



# Sequencing technologies: from short-reads to long-reads

# Intended Learning Objectives

- **Specific objectives for this session:**

1. What is whole genome sequencing
2. Steps from a sample to raw sequence data
3. Available sequencing technologies
4. Choice of the best sequencing technology for your needs
5. Short- and long-read sequencing

- **Related to other course objectives:**

1. How to generate raw sequence data > Nabil Alikhan for raw data QC
2. How does the technology chosen impact on assembly quality > Nabil Alikhan for assembly and QC

# Outline

## This session consists of the following elements:

1. From sample to raw sequence data
  - Steps to prepare your sample for DNA extraction and sequencing
2. History and evolution of sequencing technologies
  - Three generations of sequencing technologies
  - A focus on Illumina technology
  - Why do long-read sequencing?
  - A focus on Oxford Nanopore sequencing
3. Bioinformatic resources for genome assembly

# Definitions:

- **Sequencing:** determination of the DNA sequence of genes, chromosomes, or even the entire genome
- **Next-generation sequencing (NGS):** new methods for sequencing DNA and RNA faster, revolutionizing genomics and molecular biology



# The advantages of NGS

- Used to assess genomic structure and genome content
- Offers high resolution for genomic comparisons and intraspecies levels
- May contribute to epidemiological studies, including multi-scale transmission
- More molecular details don't necessarily mean more epidemiology
- Comparative genomics can contribute to studies of bacterial evolution, pathophysiology, and host adaptation

# But NGS is not epidemiology!

**Metadata** is important

*Otherwise, genomes become nothing  
more than a collection of stamps!*

**Questions asked**

# **From sample to raw sequence data**

## **PART 1**

# From sample to raw sequence data

## Steps:

- DNA extraction
- DNA quantification
- DNA quality
- Preparing the library
- Sequencing

# From sample to raw sequence data

## Steps:

### 1. DNA extraction

- From a pure culture (clonal population of a single bacterial colony)
- Plate sweep (bacterial population)
- Whole Sample (Metagenomics)
- Extraction Kits

# From sample to raw sequence data

## Steps:

### 2. DNA Quantification

- Sequencing companies ask to subject DNA to a specific concentration
- Library kit requirements
- Qubit or PicoGreen (fluorescence methods)
- Less specific Nano-drop

# From sample to raw sequence data

## Steps:

### 3. DNA quality

- Nano-drop (spectrophotometry)
- Evaluate the purity of samples
- Absorbance at wavelengths 230, 260 and 280 nm
  1. 260/280 ratio: does it have contaminating proteins?
    - DNA: 1.80
    - RNA: 2.00
  2. 260/230 ratio: does it have contaminating reagents (e.g. EDTA)?
    - DNA: >2
    - Contaminants: <2

# From sample to raw sequence data

## Steps

### 4. Preparing the Library

- Different kits available
- The choice depends on the sequencing platform, the starting material (DNA vs. RNA), the size of the genome, etc.
- This involves adding the adapters for DNA sequencing



# From sample to raw sequence data

## Steps:

### 5. Sequencing

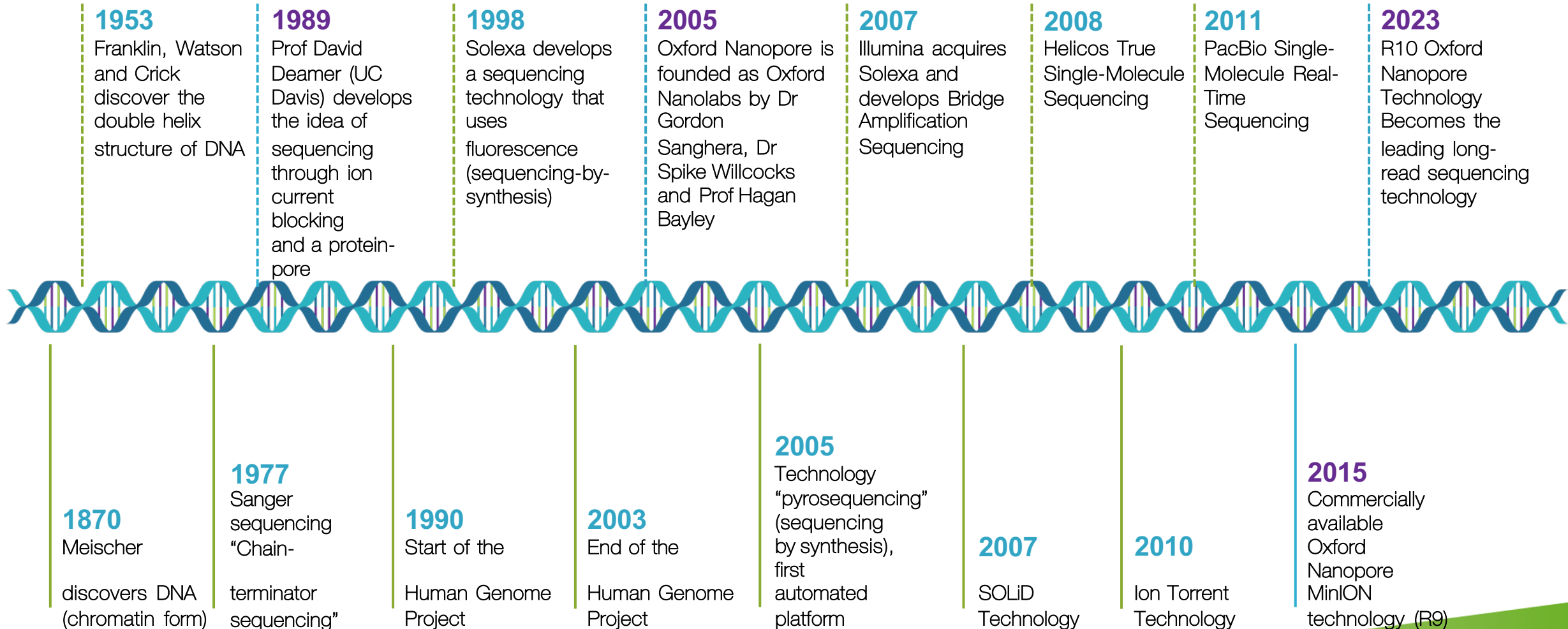
Different technologies:

- First generation (Sanger)
- Second or New Generation (NGS)
- Third generation (single molecule)

# **History and evolution of sequencing technologies**

## **PART 2**

# History and evolution of sequencing technologies



# Three generations of sequencing technologies

## First generation (Sanger)

- The first method for sequencing
- Chain-terminator sequencing
- Amplified models generated *in vivo*
- Requires plasmid preparation

## FIGURES

a)

b)

## FIGURES

### Sanger "Chain-terminator sequencing »

a) DNA to be sequenced

b) DNA polymerase

c) dNTP (deoxy nucleotides of A, C, G, T)

d) ddNTP (**d**ideoxy nucleotides) of a given type (A, C, G, T), labelled by radio or fluorescence, are included in DNA polymerization reactions at low concentrations

## Sanger "Chain-terminator sequencing" »

a) DNA to be sequenced

b) DNA polymerase

c) dNTP (deoxy nucleotides of A, C, G, T)

d) ddNTP (**d**ideoxy nucleotides) of a given type (A, C, G, T), labelled by radio or fluorescence, are included in DNA polymerization reactions at low concentrations

## FIGURES

*these ddNTPs prevent further DNA extension*

# Sanger "Chain-terminator sequencing" >>

e) The result is fragments of different random lengths; The fragments are visualized by electrophoresis on a high-resolution polyacrylamide gel (arranged by size)

c)

## FIGURES

# The Sanger chromatogram

Good

## FIGURES

Bad

Horrible



# Three generations of sequencing technologies

## Second Generation or Next Generation (NGS)

## FIGURES

- No plasmid amplification required
- The model is attached or immobilized on a solid surface or stand for amplification (flow cell)
- Numerous reactions in parallel
- Short-reads (150-300 bp)
- High accuracy (>99.9%)

# Three generations of sequencing technologies

## Second Generation or Next Generation (NGS): Technology Comparison

Technology	Roche 454 pyrosequencing	SOLiD	Ion Torrent semi-conductor	Illumina Hi/Mi-Seq
amplification	bead-support; emersion oil ("emulsion")	bead-support; emersion oil ("emulsion")	bead-support; emersion oil ("emulsion")	« bridge amplification »
sequencing	By synthesis	By ligation	By synthesis	By synthesis
detection	Light intensity (luciferase), flow cell with one dNTP added at a time (A, C, G, T)	Fluorescence	Ion sensors capture the proton released during nucleotide incorporation	Fluorescence

**A focus on Illumina technology**

**VIDEO**

**A focus on Illumina technology**

**FIGURES**

**A focus on Illumina technology**

**FIGURES**

# **A focus on Illumina technology**

## **FIGURES**

**Result: short-reads**

# Limitations of short-read sequencing

- Problems assembling short-reads with repeated sequences

# Limitations of short-read sequencing

- Problems assembling short-reads with repeated sequences

## FIGURES

Repeated regions



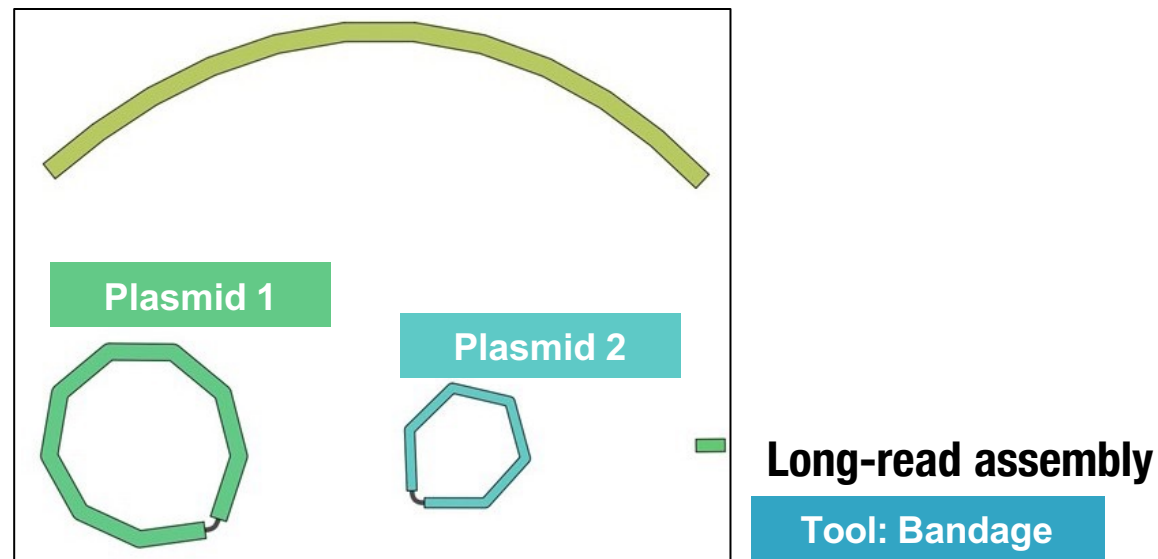
# Limitations of short-read sequencing

- Problems assembling short-reads with repeated sequences

## FIGURES

# Limitations of short-read sequencing

- Problems assembling short-reads with repeated sequences
- Assembly of plasmids or other extra-chromosomal elements (e.g. integrative conjugative elements, ICE) is mostly not possible with short-reads



