

Initiation à l'analyse de données Oxford Nanopore

Rabat, Novembre 2022



Institut de Recherche
pour le Développement
FRANCE



Université Mohammed V
Faculté des Sciences
Rabat

Bioinformatics resources

On va travailler sous Linux !

- 2 façons d'utiliser linux :

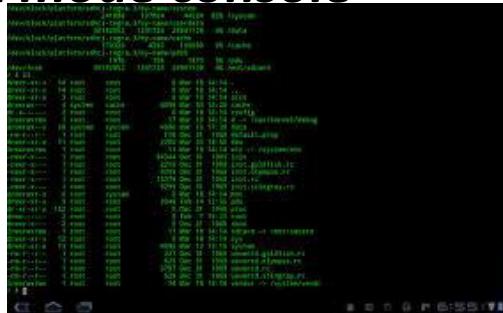
en *mode graphique*



En mode terminal

- 2 façons d'utiliser linux :

en *mode console*



avant tout !

Bases de Linux

https://github.com/SouthGreenPlatform/training_NT_teaching/Topo/linux.pdf

En mode jupyter book

- Une troisième façon d'utiliser linux :

en *mode jupyter book*



Sur le cloud IFB!



Let's discover Jupyter !

Working environment

What is jupyter book ?

- One of the most popular tool among data scientists to perform data analysis
- Provides a complete environment in which numerous programming languages can be used through a simple web browser

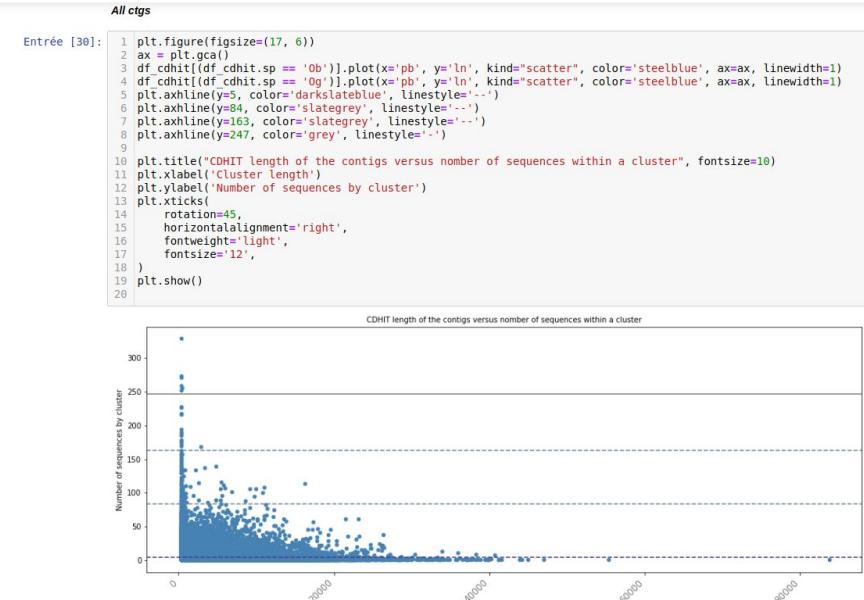
ex : Bash (Linux), Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala



Why use jupyter book ?

An unique interface/file where text,code and output codes can be mixed :

- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook



Why use jupyter book ?

An unique interface/file where text,code and output codes can be mixed :

- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook
- explanations, formulas, charts can be added

The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** jupyter parseCistr-Copy1 Dernière Sauvegarde : Il y a 8 minutes (auto-sauvegarde) Se déconnecter Python 3 O
- Toolbar:** Fichier Édition Affichage Insérer Cellule Noyau Widgets Aide
- Cell Content:**
 - Section Header:** Anchoring data analysis
 - Section 1:** 1 - CDHIT data analysis *before anchoring on genome*
 - Section 1.1:** 1.1 Removing redundancy with CDHIT
 - CDHIT Input : 1,306,676 contigs assembled from no mapped reads
 - Tests & results
 - Table:** A table showing cluster counts for different similarity thresholds (0.80 to 0.95).

	0.9	0.95
0.80	378,615	484,394
0.85	418,136	531,326
0.90	473,270	588,983
0.95	544,441	659,658
 - clusters generated after cdhit analysis : 484,394
 - Section 1.2:** 1.2 Converting cdhit file into a csv loaded as a dataframe with pandas
 - Text:** The script cdhitVsAnchoring.py creates the csv file allCtgtsIRIGIN_TOG5681.dedup8095.PANDAS.csv
 - Section 2:** Load csv file into a pandasframe
 - Code Cell [1]:** Entrée [1]:

