DTU

DTU Health Technology
Bioinformatics

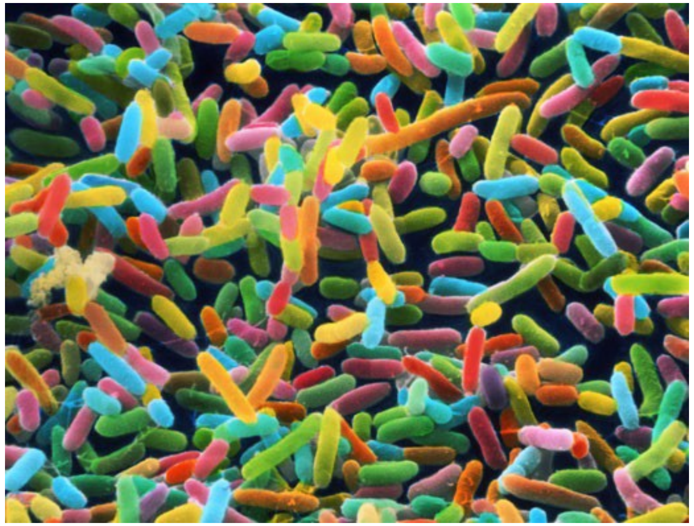
Metagenomic binning

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

Menu

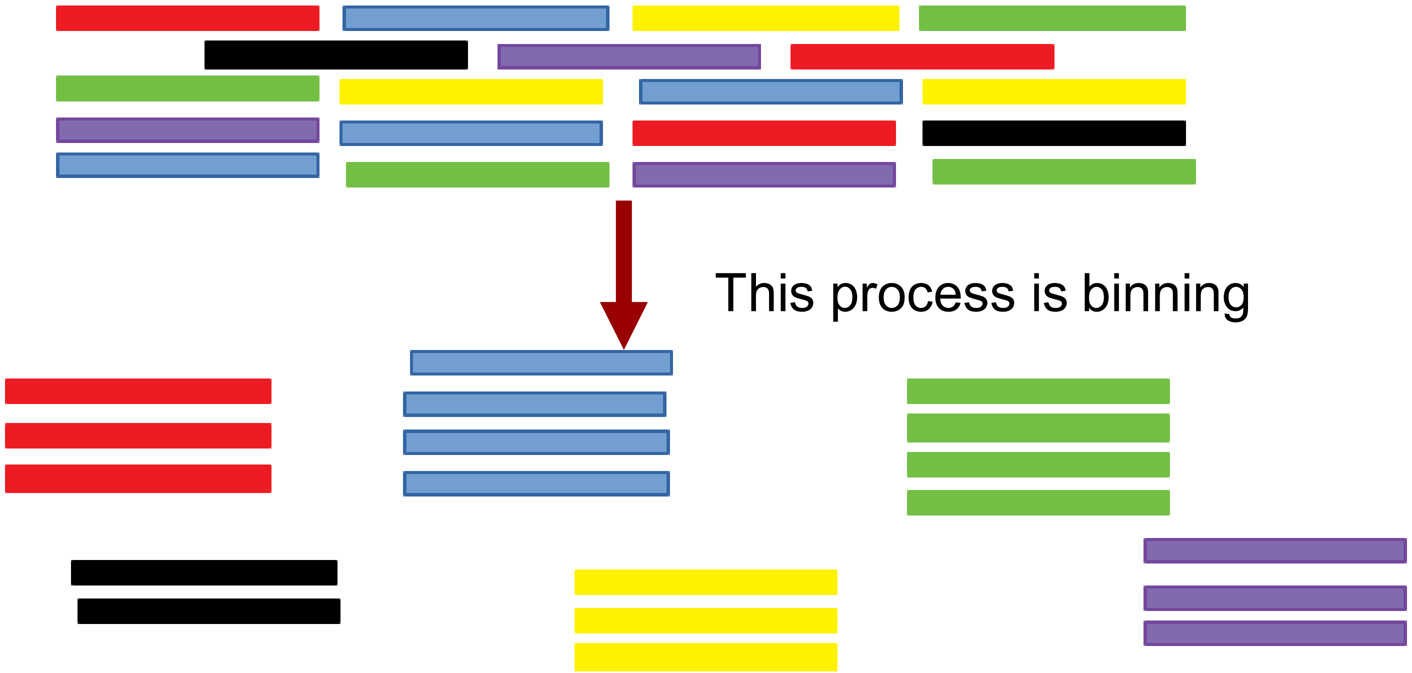
- What is binning?
- Types of metagenomic binners
- Assesing bin quality
- Dereplicating similar genomes

What is binning?