# How to get the best genome assembly?

## Short-read assembly

- high accuracy
- genome  
assembly  
fragmentation

## Long-read assembly

- variable accuracy  
(technology, kit etc.)
- complete  
genome  
assemblies
- complete assembly  
of  
extra-chromosomal  
elements

## Hybrid assembly

- high accuracy
- complete  
genome  
assemblies
- complete  
assembly of extra-  
chromosomal  
elements

# Three generations of sequencing technologies

PacBio

## FIGURES

Oxford Nanopore

### Third generation (single molecule)

- Single molecule model
- Long-reads
- Real-time sequencing
- Portability
- Low price
- Quick options for library preparation
- Now 99% > accuracy

*Loman & Pallen,  
2015*

**A focus on Oxford Nanopore Technology (ONT)**

**VIDEO**

# Stages and evolution of Nanopore technology



Chemistry

## FIGURES

Platforms

*Wang et al.,  
2021*

## FIGURES

### MinION Family: Mk1B

- First Nanopore platform
- On the market since 2015
- Small, portable, connected to computer with a USB port
- Weight: 87g!
- \$900 per flowcell (less if ordered >12 flowcells)
- Standard run: 72h
- Reusable flowcells!
- Live basecalling possible

*nanoporetech.com,  
2023*

# MinION: portability

## FIGURES

Peru

Ecuador

Antartica

Brazil

Argentina

Iceland



## FIGURES

*nanoporetech.com,  
2023*

### MinION Family: Mk1C

- Evolution of the MinION Mk1B
- Small, portable
- Integrated computer with GPU
- Live basecalling
- Live barcoding
- On the market in 2022-2023, now discontinued

# Platforms



*Ph: C. Crestani,  
2023*



# Platforms



Ph: C. Crestani,  
2023

## FIGURES

### GridION

- Since 2017
- For the bench, not suitable for the field
- Integrated computer with GPU
- Live basecalling
- Live barcoding
- 5 flowcells (MinION) at the same time
- Standard run: 72h

## FIGURES

### PromethION: P2 solo

- Since the end of 2022
- 2 flowcells (PromethION) at the same time
- Sequencer (need to be connected to a computer for basecalling and data barcoding)
- Standard run: 72h
- For large sequencing projects

## FIGURES

### PromethION

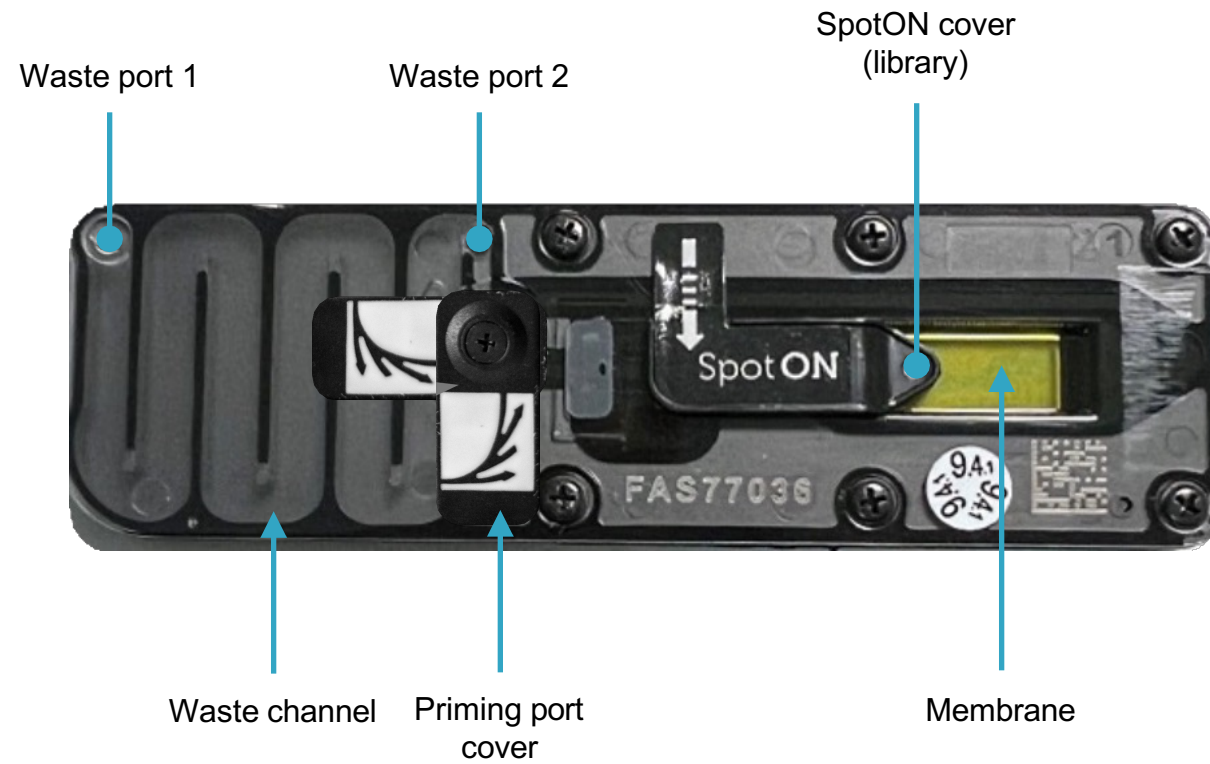
- Since 2018
- 24/48 flowcells (PromethION) at the same time
- Standard run: 72h
- For large sequencing projects (need to multiplex)