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
4
5 csv_cdhit_file = "/home/christine/Documents/These/Data/CDHIT/ALL_CGTGS_MERGE/allCtgtsIRIGIN_TOG5681.dedup8095.PANDAS.csv"
6 df_cdhit= pd.read_csv(csv_cdhit_file,names=['ctg','sp','ctg-list','sp_list'], header=0)
7 #print(df_cdhit)
8
```

Lab notebook for science data ?

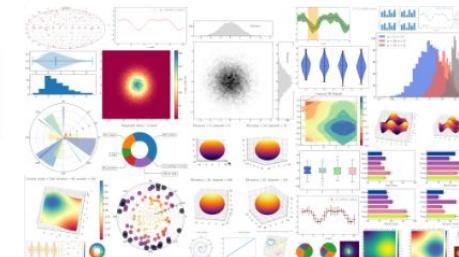
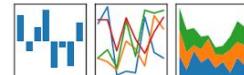


- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

How to become a super datascientist ?

- facilement importer des fichiers tabulés dans des dataframes, similaires aux dataframes sous R.
(et exporter)
- manipuler ces tableaux de données / DataFrames
- facilement tracer des graphes à partir de ces DataFrames grâce à matplotlib

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



How will you use Jupyter Notebook ?

- Launch our analyses through a jupyter book within a virtual machine launched via the IFB cloud “BIOSPHERE”



How will you use Jupyter Notebook ?

- Launch our analyses through a jupyter book within a virtual machine launched via the IFB cloud “BIOSPHERE”
- Through this virtual machine, we will create jupyter books and execute all our analysis

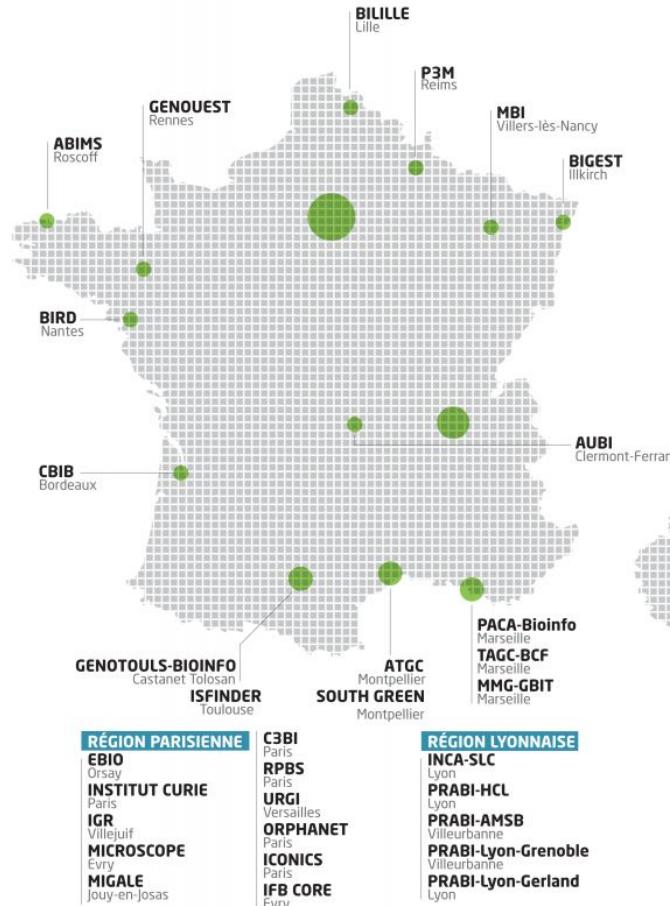


The screenshot shows a web browser window for the IFB Cloud. The address bar displays "IFB Cloud" and "mydatalocal/" with the URL "https://134.158.247.8/tree/mydatalocal". The main content area is titled "jupyter" and contains three tabs: "Files", "Running", and "Clusters". Under "Files", there is a message "Select items to perform actions on them." Below it is a file tree showing a folder named "mydatalocal". A message at the bottom says "La liste des notebooks est vide." To the right, there is a "Notebook" section with a dropdown menu showing options like "Bash", "Julia 1.5.3", "Python 3", and "R". Below this, under "Other:", are "Text File", "Folder", and "Terminal". A "New" button is visible in the top right of the "Notebook" section. The browser interface includes standard controls like back, forward, and search.

IFB ?



22 plateformes-membres
7 plateformes contributrices
8 équipes associées
>400 experts (~200 FTE)

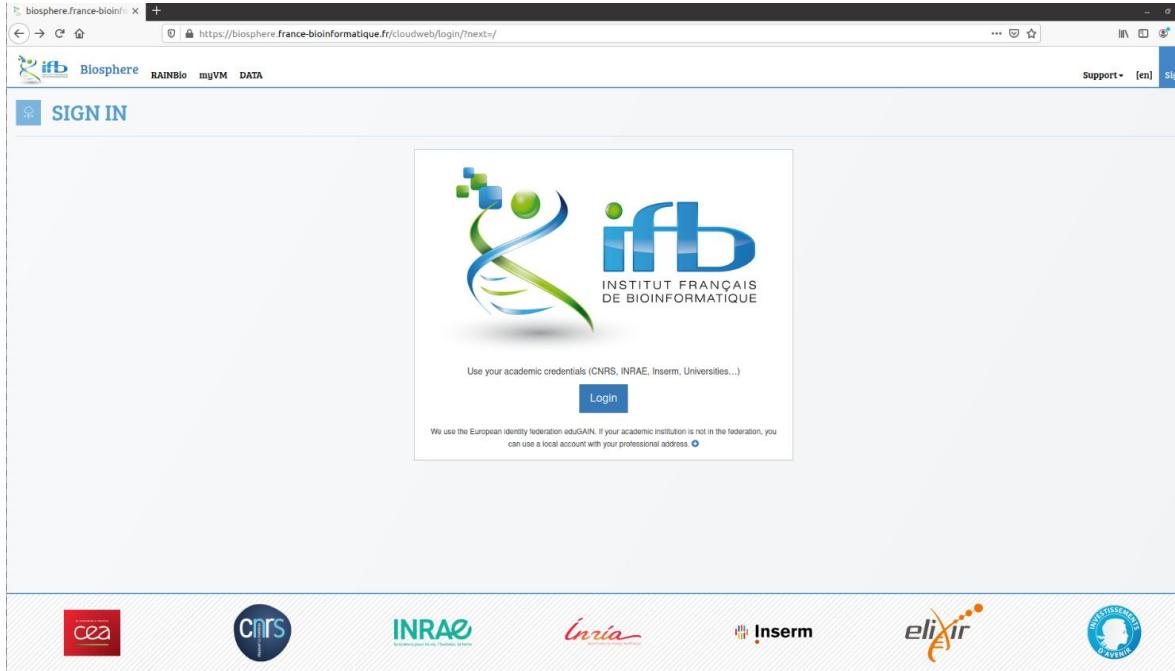


Biosphere, IFB CLOUD FOR LIFE SCIENCES

- A federation of clouds, which relies on interconnected IFB's infrastructures, providing distributed services to analyze life science data
- Access to a large set of virtual machines (computing ressources, bioinformatics tool)
- Used for scientific production in the life sciences, developments, and also to support events like cloud and scientific training sessions, hackathons or workshops.

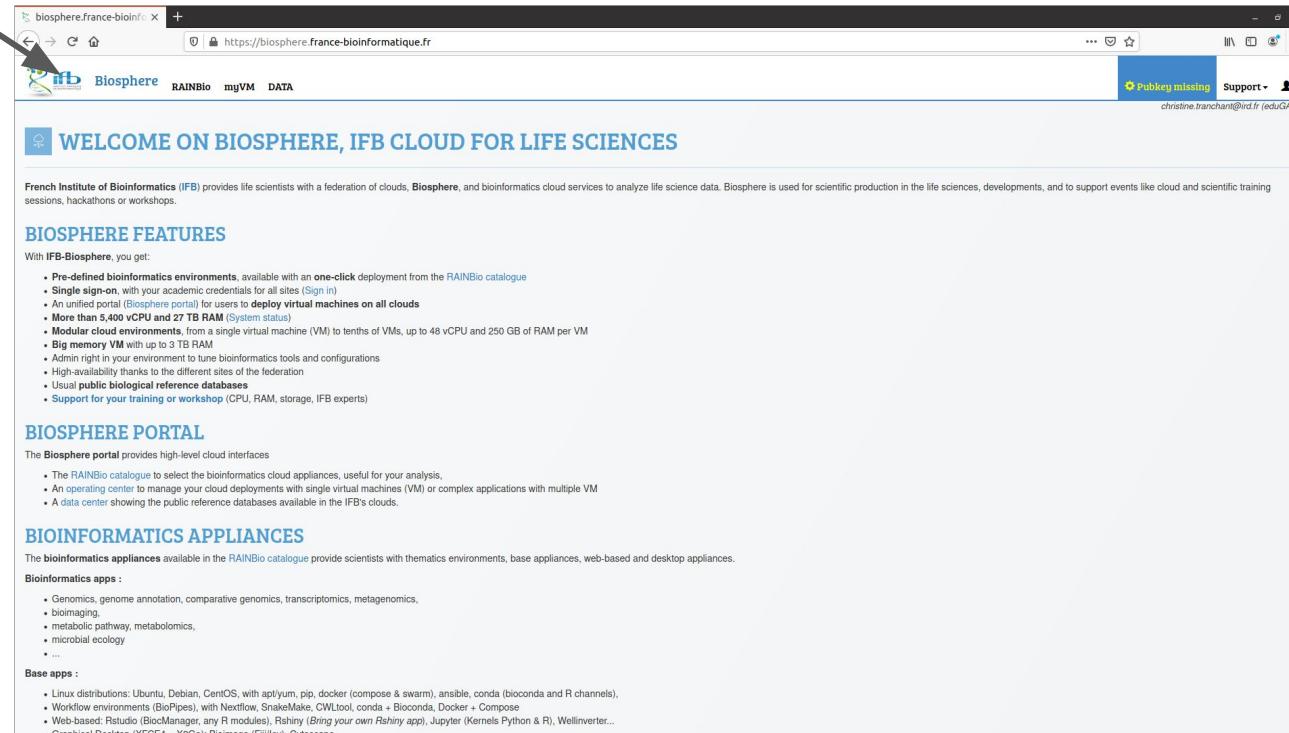
Let's start with biosphere

- Open the biosphere website : <https://biosphere.france-bioinformatique.fr/cloud/> and sign in



Connected / here we are

RAINBIO catalog to access our Virtual Machine (VM)



The screenshot shows a web browser window with the URL <https://biosphere.france-bioinformatique.fr>. The page title is "WELCOME ON BIOSPHERE, IFB CLOUD FOR LIFE SCIENCES". The page content includes sections on Biosphere Features, Biosphere Portal, and Bioinformatics Appliances, each listing various cloud services and tools available through the RAINBio catalogue.

BIOSPHERE FEATURES

With IFB-Biosphere, you get:

- Pre-defined bioinformatics environments, available with an one-click deployment from the RAINBio catalogue
- Single sign-on, with your academic credentials for all sites ([Sign in](#))
- An unified portal (Biosphere portal) for users to [deploy virtual machines on all clouds](#)
- More than 5.400 vCPU and 27 TB RAM ([System status](#))
- Modular cloud environments, from a single virtual machine (VM) to tenths of VMs, up to 48 vCPU and 250 GB of RAM per VM
- Big memory VM with up to 3 TB RAM
- Admin right in your environment to tune bioinformatics tools and configurations
- High-availability thanks to the different sites of the federation
- Usual public biological reference databases
- Support for your training or workshop (CPU, RAM, storage, IFB experts)

BIOSPHERE PORTAL

The Biosphere portal provides high-level cloud interfaces

- The RAINBio catalogue to select the bioinformatics cloud appliances, useful for your analysis.
- An operating center to manage your cloud deployments with single virtual machines (VM) or complex applications with multiple VM
- A data center showing the public reference databases available in the IFB's clouds.

BIOINFORMATICS APPLIANCES

The **bioinformatics appliances** available in the RAINBio catalogue provide scientists with thematic environments, base appliances, web-based and desktop appliances.

