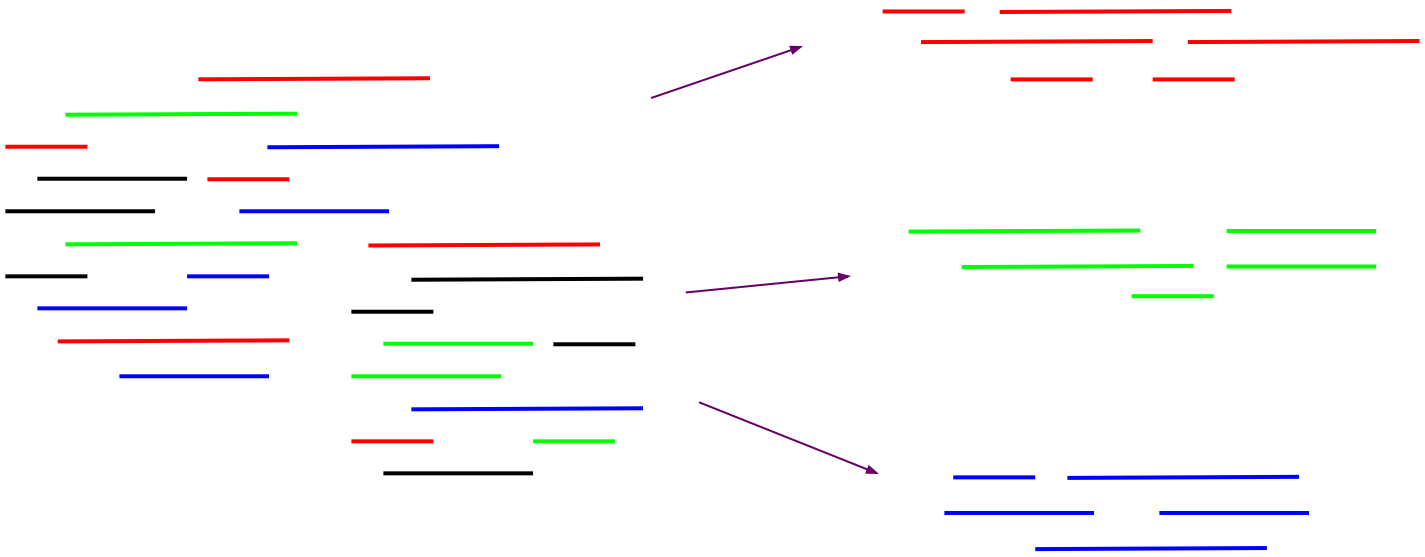
Bioinformatics workflow overview

- Host read elimination
- Read classification
- Metagenome assembly
- **Contig binning**
- Bin QC (completeness)
- Bin classification

Metagenome binning

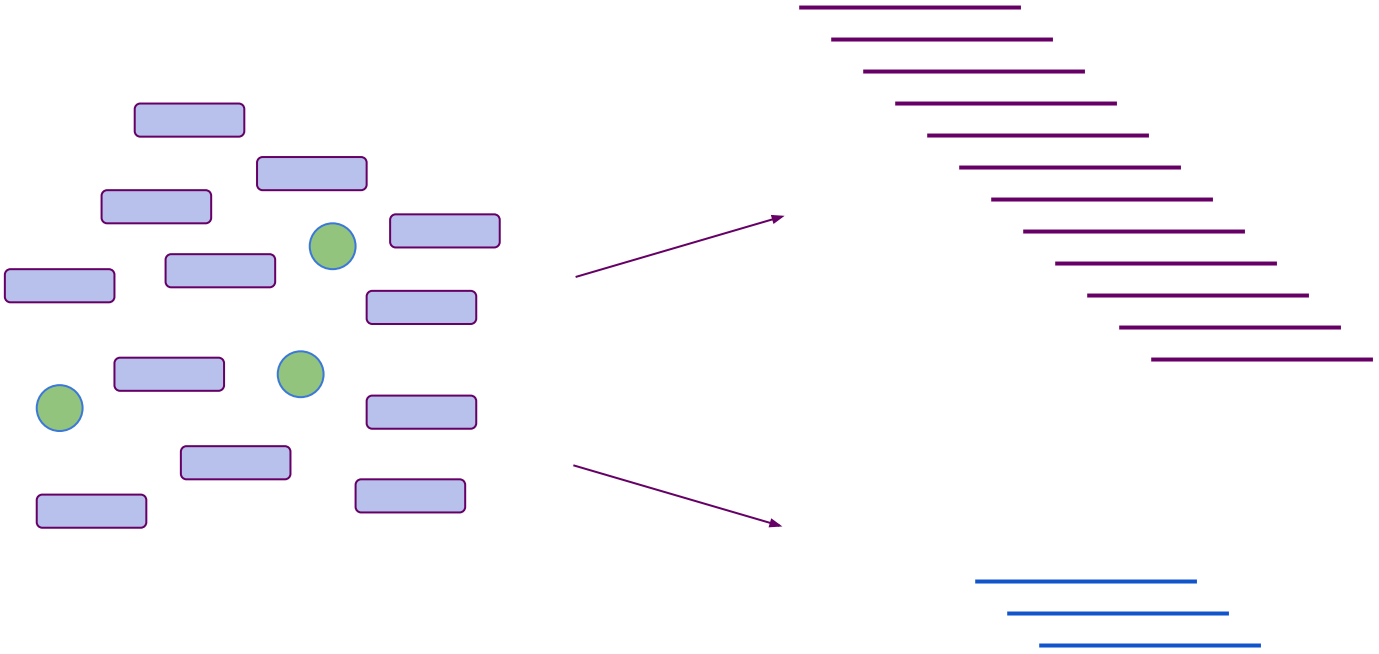
- Can we group contigs into separate genomes?



Put contigs into bins based on common characteristics

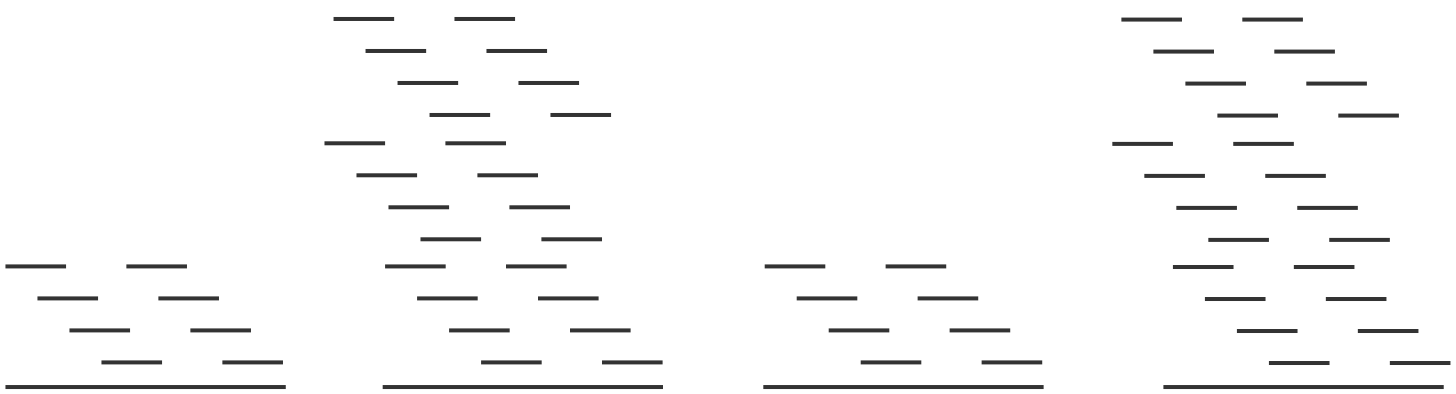
- Why bins and not reads or k-mers?
- What common traits might contigs share?