*nanoporetech.com,  
2023*

# Platforms

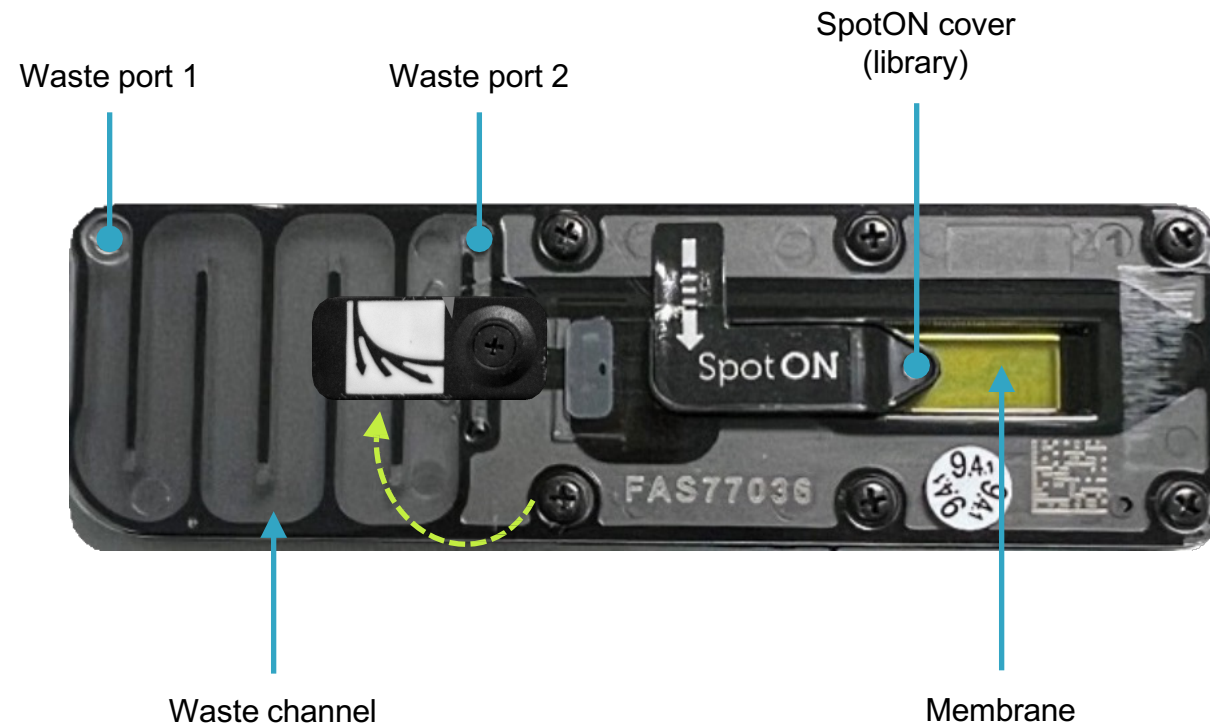
# FIGURES

# Anatomy of a MinION flowcell

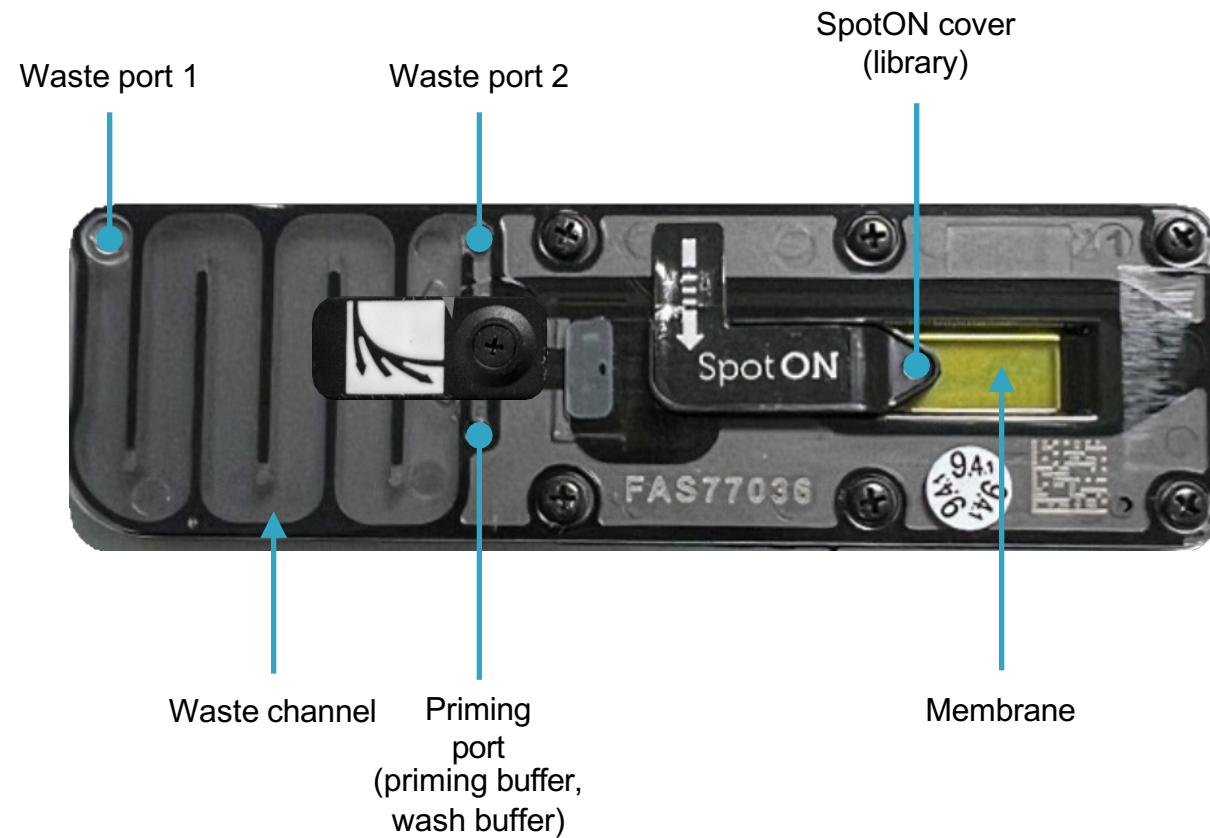




# Anatomy of a MinION flowcell



# Anatomy of a MinION flowcell



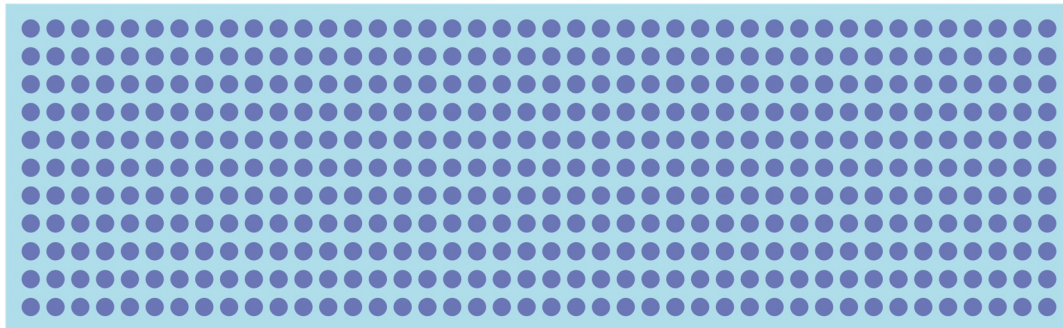
# Anatomy of a MinION flowcell



**ASIC**  
Application-specific  
integrated circuit

# Anatomy of a MinION flowcell

Synthetic membrane



Nanopores

- 512 channels
- 4 nanopores/channel
- 2048 total pores (max) - importance of quality control when receiving flow cells

# ONT MinION: the nanopores

- Ionic current disruption
- Time + disrupted current = signal
- Speed: 450bp/s for R9; 400bp/s [or 260bp/s] for R10.4.1
- Output: fast5 or pod5 (for basecalling)

## FIGURES

# FIGURES

*Wang et al.,  
2021*

# Stages and evolution of Nanopore technology

## FIGURES

2022-2023 : R10.4.1

Chemistry

Platforms

*Wang et al.,  
2021*

# Nanopore chemistry

Evolution of:

## 1) Nanopores

- **$\alpha$ -hemolysin** (diameter ~1.4-2.4 nm) *Staphylococcus* membrane channel protein, first nanopore; Increased accuracy: genetic modification of  $\alpha$ -hemolysin WT to be able to better recognize the four oligonucleotides