Bioinformatics apps :

- Genomics, genome annotation, comparative genomics, transcriptomics, metagenomics,
- biomining,
- metabolic pathway, metabolomics,
- microbial ecology
- ...

Base apps :

- Linux distributions: Ubuntu, Debian, CentOS, with apt/yum, pip, docker (compose & swarm), ansible, conda (bioconda and R channels),
- Workflow environments (BioPipes), with Nextflow, SnakeMake, CWLtool, conda + Bioconda, Docker + Compose
- Web-based: Rstudio (BioManager, any R modules), Rshiny (Bring your own Rshiny app), Jupyter (Kernels Python & R), Wellinverter...
- Graphical Desktop (XFCE4, Xfce, Bioimagine, Fiji/lov, Cytoscape)

Searching for the vm we will use

vm's name :

CoursAnalysesNanoporeSG

 **RAINBIO - APPLIANCES BIOINFORMATIQUES DANS LE CLOUD**

Catalogue des appliances bioinformatiques dans le cloud, filtrez-les en utilisant les termes présents dans l'ontologie EDAM, ou en langage naturel.

App Store (58) Appliances Outils Topics Appliance éditable Ajouter ⚙️

CoursAnalysesNanoporeSG

- ★ bcftools, BEDTools, BWA, Jupyter, Matplotlib, pandas, SAMtools
- ⚡ DNA polymorphism, Genetic variation, Genotyping experiment, GWAS study
- 🔧 Data architecture, analysis and design, Mathematics, Statistics and probability

AnalysesSV

- ★ bcftools, BEDTools, BWA, Jupyter, Matplotlib, pandas, SAMtools
- ⚡ DNA polymorphism, Genetic variation, Genotyping experiment, GWAS study

virus_ONT

- ★ Jupyter
- ⚡ Data architecture, analysis and design, Mathematics, Statistics and probability

ANF MetaBioDiv

- ★ DESeq2, ggplot2, phyloseq, RStudio
- ⚡ Transcriptomics, Microbiology, Metagenomics, Sequence analysis

Let's run your vm through the cloud

The screenshot shows the IFB Biosphere platform interface. At the top, there are links for RAINBio, myVM, and DATA. On the right, there are user profile and support links. The main area displays a virtual machine configuration for "CoursAnalysesNanoporeSG".
Description: VM used for train scientists and students from Burkina Faso and West Africa in bioinformatics analysis of data from Oxford nanopore sequencing technology with main of study viral métagenome.
Domaines associés: Computational biology, Sequence analysis.
Outils:

Jupyter	
OS	Debian 11
Recette de l'app (git)	https://github.com/SouthGreenPlatform/training_ONT_VM/tree/2022
App de base	Jupyter

Caractéristiques:

Nom long	VM used for analyse metagenomic of viruses
Version	1.0

A large dashed arrow points from the green "LANCEMENT" button in the top right corner to the green "DÉPLOIEMENT AVANCÉ" button in the "Outils" section.

Let's run your vm through the cloud

The screenshot shows the IFB Biosphere interface for deploying a virtual machine (VM). The main title is "CoursAnalysesNanoporeSG". The deployment configuration window is open, titled "Configurer le déploiement d'une appliance". The sub-titile is "Déploiement de l'appliance 'virus_ONT'". The "Name" field contains "Julie_ONT". The "Groupe à utiliser" dropdown is set to "virus_ont (Initiation à l'analyse de la séquençage de virus)" and the "tagé nome viraux" dropdown is set to "828.01". The "Cloud" dropdown is set to "ifb-core-cloudbis". The "Gabarit d'image cloud" dropdown is expanded, showing various options. An arrow points to the "ifb.m4.2xlarge (8 vCPU, 32Go GB RAM, 200Go GB local disk)" option, which is highlighted with a blue selection bar. A tooltip above the dropdown asks "Quelle gabarit d'image doit être utilisé sur ce cloud ?". The background shows the IFB Biosphere dashboard with tabs like RAINBio, myVM, and DATA, and a user profile for julie.orjuela@ird.fr.

Description

VM used for train scientists and students from Burkina Faso and West Africa sequencing technology with main of study viral métagenome.

Domaines associés

Computational biology

Sec

Annuler

Julie_ONT

virus_ont (Initiation à l'analyse de la séquençage de virus) tagé nome viraux 828.01

ifb-core-cloudbis

ifb.m4.large (2 vCPU, 8Go GB RAM, 50Go GB local disk)

ifb.m4.large (2 vCPU, 8Go GB RAM, 50Go GB local disk)

ifb.m4.xlarge (4 vCPU, 16Go GB RAM, 100Go GB local disk)

ifb.m4.2xlarge (8 vCPU, 32Go GB RAM, 200Go GB local disk) **ifb.m4.2xlarge (8 vCPU, 32Go GB RAM, 200Go GB local disk)**

ifb.m4.4xlarge (16 vCPU, 64Go GB RAM, 400Go GB local disk)

ifb.xt.e.4xlarge (BigMem) (16 vCPU, 384Go GB RAM, 600Go GB local disk)

ifb.m4.6xlarge (24 vCPU, 96Go GB RAM, 600Go GB local disk)

ifb.m4.8xlarge (32 vCPU, 128Go GB RAM, 800Go GB local disk)

ifb.xt.e.8xlarge (BigMem) (32 vCPU, 768Go GB RAM, 600Go GB local disk)

