Mini-metagenomes assembled by SPAdes

Margarita Akseshina

scientific advisor: Anton Korobeynikov

Goals

- find features for binning
- check them for reliability and robustness
- try to improve metagenomic assembly

Features

- tetranucleotide frequencies
  - 136 dimensions
- average coverage
- contig length
- codon usage
- ...

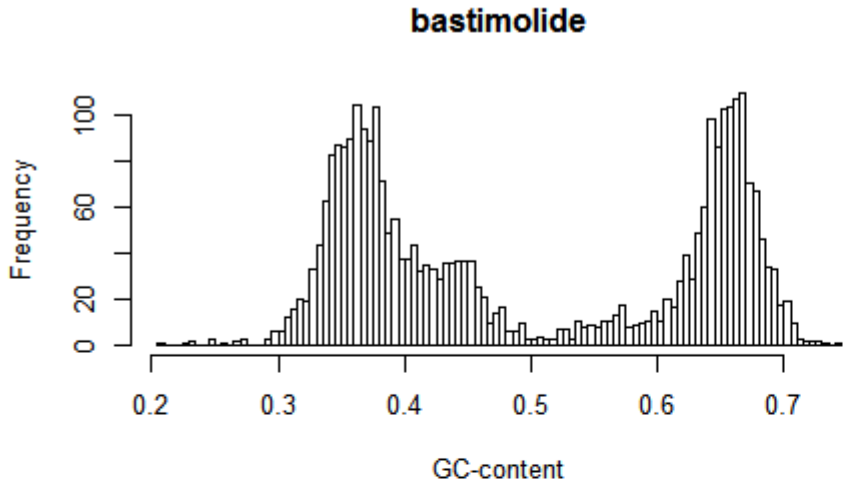
Datasets

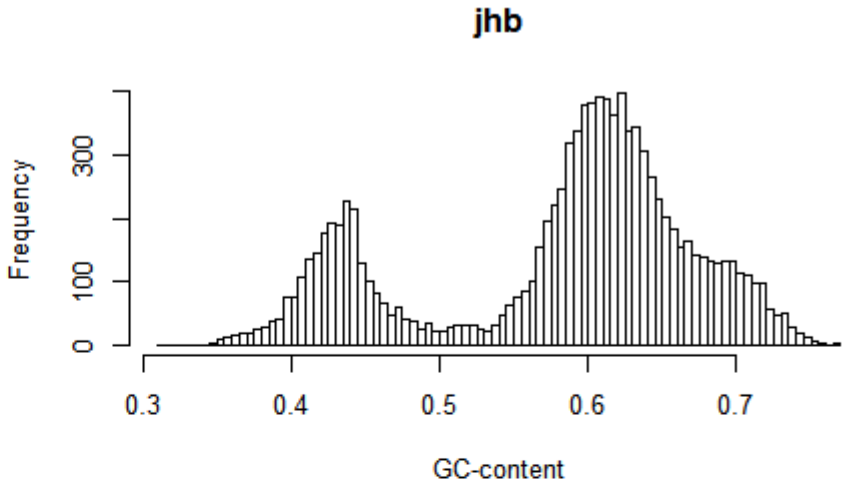
Mini-metagenomes from Lena Gerwick:

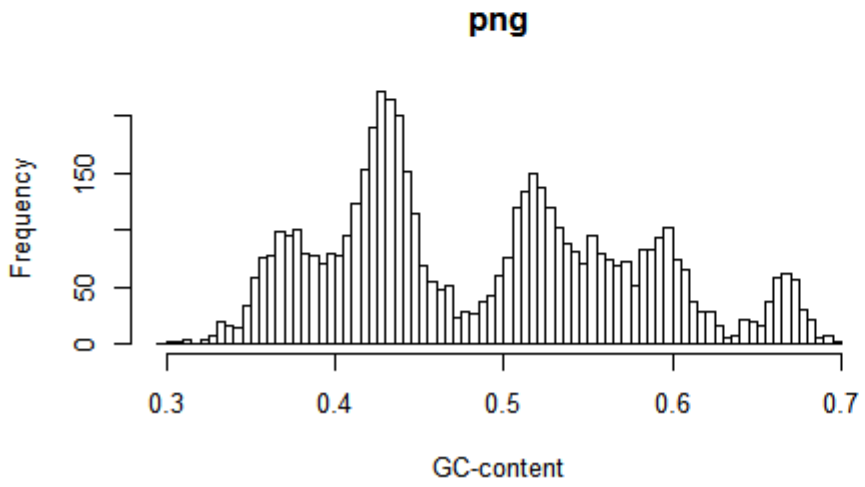
- Bastimolide producer
- Moorea Producens JHB
- Scytonema Hofmanii (UTEX B2349)
- Moorea Bouillonii PNG

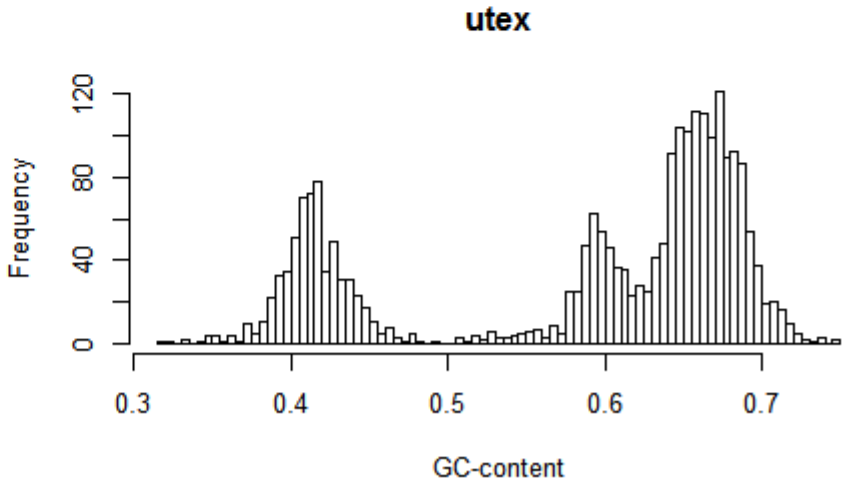
Usually contains a cyanobacteria of interest and several symbionts.

How many taxons do you see?