Abundance

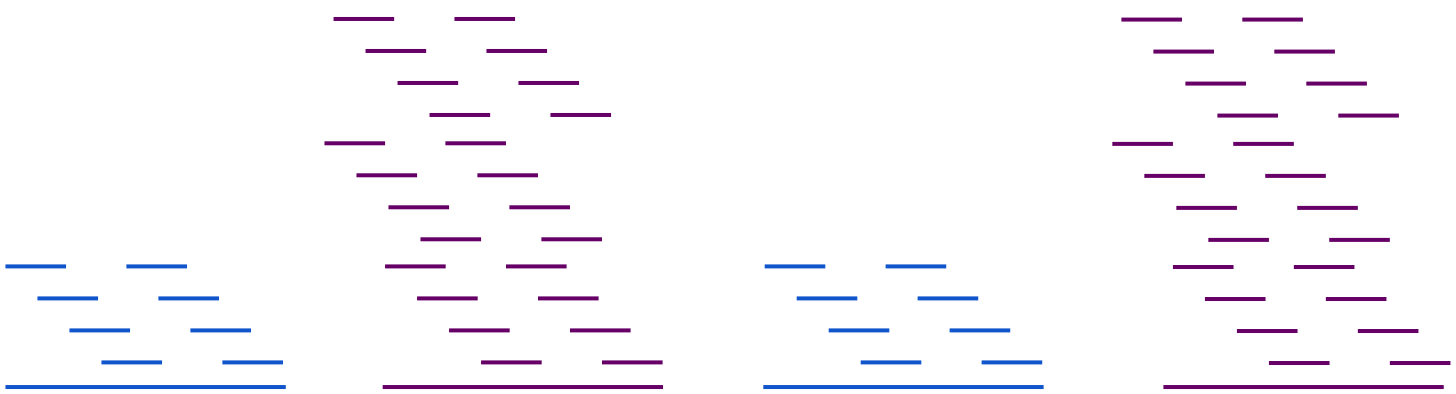


Abundance

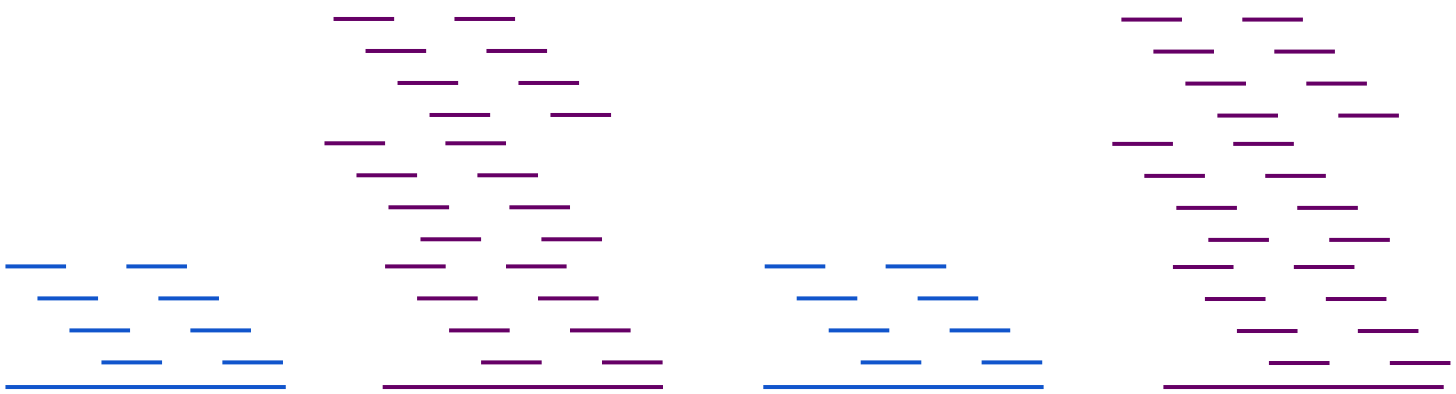
Abundance



Abundance



Abundance



- What problems do you see?

4-mers of a string

```
GTACCGTACCG
GTAC
 TACC
  TCCG
   CCGT
    CGTA
     GTAC
      TACC
       ACCG
```

4-mer frequencies

```
GTAC 2
TACC 2
TCCG 1
CCGT 1
CGTA 1
ACCG 1
```

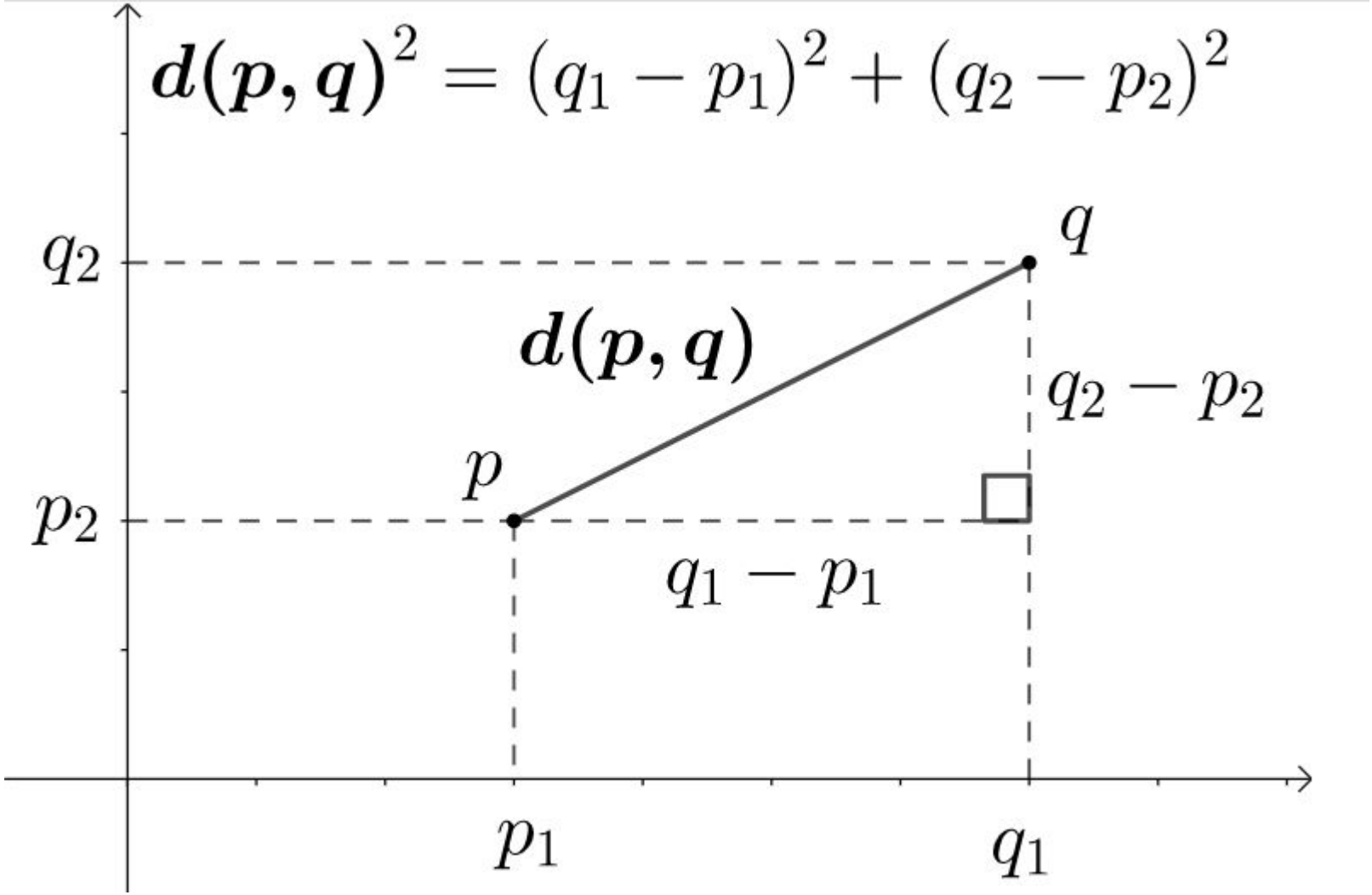
- How many possible 4-mers are there?

