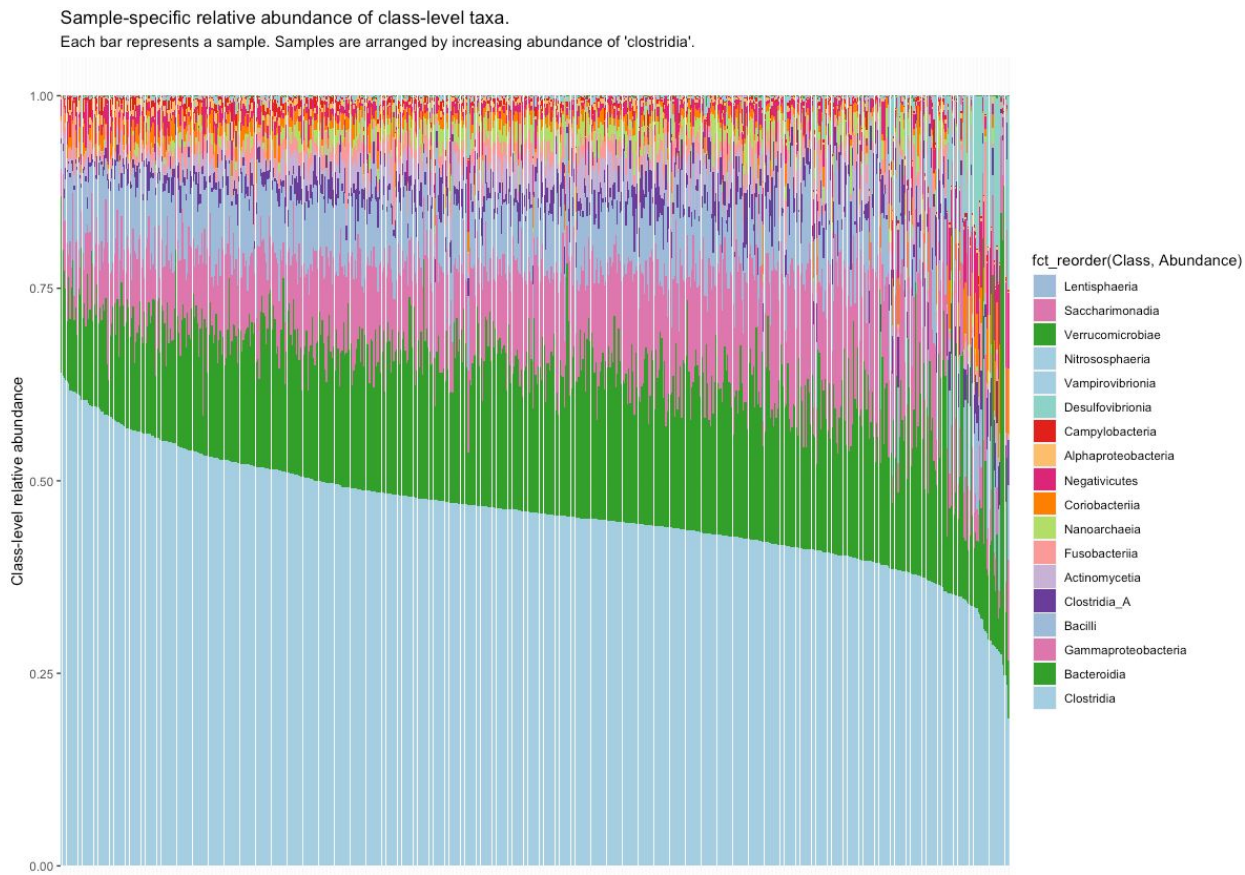
- Looks for single-copy genes
- Completeness and contamination is based on lineage-specific marker sets so NOT a universal gene-set
- Heterogeneity is an indication that contamination is caused by strain variation