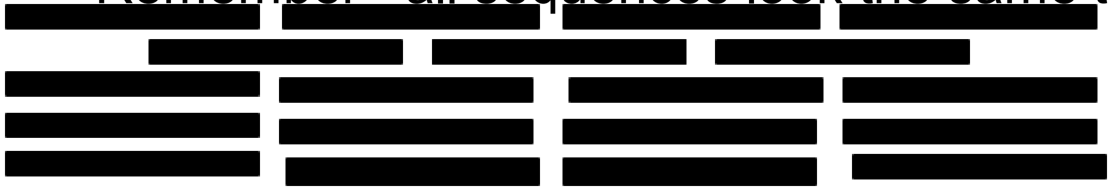
Metagenomic sample

Sequences (reads, contigs or scaffolds)



This process 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:

- 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?

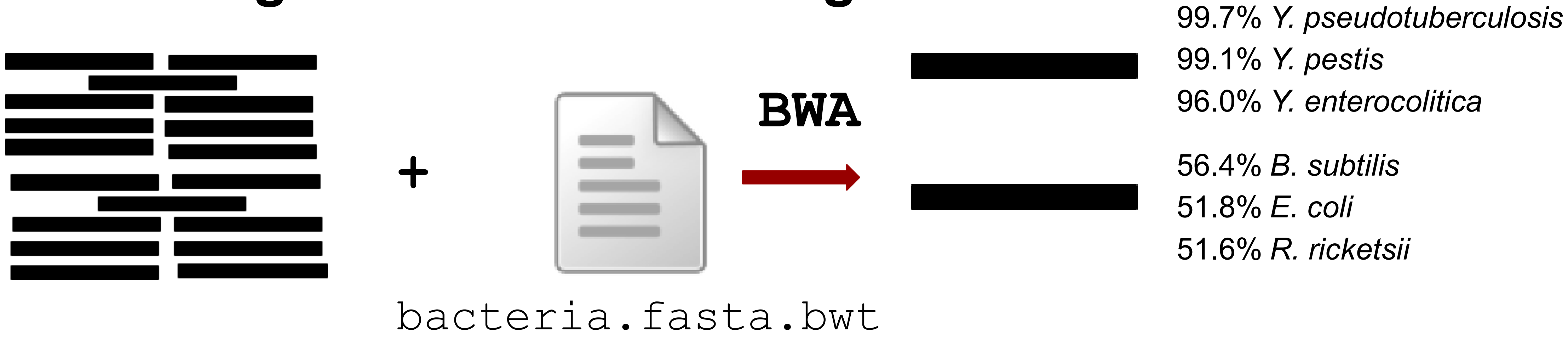
How can we do it?

- Alignment based

- Composition based

- Co-abundance based

Alignment based binning



You already know how alignment works.

Quite accurate technique, BUT:

- The current databases are *huge*, and keeps growing
- The large majority (~80% in some samples) of DNA is not in any DB
- You keep finding *E. coli* and *B. subtilis* – can you guess why?
- You cannot find *variations* from what is already known

Can of think of where this might be a good and bad technique?

Alignment based binning

To do it properly, you need good post-processing of the reads:
Lots of false positives and random hits.

These need to be filtered away by removing low-quality hits and stray hits, and by giving special attention to “unique” hits.
Obviously, we have software for this, e.g. mgmapper