4-mer frequencies from two strings

	S1	S2
GTAC	2	2
TACC	2	1
TCCG	1	1
CCGT	1	2
CGTA	1	1
ACCG	1	1

- How could you make a quantitative comparison of S1 and S2?

Euclidean Distance



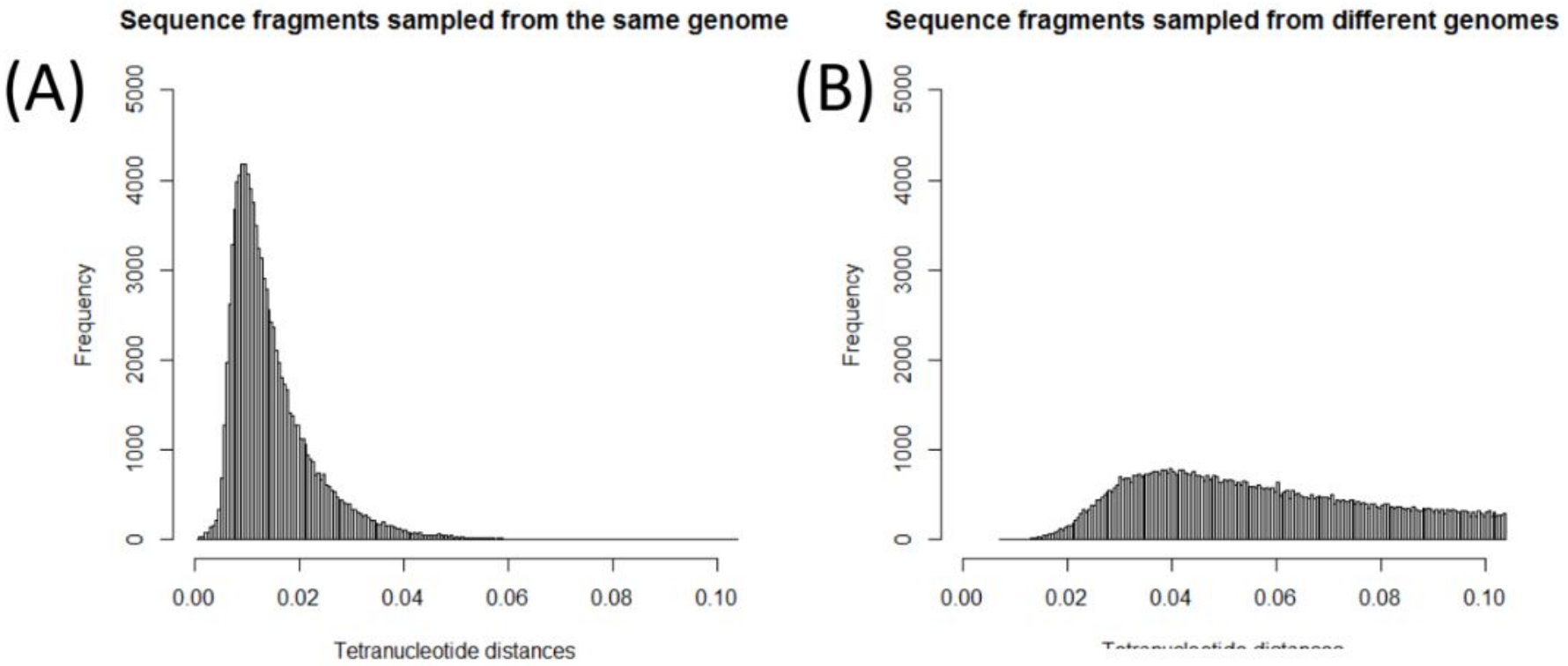
Euclidean Distance

	S1	S2
GTAC	2	2
TACC	2	1
TCCG	1	1
CCGT	1	2
CGTA	1	1
ACCG	1	1

p = (2,2,1,1,1,1)
q = (2,1,1,2,1,1)

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}$$

Euclidean Distance



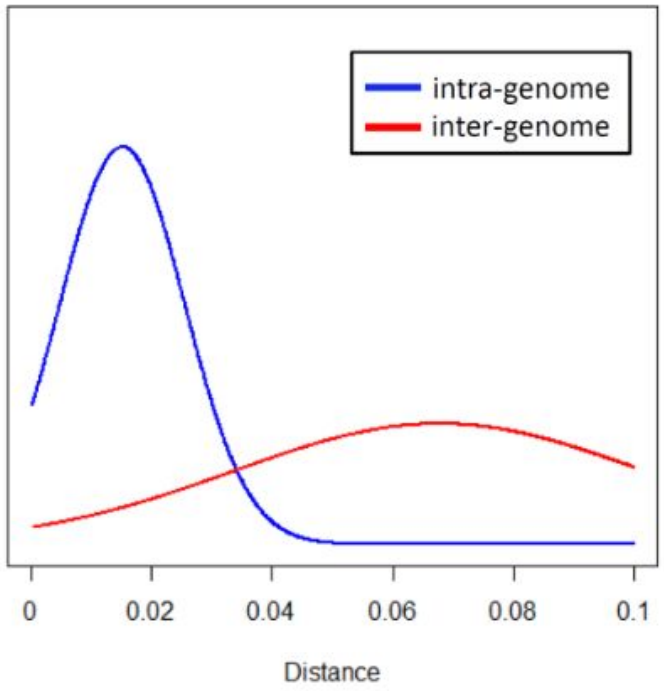
- 3,181 bacterial and archaeal genomes
- random sequences of 1kb - 1-mb
- 1,000,000 simulations