ifb.m4.12xlarge (48 vCPU, 192Go GB RAM, 1.2To GB local disk)

ifb.xt.e.12xlarge (BigMem) (48 vCPU, 1.1To GB RAM, 50Go GB local disk)

ifb.m4.14xlarge (56 vCPU, 240Go GB RAM, 1.4To GB local disk)

ifb.xt.e.16xlarge (BigMem) (62 vCPU, 1.5To GB RAM, 1.5To GB local disk)

ifb.xt.e.32xlarge (BigMem) (124 vCPU, 2.9To GB RAM, 2.9To GB local disk)

EDITER LANCER DÉPLOIEMENT AVANCÉ

VM/tree/2022

Let's run your vm through the cloud

Loading...

The screenshot shows the RAINBio interface with the following components:

- Top Bar:** IFB Biosphère, RAINBio, myVM, DATA, Support (with email julie.orjuela@ird.fr), and a user icon.
- Left Sidebar:** CLOUD
- Main Content:** A table titled "Déploiements" (Deployments) showing two entries:

ID	Nom	Début	Groupes	Spécification	Broker	Cloud	Accès
19804	virus_ONT (1.0) testontvirus	Sep 05 2022, 17h00	virus_ont	8 32 200	da98	ifb-core-cloudbis	
19759	virus_ONT (1.0)	Sep 05 2022, 10h25	DIADE	1 4 25	b680		

- Bottom Buttons:** Arrêter les déploiements (Stop deployments) and Tout voir (6) (View all 6).

A red arrow points to the "Cloud" column header in the table.

Let's run your vm through the cloud

ready !

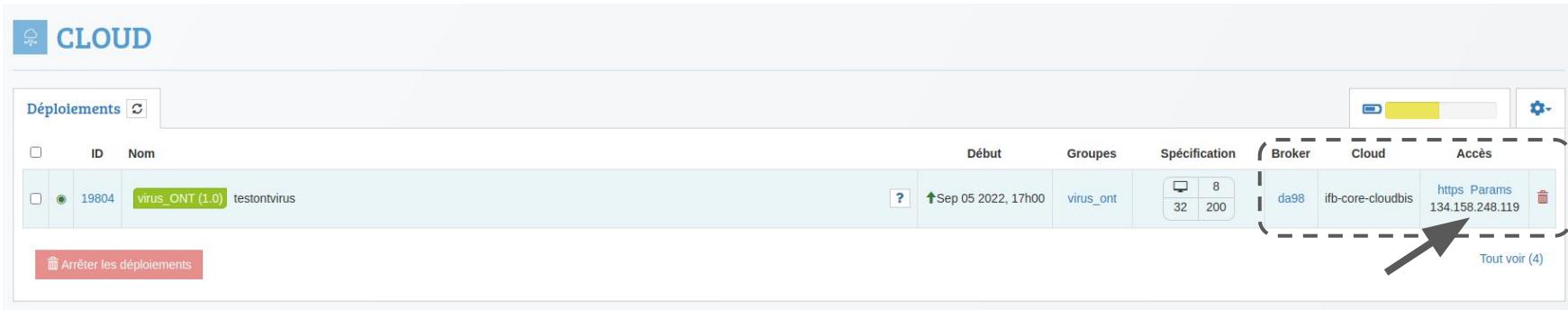
CLOUD

Déploiements

ID	Nom	Début	Groupes	Spécification	Broker	Cloud	Accès			
19804	virus_ONT (1.0) testontvirus	Sep 05 2022, 17h00	virus_ont	<table border="1"><tr><td>8</td></tr><tr><td>32</td><td>200</td></tr></table>	8	32	200	da98	ifb-core-cloudbis	https Params 134.158.248.119
8										
32	200									

Arrêter les déploiements

Tout voir (4)



Let's run your vm through the cloud

get the url... link “https”

The screenshot shows a cloud deployment interface with the following details:

- Déploiements**: A table listing one deployment.
- ID**: 19804
- Nom**: virus_ONT (1.0)
- Début**: Sep 05 2022, 17h00
- Groupes**: virus_ont
- Spécification**: Broker (8 cores, 32GB RAM), Cloud (32 cores, 200GB RAM), Accès (da98, ifb-core-cloudbis, https Params 134.158.248.119)
- Actions**: Arrêter les déploiements (Stop deployments) button.

A dashed box highlights the "Accès" section, and a large arrow points from the bottom right towards the "https Params" field.

Let's run our vm through the cloud

Get the token identifiant... link “Params”

The screenshot shows a cloud management interface with a modal dialog titled "Paramètres". Inside the dialog, there is a table with two columns: "nom" and "valeur". A single row is visible, containing "JUPYTER_TOKEN" in the "nom" column and "28f9a32ae92eaecbc816880489c9217e3263f9fd4614352" in the "valeur" column. The background of the interface displays a list of tasks or jobs. One job, named "virus", is shown in detail. The "Accès" (Access) section for this job includes a URL: "https://134.248.119.134". A yellow arrow points from the text "link ‘Params’" in the previous slide to this URL in the screenshot.

nom	valeur
JUPYTER_TOKEN	28f9a32ae92eaecbc816880489c9217e3263f9fd4614352

Début	Groupes	Spécification	Broker	Cloud	Accès
Sep 05 2022, 17h00	virus_ont	8 32 200	da98	ifb-core-cloudb1s	https://134.248.119.134

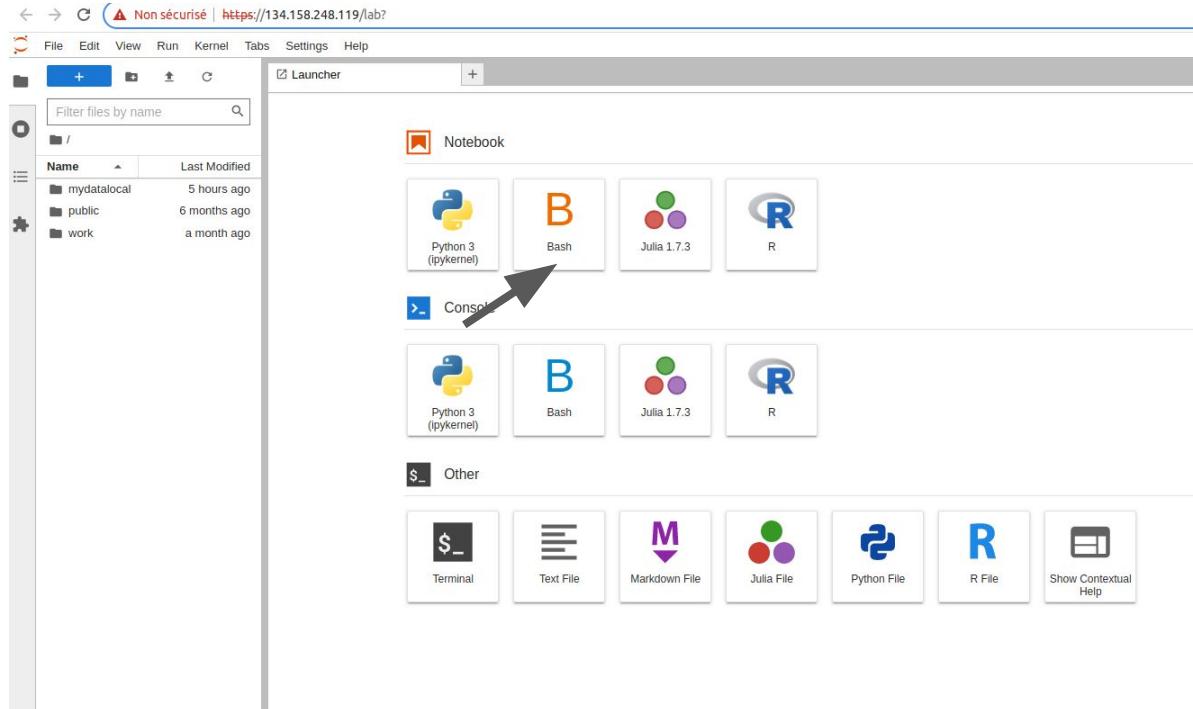
Let's run our vm through the cloud

Open your vm ([https link](https://134.158.247.8/tree)) to access to your own jupyter lab

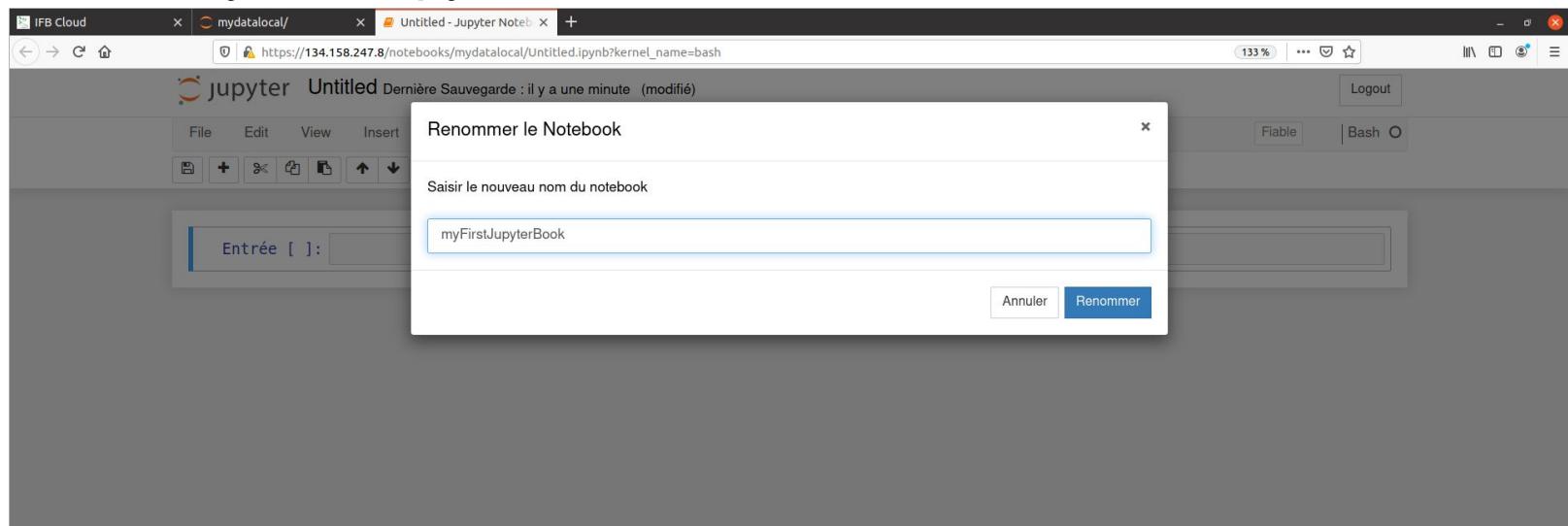


Create your first jupyter book

Go into the directory “work” and create a new jupyter book
-> kernel : bash



Rename your first jupyter book



Run your first bash command - *git clone*

All jupyterbook used for practice are here :

https://github.com/SouthGreenPlatform/training_ONT_teaching

Download all the jupyter books with the command *git clone*

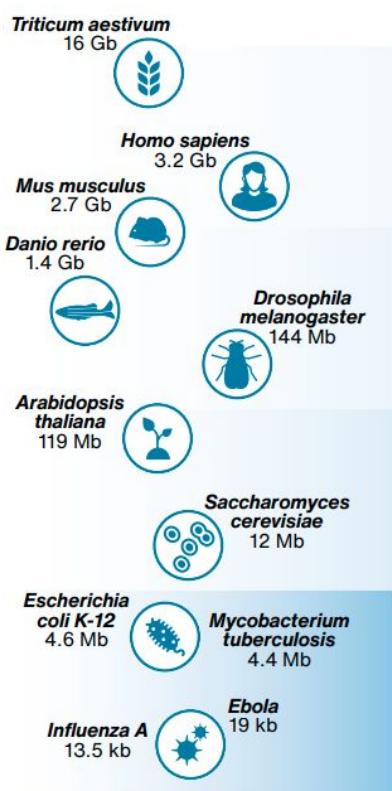
`git clone https://github.com/SouthGreenPlatform/training_ONT_teaching.git`

The screenshot shows a Jupyter Notebook interface. On the left, there is a file browser window titled 'work' showing two files: 'training_SV_teaching' and 'MyFirstJupyterBook.ipynb'. The main area displays a notebook titled 'My first Juptyper book - Training SG SV'. Below it, another section titled 'My first linux command - pwd' shows the output of the command 'pwd' in a code cell, which is '/home/jovyan/work'. A third section at the bottom contains the text 'Download all jupyter book we will use for this week - git clone' followed by the URL 'url https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022'. Below this, a code cell shows the command 'git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git' being run, with its output indicating the cloning process into a directory named 'training_SV_teaching'.

