DTU

DTU Health Technology
Bioinformatics

Metagenomics & binning

Gisle Vestergaard
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gisves@dtu.dk

Menu

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

Break time

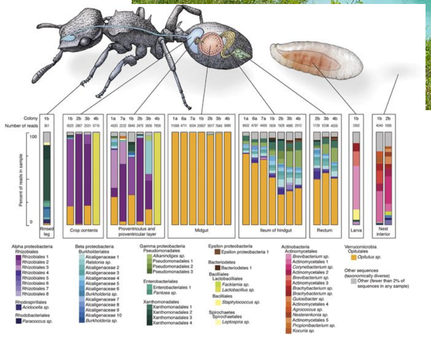
- What is binning?
- Types of metagenomic binners
- Assesing bin quality
- Dereplicating similar genomes

What is metagenomics?

"Metagenomics (Environmental Genomics, Ecogenomics or Community Genomics) is the study of genetic material recovered directly from environmental samples.”

A Microbiome is all the microbes in a defined habitat

What is the difference between a microbiome and a metagenome?

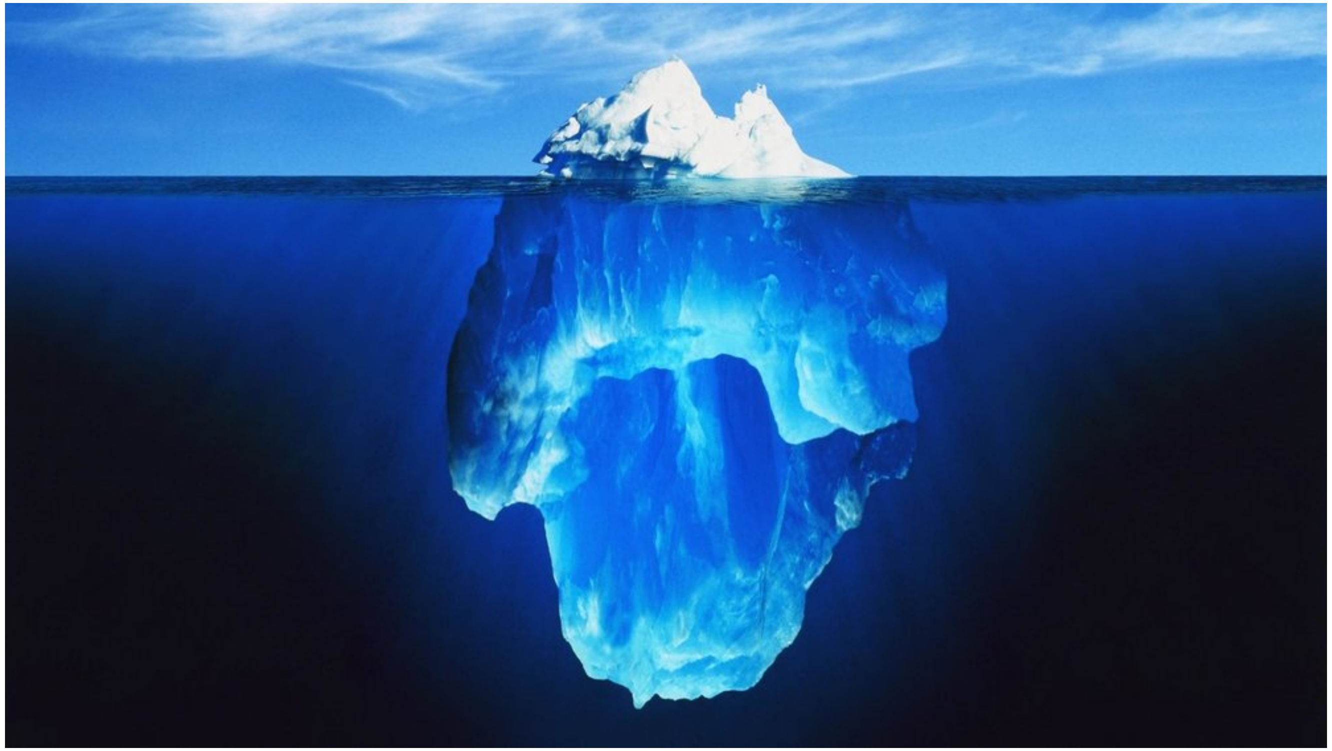


Microbiome research in pre-sequencing days

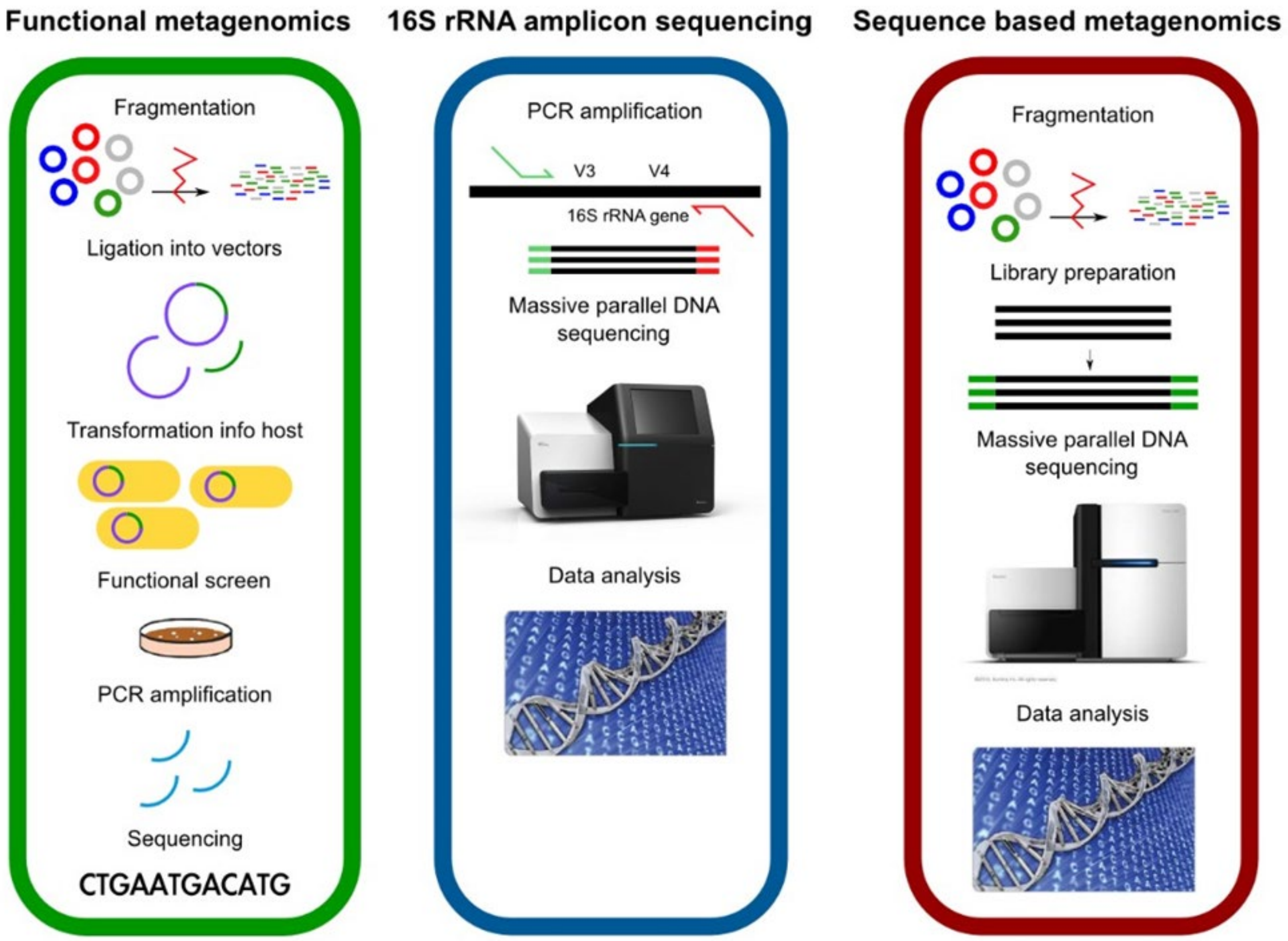
- Culturable organism chosen as models
- Might not be representative even for close relatives

From culturing to sequencing

- Microbiome research previously limited to culturable organisms
- +99% of prokaryotes in the environment cannot be cultured



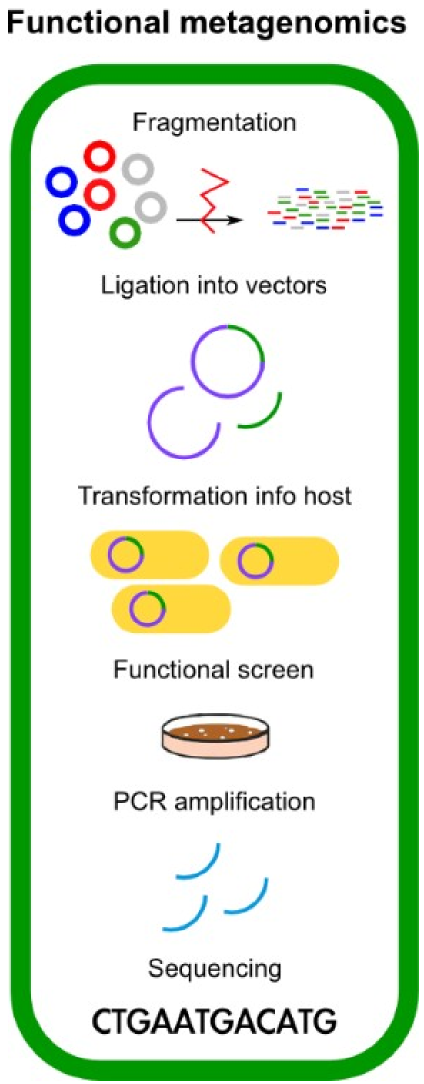
Methods


Functional metagenomics
Fragmentation
Ligation into vectors
Transformation info host
Functional screen
PCR amplification
Sequencing
CTGAATGACATG
16S rRNA amplicon sequencing
PCR amplification
V3
V4
16S rRNA gene
Massive parallel DNA sequencing
Data analysis
Sequence based metagenomics
Fragmentation
Library preparation
Massive parallel DNA sequencing
Data analysis