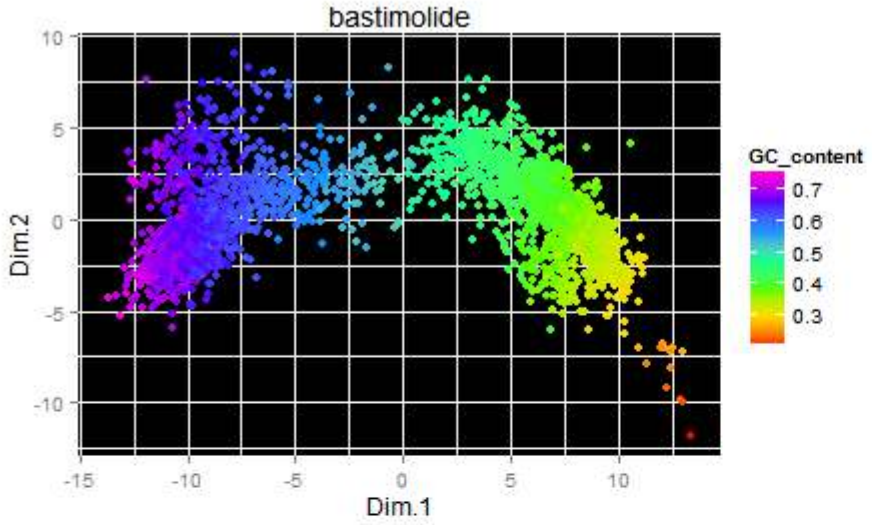


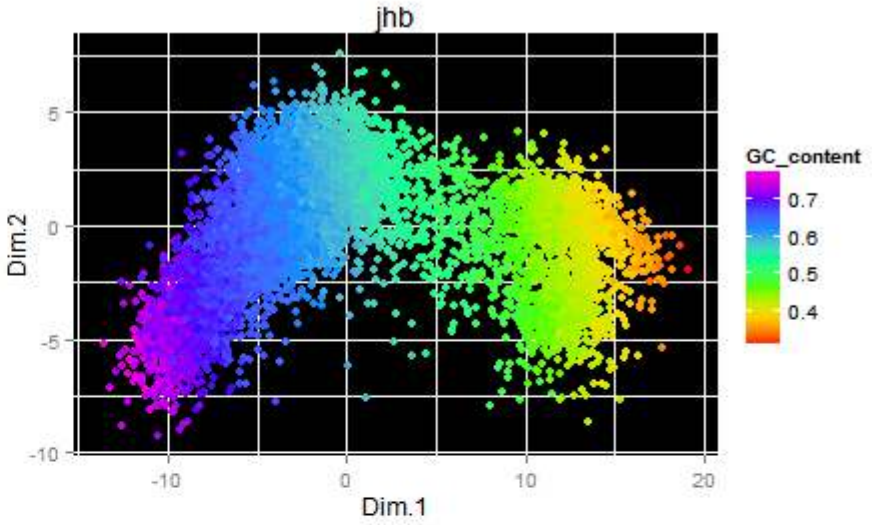


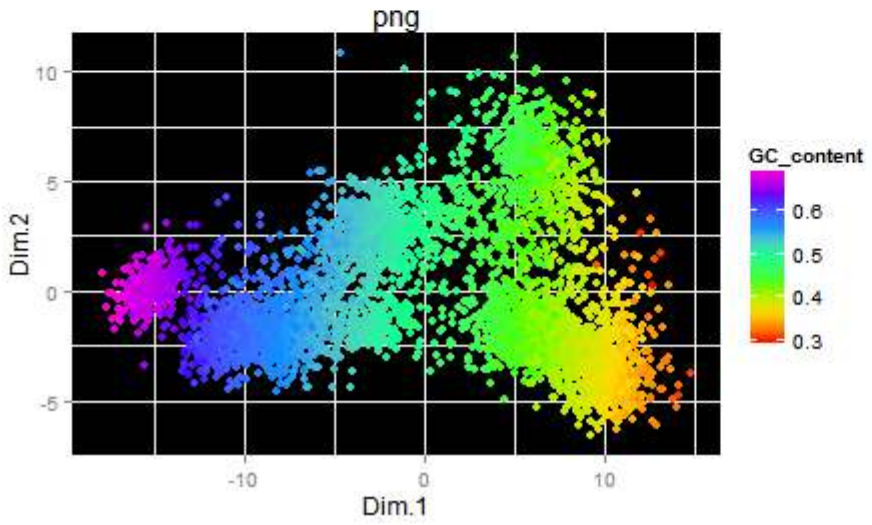


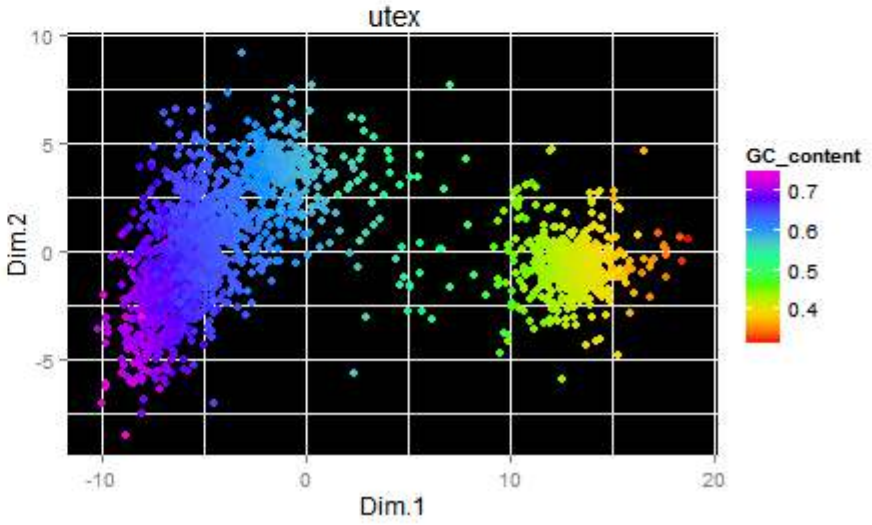


PCA with GC-content

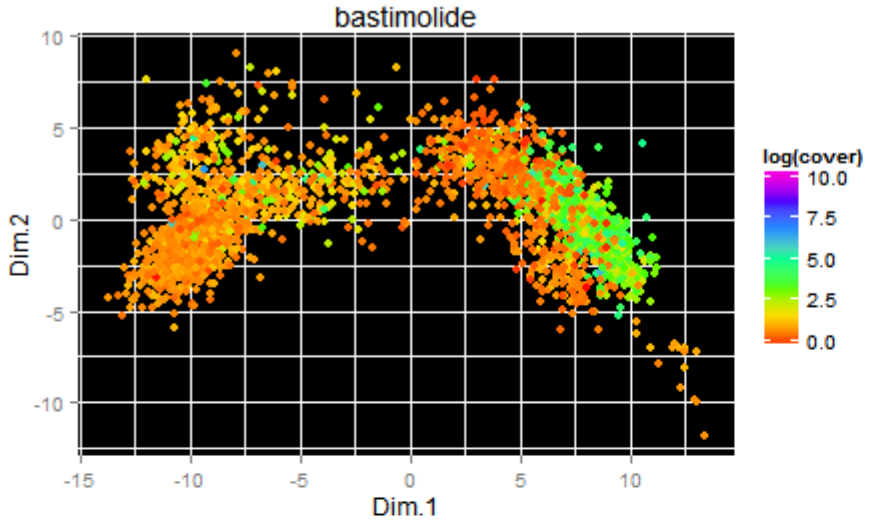


