Introduction to Nanopore: Principles, Updates, and Bioinformatics

Practical session (Nanopore assembly)

Rania Ouazahrou, April 11th 2025
Biomics platform, Institut Pasteur Paris
Bioinformatics and biostatistics hub, Institut Pasteur Paris

Intended Learning Objectives

Specific objectives of this session:

1. Install tools using conda
2. Run an assembly

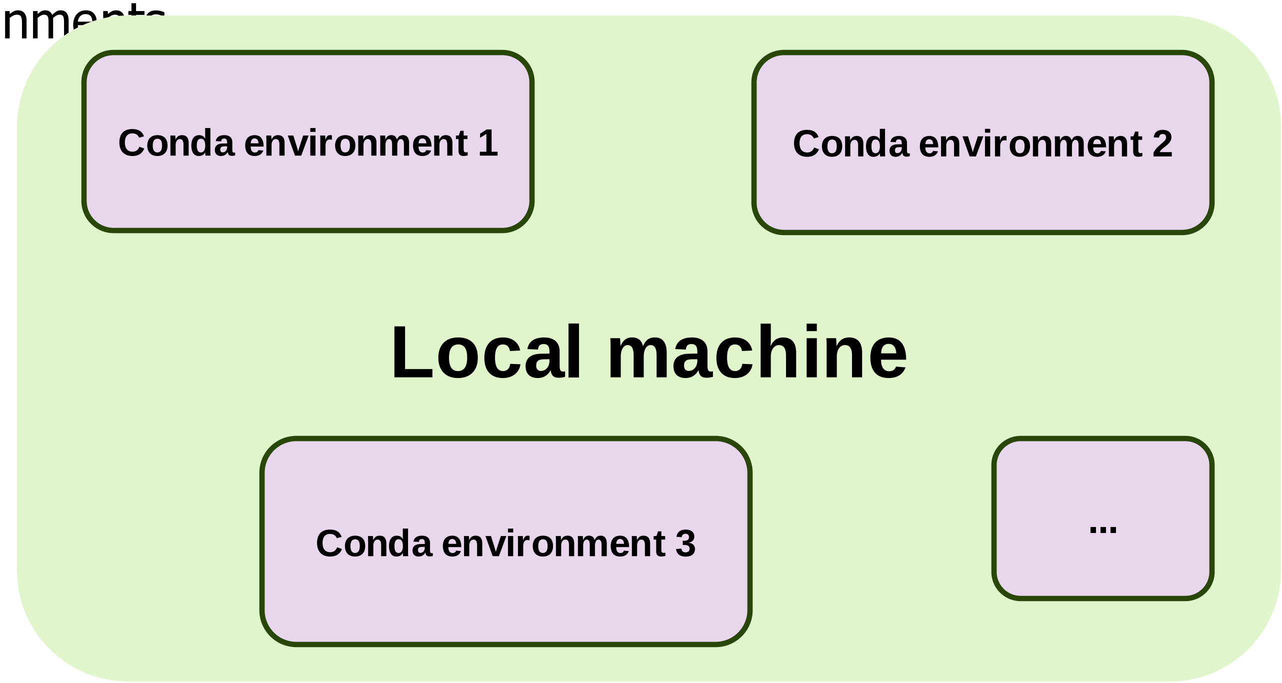
Outline

This session consists of the following elements

1. Review of the installation part
2. Review of the command lines for each tool
3. Analyse the results

Installation with conda

- Conda is an environment manager that allows you to create isolated environments on your machine. Within each environment, you can install tools and their dependencies without affecting your system-wide installations.
- For example, Conda makes it possible to install and manage different versions of the same software in different environments.



Environment management

- Conda provides three different channels: forge, bioconda and default.

Conda default channel is not free

Alternative:

Micromamba

https://mamba.readthedocs.io/en/latest/installation/mamba-installation.html

Miniforge

https://github.com/conda-forge/miniforge

For those unfamiliar with command lines:

- can enroll in the linux webinar "Unix for beginners & Introduction to the Conda ecosystem"

https://learning.ecdc.europa.eu/course/view.php?id=888

Pre-requisites

The minimal prerequisites for participants to meet the learning objectives of the practical exercise are:

- Ability to navigate a Unix terminal and use bioinformatic command line tools.
- Access to a Unix terminal (or emulator) during practical exercises.

Participants not having these skills can still **attend the virtual training as observers**.