Functional metagenomics

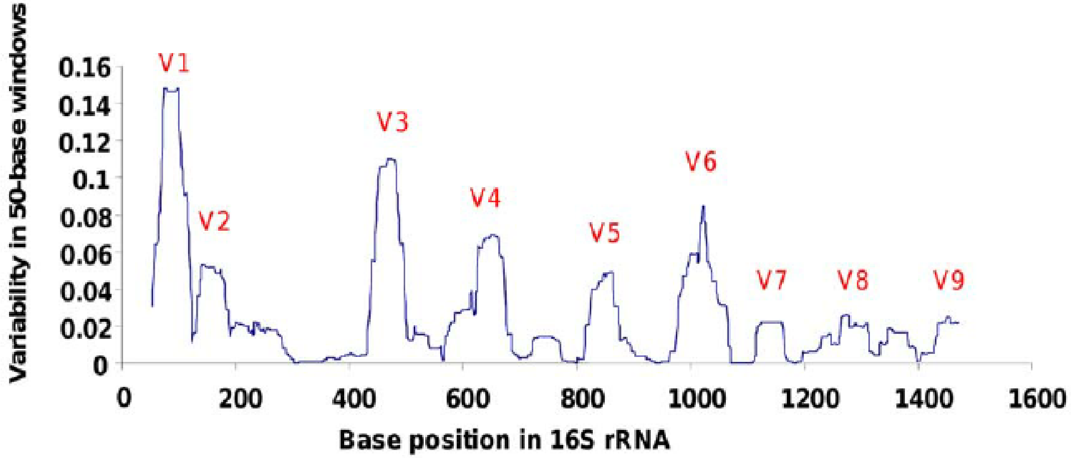
- Industrial large scale screening
- Time consuming due to functional screens
- Setup determines screening specificity

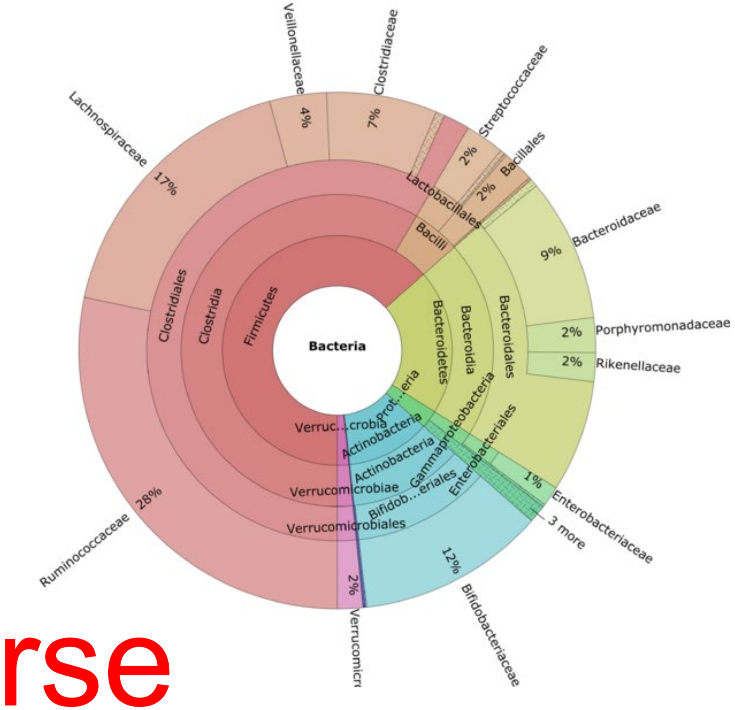
Not part of this course



16S rRNA

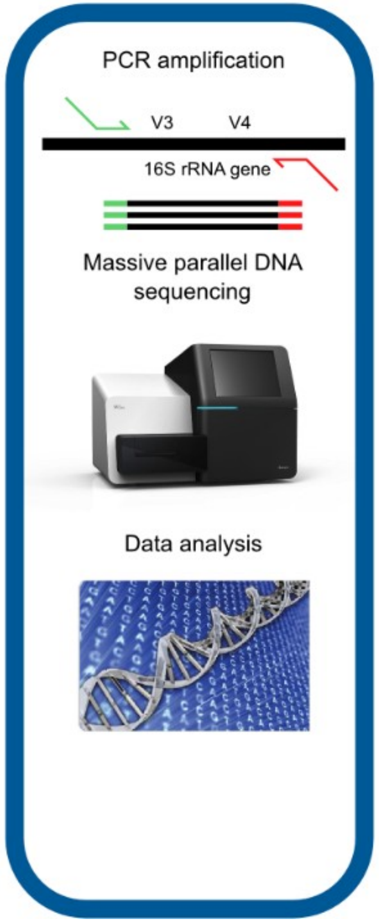
- Ribosomal RNA gene conserved in Bacteria and Archaea
- Conserved areas interspaced with hypervariable regions
- Allows the determination of the phylogenetic composition of a microbiome
- Bias by primers selected