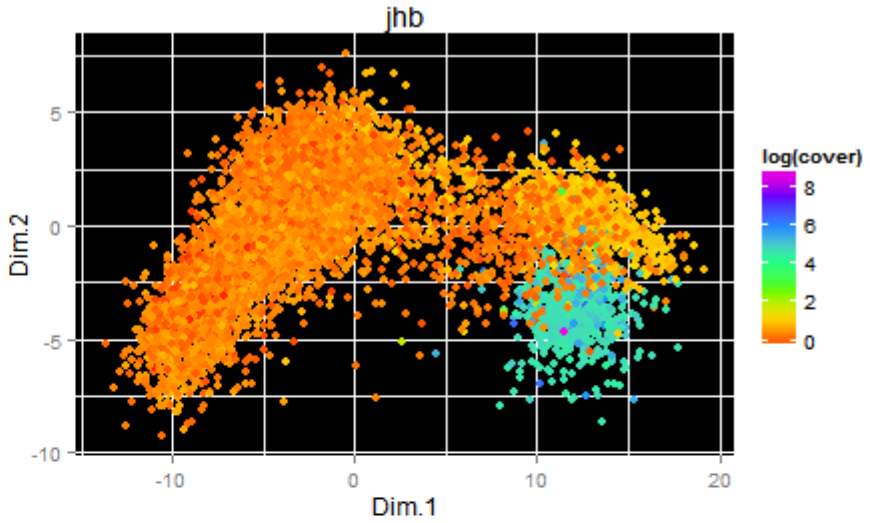


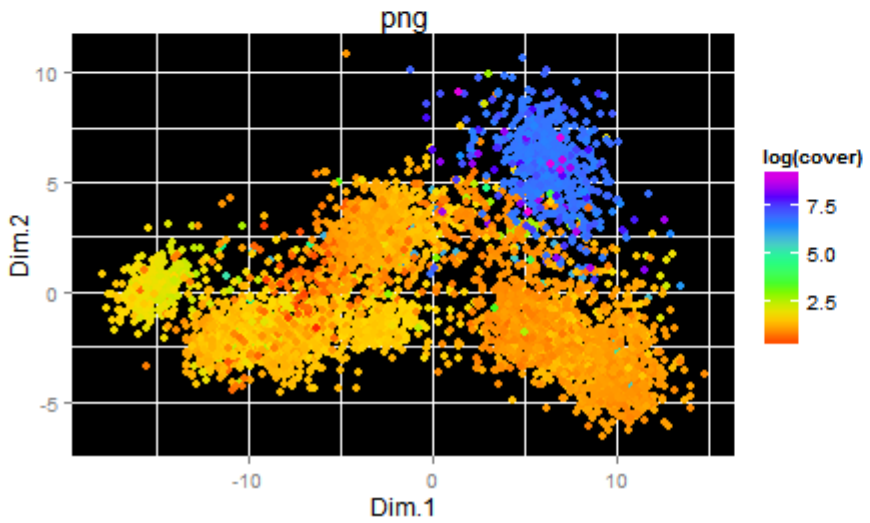


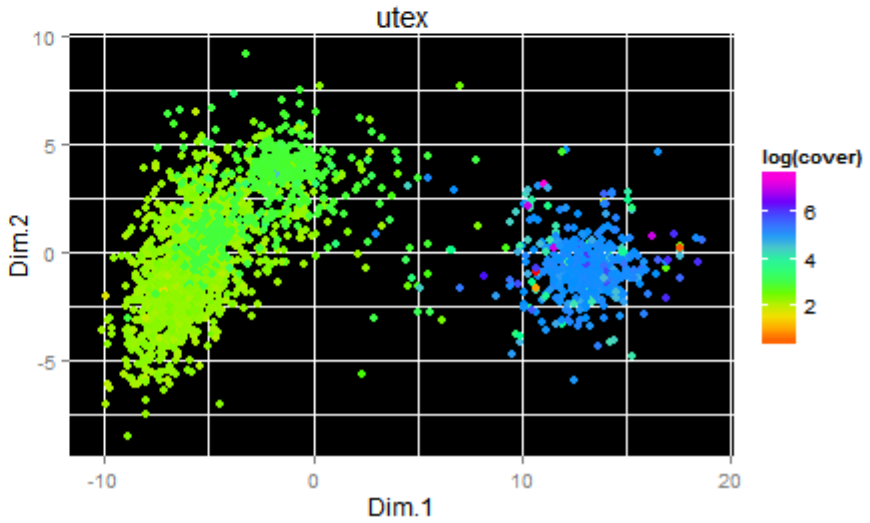


PCA with coverage

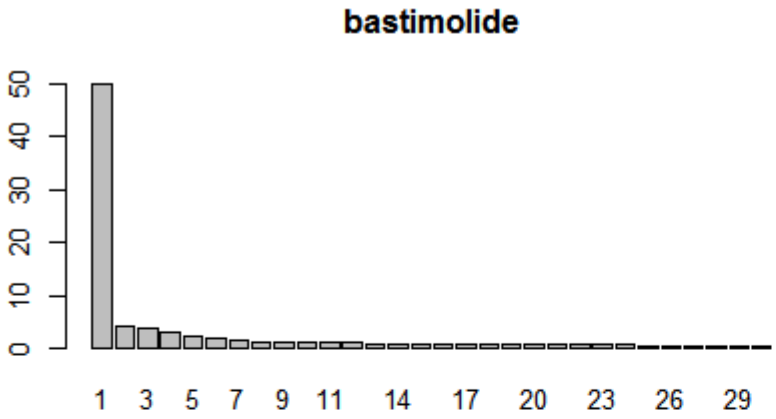


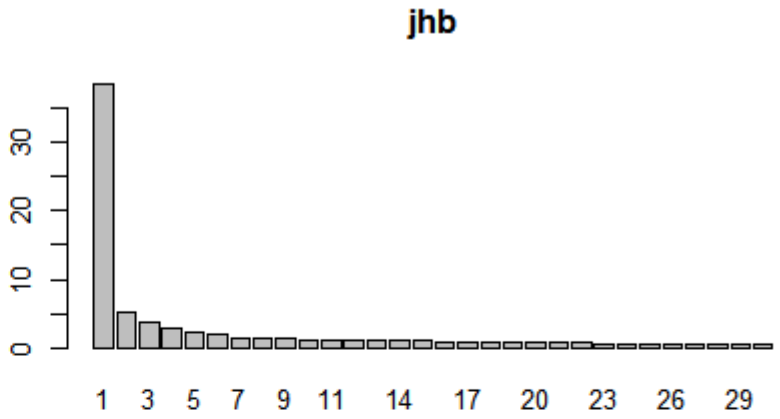