- **MspA** (*Mycobacterium smegmatis* porin A) (diameter ~1.2 nm): results similar to  $\alpha$ -hemolysin

## 2) Motor protein

- **phi29 DNA polymerase**: better performance of DNA translocation through the nanopore (control of DNA translocation speed, signal enhancement)



# Nanopore chemistry: R9.4 vs R10.4.1



## FIGURES

*nanoporetech.com,  
2023*

# Nanopore chemistry: R9.4 vs R10.4.1

## R9.4

- Nanopore: CsgG, curlin sigma S-dependent growth subunit G from *Escherichia coli*
- New motor protein (E8) vs. R9 (E7) (secret origin, increased accuracy)
- Translocation Speed: 450 bp/s
- Sequencing accuracy: ~85-94%

## R10.4.1

- Nanopore: double reader head(a "wider" window) to improve the signal accuracy of homopolymer regions
- Novel motor protein (E8.2.1) (V14 Sequencing Kits)- higher accuracy and raw read quality
- Translocation speed: 400 bps (>quantity) [or 260 bps (>quality)]
- Sequencing accuracy: 99.6% (simplex), 99.92% (duplex)
- Raw data either in fast5 (Guppy basecaller) or in pod5 (Dorado basecaller)

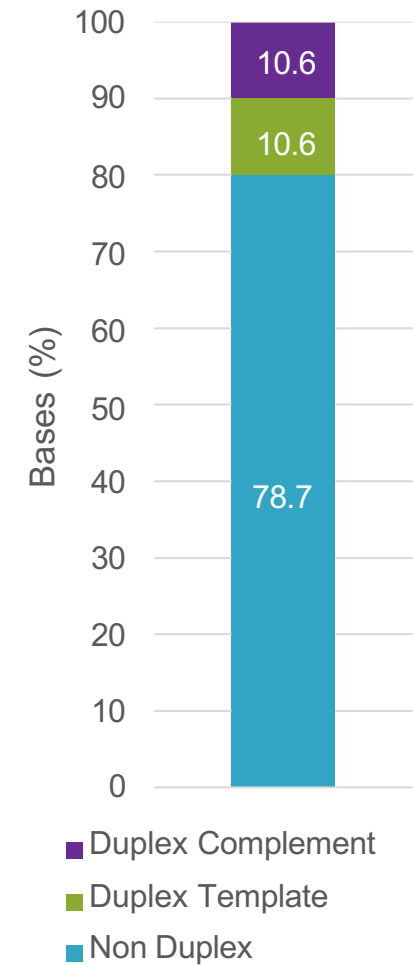
# Nanopore chemistry: R9.4 vs R10.4.1

## FIGURES

# Nanopore chemistry: R9.4 vs R10.4.1

## FIGURES

Sequencing  
accuracy: 99.6%  
(simplex)  
vs  
99.92% (duplex)