If you are unsure about your command line skills, or if you wish to get some reminders about Unix command line, we encourage you to watch the recordings of the following virtual trainings:

- GenEpi-BioTrain - Virtual training 02: From sequencer to polished reads for bacteria
- GenEpi-BioTrain - Virtual training 09 & 10 - Unix for beginners & Introduction to the Conda ecosystem

- use Galaxy (web-based platform for data analysis)

https://usegalaxy.org/

Create a new environment

1- Create an environment called "assembly"

```
conda create -n assembly
```

2- Activate the environment

```
conda activate assembly
```

3- Install the tools

```
conda install flye
conda install bioconda::quast
conda install -c bioconda nanoplot
```

4- Test the installation of the tools

```
flye --help
quast --help
NanoPlot --help
```

Busco installation

For busco we're going to need another environment, because we have dependencies conflicts between the tools.

For example, flye or quast needs a specific version of one package, and busco needs a different version, so we cannot install different versions of one package or tool within the same environment. That' s why we need a separate environment for busco.

1- creation of the environment

```
conda create -n busco
```

2- activate the env

```
conda activate busco
```

3- install busco

```
conda install -c conda-forge -c bioconda busco=5.8.2
```

Check the characteristics of your sample

https://pmc.ncbi.nlm.nih.gov/articles/PMC6927303/#abstract1

- *Bacteroides fragilis*
  - Genome size: 5.5 Mb
  - Plasmid(s): yes
  - Circular genome: yes
  - No heterozygosity (bacterial genome)
  - Number of chromosomes: 1
  - Multiple repeat regions
  - Coverage: ~58X (Number of reads * mean read length / genome size)
- Type of reads:
  - Non corrected Nanopore reads
  - Flowcell: R9.4 (not commercialized anymore)
  - Basecalled with albacore (replaced by dorado)

Sequencing (flowcell) and **basecalling** made with **older versions**, so we expect high error rates and therefore **low-quality assembly**

Running the assembly

1- activate the environment (assembly)

```
conda activate assembly
```

2- run the quality control with Nanoplot

```
NanoPlot --fastq CCUG4856T.fastq.gz -o nanopolot_output
```

3- run the assembly with Flye

```
flye --nano-raw CCUG4856T.fastq.gz --out-dir flye_output --genome-size 5m
```

4- assess the assembly quality

```
quast flye_output/assembly.fasta -o quast_results
```

Busco

1. Check the datasets available within Busco

```
busco --list-datasets
```

2. Choose the adequate dataset for your sample

- Here we used the general dataset for *bacteria* (domain), which contains only 116 genes that are shared between bacteria

```
busco -i flye_output/assembly.fasta -m genome -l bacteria_odb12 -o busco_bacteria
```

- We ca also use more specific databases, like "*bacteroidales*" which corresponds to the order, or "*bacteroides*" database wich corresponds to the genus

```
busco -i flye_output/assembly.fasta -m genome -l bacteroidales_odb12 -o busco_bacteroidales
```

```
busco -i flye_output/assembly.fasta -m genome -l bacteroides_odb12 -o busco_bacteroides
```

https://busco-data.ezlab.org/v5/data/lineages/

https://busco.ezlab.org/list_of_lineages.html

Quality control of the reads

Nanoplot:

Nanoplot-report.html

Summary Statistics | Plots | Report issue on Github

NanoPlot reports

Summary statistics

General summary	
Mean read length	5,897.2
Mean read quality	8.5
Median read length	2,959.5
Median read quality	8.9
Number of reads	49,980.0
Read length N50	12,053.0

Assembly

Flye output:

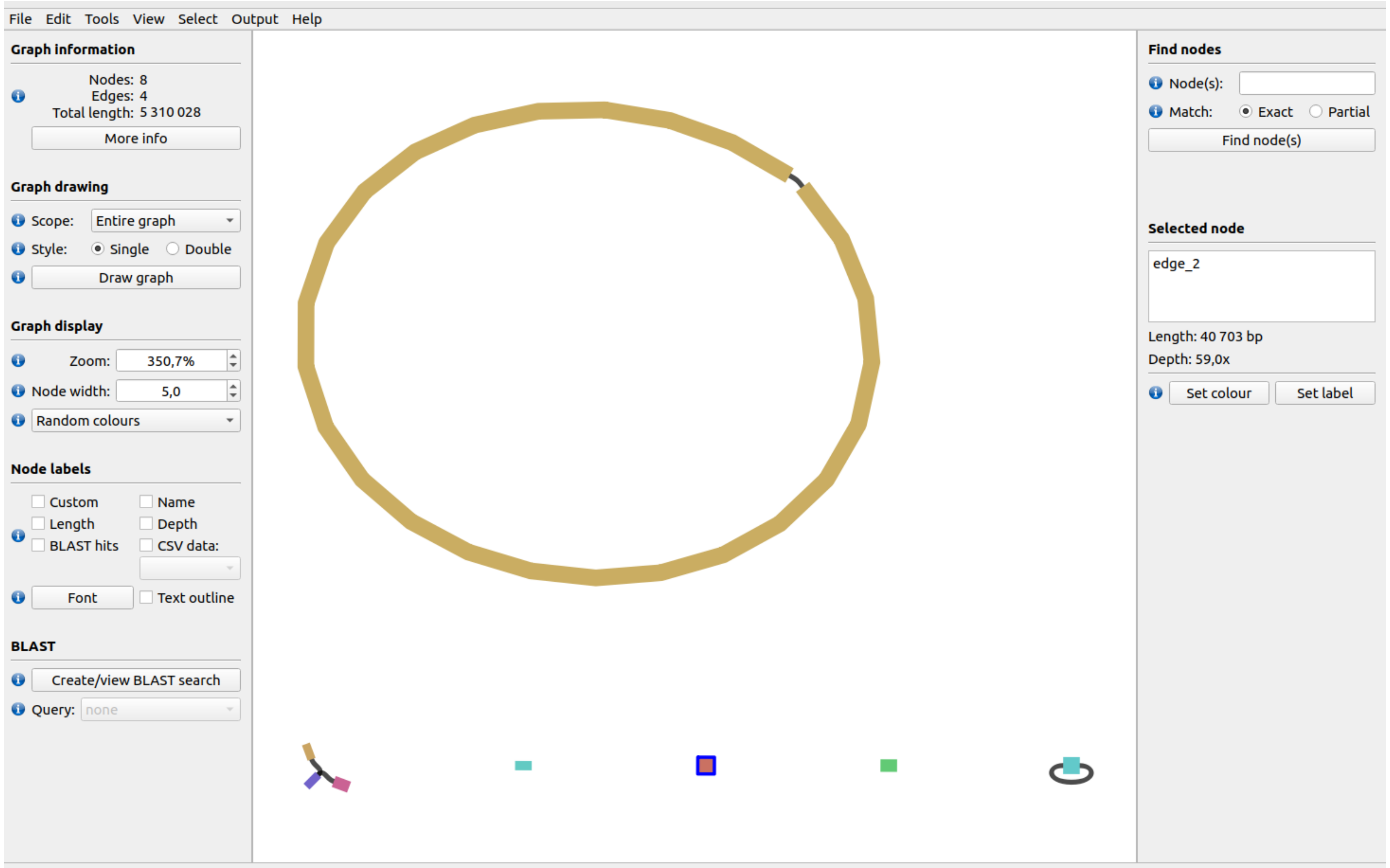
```
00-assembly   30-contigger     assembly.fasta.fai  assembly_info.txt  contig_8.fasta
10-consensus  40-polishing     assembly_graph.gfa  contig_2.fasta     flye.log
20-repeat     assembly.fasta   assembly_graph.gv   contig_7.fasta     params.json
```

See the file "assembly_info.txt" to have more info about the assembled contigs.

4 contigs of different length:

#seq_name	length	cov.	circ.	repeat	mult.	alt_group	graph_path
contig_3	5173999	58	Y	N	1	*	3
contig_2	40703	59	N	Y	1	2	*,2,*
contig_8	36425	63	Y	N	1	*	8
contig_7	13845	15	N	N	15	*	*,7,*