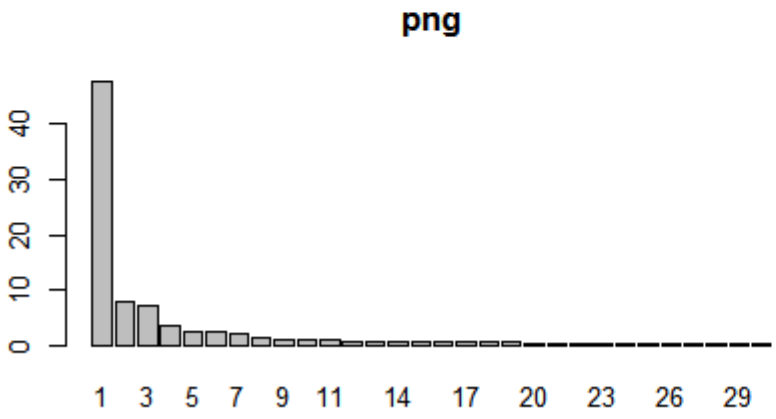


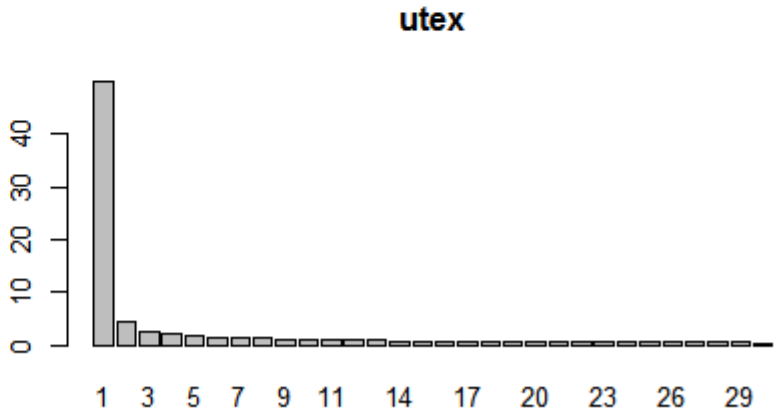


How many PC to keep?
Scree plots

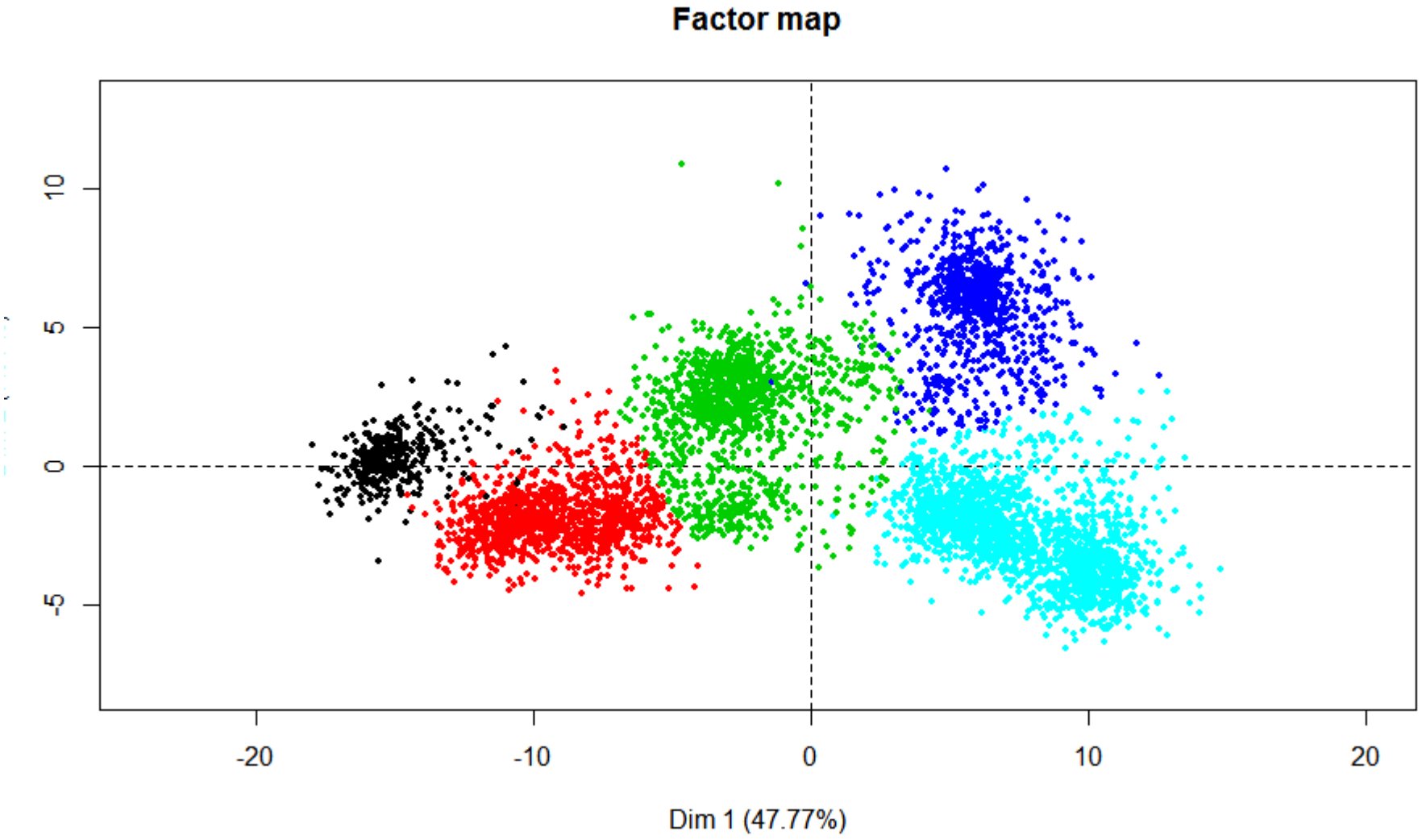




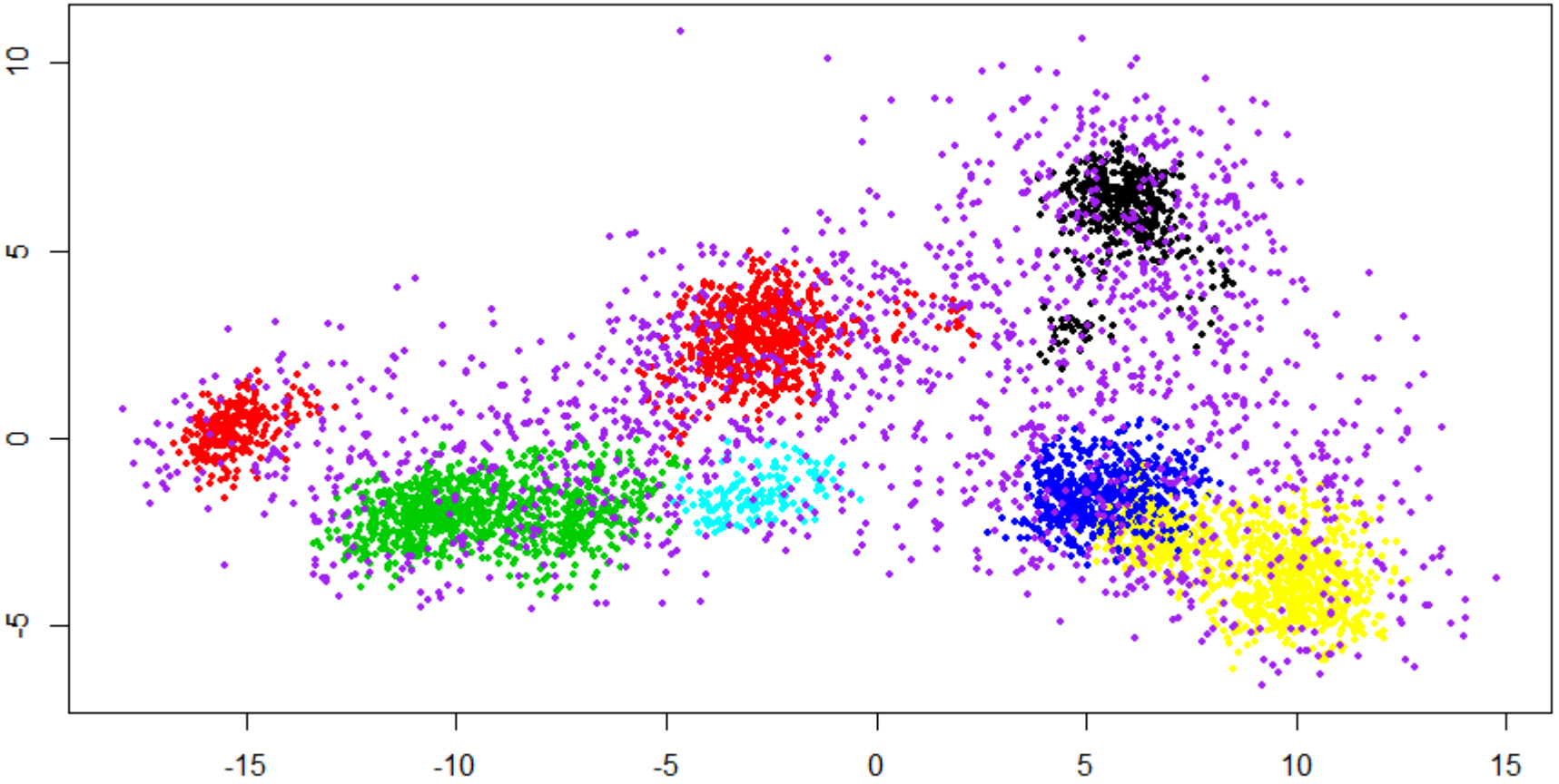