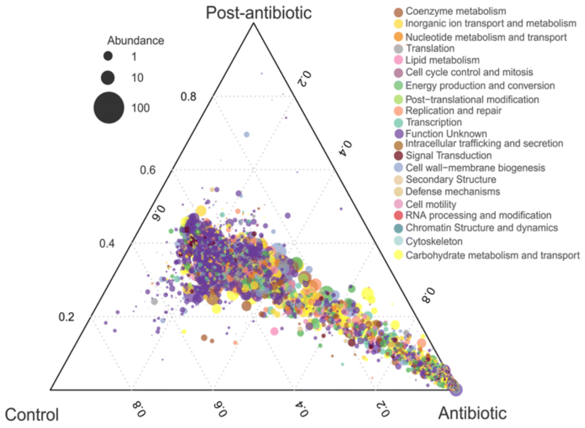
Not part of this course

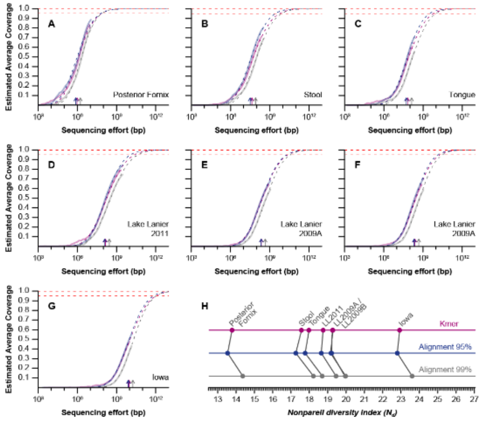
16S rRNA amplicon sequencing

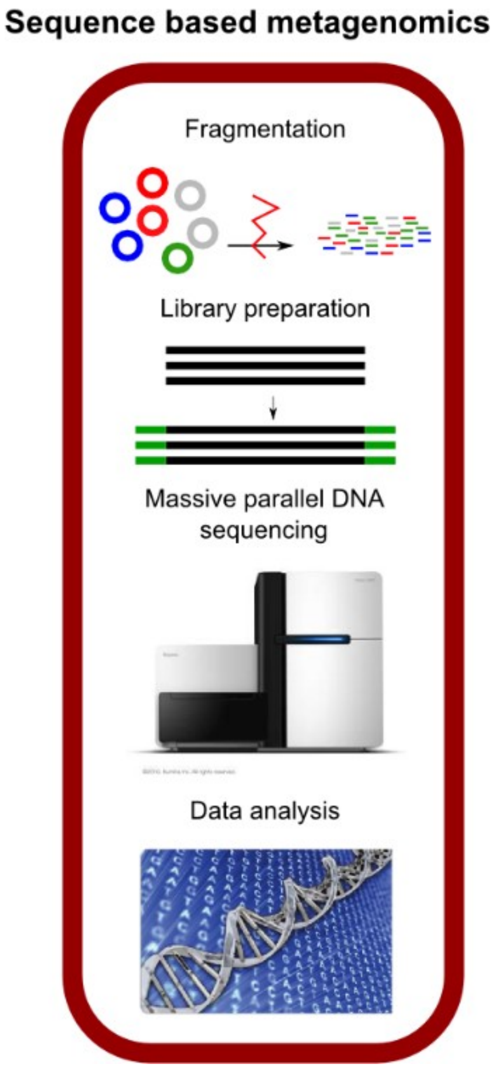


Metagenomics

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



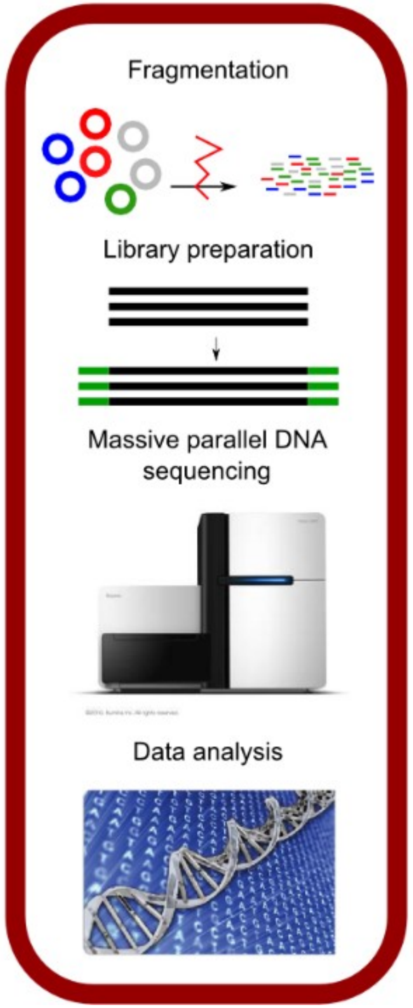




Metagenomics – the downside

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

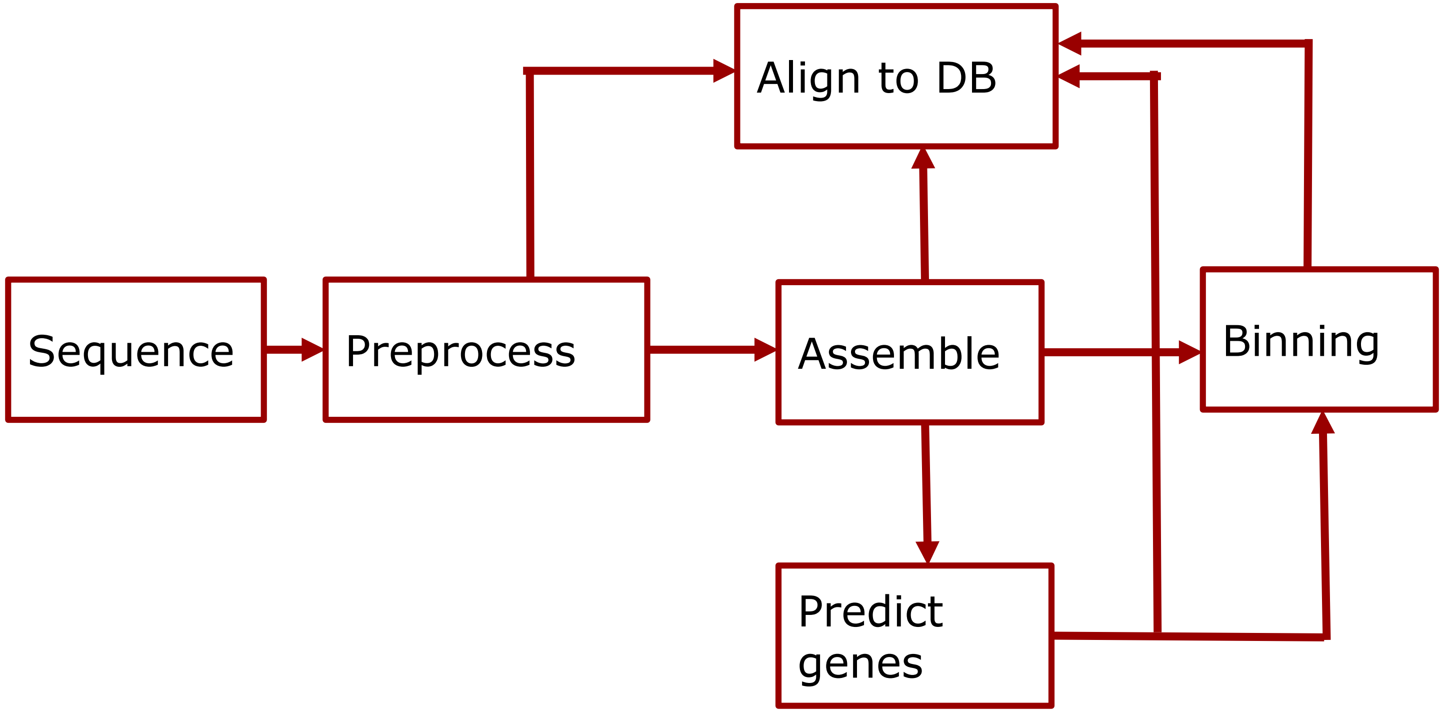
Sequence based metagenomics



The most common problem with metagenomics

- Many projects are data-driven and exploratory
- Hypotheses are essential to stay on track

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

The nature of metagenomics cause problems in all elements of a normal workflow

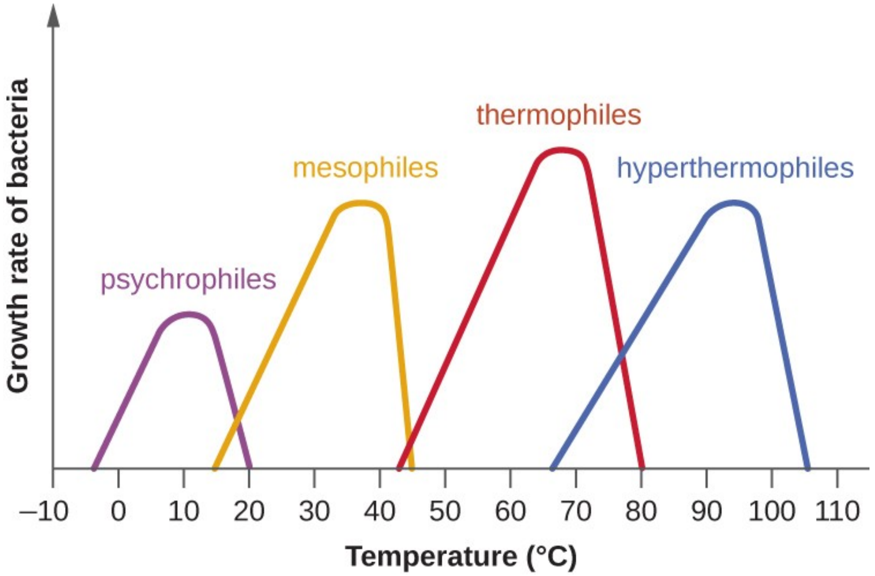
Sequence

Sampling & storage

- Host contamination
- Storage at 4˚ Celsius will affect composition

Sample preparation and sequencing method used

- Library preparation kits introduce biases

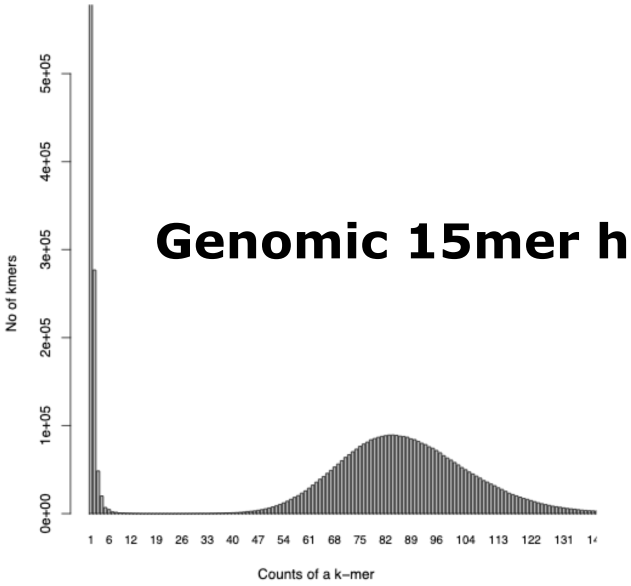


Preprocessing problems

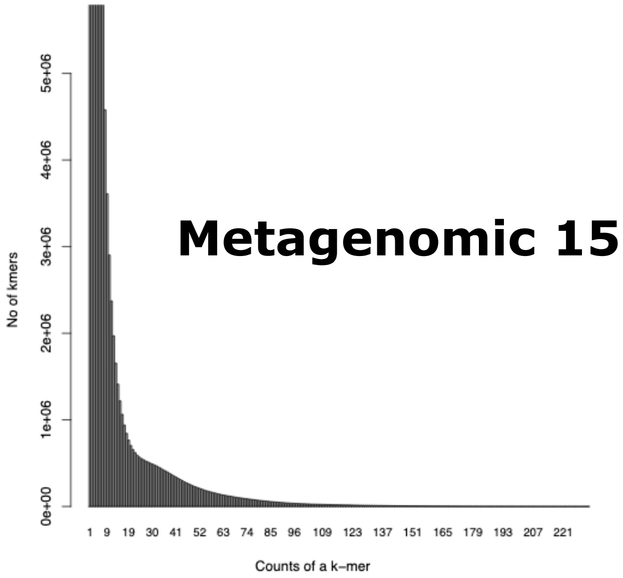
Preprocess

We can still trim reads for low quality and adapters but:

- Kmer correction is close to impossible. Can you think of why?



Genomic 15mer histogram

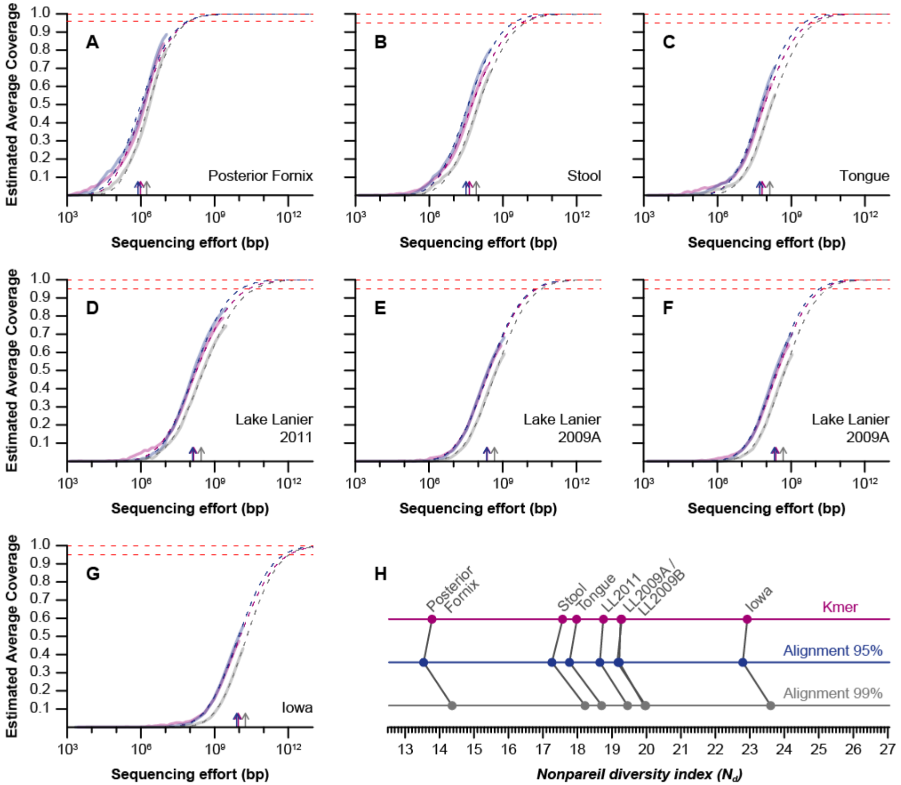


Metagenomic 15mer histogram

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?