```
git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git
Cloning into 'training_SV_teaching'...
remote: Enumerating objects: 70, done.
remote: Counting objects: 100% (70/70), done.
remote: Compressing objects: 100% (48/48), done.
remote: Total 70 (delta 35), reused 49 (delta 20), pack-reused 0
Unpacking objects: 100% (70/70). 134.35 KiB | 1.62 MiB/s. done.
```

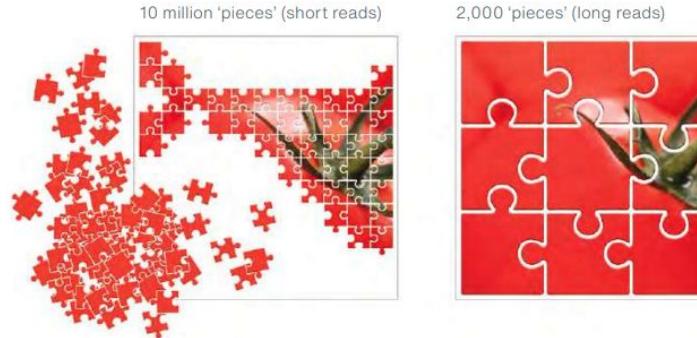
Let's start !

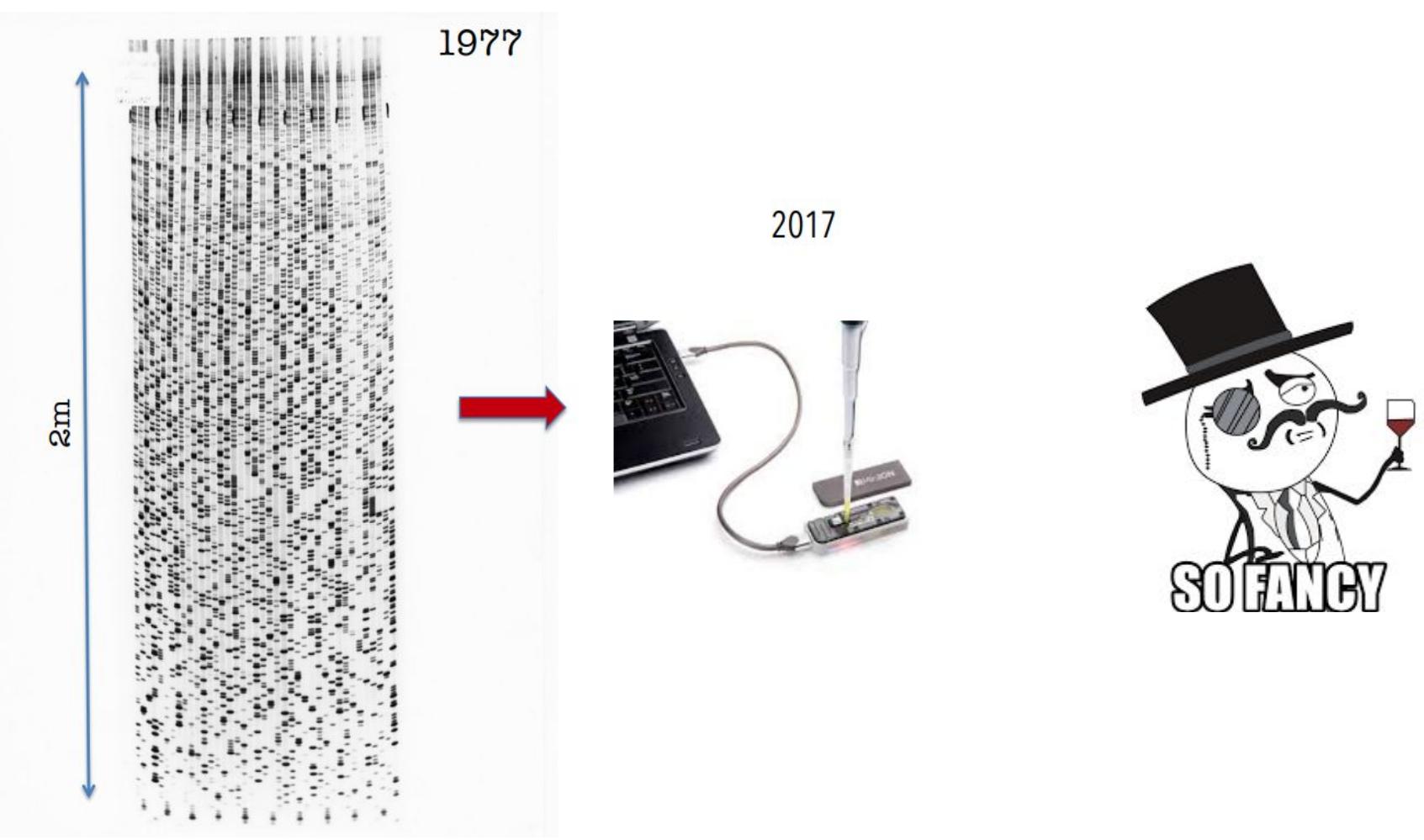
Why use Long reads ?



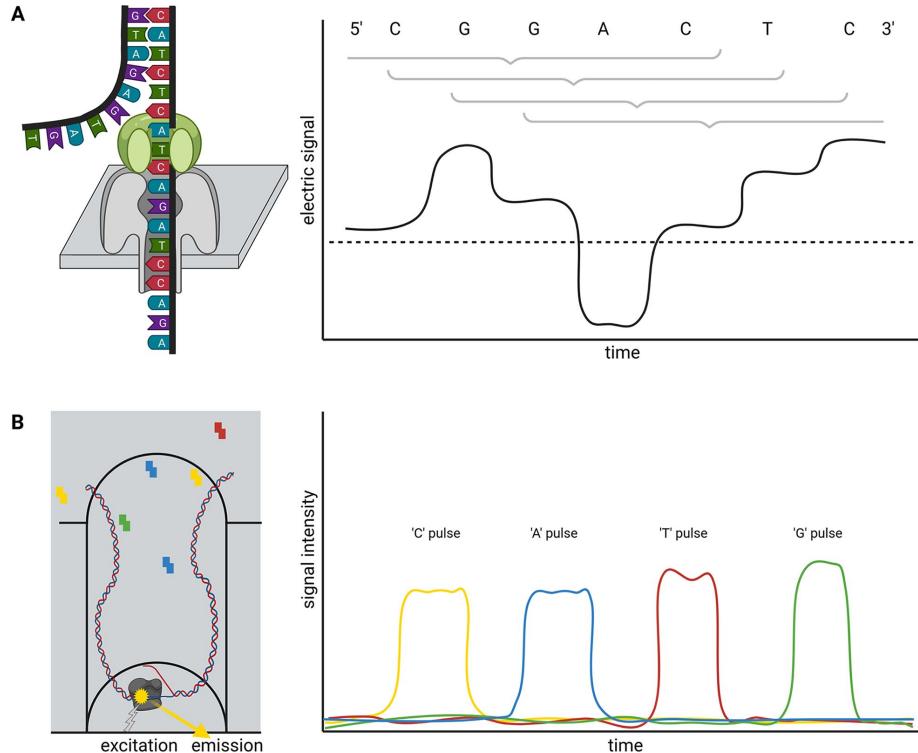
Microbial genomes	Human genomes	Animal genomes	Plant genomes
-------------------	---------------	----------------	---------------

- Simplify de novo assembly and correct existing genomes
- They bridge repetitions and build less fragmented genomes. SV, repeats, phasing
- They come from technologies which do not amplify the DNA fragments and therefore have less coverage bias.
- They are affordable.
- Detecting base modifications : they provide methylation information
- Analysing long-read transcriptomes





Long Reads



ONT is based on the translocation of a DNA or RNA strand through a nanopore located in an artificial membrane. Multiple nucleotides located in the nanopore determine the flow of ions through this nanopore in a specific way by physically blocking the space. This change in ion flux is recorded as an electric signal and further converted into sequence information.

Single-Molecule Real Time (**SMRT**) sequencing detects fluorescent light emitted from nucleotides upon incorporation into a DNA strand. The DNA polymerase is located at the bottom of a well and synthesises a new DNA strand. The integration into the new DNA strand keeps the nucleotide for a sufficiently long time in the well to allow detection.

Two technologies

Oxford Nanopore



MinION

GridION



PromethION

Pacific BioScience



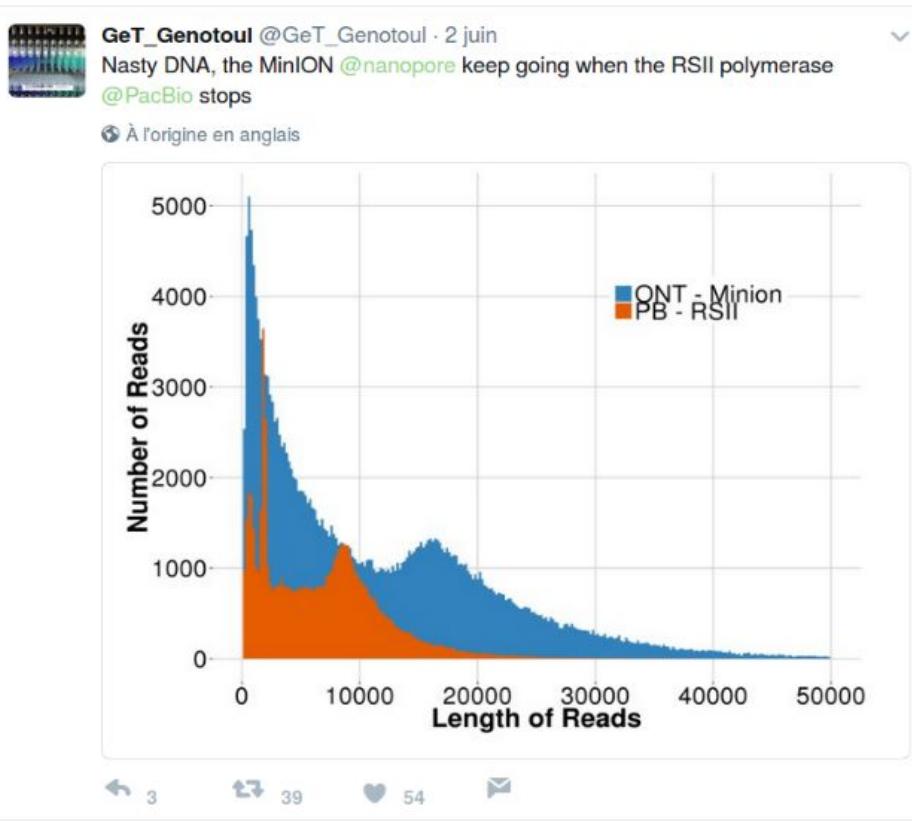
RSII



Sequel

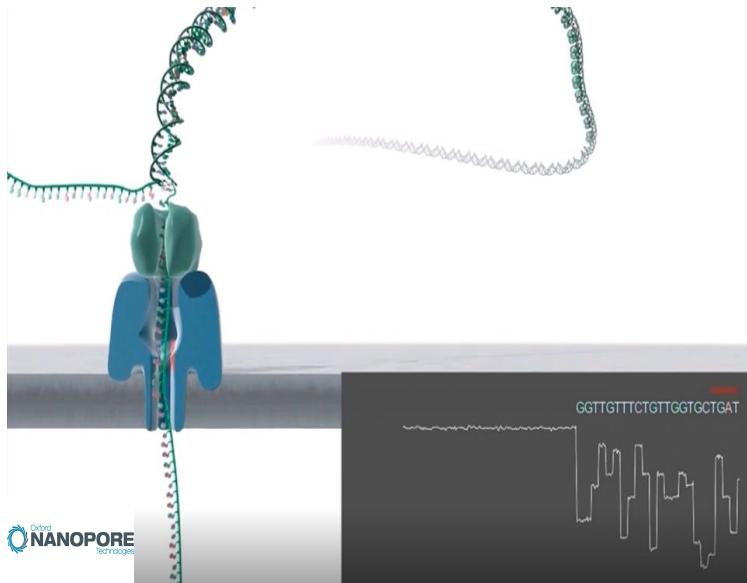
from Elixir GAAS 2018

Same sample / RSII vs MinION



SMRT limited by the longevity of the polymerase. A faster polymerase for the Sequel sequencer (chemistry v3, 2018) increased the read lengths to an average 30-kb polymerase read length.