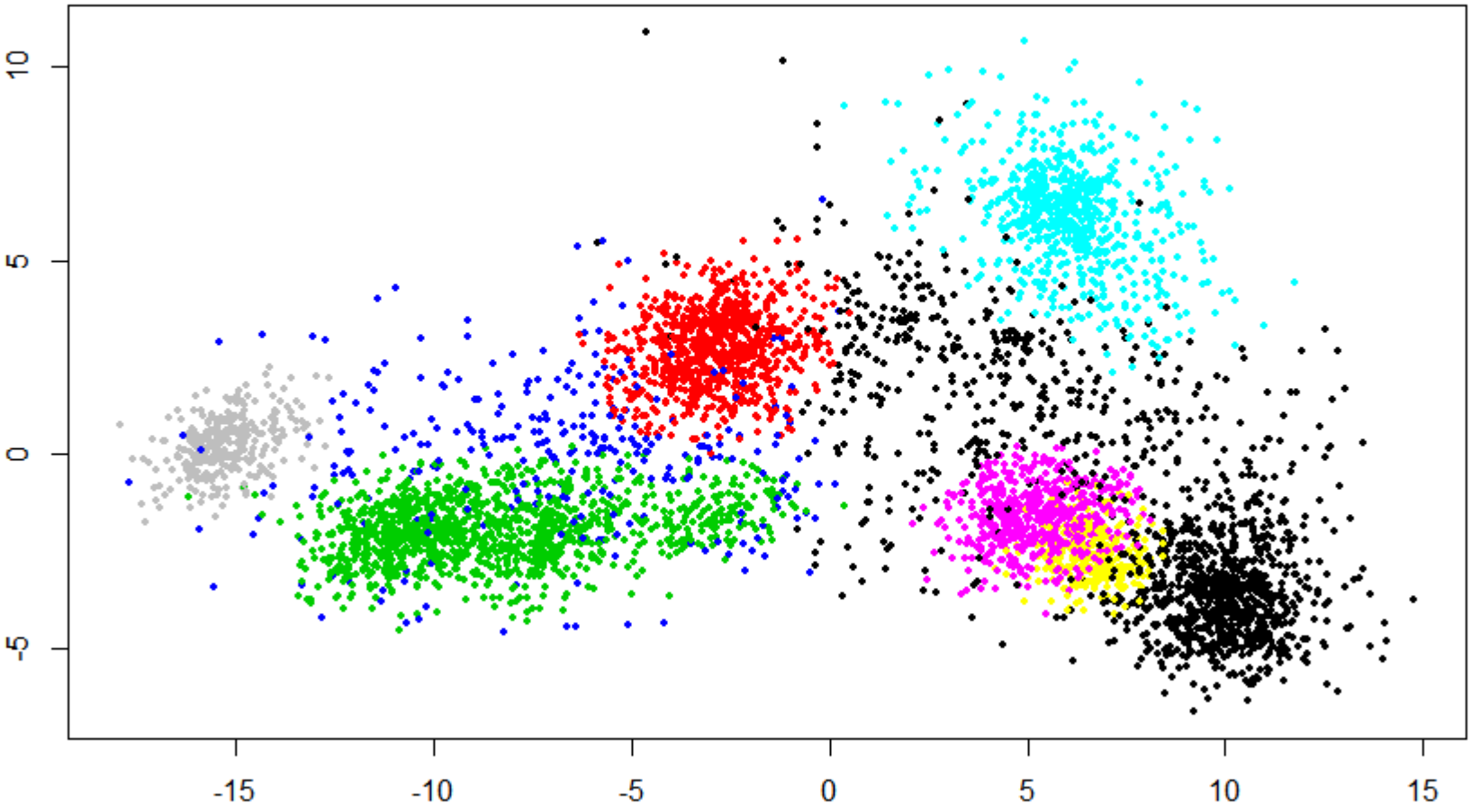
Hierarchical clustering



DBSCAN



Mclust



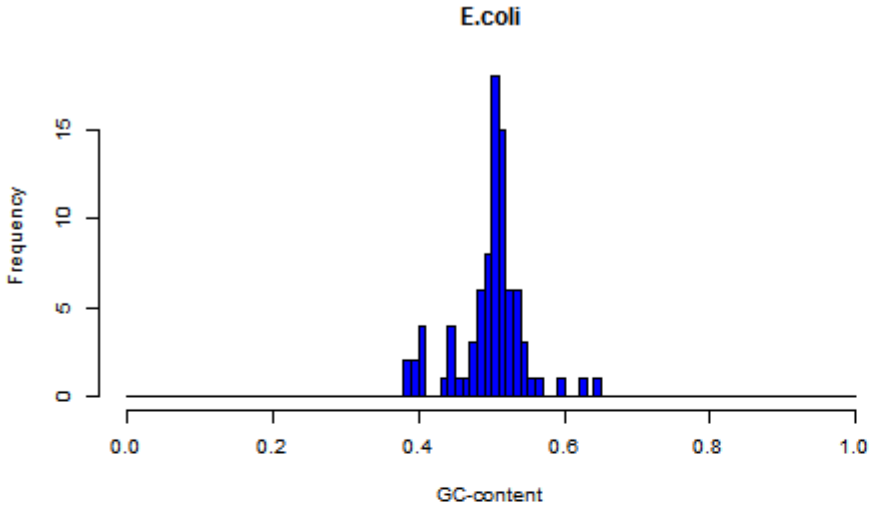
Further questions

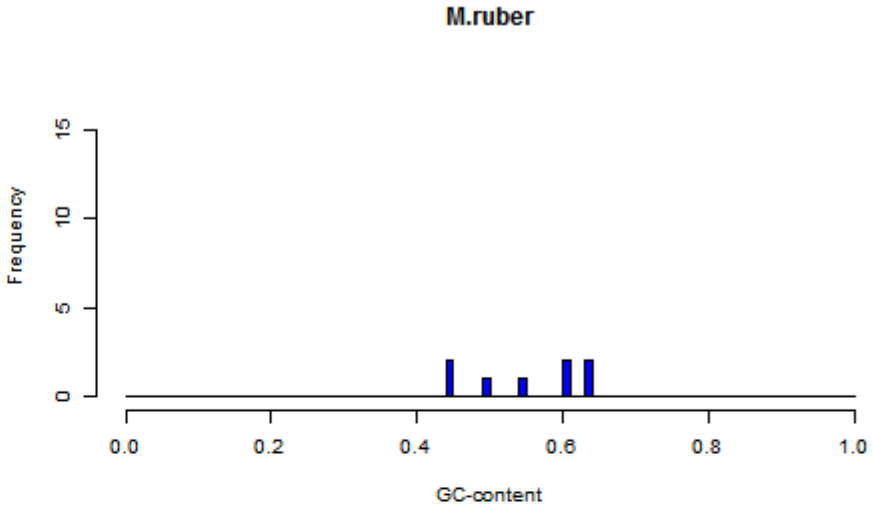
- a suitable length filter
- a method for clustering