Alignment problems

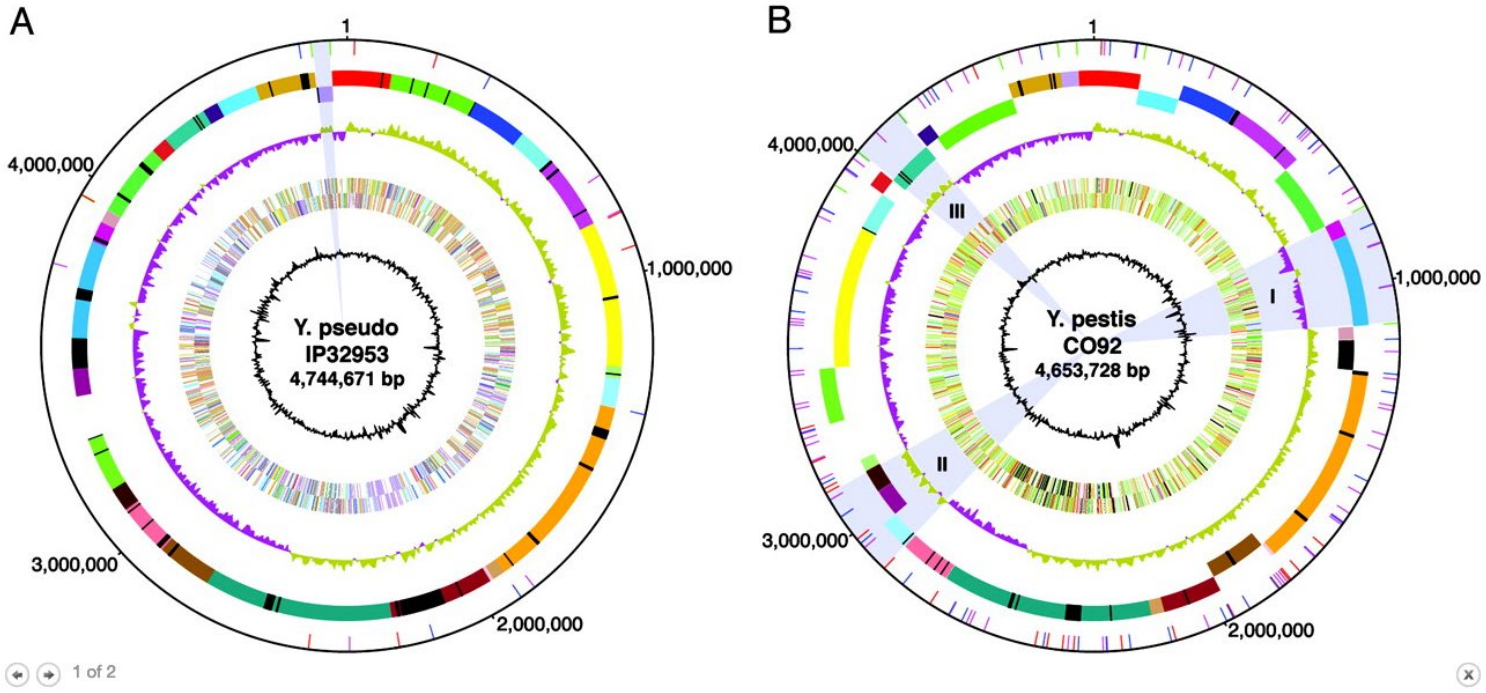
Align to DB

- 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 are able to 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.*
3000 / 4000 genes 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!

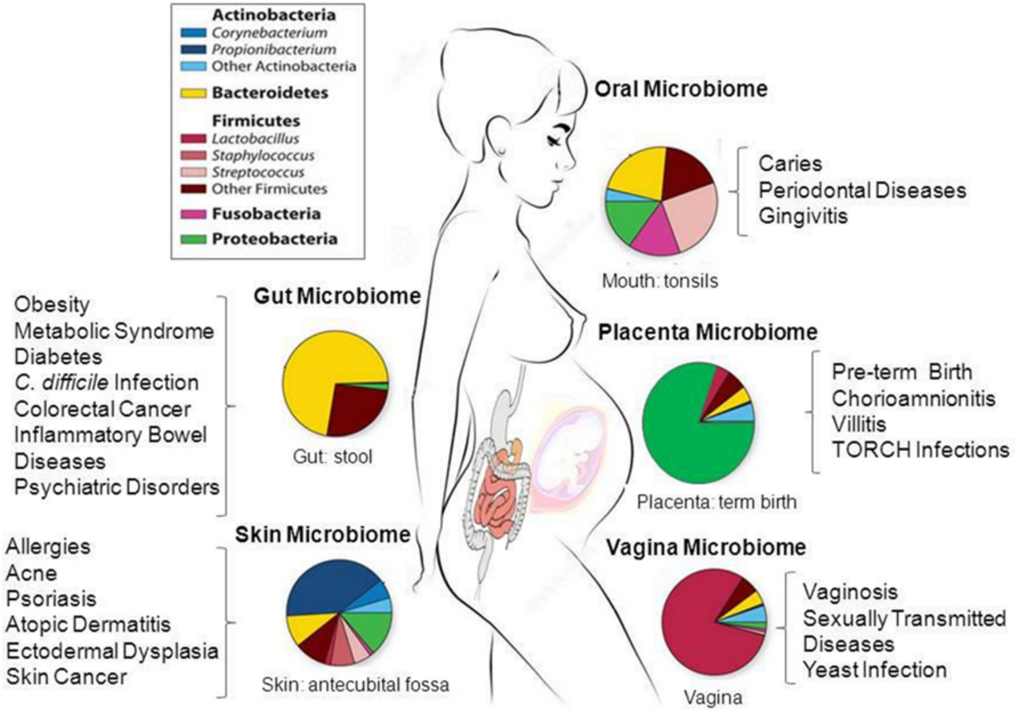


Uses for metagenomics

- Find new enzymes for industrial use
- Find new antibiotics
- Interactions between host and microbiome

Gut microbiome from a fat and a skinny human transplanted into genetically identical mice (Turnbaugh 2006 Nature) Gut microbiome has also been linked to asthma, diabetes, inflammatory bowel disease and many more.

- Many examples of transitions from fundamental research to application such as CRISPR-Cas systems