# Improving the quality of your long-read genomes

## Factors external to the lab:

- Nanopore Chemistry: novel nanopores and motor proteins (R9.4 vs. R10.4.1)



## Factors internal to the lab:

- DNA quality (long fragments, contaminants)
- Choice of flowcells and kit (e.g. more precise Ligation Kit from the Rapid Barcoding Kit – still true?)
- Choice of translocation speed (R10.4.1 [260 bps >quality or] 400 bps >quantity)
- Amount of data generated (minimum coverage! At least 40X)
- Choice of basecalling tool (Guppy standard; new Dorado)
- Choice of algorithm for basecalling (HAC or SUP)
- Optional step of filtering reads according to their quality (Qscore), choice of tool
- Choice of assembly tool (e.g. Unicycler, Flye, other)
- Optional polishing step (e.g. Racon, Medaka; be careful, Medaka can introduce errors!)

# MinION: the cheapest sequencing technology



## FIGURES

# **Bioinformatic analyses**

## **PART 3**

# Command line

```
ccrestan — ccrestan@maestro-submit:/pasteur/zeus/projets/p01/Corynebacterium-ngs/ChiaraCrestani/Nanopore/20230405_Kleb_R10_NCC — One Dark — 164x31
[(base) ccrestan ~ >> ssh ccrestan@maestro.pasteur.fr
ccrestan@maestro.pasteur.fr's password:
Permission denied, please try again.
ccrestan@maestro.pasteur.fr's password:

Nanopore Submit

Storage
=====

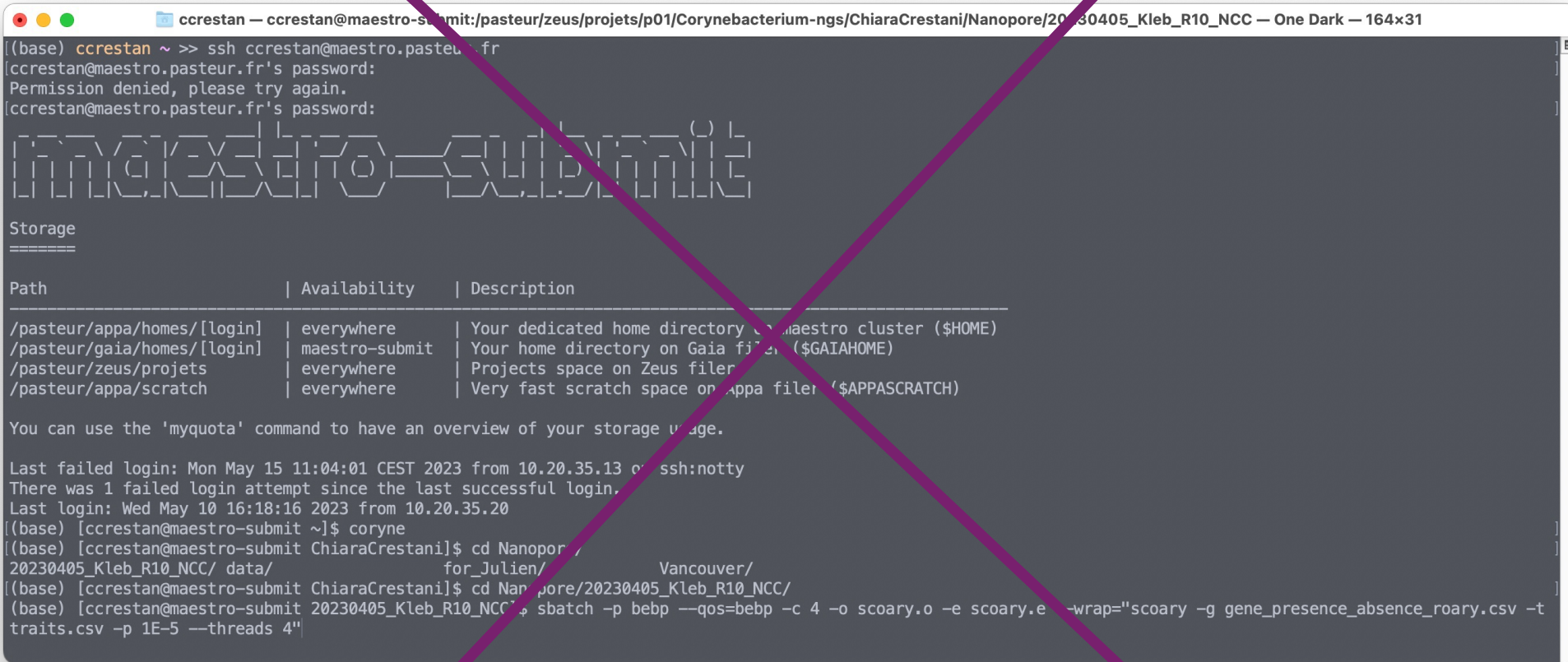
Path | Availability | Description
-----
/pasteur/appa/homes/[login] | everywhere | Your dedicated home directory on maestro cluster ($HOME)
/pasteur/gaia/homes/[login] | maestro-submit | Your home directory on Gaia filer ($GAIAHOME)
/pasteur/zeus/projets | everywhere | Projects space on Zeus filer
/pasteur/appa/scratch | everywhere | Very fast scratch space on Appa filer ($APPASCRATCH)

You can use the 'myquota' command to have an overview of your storage usage.

Last failed login: Mon May 15 11:04:01 CEST 2023 from 10.20.35.13 on ssh:notty
There was 1 failed login attempt since the last successful login.
Last login: Wed May 10 16:18:16 2023 from 10.20.35.20
[(base) [ccrestan@maestro-submit ~]$ coryne
[(base) [ccrestan@maestro-submit ChiaraCrestani]$ cd Nanopore/
20230405_Kleb_R10_NCC/ data/ for_Julien/ Vancouver/
[(base) [ccrestan@maestro-submit ChiaraCrestani]$ cd Nanopore/20230405_Kleb_R10_NCC/
(base) [ccrestan@maestro-submit 20230405_Kleb_R10_NCC]$ sbatch -p bebp --qos=bebp -c 4 -o scoary.o -e scoary.e --wrap="scoary -g gene_presence_absence_roary.csv -t
traits.csv -p 1E-5 --threads 4"
```