- number of principal components
- validation of the results
- ...

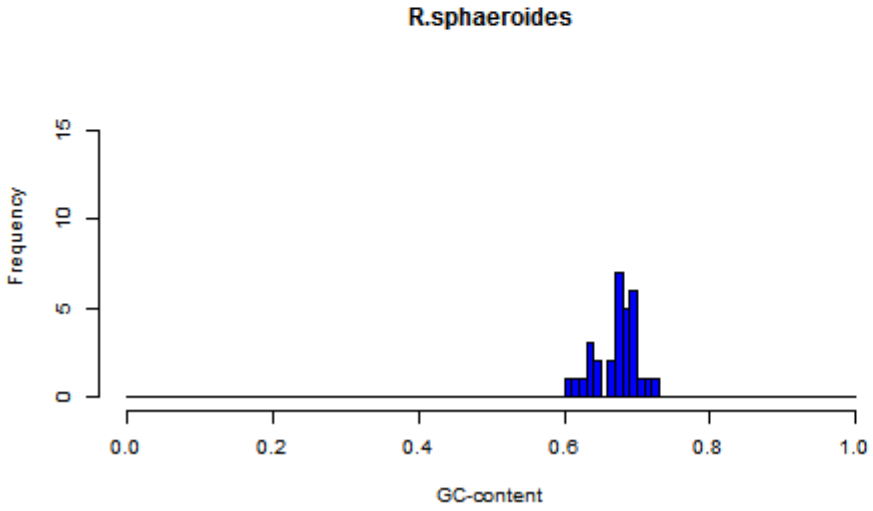
How to model data

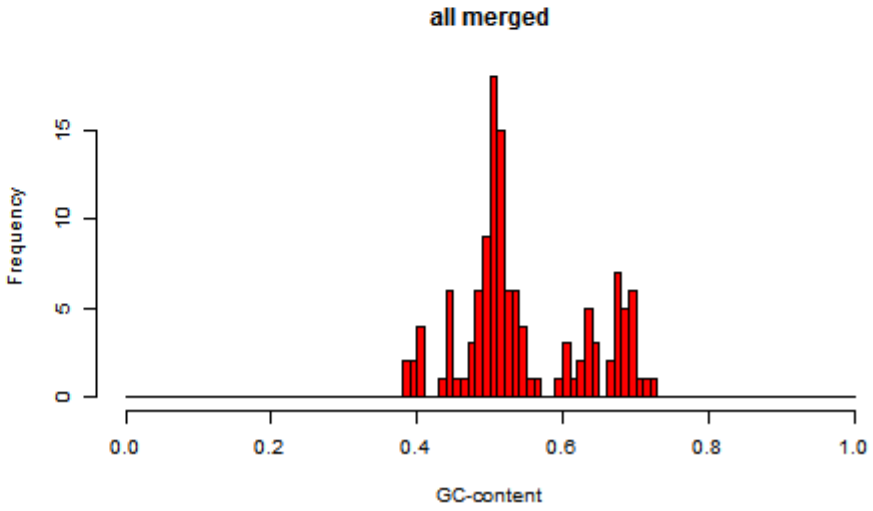
- merge several assemblies
- merge different reads and then assemble