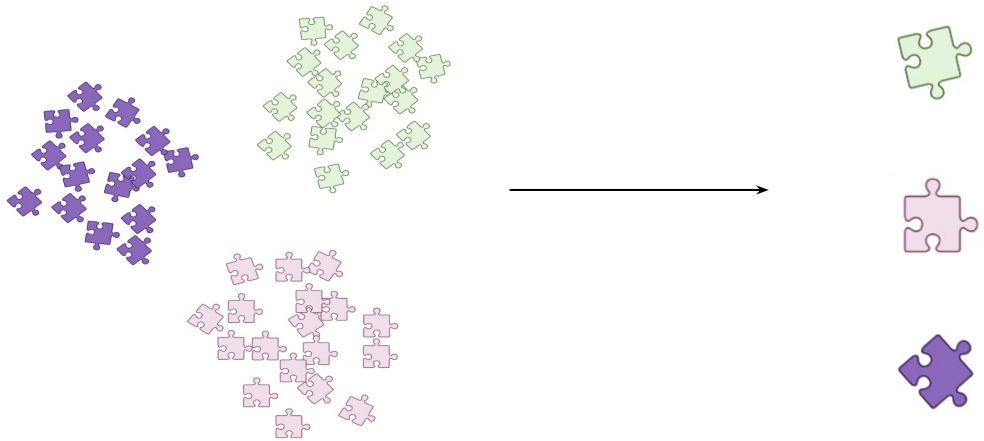
Abundance of reconstructed genomes

- Map all reads to all contigs or genes
  – Does generally not work

- Gene-based binning can use signature genes
  – Map reads to the 100 genes with highest bin correlation



Signature genes – the idea

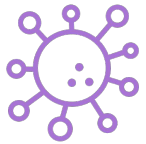
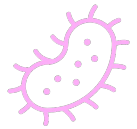


Signature genes

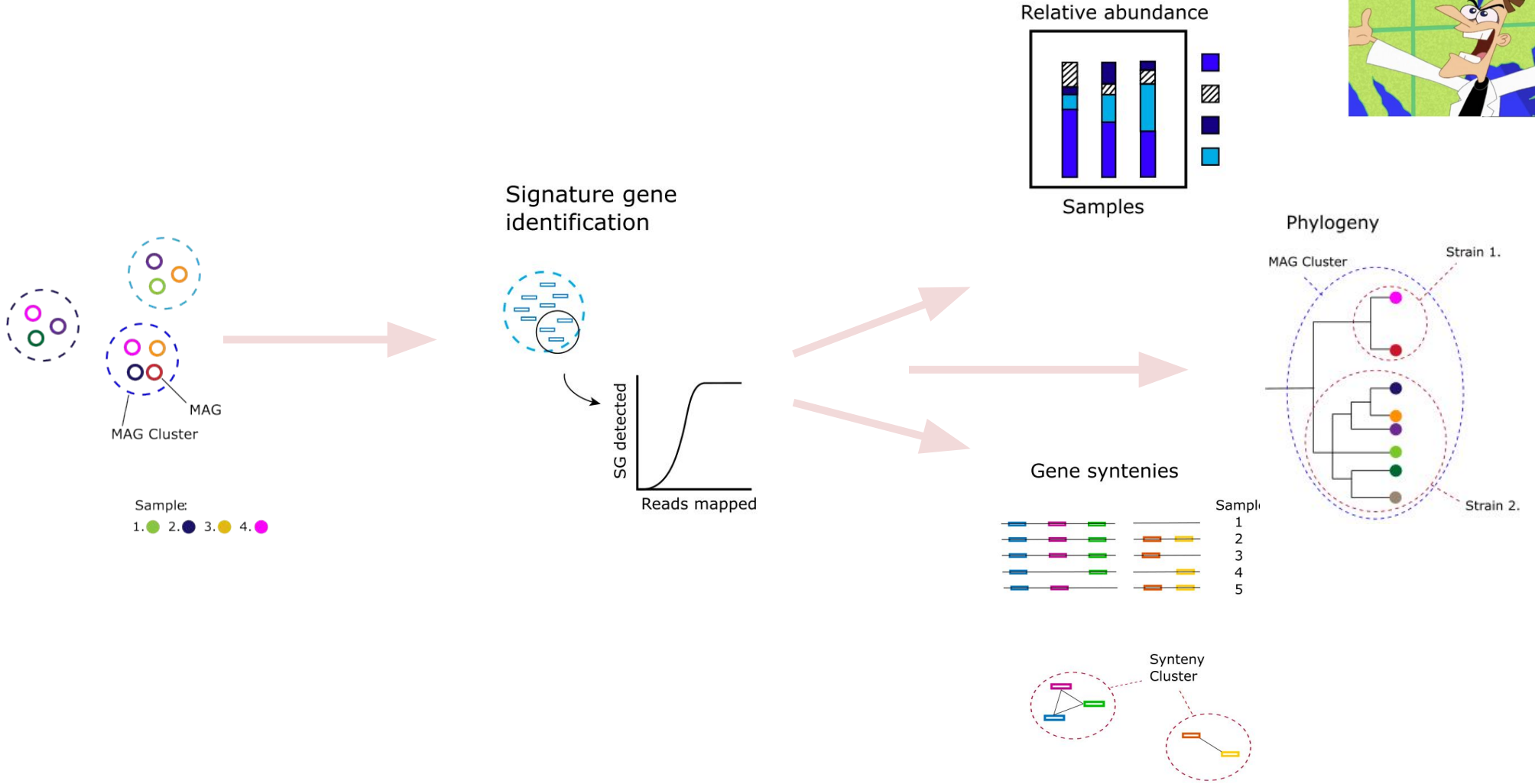
- Unique for the species
- Found in all members of the species
- Found with the same frequency in the species