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

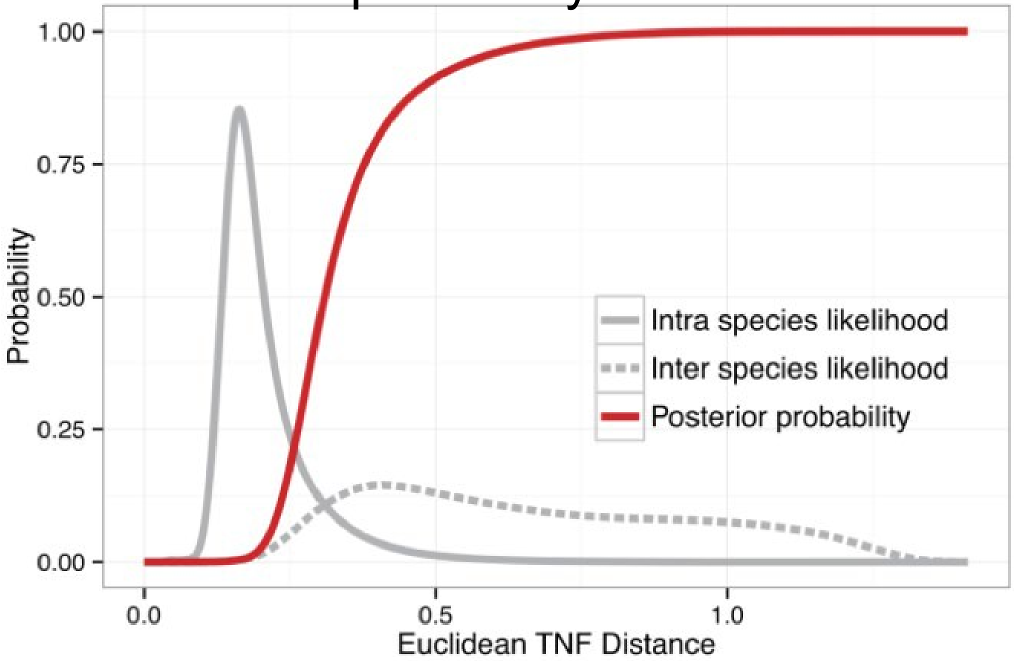
*People have looked into it. See: http://genome.cshlp.org/content/13/2/145.full, https://doi.org/10.1093/molbev/msp032

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?

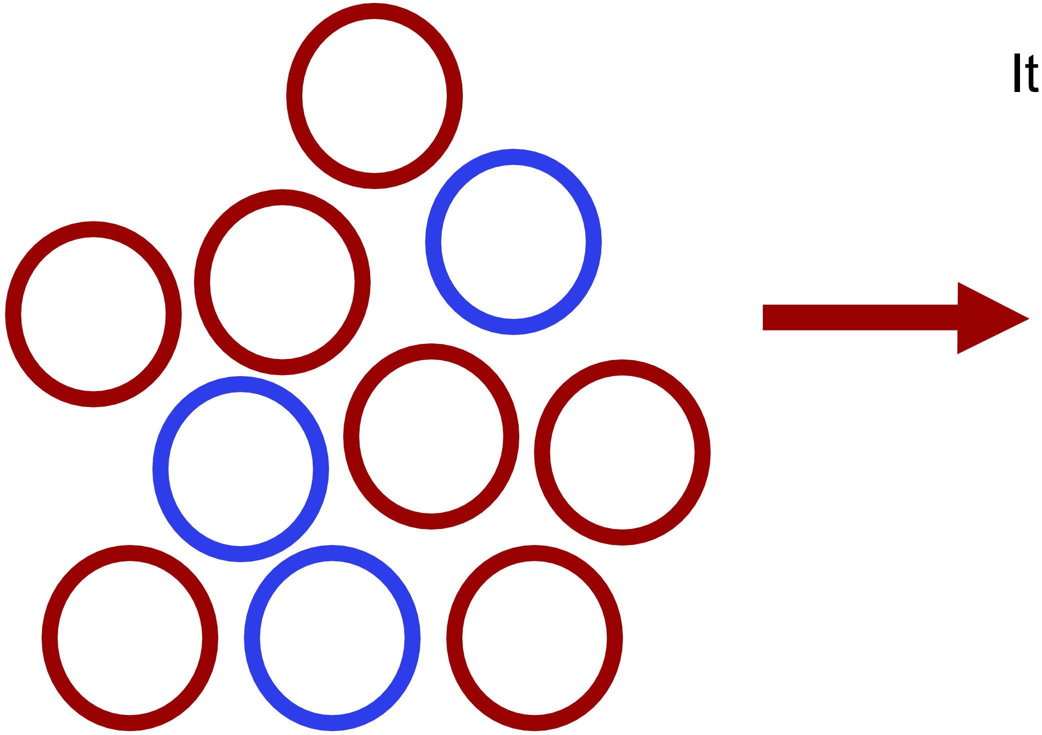
It works by comparing frequencies of kmers between contigs.
In short sequences, there are few kmers, frequencies are inaccurate.
It’s like trying to compare two parliament elections asking only 100 people!