Assemblies

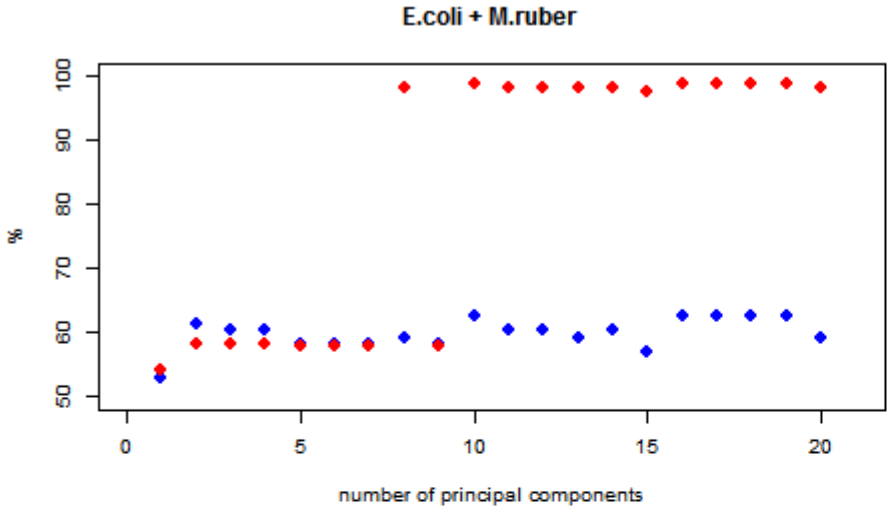

E.coli
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
GC-content

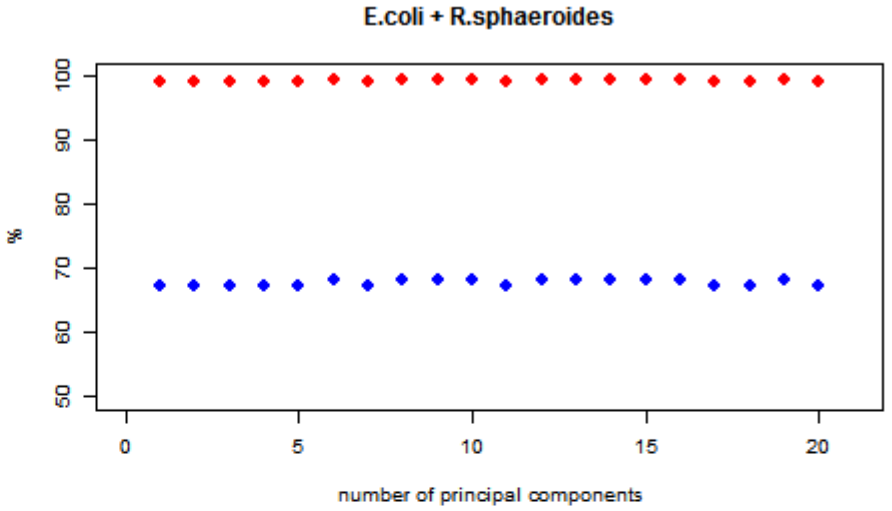

M.ruber
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
GC-content

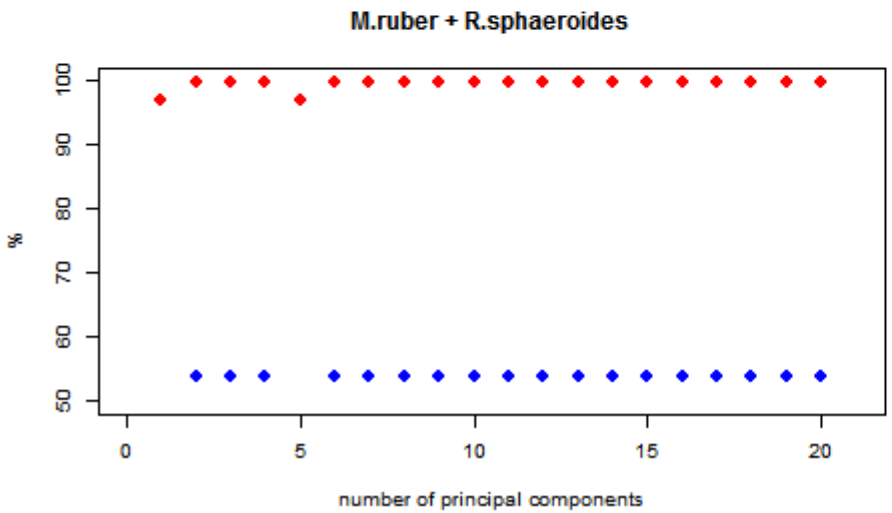

R.sphaeroides
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
GC-content

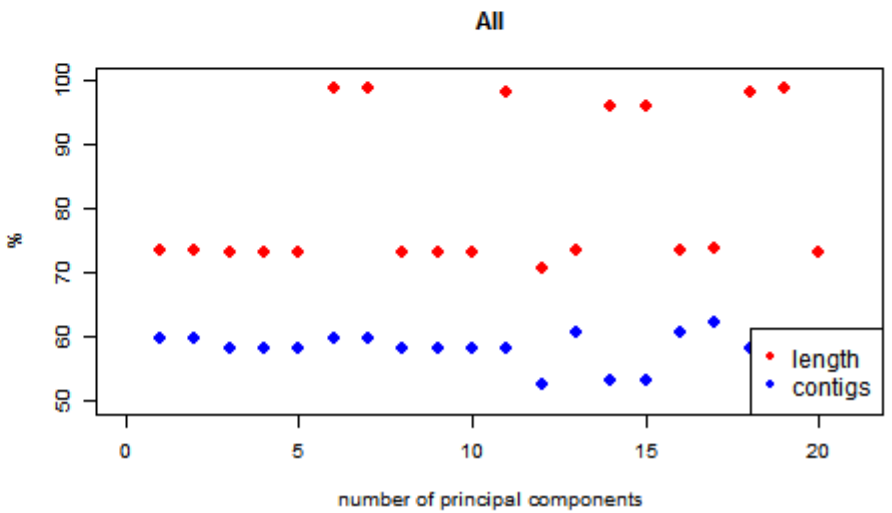

all merged
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
GC-content

Number of PC (hclust)