New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

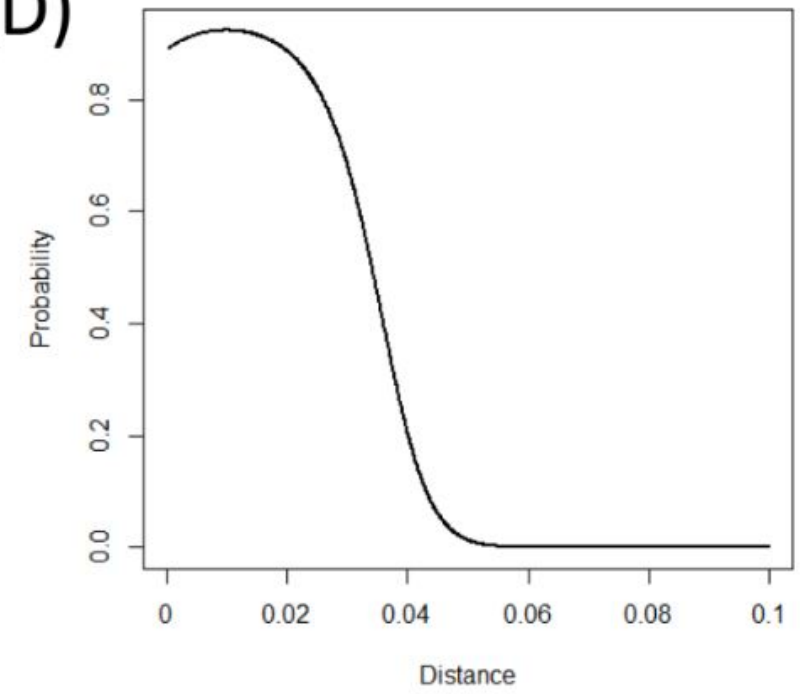
NCGR
National Center for Genome Resources

Euclidean Distance

(C)



(D)



maxBin

Methodology | Open Access | Published: 01 August 2014

MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm

Yu-Wei Wu ✉, Yung-Hsu Tang, Susannah G Tringe, Blake A Simmons & Steven W Singer

Microbiome **2**, Article number: 26 (2014) | Cite this article

https://doi.org/10.1186/2049-2618-2-26

maxBin

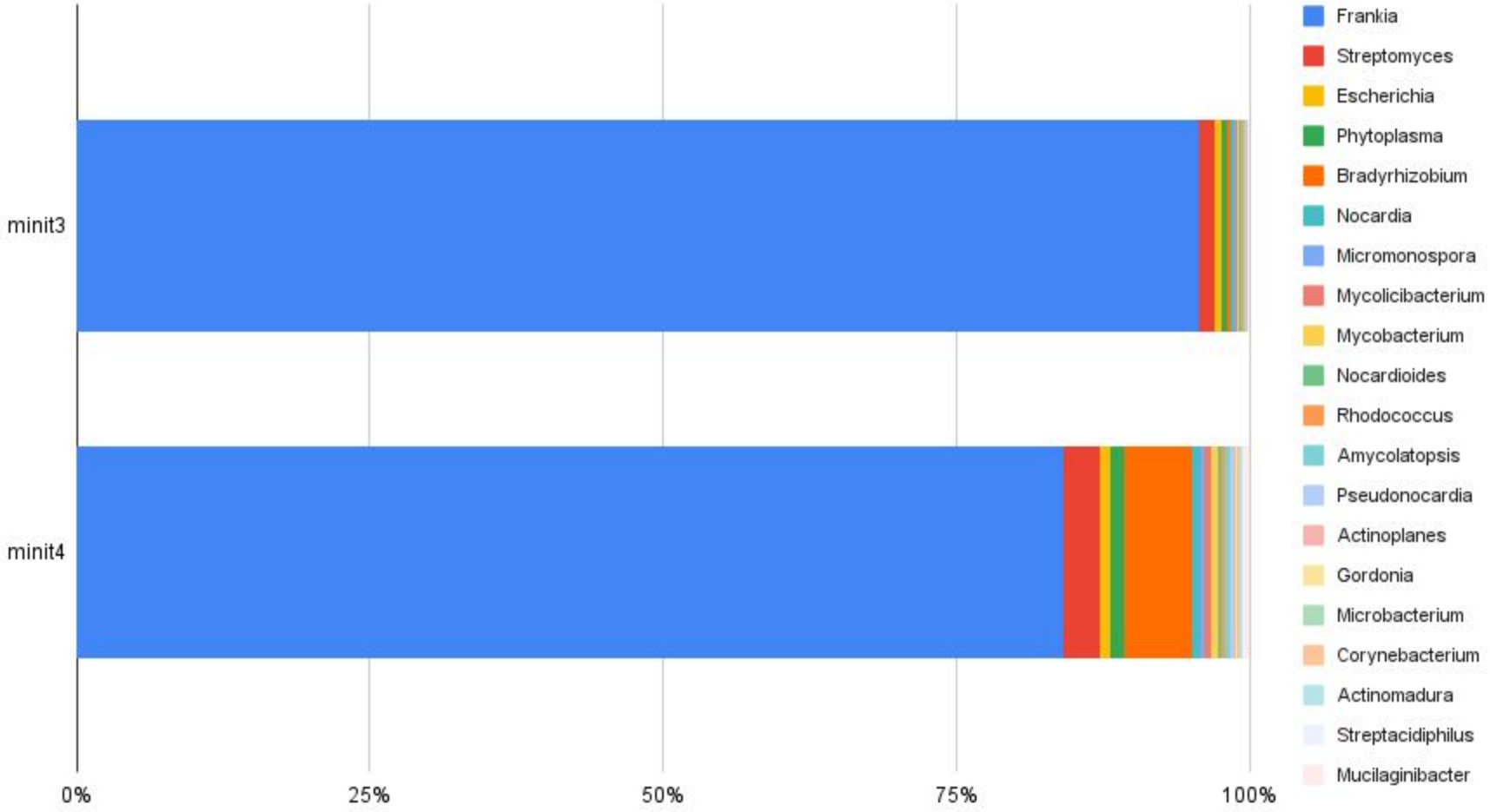
- Computes probability that two contigs are from the same genome based on 4-mer frequencies
- Computes probability that two contigs are from the same genome based on coverage depth
- Combines these probabilities using an Expectation Maximization algorithm
- Assigns contigs to a bin if $P > 0.8$
- Attempts to estimate bin completeness based on 107 marker genes