Graph visualization with bandage



https://github.com/rrwick/Bandage/wiki

Assembly quality assessment

Quast:
- html report
- report.txt

QUAST

Quality Assessment Tool for Genome Assemblies

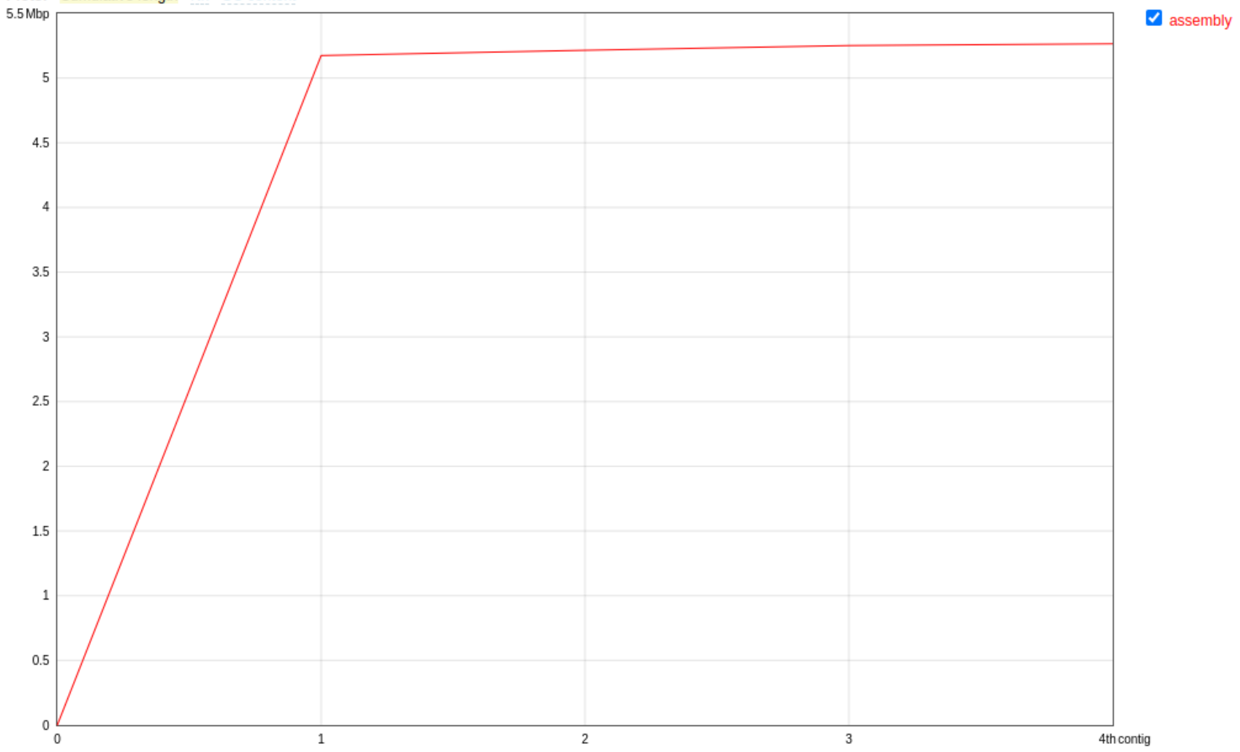
13 March 2025, Thursday, 16:53:56

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Statistics without reference	assembly
# contigs	4
# contigs (>= 0 bp)	4
# contigs (>= 1000 bp)	4
# contigs (>= 5000 bp)	4
# contigs (>= 10000 bp)	4
# contigs (>= 25000 bp)	3
# contigs (>= 50000 bp)	1
Largest contig	5 173 999
Total length	5 264 972
Total length (>= 0 bp)	5 264 972
Total length (>= 1000 bp)	5 264 972
Total length (>= 5000 bp)	5 264 972
Total length (>= 10000 bp)	5 264 972
Total length (>= 25000 bp)	5 251 127
Total length (>= 50000 bp)	5 173 999
N50	5 173 999
N90	5 173 999
auN	5 085 201
L50	1
L90	1
GC (%)	43.11
Per base quality	
# N's per 100 kbp	0
# N's	0

Plots: Cumulative length Nx GC content



Contigs are ordered from largest (contig #1) to smallest.

Busco

```
---------------------------------------------------
|Results from dataset bacteria_odb12               |
---------------------------------------------------
|C:42.2%[S:42.2%,D:0.0%],F:44.8%,M:12.9%,n:116     |
|49     Complete BUSCOs (C)                        |
|49     Complete and single-copy BUSCOs (S)        |
|0      Complete and duplicated BUSCOs (D)         |
|52     Fragmented BUSCOs (F)                      |
|15     Missing BUSCOs (M)                         |
|116    Total BUSCO groups searched                |
---------------------------------------------------
```

```
---------------------------------------------------
|Results from dataset bacteroidales_odb12          |
---------------------------------------------------
|C:35.8%[S:35.2%,D:0.6%],F:45.1%,M:19.1%,n:517     |
|185    Complete BUSCOs (C)                        |
|182    Complete and single-copy BUSCOs (S)        |
|3      Complete and duplicated BUSCOs (D)         |
|233    Fragmented BUSCOs (F)                      |
|99     Missing BUSCOs (M)                         |
|517    Total BUSCO groups searched                |
---------------------------------------------------
```

```
---------------------------------------------------
|Results from dataset bacteroides_odb12            |
---------------------------------------------------
|C:30.2%[S:30.2%,D:0.0%],F:34.5%,M:35.3%,n:1173    |
|354    Complete BUSCOs (C)                        |
|354    Complete and single-copy BUSCOs (S)        |
|0      Complete and duplicated BUSCOs (D)         |
|405    Fragmented BUSCOs (F)                      |
|414    Missing BUSCOs (M)                         |
|1173   Total BUSCO groups searched                |
---------------------------------------------------
```

The more general the database, the more genes are detected in the assembly.

However, for the most accurate assessment, it is best to use the most specific database possible.

https://busco-data.ezlab.org/v5/data/lineages/

In summary

- If the input data is of poor quality, it will negatively affect the assembly results.
- The results obtained with Canu highlight the importance of using multiple assemblers to compare and choose the best assembly.

Acknowledgements

The creation of this training material was commissioned by ECDC to Institut Pasteur with the direct involvement of Rania Ouazahrou

Q&A session