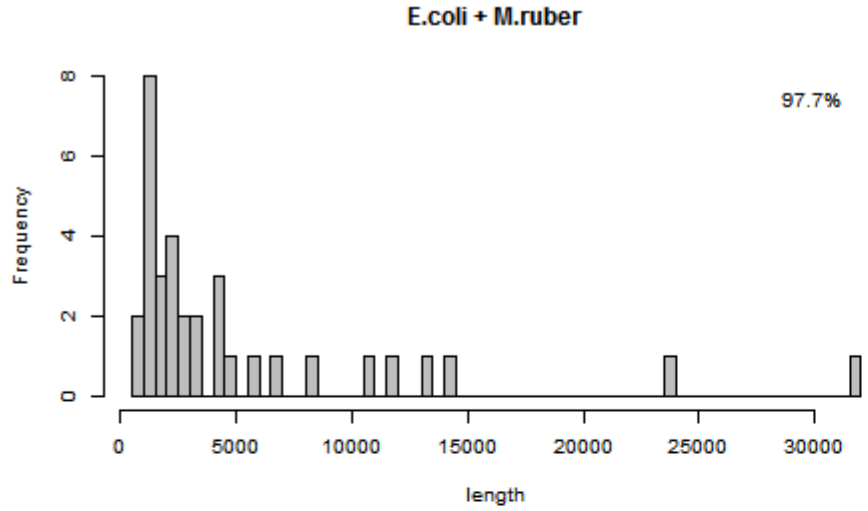


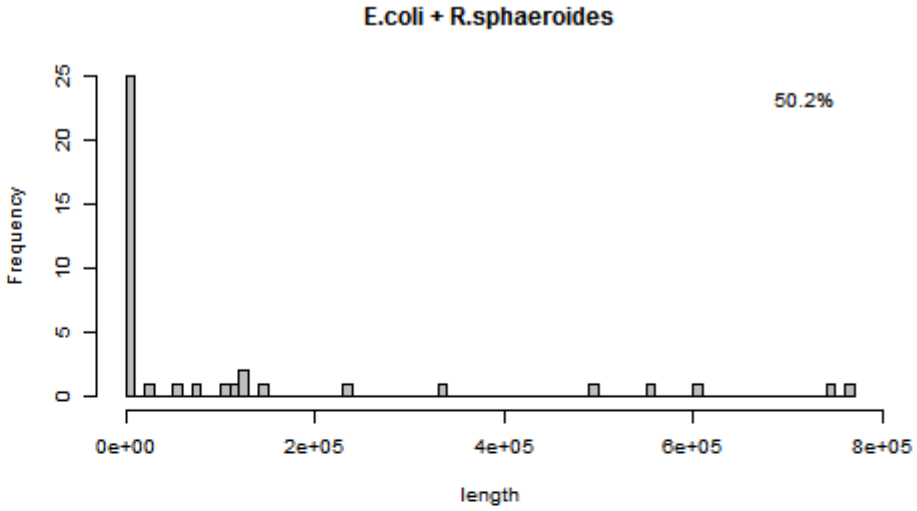


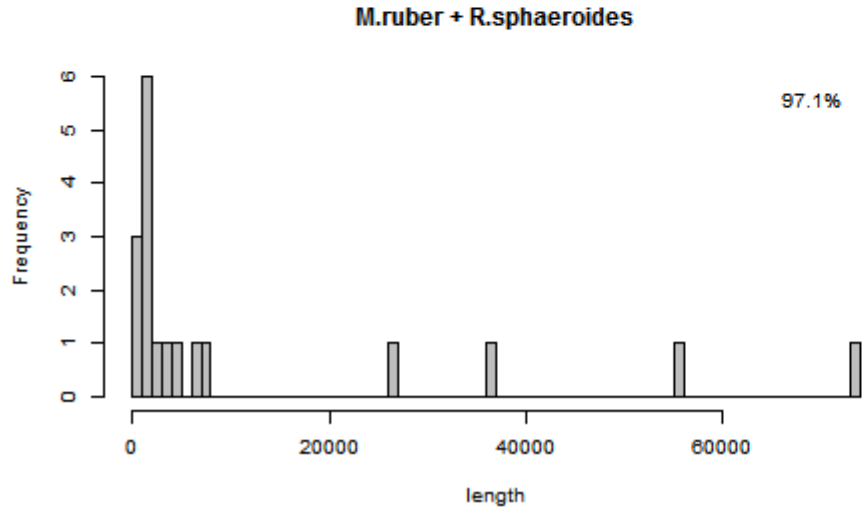


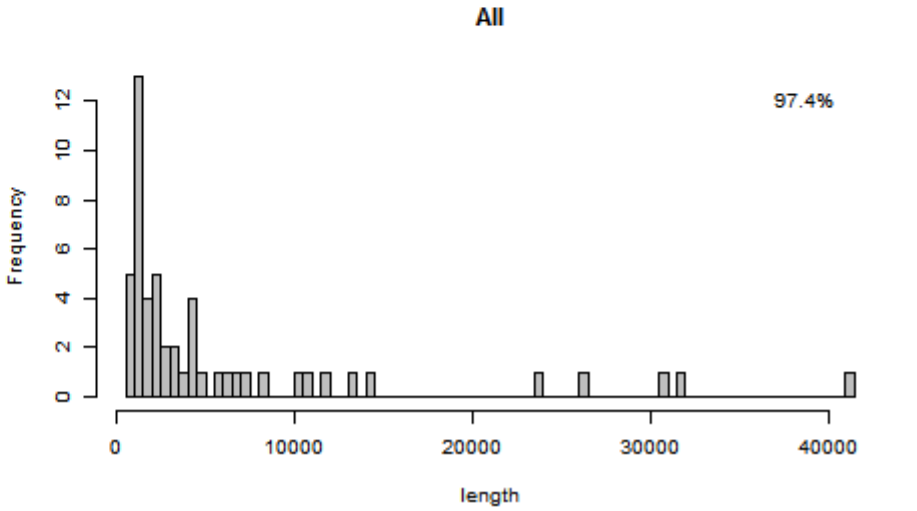


False contigs length (mclust, 10 PC)









Thank you!