Oxford Nanopore Technology



Involves passing a DNA molecule through a nanoscale pore and then measuring changes in electrical field surrounding the pore.

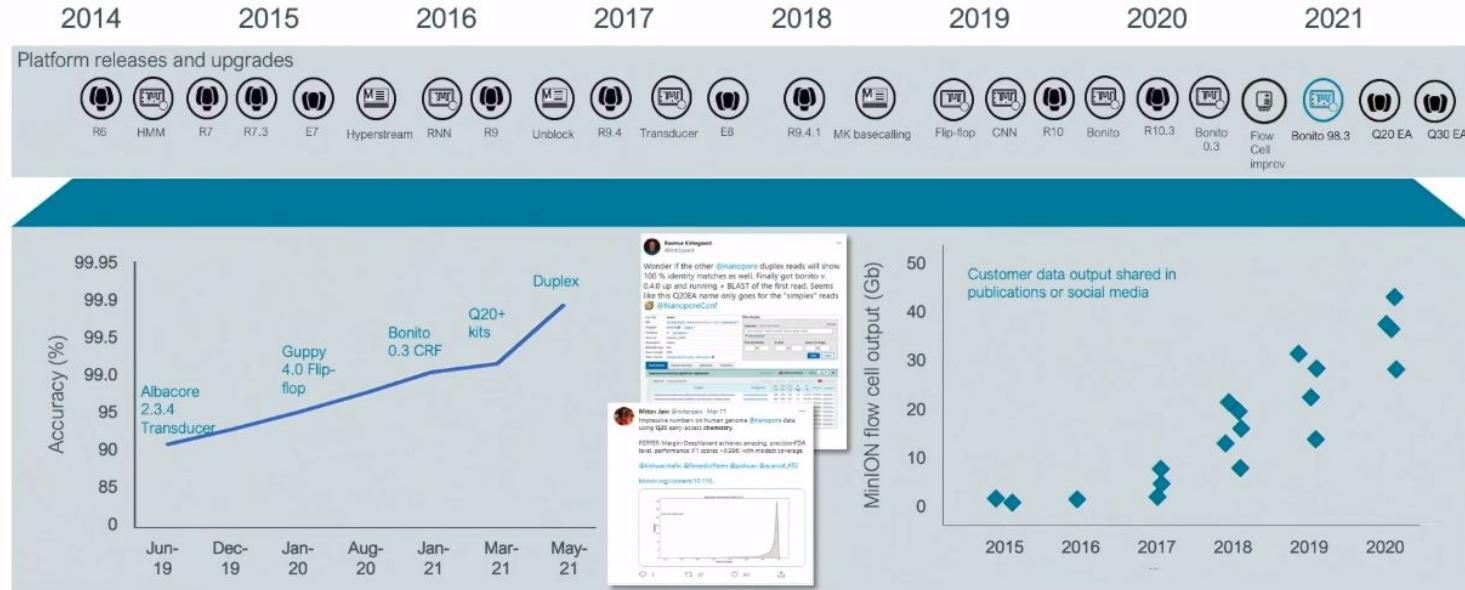
- + Long reads 2-300 kb++ (record 4Mb!!)
- + Portability and sequencing speed
- Error rate (0.5-5% as compared to 0.5% for Illumina)
- Homopolymers in reads : Follow caller version updates !
- Some DNAs are harder to sequence because they do not go easily through the pores : Lab!

Plant genome project workflow from DNA extraction over ONT sequencing to data submission

	task	consumed time	hands-on time	equipment	estimated costs of consumables	estimated costs of lab equipment
A	 plant incubation in darkness	2-3d	1h			
B	 non-destructive sampling	-	1h			
C	 DNA extraction	1d	8h	waterbath, centrifuge	\$50	\$1000 \$8000
D	 quality control	1h	1h	NanoDrop, Qubit	\$20	
E	 short fragment depletion	2h	1h	centrifuge	\$50	
F	 quality control	1h	1h	NanoDrop, Qubit	\$20	\$5000 \$5000
G	 library preparation & sequencing	1-5d	4-16h	centrifuge, magnetic rack, sequencer	\$3000	\$250 \$1000
H	 basecalling	1d	1h	computer with GPU		\$3000
I	 assembly	1-15d	1h			
J	 polishing	1-5d	1h	compute cluster / cloud		
K	 annotation	1-5d	1h			
L	 data submission	2h	2h	fast internet connection		

Upgrades drive performance enhancements

...and core ones ship in consumables and software



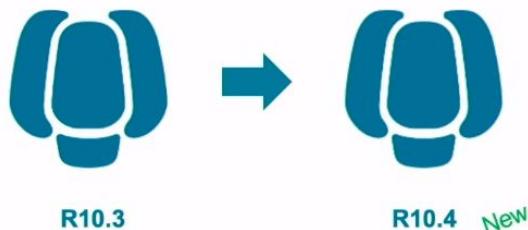
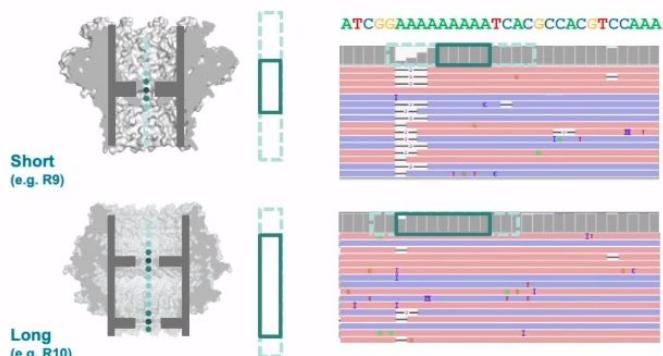
Last upgrades !

Oxford's Nanopores

R9 and R10

Short and long "reader heads"

- Length of the main discrimination site ("read head") affects accuracy
- Short read heads allow easier decoding of individual bases (R9 series)
- Longer read heads see more range and are more information rich (R10 series)



Improving the R10 series

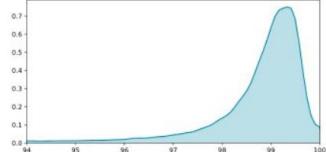
- We are continuously seeking to improve our nanopores
- R10 series of pores are still being iterated on – new R10.4 version
 - Extended discrimination profile, more sensing range
 - Higher flow cell yield of nanopores

← Tweet

 Oxford Nanopore @nanopore ...