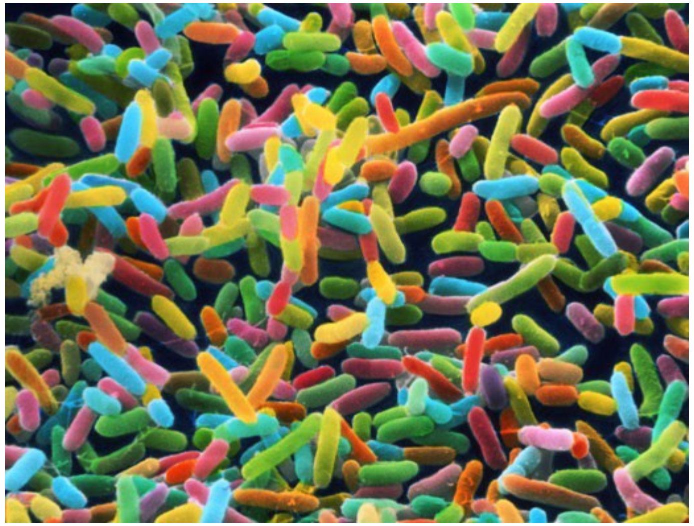
Next up…

Metagenomic binning

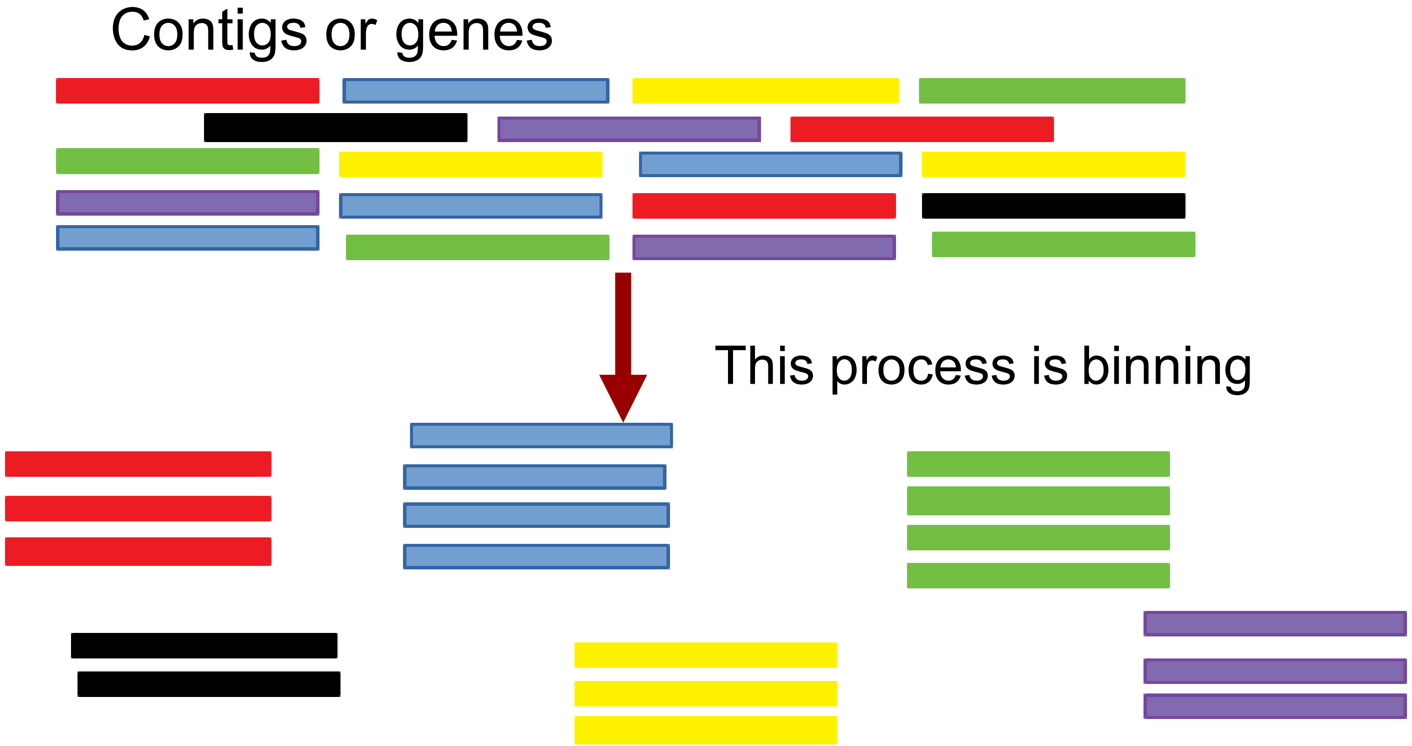
Menu

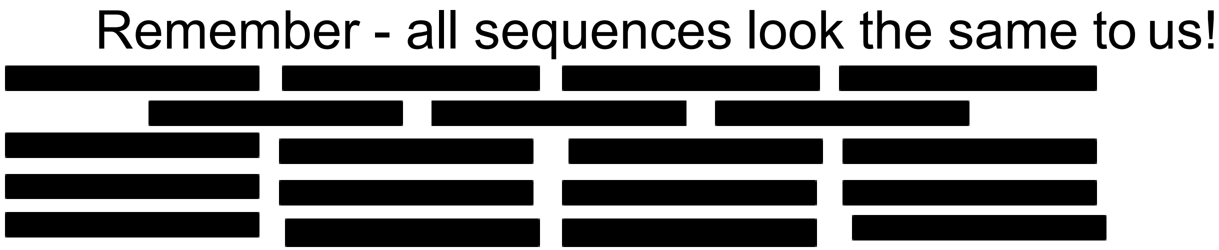
- What is binning?
- Types of metagenomic binners
- Assesing bin quality
- Dereplicating similar genomes

What is binning?



Metagenomic sample





Why do we care?

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

Sometimes, this is okay:

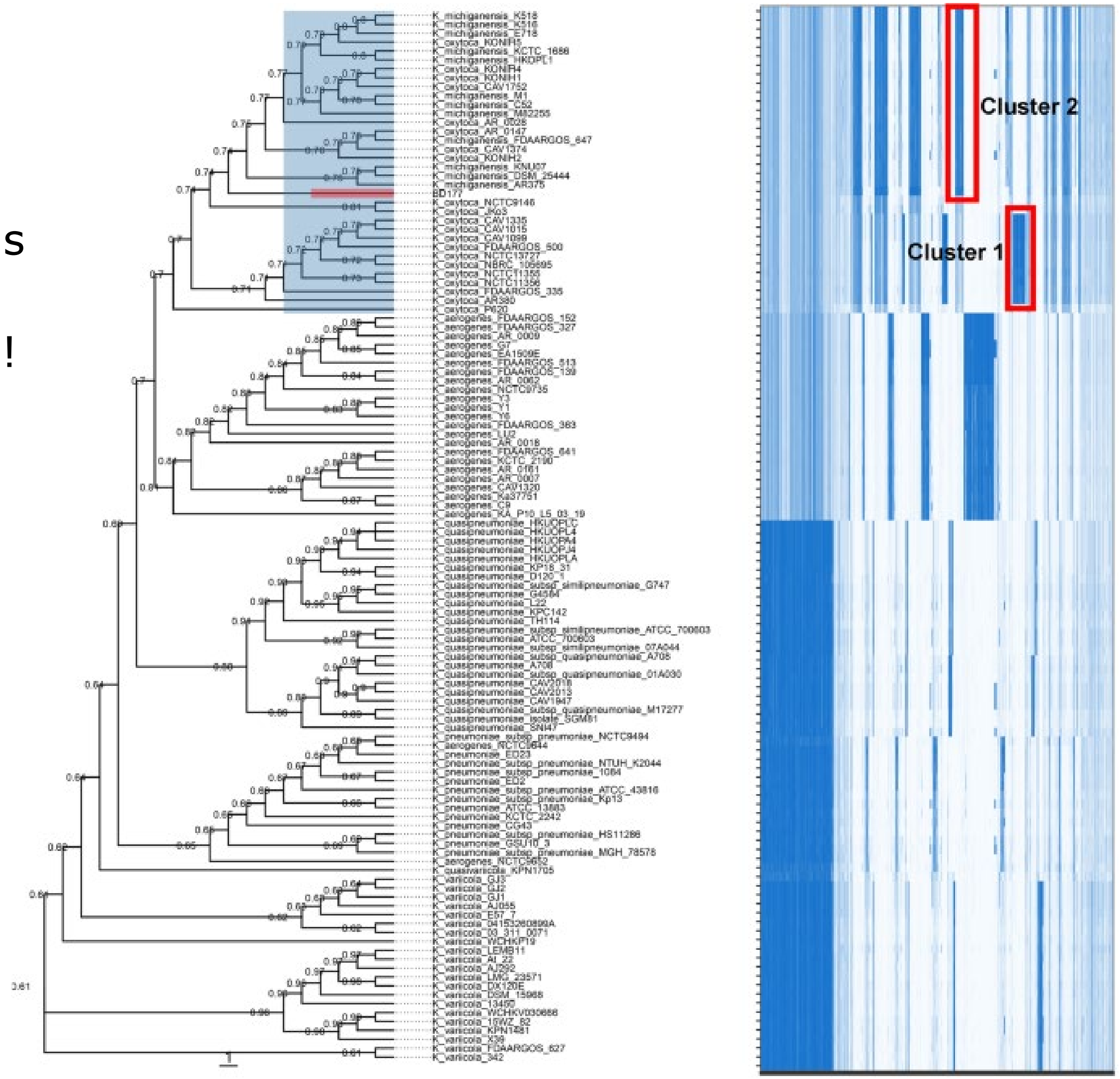
- If we find 16s DNA 99.9% identical to *B. subtilis*, it's there.
- When scanning for interesting genes with a certain signature

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?

Klebsiella Pangenome

- A total of 49305 genes
- Only 858 core genes
- Clusters 1 and 2 distinguish two very similar strains one of which is probiotic
- This also allows the investigation of defining genes!



Main methods for binning

- Composition-based
- 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.

No-one knows why.* It's ONLY not due to GC content and codon bias. Maybe host restriction enzymes and different biases DNA replication/repair errors?

No matter why, it means 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
- Annoying to rely on something no-one knows how works

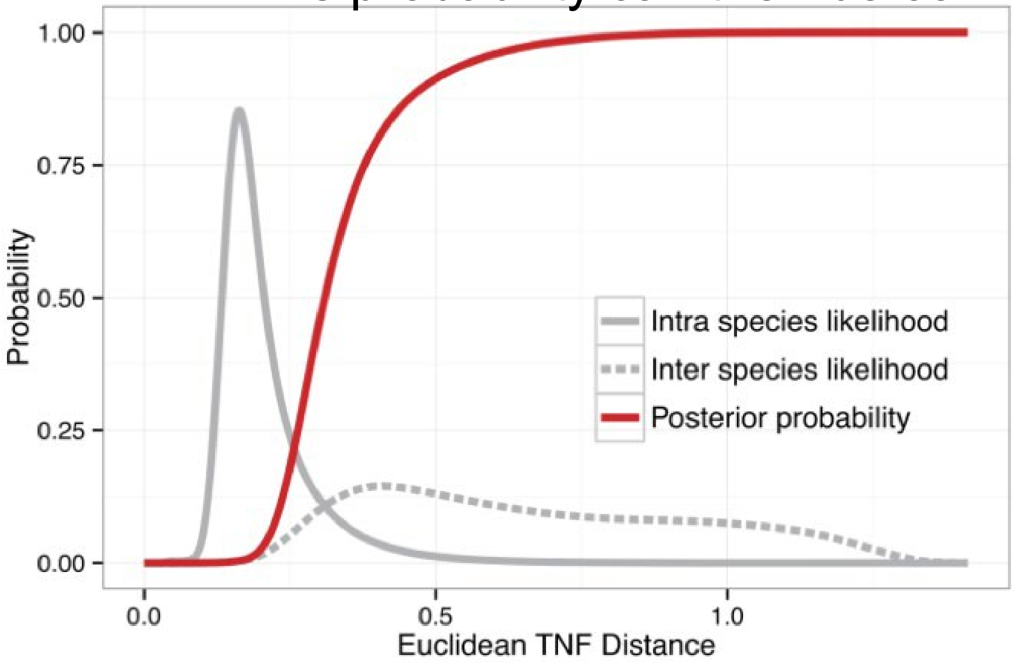
*People have looked into it. See: http://genome.cshlp.org/content/13/2/145.full, https://doi.org/10.1093/molbev/msp032

Composition-based binning: Example (MetaBAT)*

*MetaBAT combines compositional and co-abundance binning.

They sampled 1 billion contig pairs from known genomes.
Calculated 4-mer 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?