Running maxBin

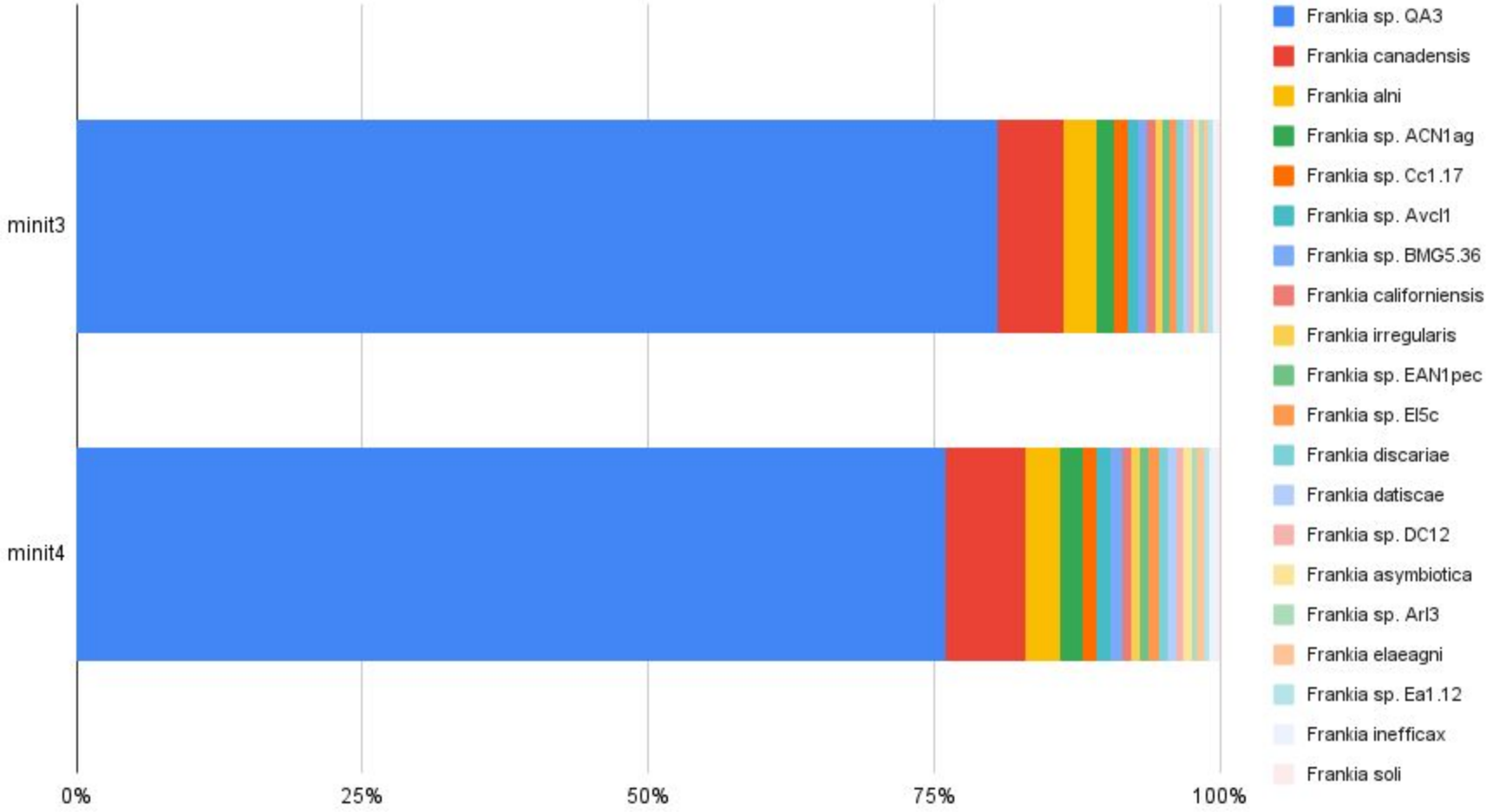
- We'll used pooled minit1 and minit2 data (why?)
- Look at /home/cjb/copy/maxbin
- Two people can run it simultaneously

```
run_MaxBin.pl \
-thread 64 \
-contig /path/to/assembled/contigs.fasta \
-reads /path/to/minit3/reads/something.fasta \
-reads2 /path/to/minit4/reads/something.fasta \
-out maxbin_bins
```

Genus proportions



Frankia strain proportions



Running maxBin

- Took 8 hours, 44 minutes
- I have precomputed data
- /home/cjb/analysis/070822/maxbin

Bioinformatics workflow overview

- Host read elimination
- Read classification
- Metagenome assembly
- Contig binning
- **Bin QC (completeness)**
- Bin classification

Bin QC (completeness)

- How complete is each genome bin?
- What percentage of each bacterial genome does it represent?

CHECKM

GENOME RESEARCH
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2015 Jul; 25(7): 1043–1055.
doi: 10.1101/gr.186072.114

PMCID: PMC4484387
PMID: 25977477

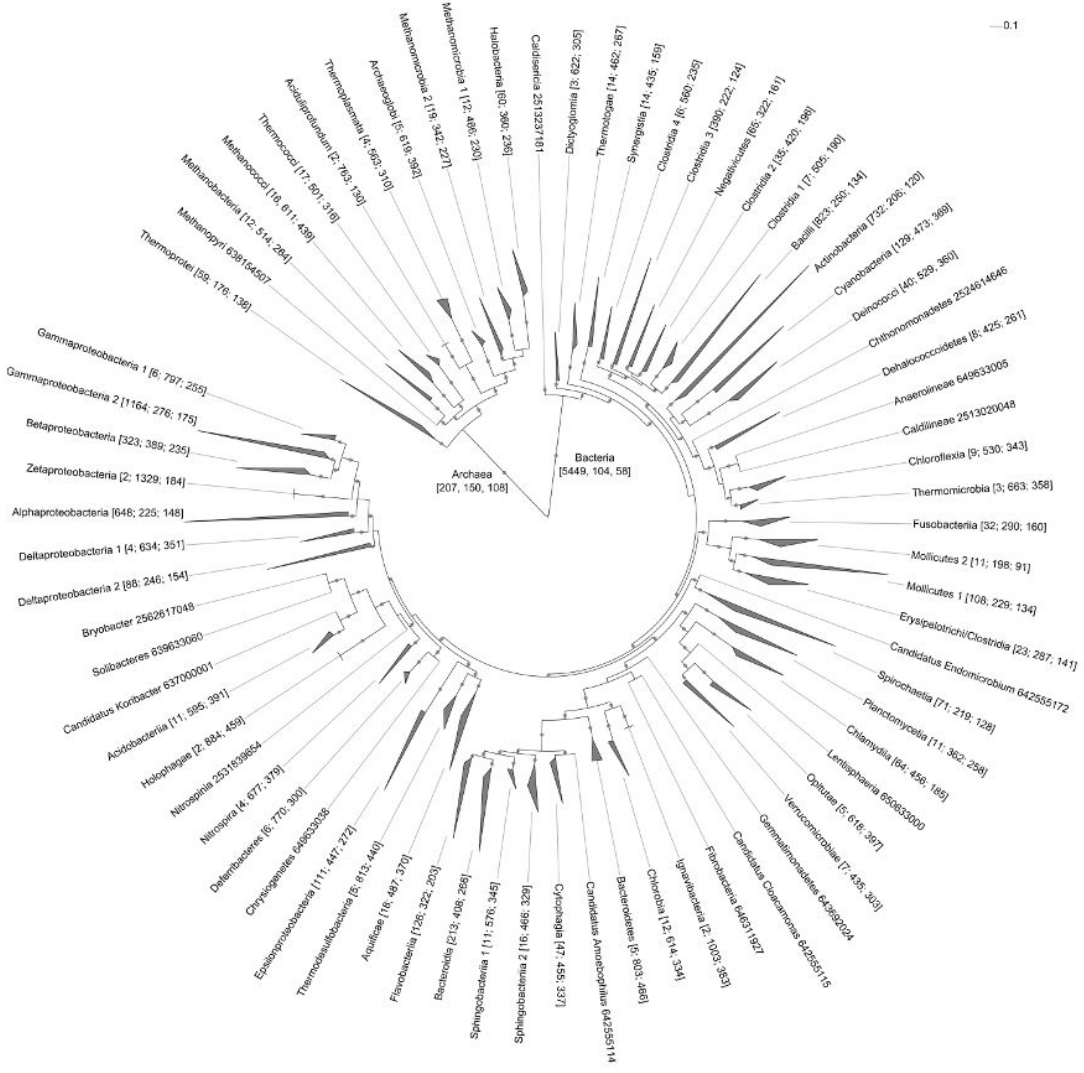
CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