Flow cells using our latest pore — R10.4 — can now be trialled through the expanding Q20+ Early Access Programme, which is now open to all applicants. Find out more about Q20+ and R10.4, and register to take part in the programme, here: bit.ly/3CEIJl9

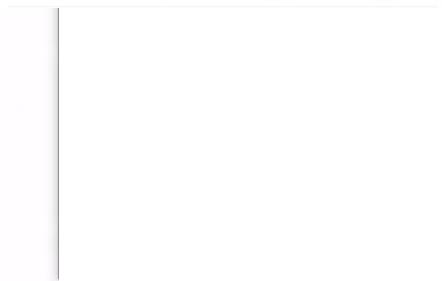
Raw read modal 99.3%, >Q20



A histogram showing the distribution of raw read accuracy percentages. The x-axis ranges from 94 to 100, and the y-axis ranges from 0.0 to 0.7. The distribution is highly skewed to the right, with the highest frequency (around 0.7) occurring at approximately 99.3% accuracy. The area under the curve is shaded light blue.

8:30 AM · Sep 23, 2021 · HubSpot

33 Retweets 1 Quote Tweet 62 Likes



© 2021 Oxford Nanopore Technologies Limited.
Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

 Oxford
NANOPORE
Technologies

<https://community.nanoporetech.com/posts/q20-early-access-group-br>

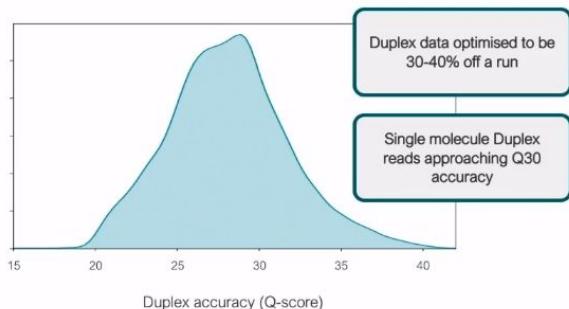
Last upgrades !

Nanopore accuracy

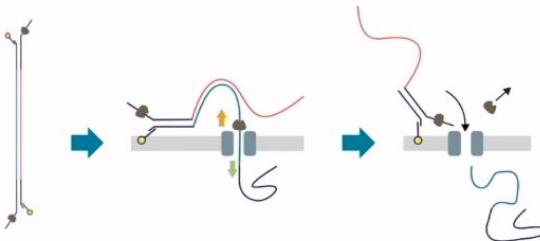
When we last spoke...

Duplex reads

- Possible when complement strand is sequenced immediately after template
- High duplex accuracy delivered by combining data template and complement
- New algorithms have been developed specifically for data combination
- Recent chemistries have optimised the amount of duplex data generated



© 2021 Oxford Nanopore Technologies Limited.
Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.



Generating duplex data

- Chances of seeing the complement follow template increased with Q20+ chemistry
- Early protocols available in EA community
- Longest Duplex Q30 read to date: 156 kbase



← Tweet

Oxford Nanopore @nanopore ...

Flow cells using our latest pore — R10.4 — can now be trialled through the expanding Q20+ Early Access Programme, which is now open to all applicants. Find out more about Q20+ and R10.4, and register to take part in the programme, here: bit.ly/3CEIJl9

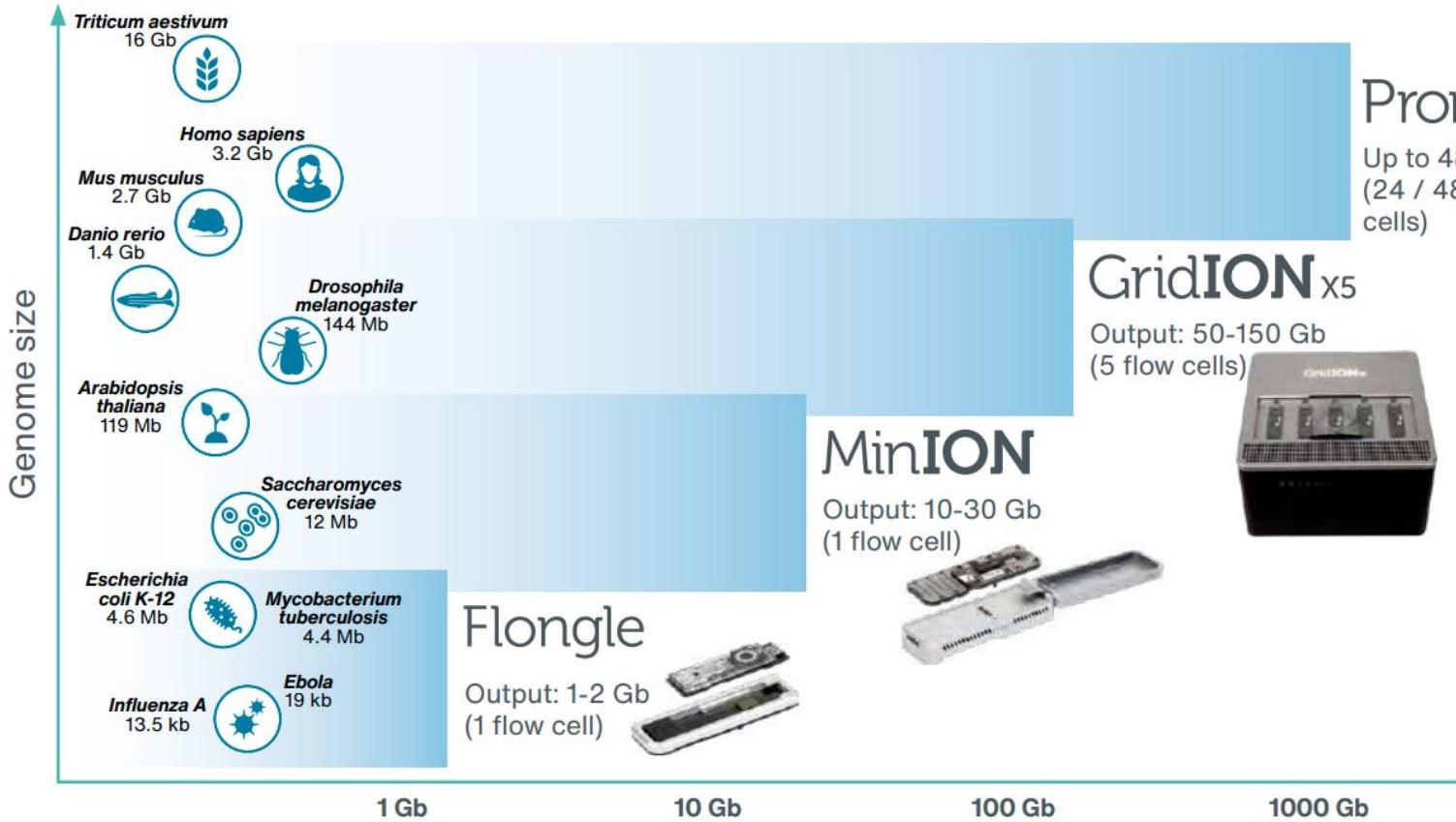
Raw read modal 99.3%, >Q20

Raw read accuracy (%)

8:30 AM · Sep 23, 2021 · HubSpot

33 Retweets 1 Quote Tweet 62 Likes

A lot of data !



