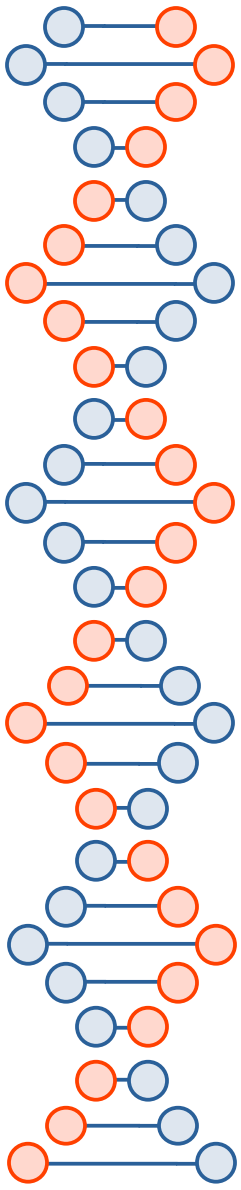
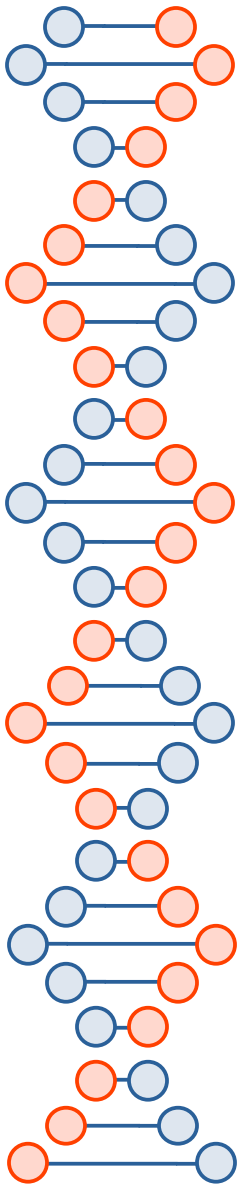


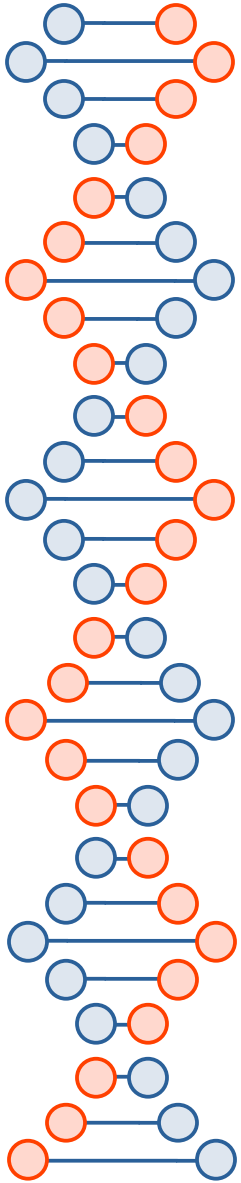
Bioinformatics Data Processing

Final assignment



Final assignment – Overview

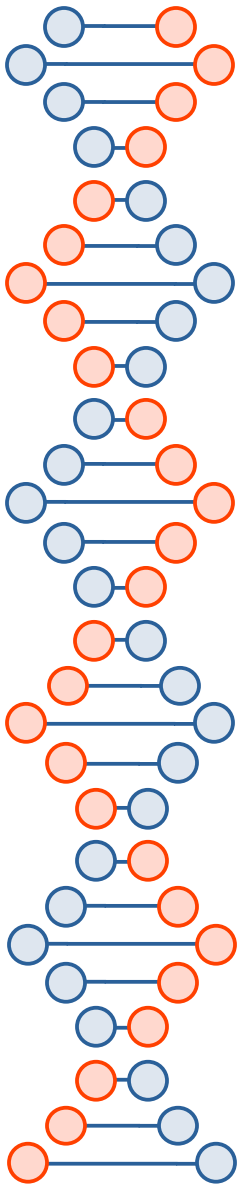
- Metagenomics is the study of genetic material recovered directly from environmental samples without the need to isolate or culture individual organisms.
- The typical workflow involves: 1) DNA extraction, 2) sequencing, 3) assembly, 4) genome binning, and 5) taxonomic and functional analyses.
- You are given results from two metagenomic assemblers, **metaMDBG** and **myloasm**, applied to nanopore sequencing data from a hot spring microbial community.