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

Co-abundance based binning

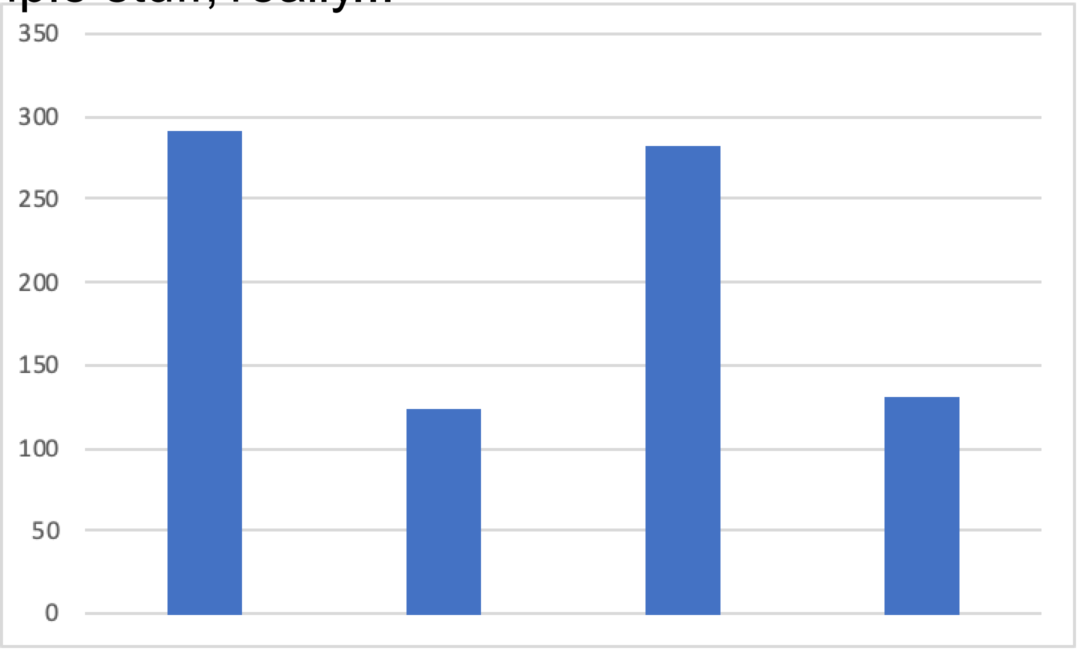
Principle: If read/contig A and B both come from the same genome/plasmid, then they should exists in approximately equal amount in all samples. Therefore, they should have similar depth.

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?

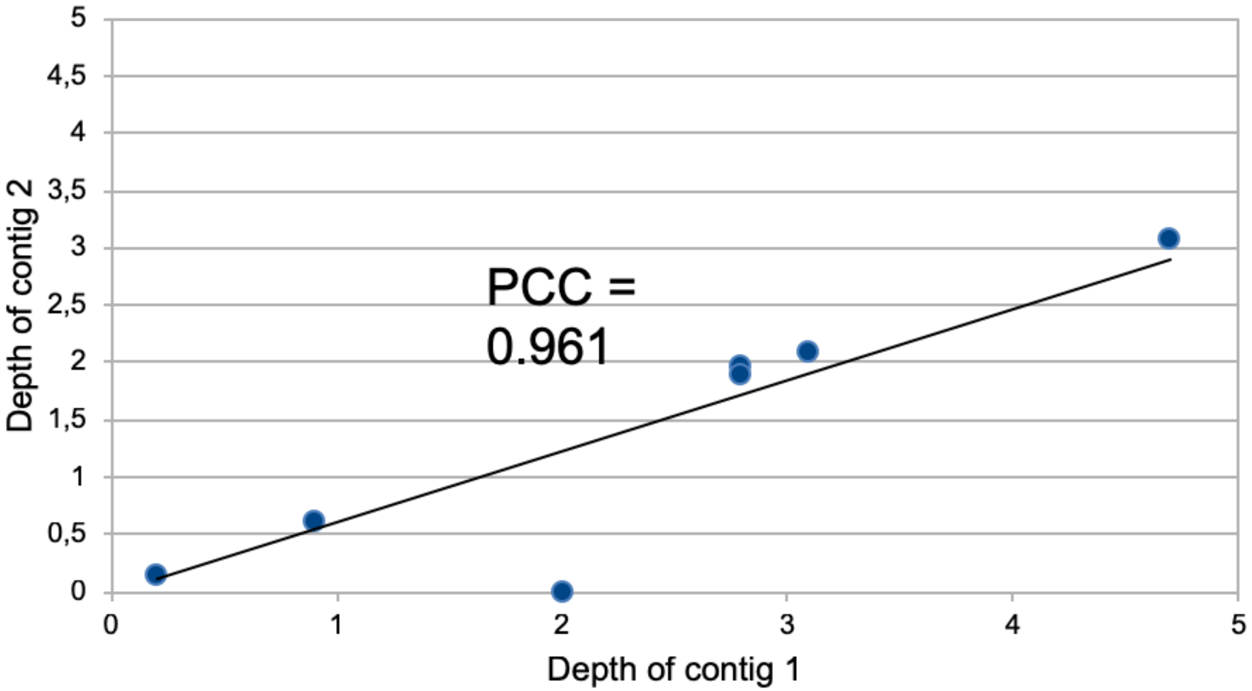
It's simple stuff, really...