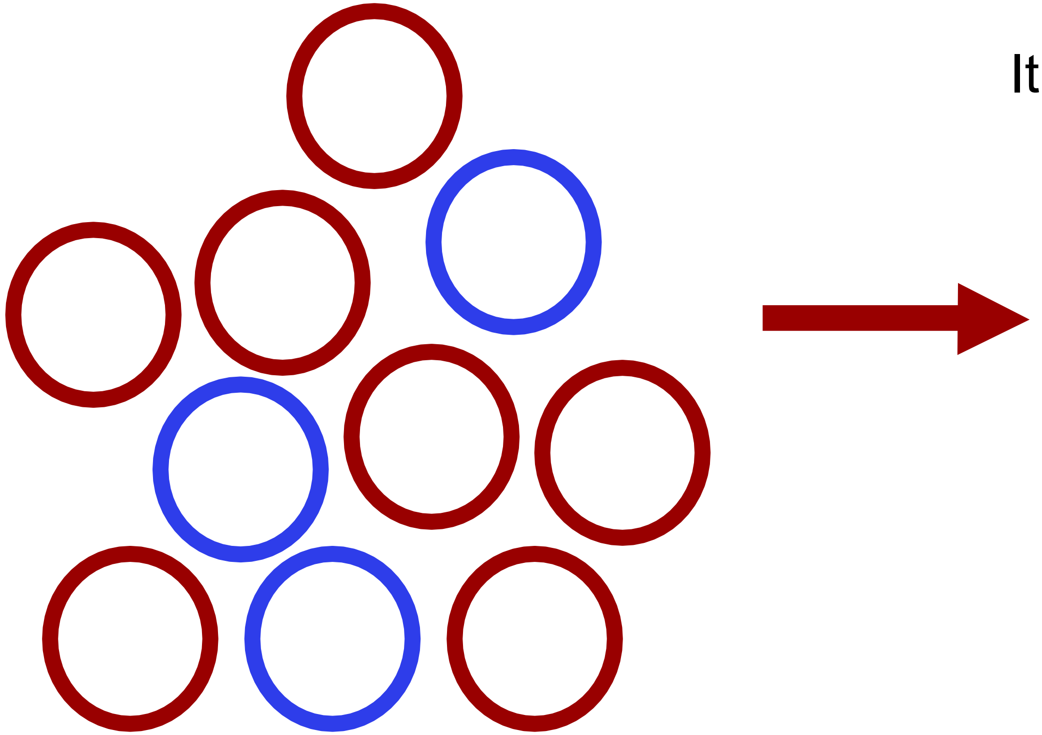
It works by comparing frequencies of kmers between contigs.
In short sequences, there are few kmers, frequencies are inaccurate.
It’s like trying to compare two parliament elections asking only 100 people!

Experiments show that 500 bp is enough to gain *some* information, 1,500 bp is enough to do binning, and the accuracy still increases up to about 25,000 bp!

Co-abundance-based binning

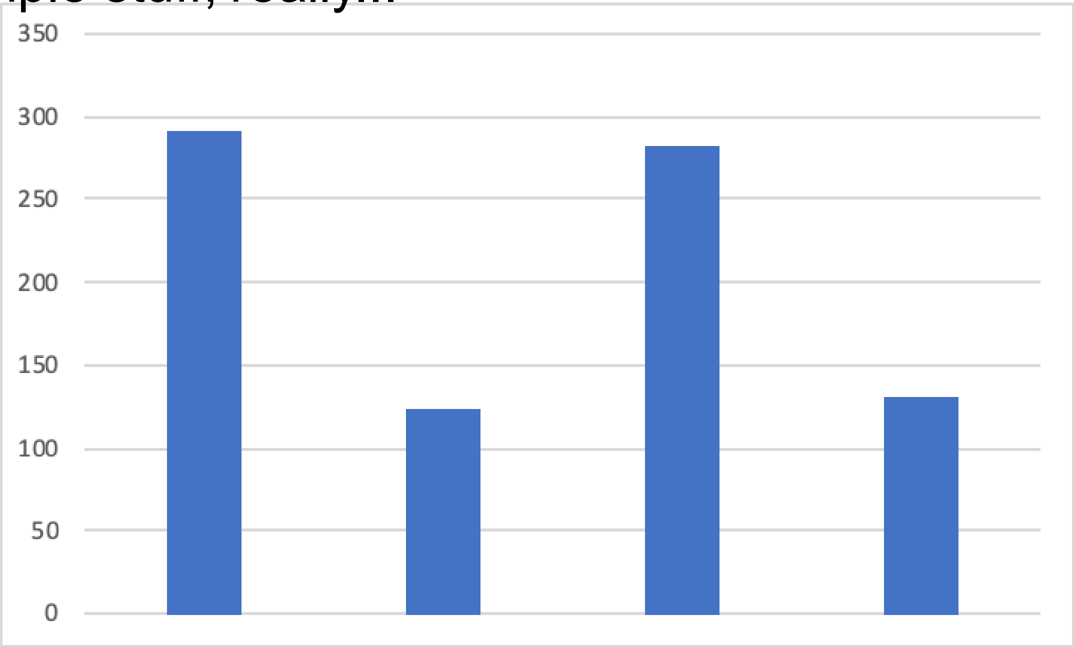
Principle: If read/contig A and B both come from the same genome/plasmid, then they should exists in approximately equal amount in all samples. Therefore, they should have similar depth.

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?

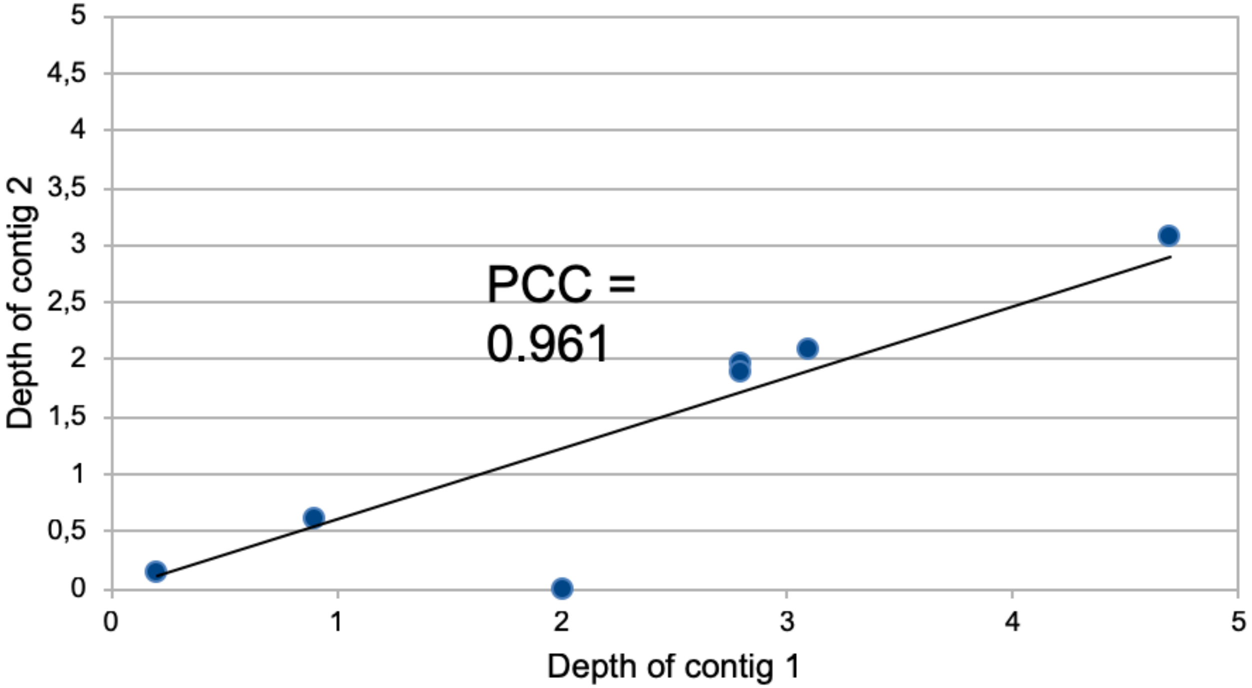
It's simple stuff, really...



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 it is only reliable with many samples!
- Also, BWA MEM sucks at mapping against metagenomic data. Someone needs to create an aligner better suited for metagenomics!

Co-abundance-based binning

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

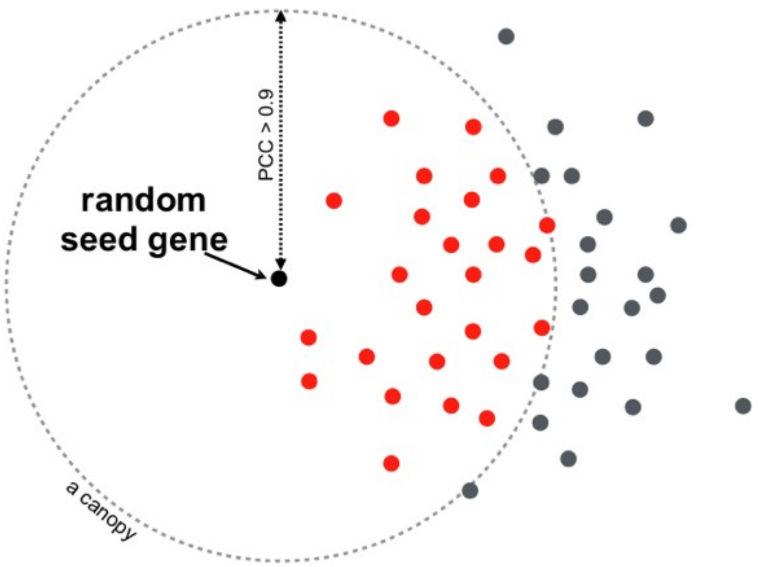
- It takes a long time to do it (LOTS of correlations to calculate).
- It's better with many samples with different abundances.
- You need to have a minimum level of depth.