# Command line



```
ccrestan — ccrestan@maestro-submit:/pasteur/zeus/projets/p01/Corynebacterium-ngs/ChiaraCrestani/Nanopore/20230405_Kleb_R10_NCC — One Dark — 164x31
(base) ccrestan ~ >> ssh ccrestan@maestro.pasteur.fr
ccrestan@maestro.pasteur.fr's password:
Permission denied, please try again.
ccrestan@maestro.pasteur.fr's password:

[base] ccrestan@maestro-submit ~$ coryne

Storage
=====

Path | Availability | Description
-----|-----|-----
/pasteur/appa/homes/[login] | everywhere | Your dedicated home directory on Maestros cluster ($HOME)
/pasteur/gaia/homes/[login] | maestro-submit | Your home directory on Gaia file server ($GAIAHOME)
/pasteur/zeus/projets | everywhere | Projects space on Zeus file server
/pasteur/appa/scratch | everywhere | Very fast scratch space on Appa file server ($APPASCRATCH)

You can use the 'myquota' command to have an overview of your storage usage.

Last failed login: Mon May 15 11:04:01 CEST 2023 from 10.20.35.13 on ssh:notty
There was 1 failed login attempt since the last successful login.
Last login: Wed May 10 16:18:16 2023 from 10.20.35.20
(base) [ccrestan@maestro-submit ~]$ coryne
(base) [ccrestan@maestro-submit ChiaraCrestani]$ cd Nanopore/20230405_Kleb_R10_NCC/
20230405_Kleb_R10_NCC/ data/ for_Julien/ Vancouver/
(base) [ccrestan@maestro-submit ChiaraCrestani]$ cd Nanopore/20230405_Kleb_R10_NCC/
(base) [ccrestan@maestro-submit 20230405_Kleb_R10_NCC]$ sbatch -p bebp --qos=bebp -c 4 -o scoary.o -e scoary.e --wrap="scoary -g gene_presence_absence_roary.csv -t
traits.csv -p 1E-5 --threads 4"
```