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:

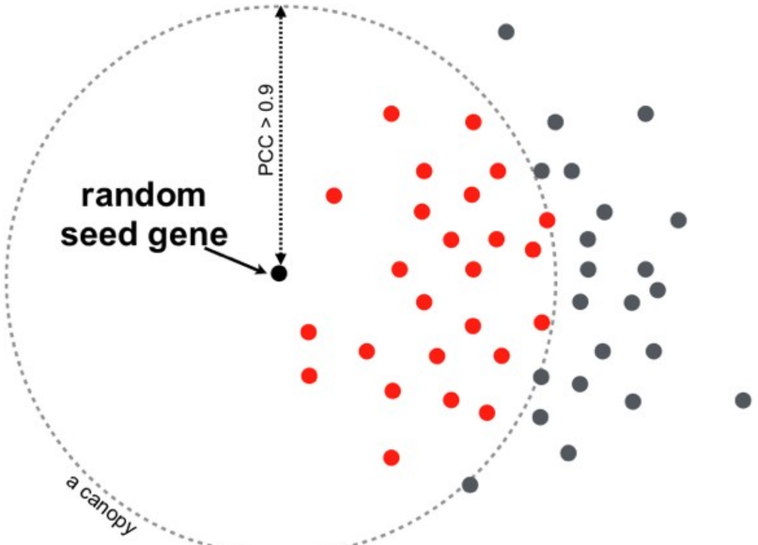
1) Pick random seed contig

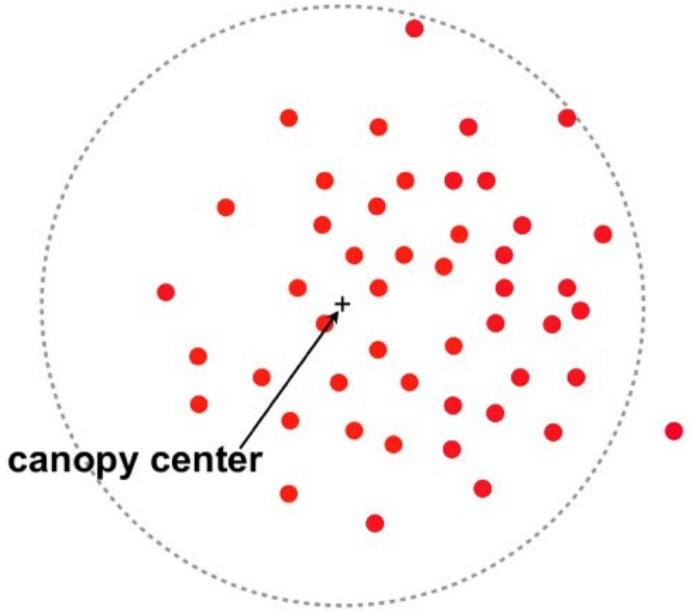
2) Pick all contigs with Pearson correlation > 0.9