Donovan H. Parks,[1] Michael Imelfort,[1] Connor T. Skennerton,[1] Philip Hugenholtz,[1,2] and Gene W. Tyson[1,3]

https://doi.org/10.1101%2Fgr.186072.114

CHECKM

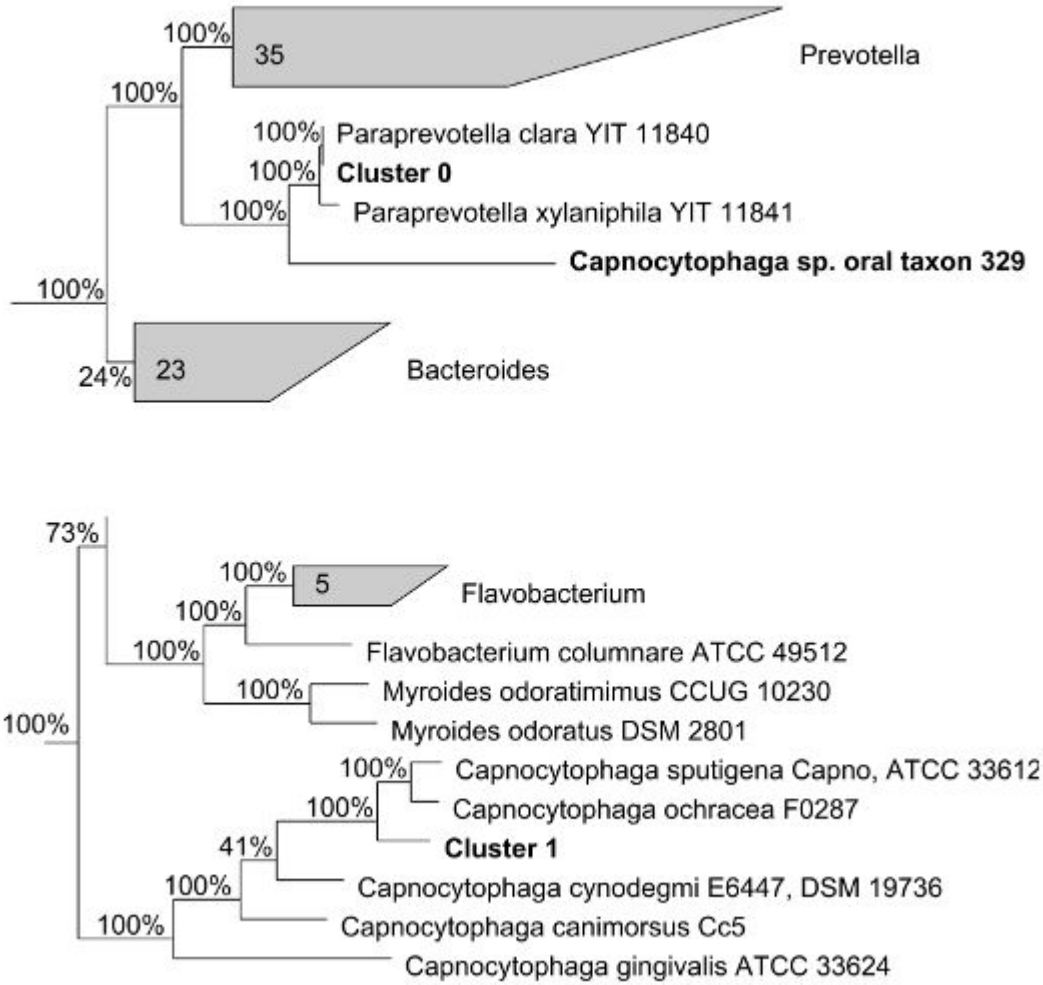
Phylogeny inferred from 43 single copy genes