# **Bioinformatic resources**

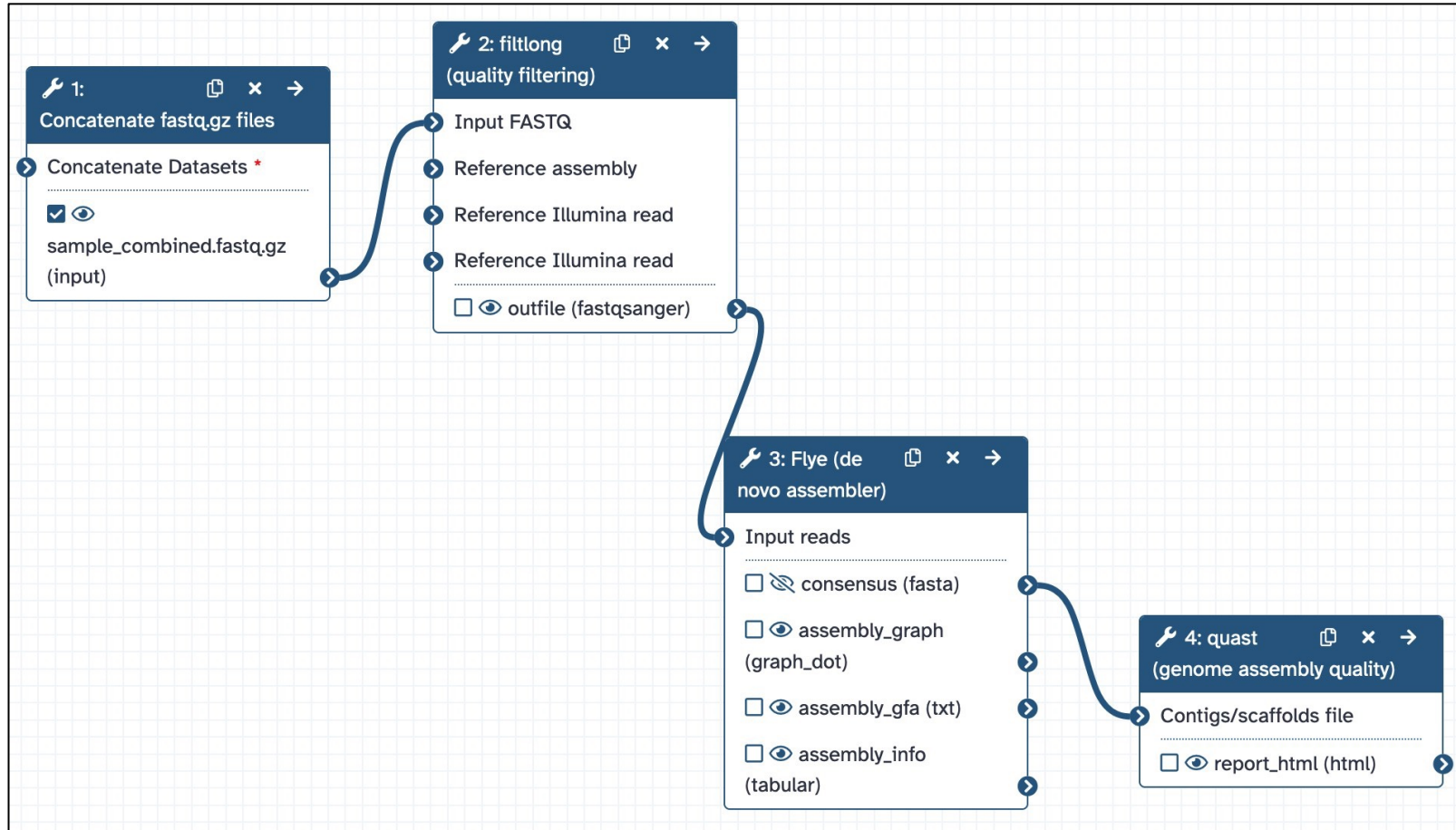
# Bioinformatic resources



## Galaxy

- Free tool
- Browser application (Firefox, Safari, Google Chrome, etc.)
- Private account
- Upload/download sequencing data
- Access to bioinformatics tools for analyses
- Ability to create workflows (pipelines) for frequent or repeated scans
- Problems: internet connection and speed depending on server queue

# Galaxy: platform for online bioinformatic analyses



# Galaxy: platform for online bioinformatic analyses

The screenshot displays the Galaxy web interface. At the top, the header includes the Galaxy logo, the text "Galaxy / Africa Europe", and navigation links for "Workflow", "Visualize", "Shared Data", "Help", "User", and a grid icon. A status bar on the right indicates "Using 0%".

On the left, a "Tools" sidebar contains a search bar, an "Upload Data" button, and a list of tool categories: "Get Data", "Send Data", "Collection Operations", "GENERAL TEXT TOOLS", "Text Manipulation", "Convert Formats", "Filter and Sort", "Join, Subtract and Group", "GENOMIC FILE MANIPULATION", "Convert Formats", "FASTA/FASTQ", "Quality Control", "SAM/BAM", "BED", "VCF/BCF", and "Nanopore".

The main workspace features a green notification banner with a checkmark icon, stating: "Successfully invoked workflow Nanopore assembly pipeline . You can check the status of queued jobs and view the resulting data the History panel." Below this, a progress bar shows "4 of 4 steps successfully scheduled." and "1 of 4 jobs complete....".

On the right, the "History" panel displays a search bar and a list of datasets. The "Unnamed history" section shows a list of datasets with their sizes, IDs, and icons for viewing, editing, and deleting. The datasets listed are:

- 366 MB, ID 33, 265, 7, 7
- sembly info
- 297 : Flye on data 294: graphical fragment assembly
- 296 : Flye on data 294: assembly graph
- 295 : Flye on data 294: consensus
- 294 : fittlong on data 293: Filtered FASTQ
- 293 : sample\_combined concatenated
- 285 : fastq\_runid\_4f12a6a1ffcfc3f5a1772ff52994d7811f551fae\_25.fastq.gz

The URL at the bottom of the browser window is <https://africa.usegalaxy.eu/datasets/4838ba20a6d86765147a4546efb75c72/edit>.

# Bioinformatic resources



## EPI2ME

- Tool developed by Oxford Nanopore (free)
- Workflows (pipelines) already available (e.g. wf-bacterial-genomes for genome assembly with Flye)
- Browser application (Firefox, Safari, Google Chrome, etc.)
- Downloadable application:
  1. EPI2ME Command Line (cloud-based analytics)
  2. EPI2ME Desktop (cloud-based analytics)
    - Analytics on Amazon Web Services (AWS) server, paid for by Nanopore
    - Issues: speed based on server queue; the wf-bacterial-genomes workflow is not available
  3. EPI2ME Labs (local analyses)
    - Problem: Requires a very powerful computer (RAM, GPU, etc.)
- Future developments planned by Nanopore

# FIGURES

## And then...

## Institut Pasteur GenEpi-BioTrain teaching staff

Sylvain Brisse  
Carla Parada-  
Rodrigues José  
Delgado-Blas Carolina  
Silva-Nodari François-  
Xavier Weill Alexandra  
Moura  
Marc Lecuit  
Julien Guglielmini  
Fabien Mareuil  
Remi Planel

## GenEpi-BioTrain teaching staff from other Institutions

Nabil-Fareed  
Alikhan Martin  
Maiden Nathalie  
Jourdan François  
Lebreton Alessandra  
Carattoli Gabriele  
Arcari  
Mathieu  
Tourdjman  
Eleonora Sarno  
Mirko Rossi  
Cecilia Jernberg  
Priyanka  
Nannapaneni



# Acknowledgements

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