Signature genes – the idea

The signature genes are used for quantification

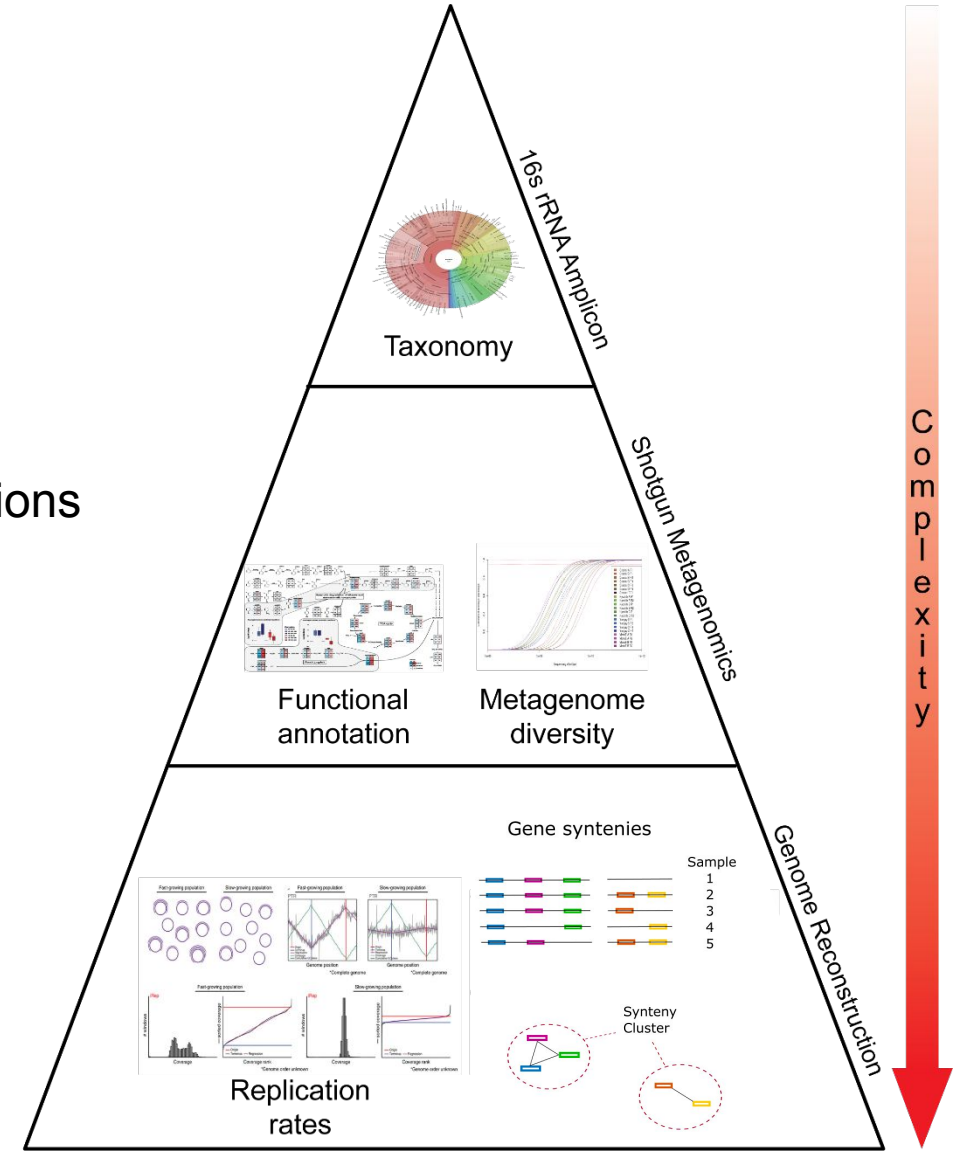
	Sample 1	Sample 2
	10	1
	0	3
	10	0

MAGinator – using signature genes for microbiome profiling



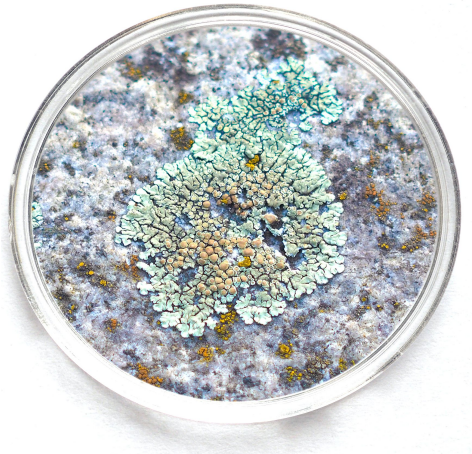
Horses for courses

- Question dictates the appropriate tool
- Amplicon sequencing is easy & cheap
- Metagenomics allows more questions
  – Discovery of novel proteins, antibiotics etc.
- Binning provides context and allows even more questions



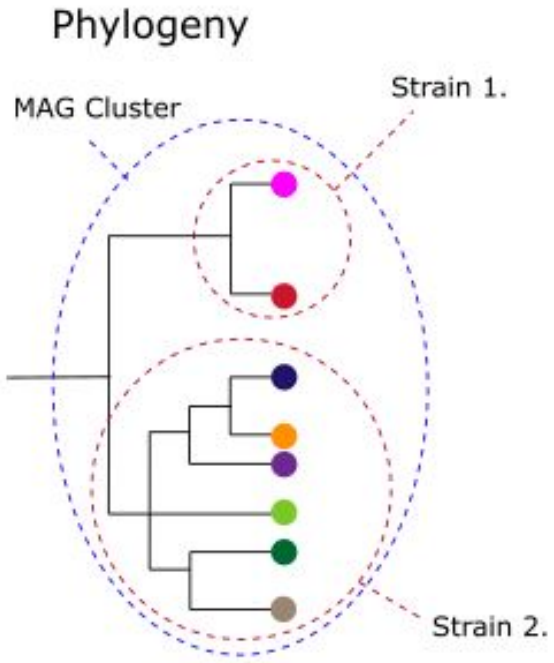
Summary

- Binning is a way of separating sequences(often contigs or genes) into genomes
- Adds additional context connecting genetic content and functions with taxonomy
- Computationally heavy
- Some methods are good for reconstructing genomes while others can pick up smaller elements such as plasmids and accessory genes
- Each have strengths and weaknesses



Student project:

Do the functional profiles of the strains follow the phylogeny?



- Computational project
  - Shell scripting
  - Python

- Collaboration with KU (me) and COPSAC

COPSAC

COPENHAGEN PROSPECTIVE STUDIES ON ASTHMA IN CHILDHOOD