Final assignment – Overview

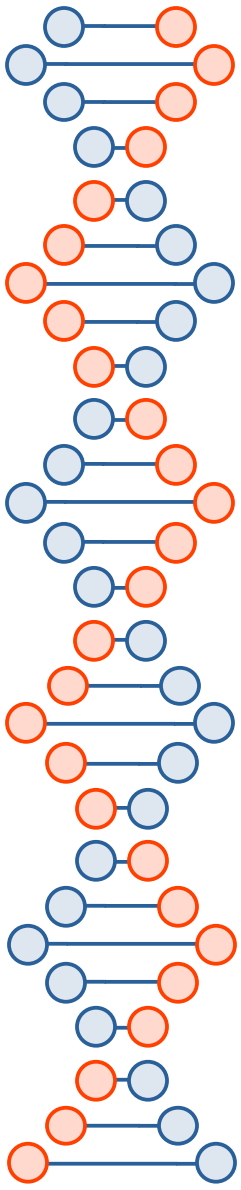
- Your goal is to **compare the performance of both assemblers** using only the text-based metadata available from:
  - **Contig FASTA headers**
  - **CheckM2** (genome completeness & contamination)
  - **GTDBtk** (taxonomy assignment)

- It is a good practice to open the files as plain text first and see their structure.



Final assignment – Overview

- You will perform all data processing, visualization, and summarization ~~**entirely in R and the tidyverse**~~.
- You will produce:
  - A ~~**Quarto**~~ **report** documenting your entire workflow
  - A **GitHub repository** containing your, code, analysis, and figures
- No extra computational work is required.



Final assignment – data provided

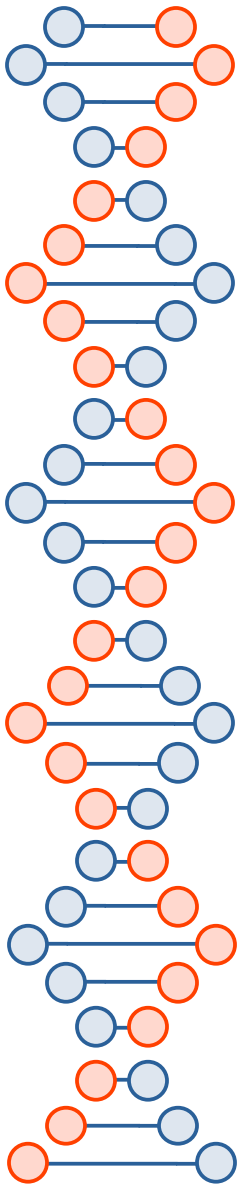
- Analyses were performed on the MetaCentrum grid. The resulting files are located at::

```
/storage/praha1/home/strejcem/results
```

- Both assemblers output the contig statistics in their FASTA headers. The headers were extracted using:

```
grep ‘>’ myloasm.assembly.fasta > myloasm_headers.txt
```

```
grep ‘>’ metamdbg.assembly.fasta > metamdbg_headers.txt
```



Final assignment – FASTA headers

- Assembly is a process where short (100–10kbp) fragments are assembled into contigs (continuous sequences).
- Both assemblers report similar information in the contig headers.
- myloasm:

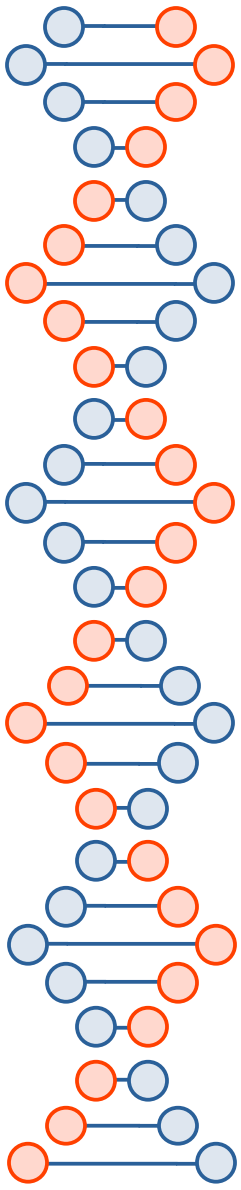
```
>u3840050ctg_len-10849_circular-no_depth-2-2-2_duplicated-no mult=1.00
```