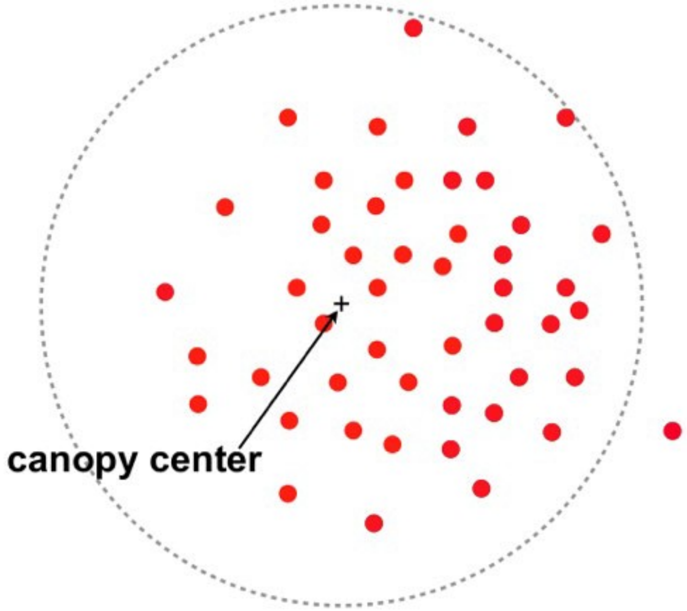
  Can you think of what happens to the correlation if most contigs have 5-10 reads mapping to them?
- You assume that each genome is present in many of your samples.
- Sensitive to having too many contigs to map against
  (random hits, reads attracted to the contigs they are part of)
  (also, let me repeat: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

Example of co-abundance-binning: Canopy

Algorithm:

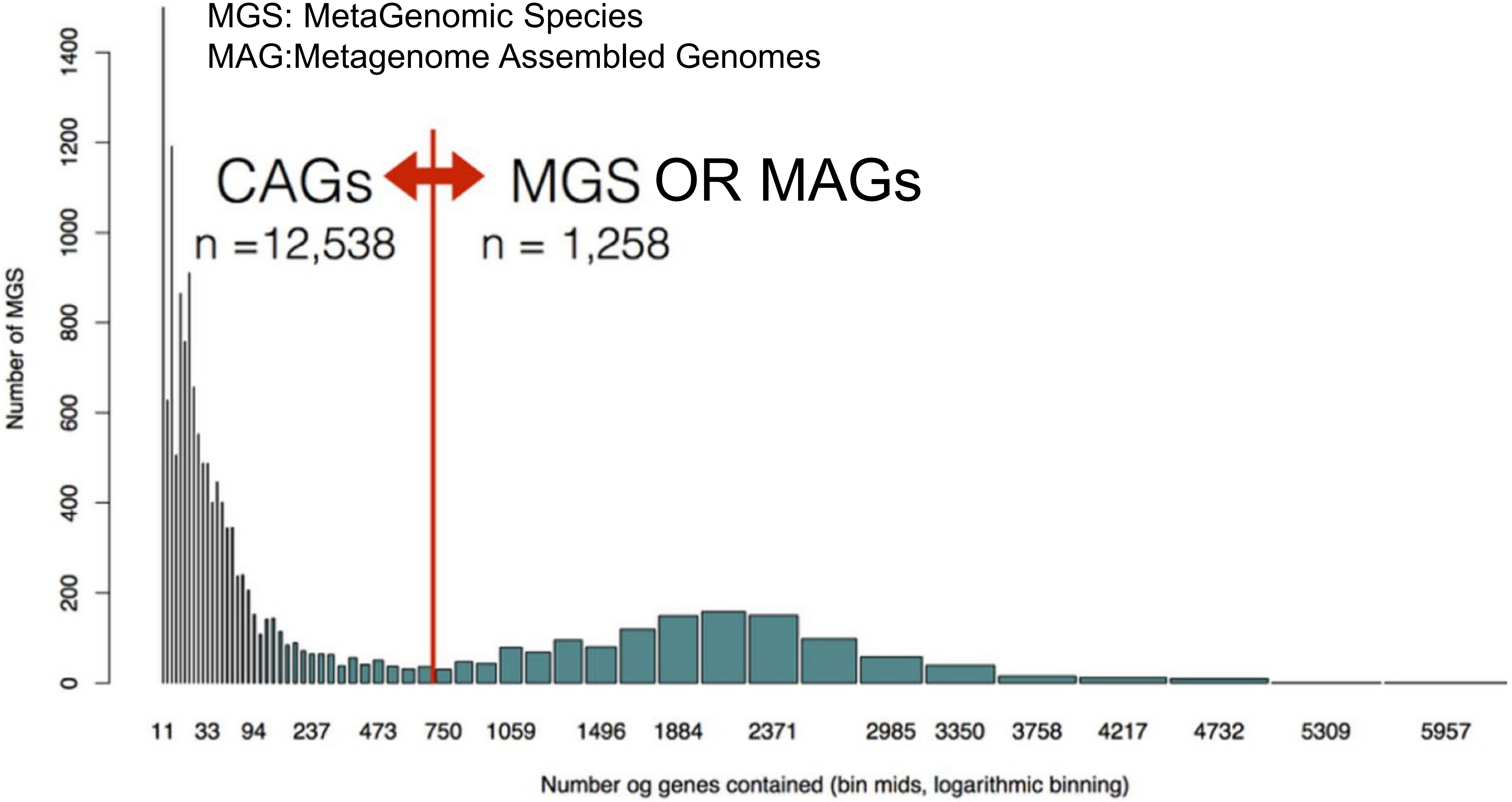
1) Pick random seed contig
2) Pick all contigs with Pearson correlation > 0.9
3) Select centre of cluster
4) Repeat 2 and 3 until centre is stable
5) Continue until all contigs have been assigned to a cluster





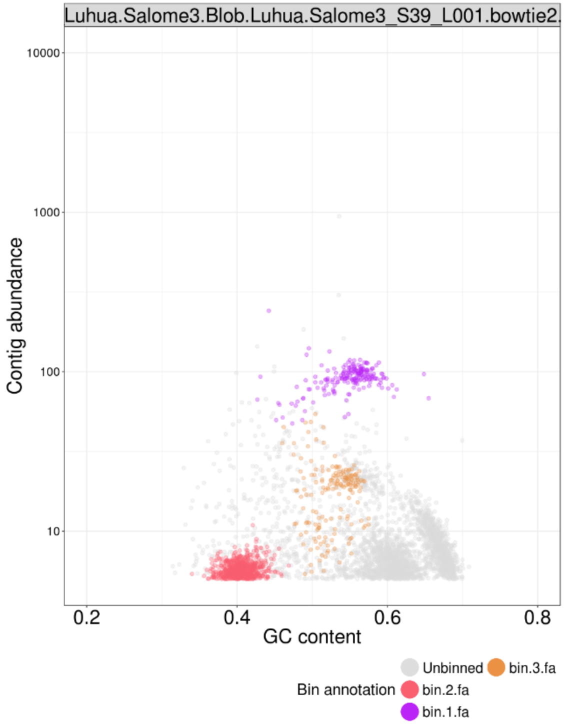
Not just microbial genomes gets binned….

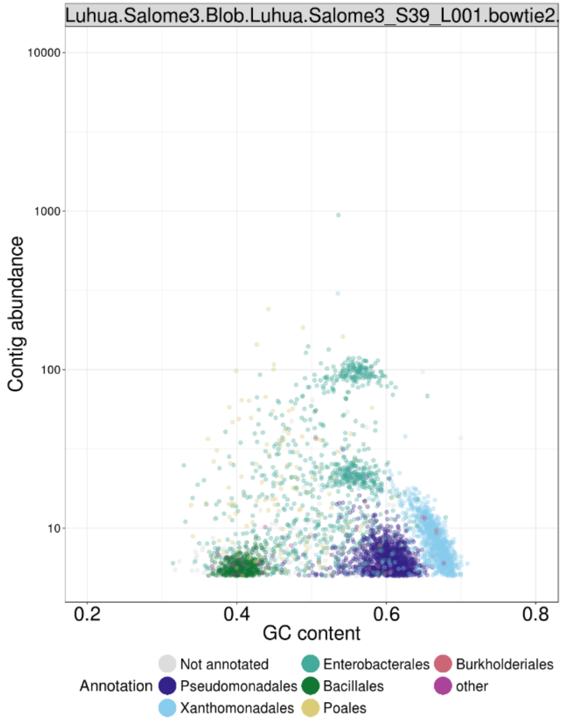
CAG: Cluster of Associated Genes
MGS: MetaGenomic Species
MAG:Metagenome Assembled Genomes



Assesing bin quality - Blobology

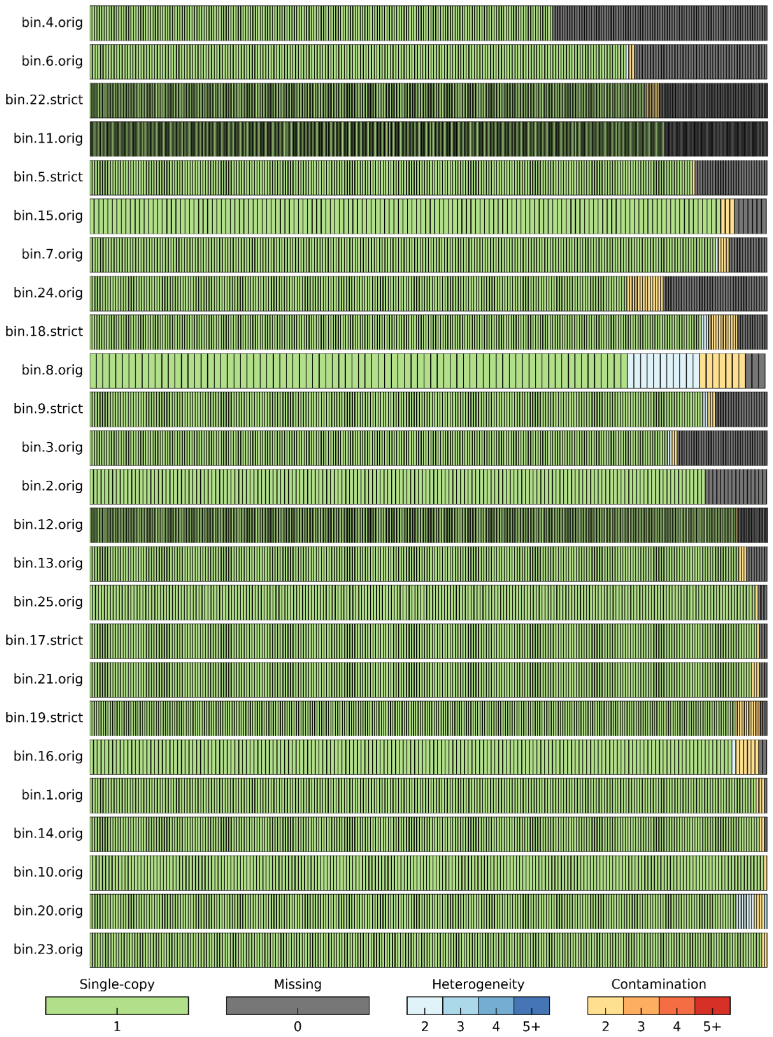
- Blobology plot shows contig abundance vs taxonomy or bin