3) Select centre of cluster

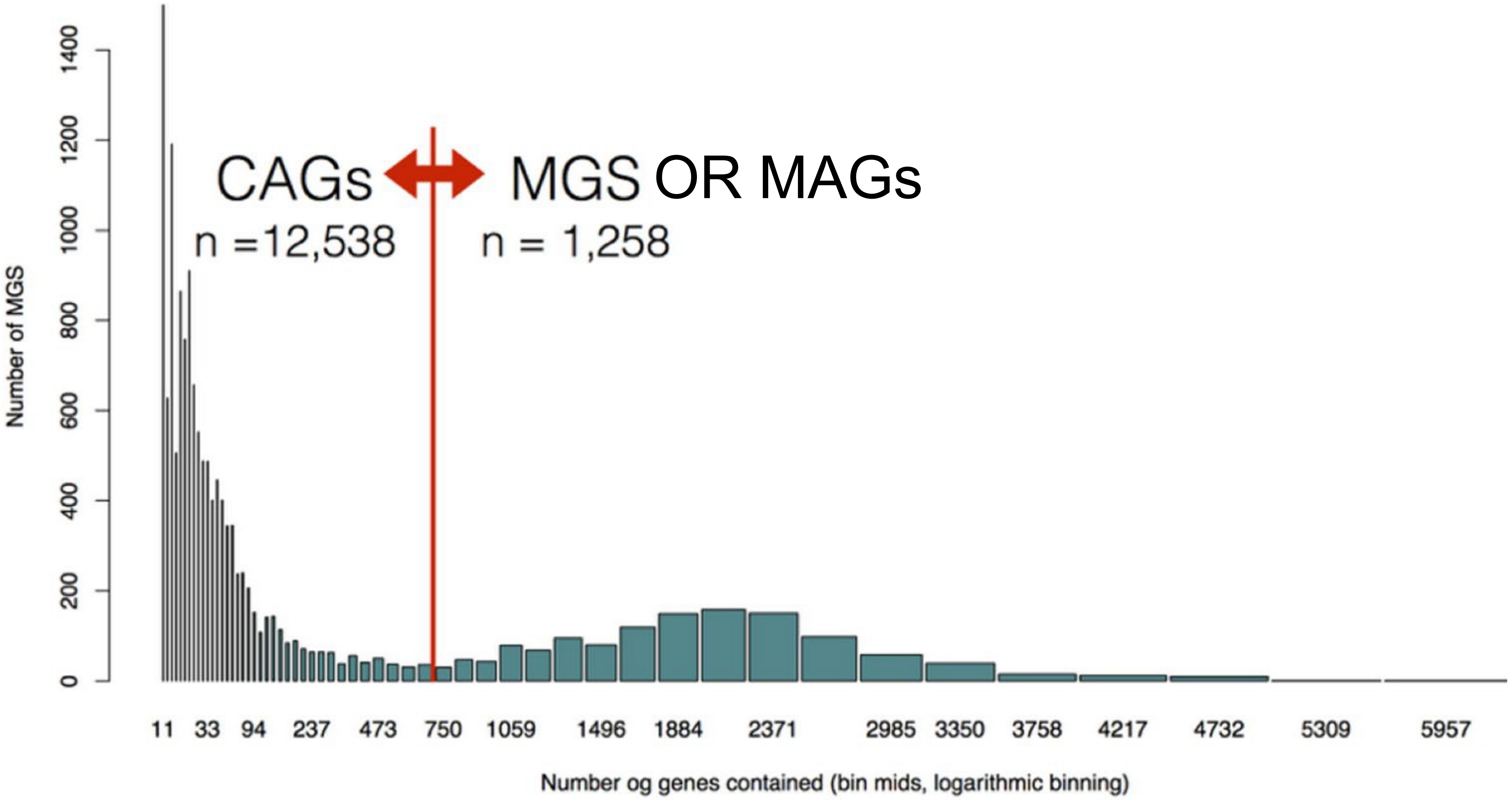
4) Repeat 2 and 3 until centre is stable