contig name

contig length

contig circularity

coverage: 99% - 99.75% - 100% identity mapping

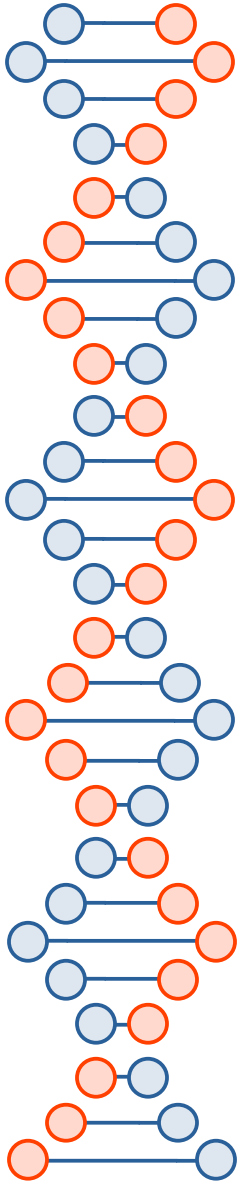


Final assignment – FASTA headers

FASTA is a common file format for sequencing data. It consist of a sequence header that starts with ">" followed by the actual sequence.

From the header files, you can extract:

- **assembler** (read_tsv(id = "assembler"))
- **contig ID**
- contig **length**
- **circular** vs **non-circular** (for myloasm, consider “possible” circularity as full circularity)
- **depth** statistics = **coverage**, i.e. how many times, was the genomes sequenced in average (for myloasm use the first number of the three)

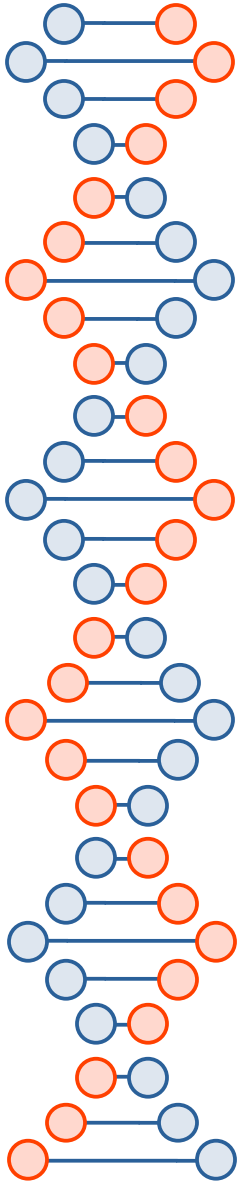


Final assignment – CheckM2

- Assembly is usually followed by genome binning. In metagenomics, the outputs are genome bins, also known as MAGs (Metagenome-Assembled Genomes). In this project, genome binning was not performed. We are interested only in (almost) complete MAGs, i.e., circular single contigs.
- CheckM2 evaluates genome/MAG completeness and contamination using set of markers.
- Use the **completeness** and **contaminations** columns

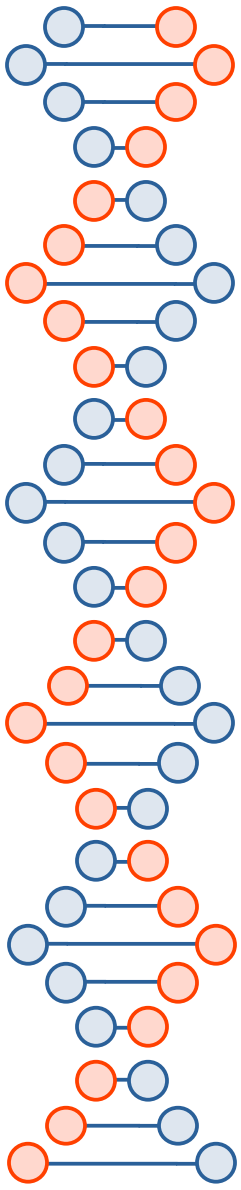
Final assignment – GTDB-Tk

- GTDB-Tk's taxonomic classifications of archaeal and bacterial MAGs are located in the GTDB-Tk directory, in the files labeled ar50 and bac120, respectively.
- The taxonomy information is in the column **classification.**
- You might need to extract the Phylum information.
- Ignore "Unclassified Archaea/Bacteria"