New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

NCGR
Center for Genome Resources

CHECKM

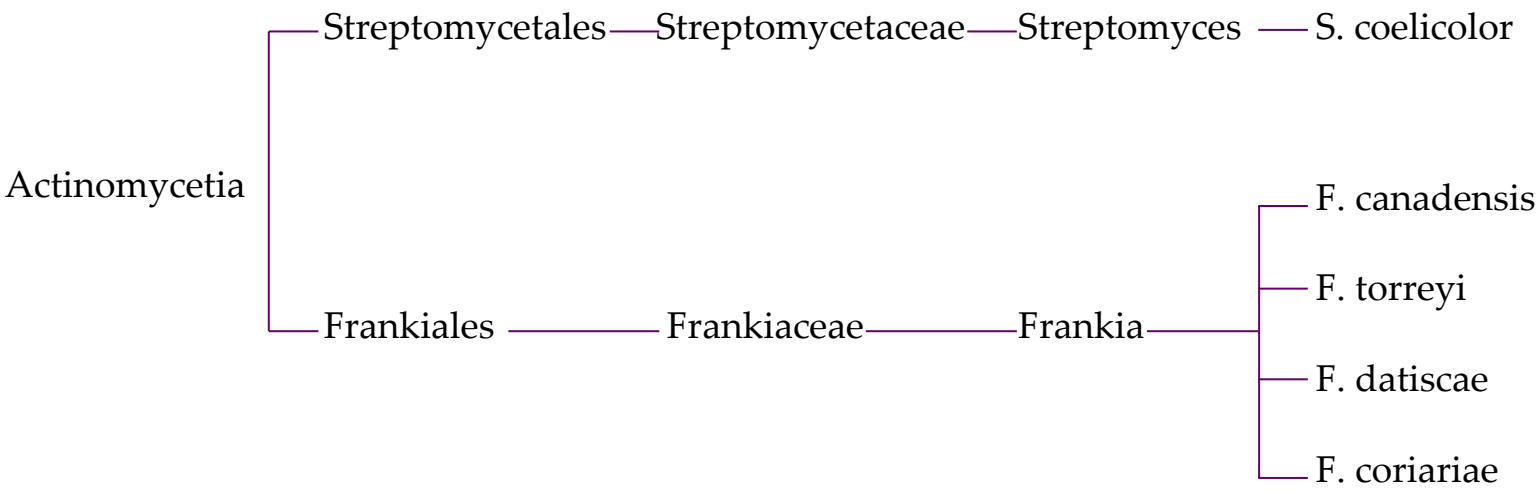


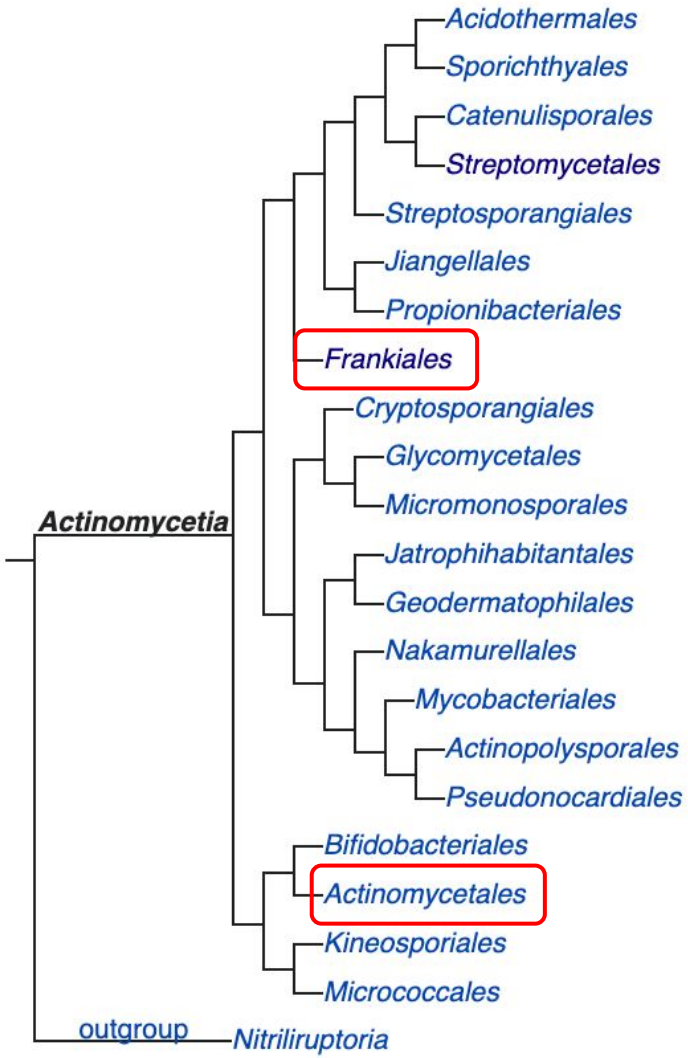
CHECKM

- Places input genome in the phylogenetic tree
- Identifies the appropriate set of marker genes
- Searches the input genome for those genes
- Evaluates percentage completeness

https://github.com/Ecogenomics/CheckM/wiki/Reported-Statistics

Example taxonomy





Checkm Frankia entries

IMG_2506783011
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Candidatus_Frankia;s__Candidatus_Frankia_datiscae

IMG_2522125134 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_2515154087 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_2517093007 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_649633045 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_2506381019 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_2509887025 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_637000116 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_641228492 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_2517572101 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_2508501039 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__

IMG_637000115 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Frankiaceae;g__Frankia;s__Frankia_alni

Running CHECKM

- /home/cjb/copy/checkm.txt

https://github.com/Ecogenomics/CheckM/wiki/Reported-Statistics

Bioinformatics workflow overview

- Host read elimination
- Read classification
- Metagenome assembly
- Contig binning
- Bin QC (completeness)
- **Bin classification**

Bin classification

- MaxBin identified bins and grouped contigs
- Checkm evaluated the quality of each bin
- But what bacteria do the bins represent?

CAT/BAT

von Meijenfeldt *et al. Genome Biology* (2019) 20:217
https://doi.org/10.1186/s13059-019-1817-x

Genome Biology

METHOD Open Access

Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT

F. A. Bastiaan von Meijenfeldt[1†], Ksenia Arkhipova[1†], Diego D. Cambuy[1], Felipe H. Coutinho[2,3,4] and Bas E. Dutilh[1,2*]

https://doi.org/10.1186/s13059-019-1817-x

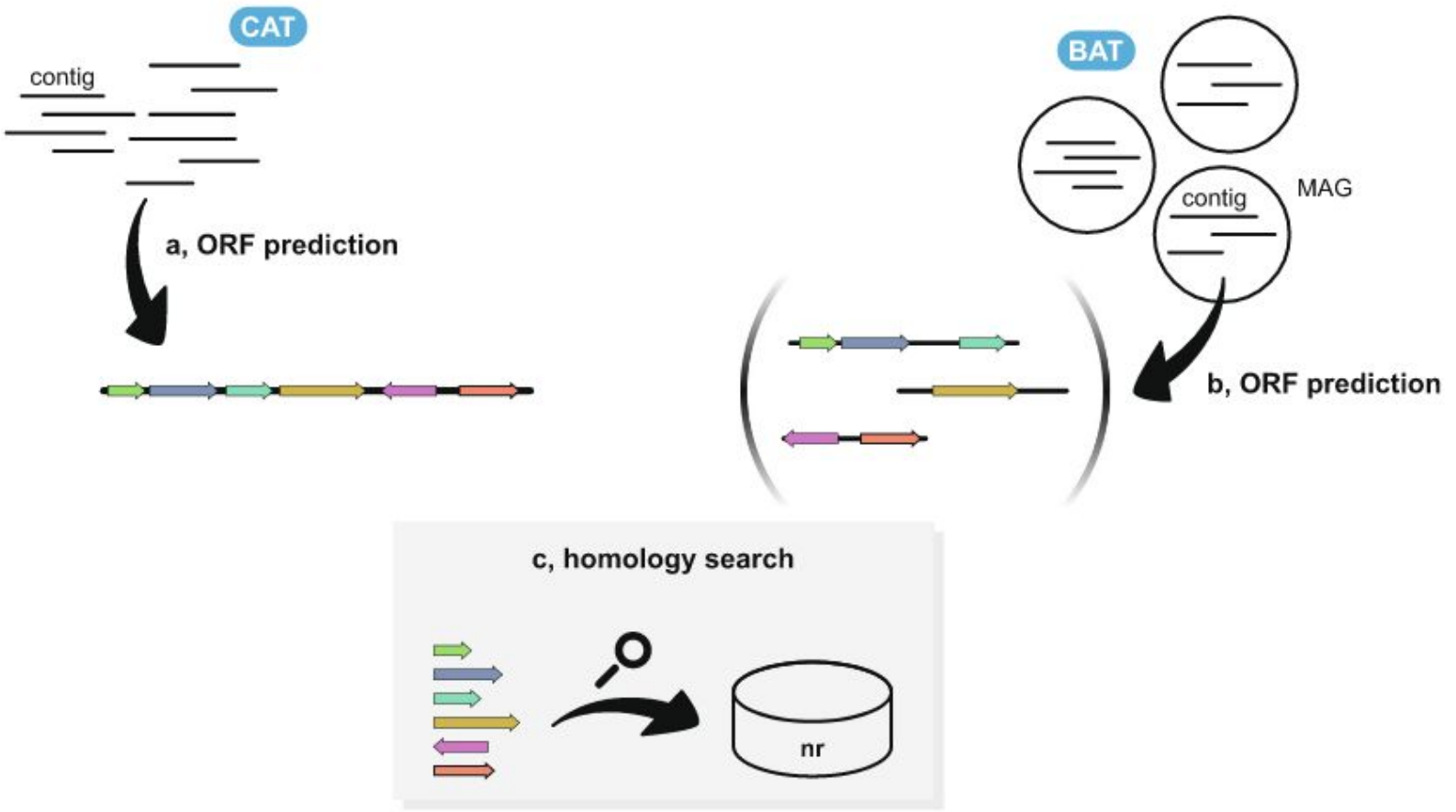
CAT/BAT

- The CAT program attempts to classify our bins
- Meaning giving us the best estimate of what organism each bin represents
- Wait a minute, didn't Checkm do that?
- Checkm currently uses a tree made with 5656 organisms