5) Continue until all contigs have been assigned to a cluster



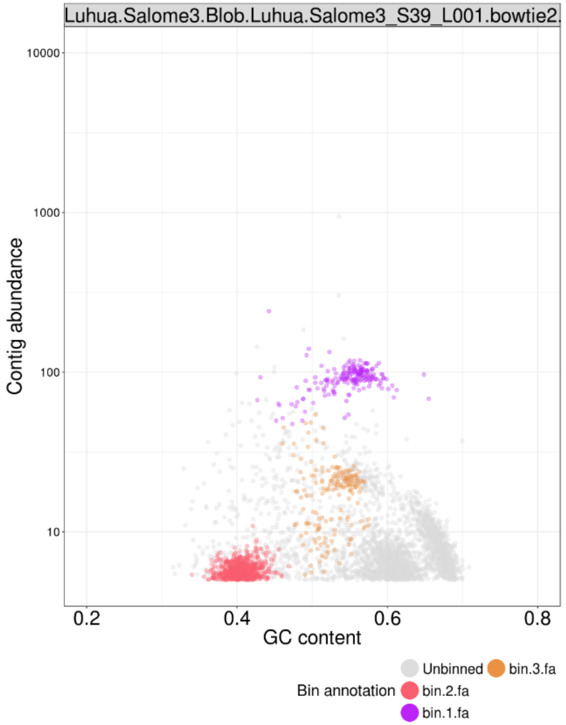


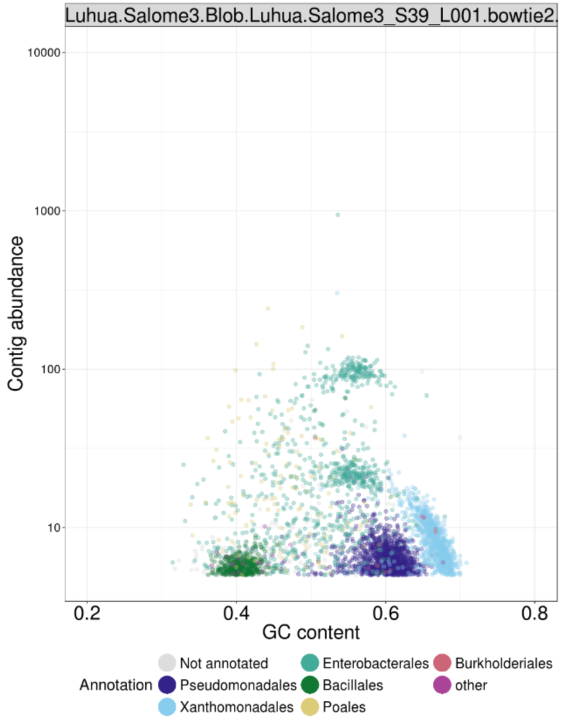
Not just microbial genomes gets binned….



Assesing bin quality - Blobology

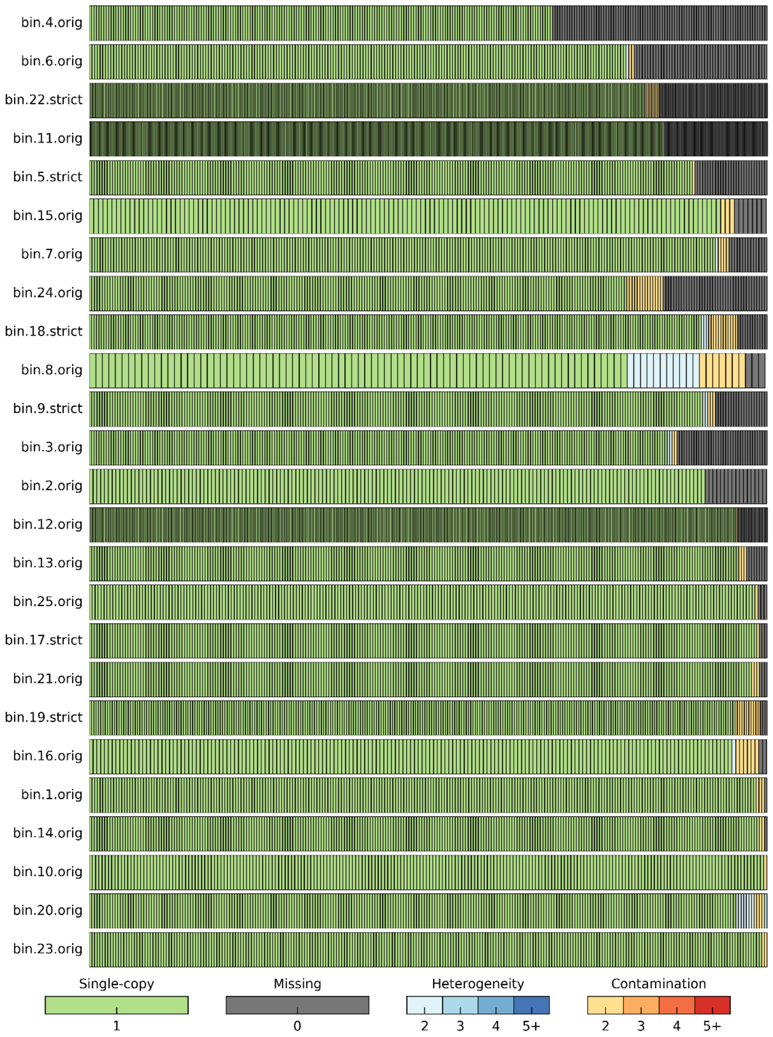
- Blobology plot shows contig abundance vs taxonomy or bin





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
- Contamination are marker genes found in multiple copies
- Heterogeneity is an indication if the contamination is from very different organisms or very similar organisms



Problem: What is a good binning?