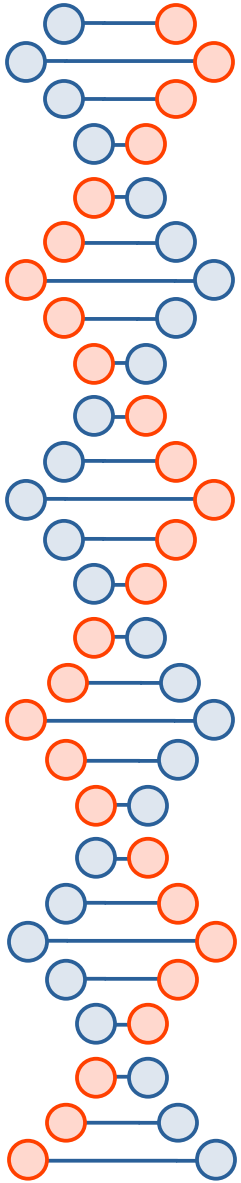
Final assignment – Contigs

- Plot the contig length distributions for each assembler, distinguishing between circular and non-circular contigs.
- Evaluate how contig length correlates with sequencing coverage.



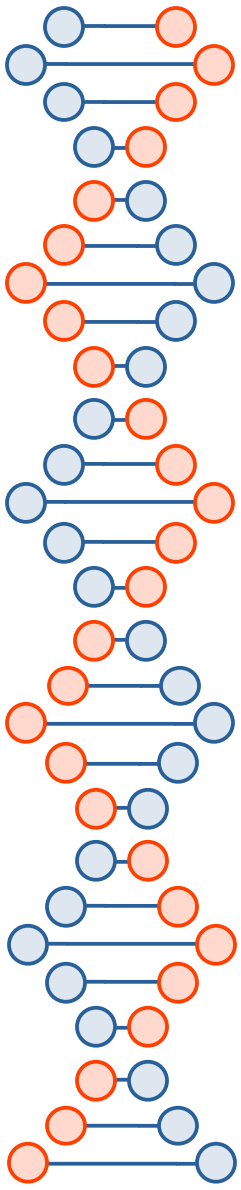
Final assignment – The assemblers

- Compare the two assemblers based on the number and quality of the large (>500 kb) circular contigs they reconstruct. Use appropriate visualizations.
- You may group MAGs according to the following quality thresholds (or display these thresholds as reference lines in scatter plots):
  - High quality: >90% completeness and <5% contamination
  - Medium quality: >50% completeness and <10% contamination
  - Low quality: below the thresholds above
- Additionally, report how many large circular contigs were reconstructed per phylum.
- Finally, provide a brief summary indicating which assembler appears to perform better.



Final assignment – Don’t’ forget …

- Explore a variety of visualizations, such as histograms or density plots, barplots, scatter plots, boxplots, and Venn diagrams, etc.
- Make sure to label all figures, axes, and legends clearly.
- Use faceting (e.g., facet_wrap or facet_grid) where appropriate.
- Incorporate colors or shapes to distinguish assemblers, quality categories, or circular vs. non-circular contigs.
- Consider applying logarithmic axis scaling when it improves readability or highlights relevant patterns.



Final assignment – handout

- Create a protocol using Quarto (or an alternative) that documents all your steps, including code and figures.
- Render the notebook in GitHub-Flavored Markdown (GFM) using the following YAML header:

```
---
format: gfm
---
```

- Rename the output .md file to README.md. Upload this file along with the directory containing the figures to your GitHub repository.
- Finally, send an email to strejcem@vscht.cz with the link to your repository.
- Good luck!