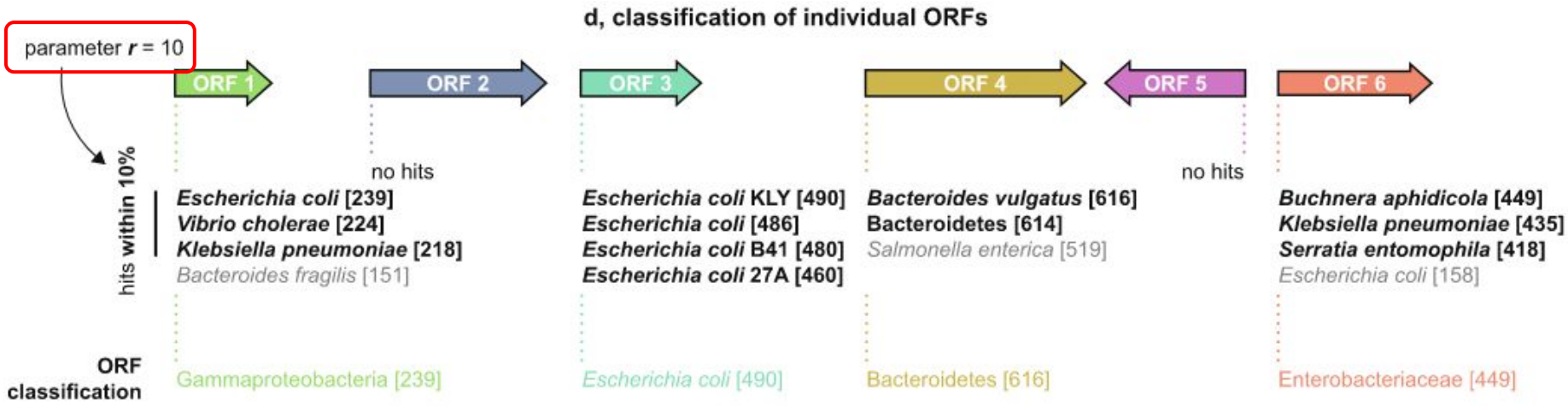
CAT/BAT

- The CAT program attempts to classify our bins
- Wait a minute, didn't Checkm do that?
- Checkm currently uses a tree made with 5656 organisms
- CAT looks at all non-redundant proteins in NCBI
- That samples a lot of organisms spanning many taxa

CAT algorithm



CAT algorithm



- *r* is the range, meaning the % of the top bit score within which hits will be considered
- e.g. if the top bit score is 500, *r* = 10 means that taxa having scores above 450 will be considered

Bit Score

- The size of the database needed to find a hit of equal value by chance
- Equal to $2^{\text{bit-score}}$ bp

CAT algorithm

e, contig / MAG classification

	ORF 1	ORF 3	ORF 4	ORF 6	sum	fraction of B_{sum}	>mbs
Superkingdom							
Bacteria	**239**	**490**	**616**	**449**	**1794**	**1.0**	**yes**
Phylum							
Proteobacteria	**239**	**490**		**449**	**1178**	**0.66**	**yes**
Bacteroidetes			616		616	0.34	no
Class							
Gammaproteobacteria	**239**	**490**		**449**	**1178**	**0.66**	**yes**
Order							
Enterobacteriales		**490**		**449**	**939**	**0.52**	**yes**
Family							
Enterobacteriaceae		**490**		**449**	**939**	**0.52**	**yes**
Genus							
Escherichia		490			490	0.27	no
Species							
Escherichia coli		490			490	0.27	no

B_{sum} = 239 + 490 + 616 + 449 = 1794

mbs = 0.5 x 1794 = 897

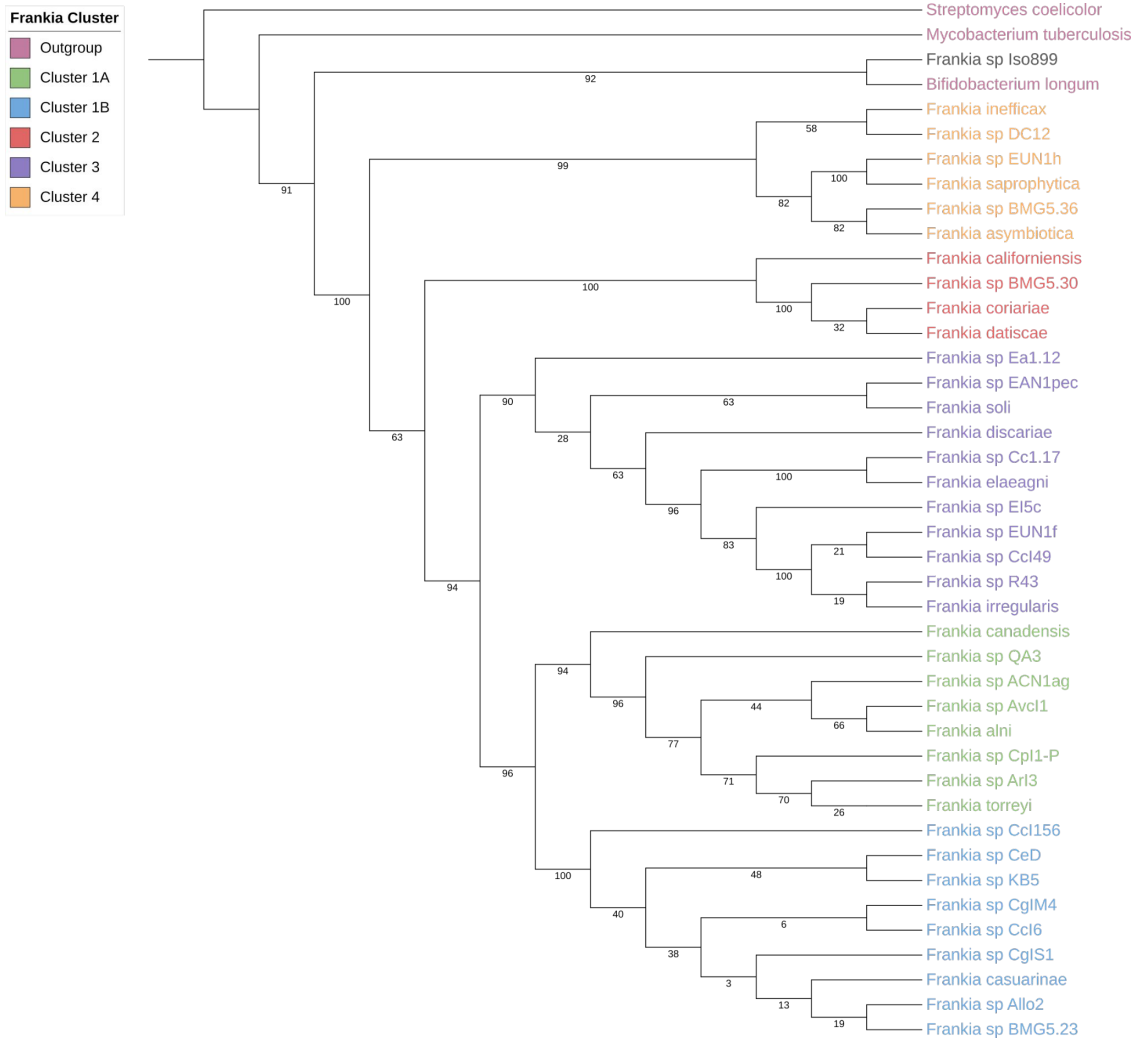
parameter ***f*** = 0.5

Final classification:
Bacteria (1.0)
Proteobacteria (0.66)
Gammaproteobacteria (0.66)
Enterobacteriales (0.52)
Enterobacteriaceae (0.52)

CAT performance on a known Frankia genome

- Frankia sp. ArI3
- Very close relative of other cluster 1A genomes
- How does CAT perform?
- This might be informative for our mixed Frankia metagenomes
- Varying parameter *r* from 1-10 %

glnA1 (protein) phylogeny



New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

NCGR
National Center for Genome Resources