- 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?

Canopy (2014)
MaxBin (2014)
VizBin (2014)
GroopM (2014)
MetaBAT (2015)
MyCC (2016)
Vamb (2019) ← Made here at DTU

.. and several more

There are no comprehensive benchmarks, which is a big problem. However, some people* are doing **great** work to change this.

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

Ensembl methods

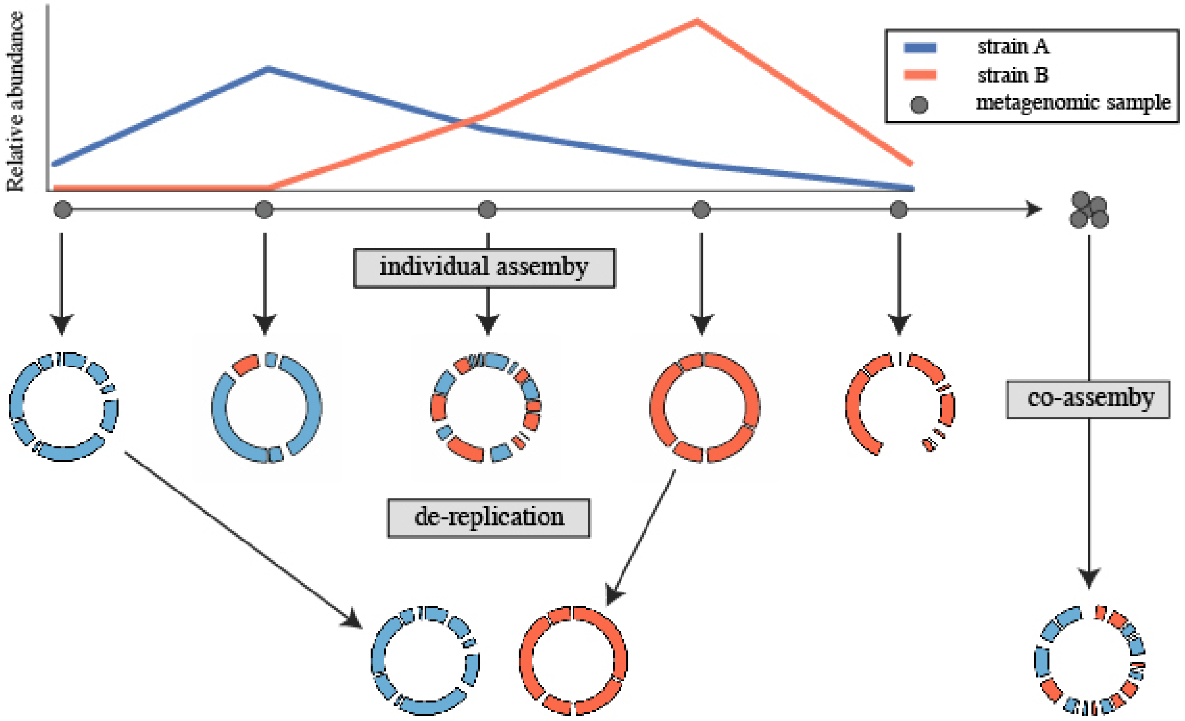
- DAS Tool
  - Ranks bins from multiple binners based on single copy genes
- MetaWRAP
  - Generates hybrid bin sets from each possible combination of binners
  - Each original and combination set is compared and best is chosen
  - Dereplication is done sp each contig only is part of one bin
  - Re-assembly of each individual bin

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



Summary

- Binning is a way of separating sequences(often contigs) into genomes
- Adds additional context connecting genetic content and synteny with taxonomy
- Various ways of doing it
  - Alignment
  - Abundance
  - Composition
- Each have strenghts and weaknesses
- Ensemble methods combines methods
- Dereplication removes redundancy