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



Benchmarking reconstructed genomes

- Which taxonomic level is the right level to bin at?
- Are plasmids, prophages etc. considered a necessary part of a bin?
- Do we allow sequences to be in multiple bins? If so, a binner can cheat!
- What postprocessing can you assume people are doing after the binning

As far as I can tell, literally no-one has a good answer to this. Best we have is the work of Sczyrba, 2017 and Bowers et al., 2017

What programs are available for binning?

Gene-based:

Canopy (2014)

MSPminer (2019)

Contig-based:

MetaBAT2 (2019)

Vamb (2019) ← Made here at DTU

.. and several more

- Gene-based binners are very good at identifying operons and other functional units
- Contig-based binners are better at reconstructing genomes

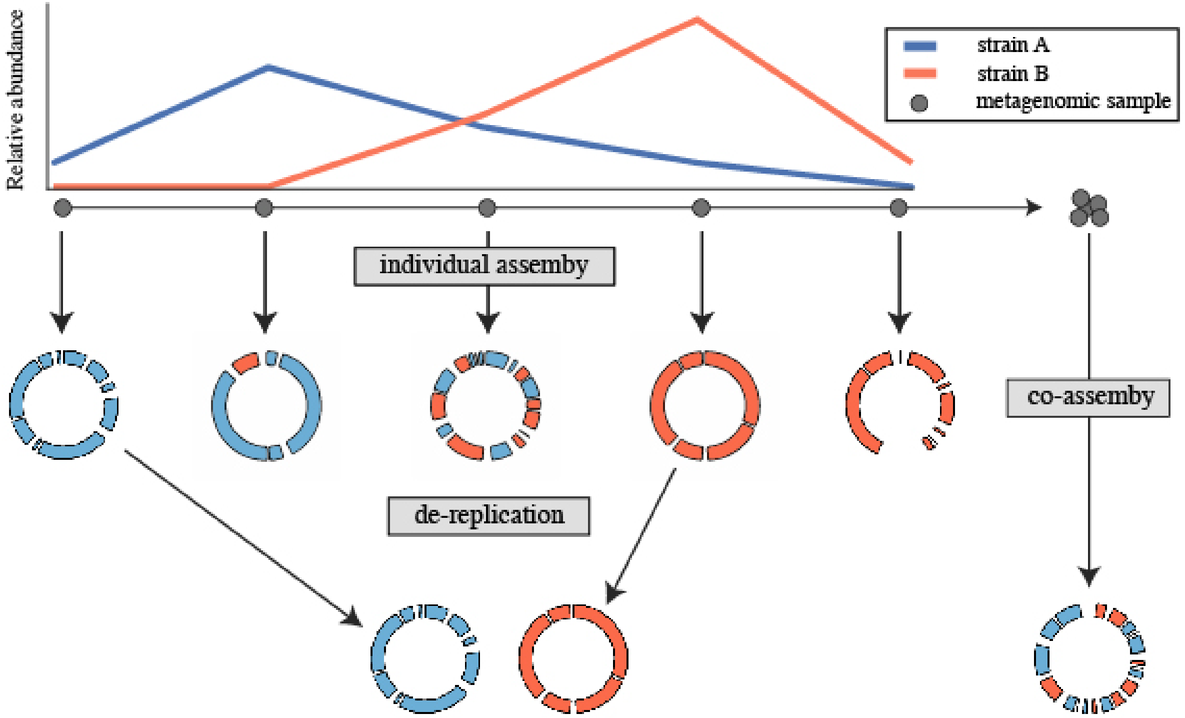
MetaBAT2 is probably the best one (except for Vamb, of course!). So you’re going to use MetaBAT in the exercises.

Dereplication

- Co-assembly is not feasible in most cases
  - Repetitive regions causes fragmented assemblies
  - Running out of memory
- Independent assembly
  - Identical or similar genomes are now redundantly present
  - Dereplication means identifying similar genomes from a larger set and picking the best ones
- VAMB solves this by separating each bin per sample

dRep

- All-vs-All alignments are time consuming so we do?
- Seed and extend!
  - Fast algorithm: Mash
    - Finds similar bins
  - Precise algorithm: ANIm
    - Robust to genomes incompleteness and accurate
  - Pick the best bins

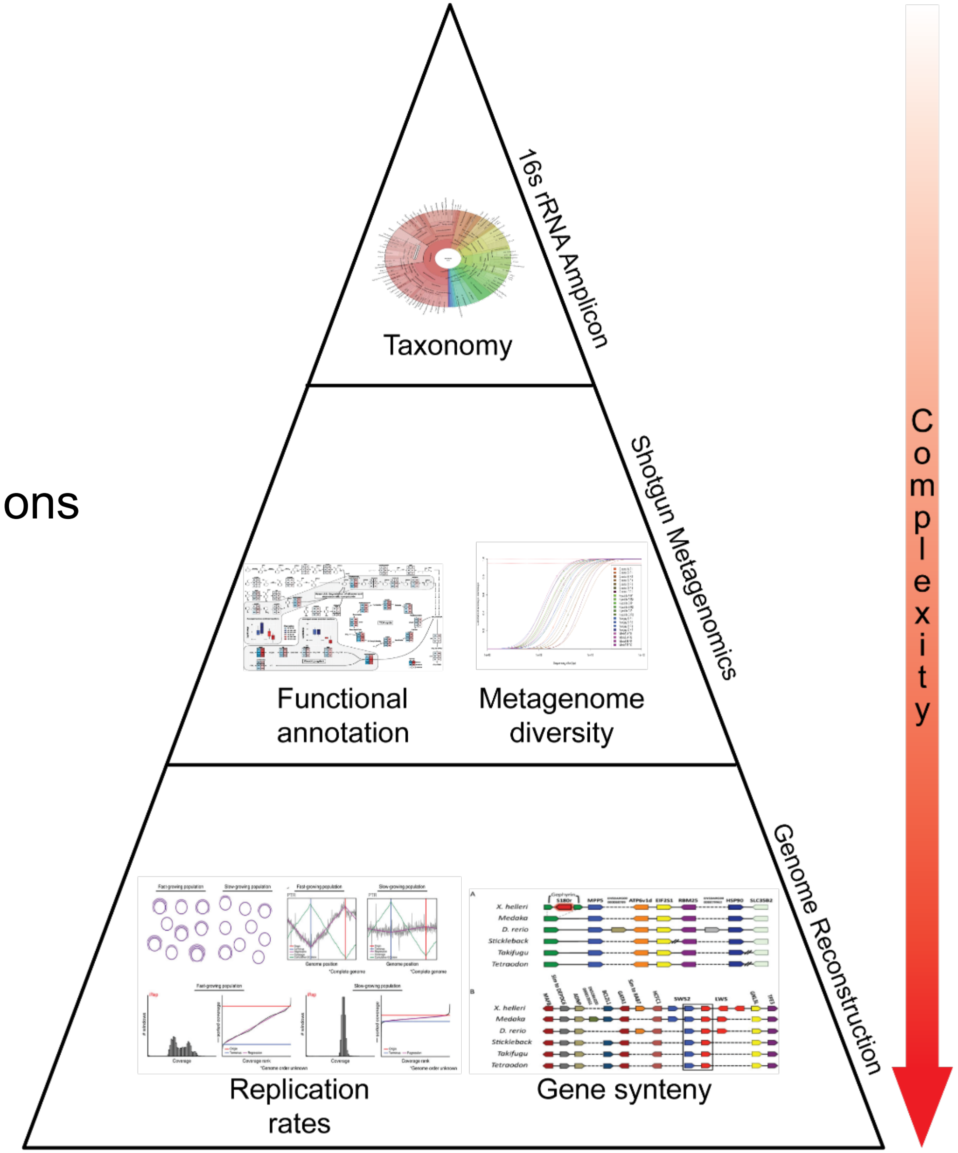


Abundance of reconstructed genomes

- Map all reads to all contigs or genes
  - – Does not work very well. Guess why?
- Gene-based binning can use signature genes
  - – Map reads to the 100 genes with highest bin correlation

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



16s rRNA Amplicon

Taxonomy

Shotgun Metagenomics

Functional annotation

Metagenome diversity

Genome Reconstruction

Replication rates

Gene synteny

Complexity

Summary

- Binning is a way of separating sequences(often contigs or genes) into genomes
- Adds additional context connecting genetic content and synteny with taxonomy
- Different ways of binning such as:
  - Alignment
  - Abundance
  - Composition
- Some methods are good for reconstructing genomes while others can pick up smaller elements such as plasmids, viruses and operons
- Each have strenghts and weaknesses
- Dereplication removes bin redundancy from several independent assemblies
- Horses for courses!