A lot of data !

MinION



MinION Mk1C



GridION



P2 Solo



P2



**PromethION
24**



**PromethION
48**



MinION and Flongle Flow Cell compatible

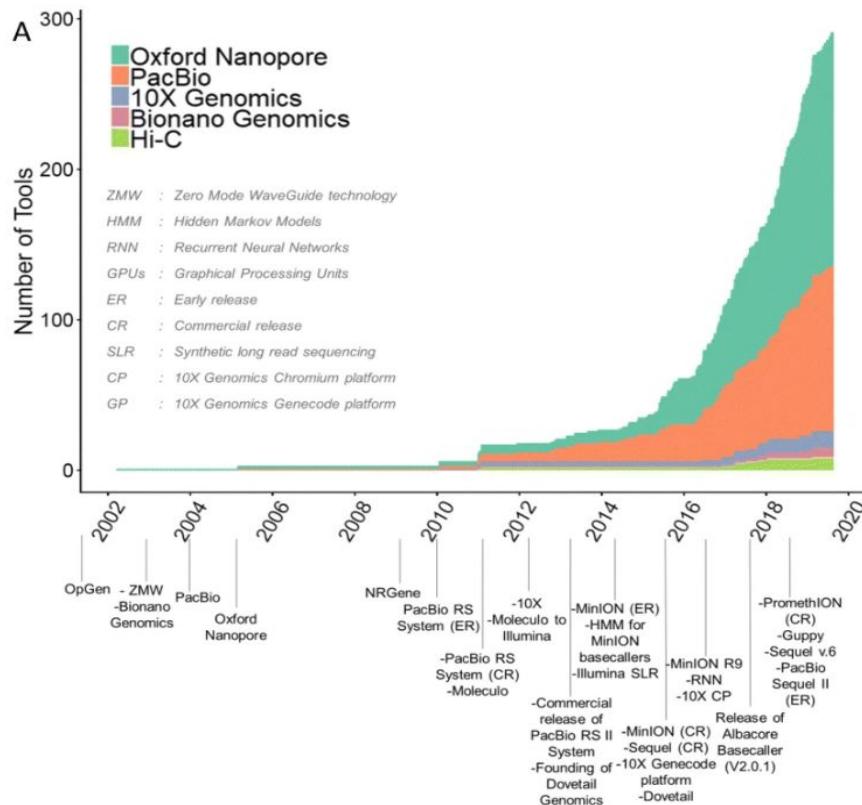
PromethION Flow Cell compatible

Configuration	Platform			Techniques		Tech specifications	
Number of flow cells per device	1	1	5	2	2	24	48
Maximum number of channels per flow cell	512	512	512	2,675	2,675	2,675	2,675
Run time	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours
Device TMO [†]	50 Gb	50 Gb	250 Gb	580 Gb	580 Gb	~7 Tb	~14 Tb
Maximum number of flow cells per year*	104	104	520	208	208	2,596	4,992
Offer sequencing as a service	No	No	Yes	Yes	Yes	Yes	Yes

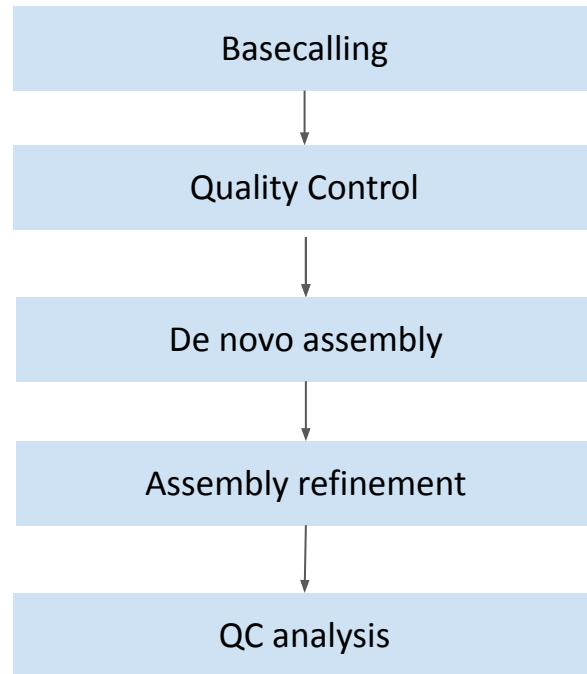
The data that these platforms produce differ qualitatively from second-generation sequencing, thus necessitating tailored analysis tools



A lot of tools are being developed and upgraded frequently !

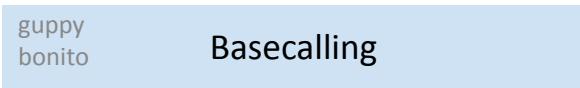


Typical long-read analysis pipelines for ONT data



Typical long-read analysis pipelines for ONT data

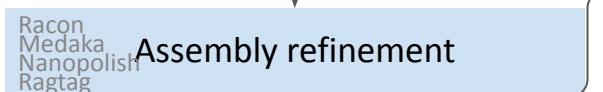
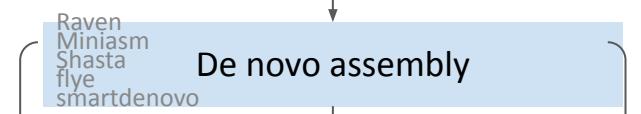
Demo



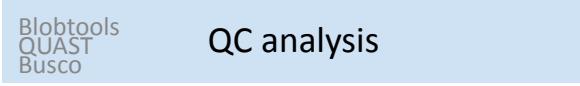
Practical 1



Practical 2



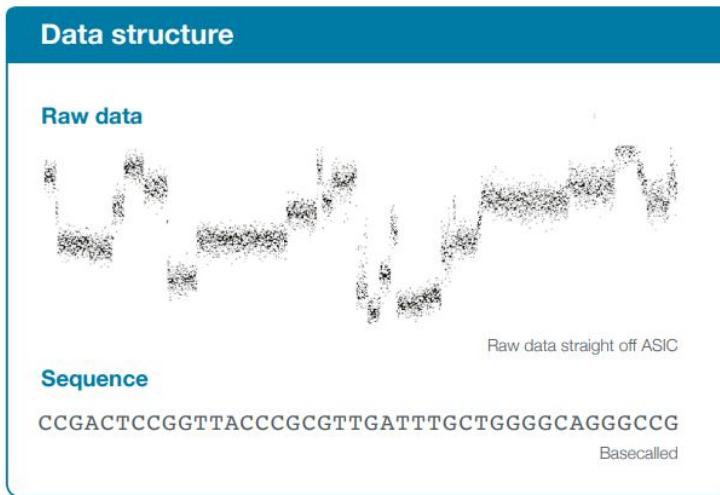
Practical 3



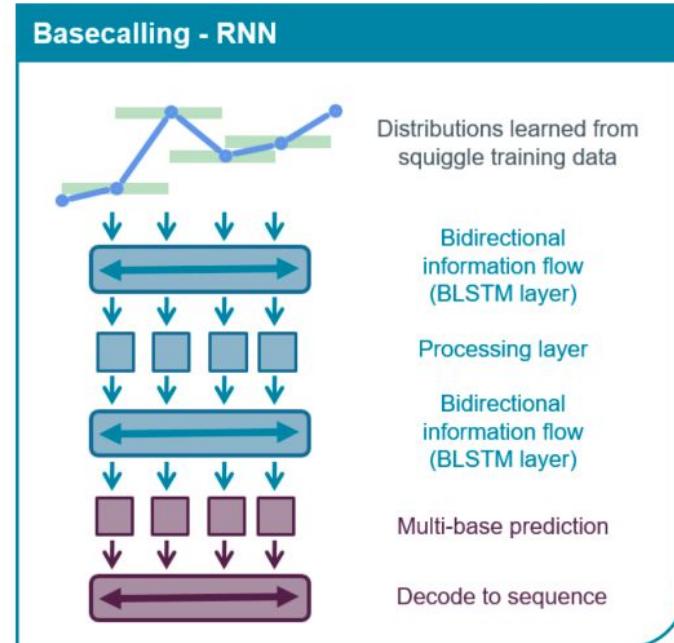
Chapitre 1

Reads Quality Control

ONT Read calling

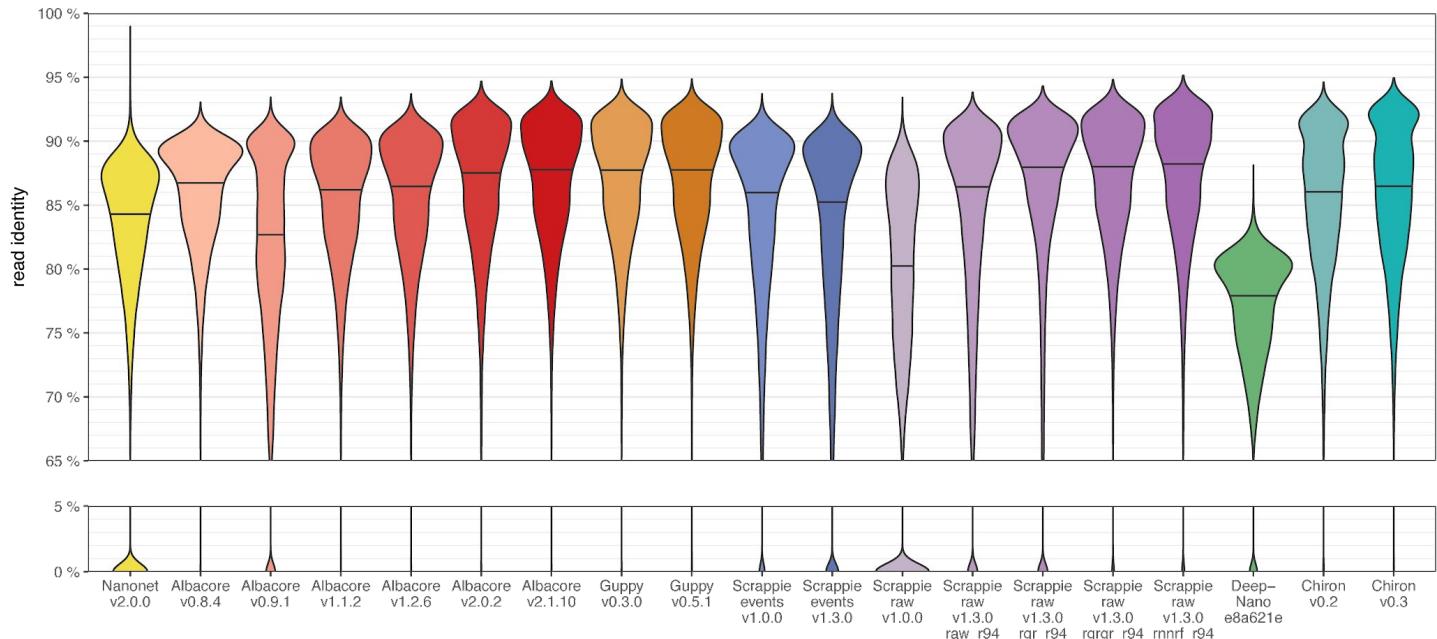


Reccurent Neural Network (RNN) – works like your brain! It can learn on the previous data and improve its performance on new data

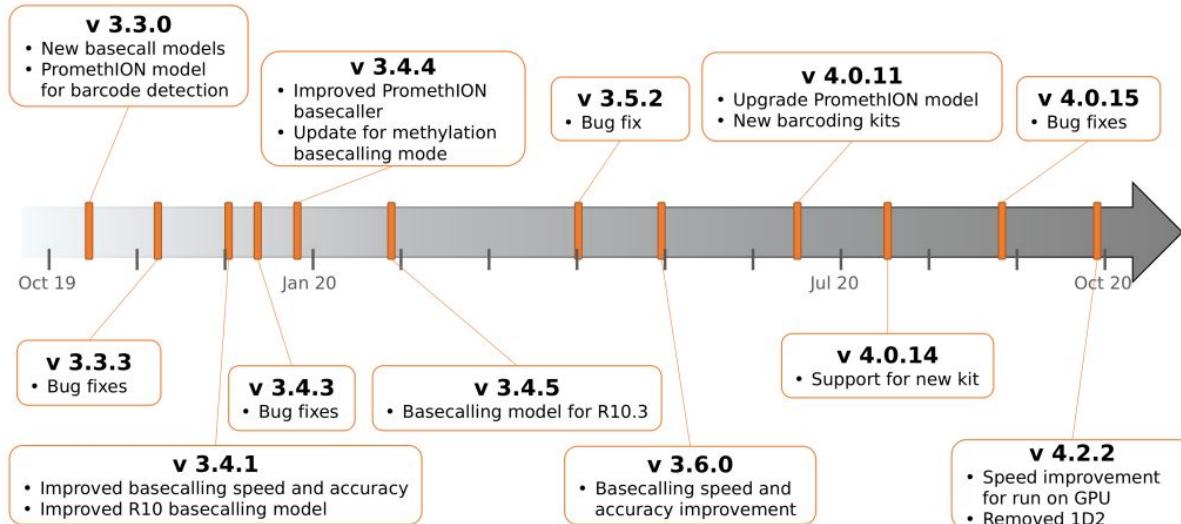


Nanopore basecallers are trained on many sequenced data, so you can run it on your data even if you are sequencing first time

ONT Read calling

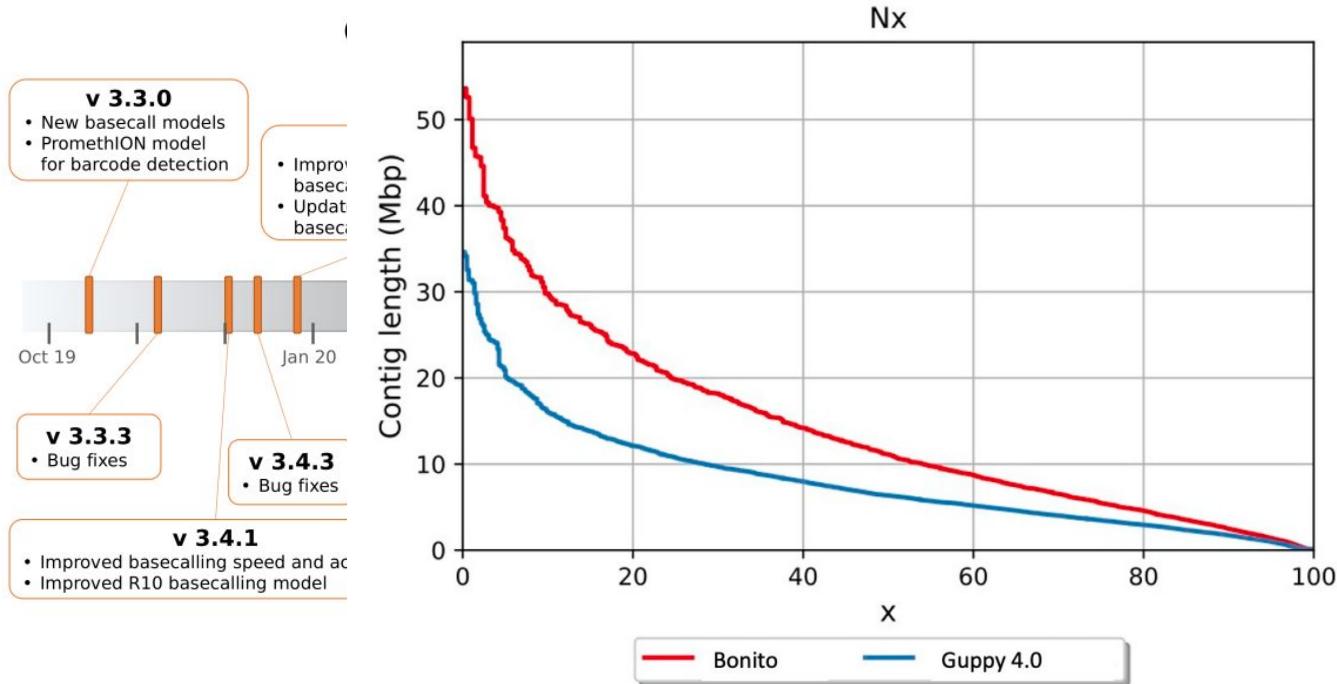


Guppy basecaller releases



V 6.0.1
includes sup
models and now
can be specific
to some taxons !

(+ Many other basecallers prior to Guppy [1] and to come.)



(+ Many other basecallers prior to Guppy [1] and to come.)

Harmeet Singh @GiessenSingh ...

Contiguity comparison between Wheat @nanopore assemblies using Guppy and Bonito base calling. Looks like Bonito increases N50. #Bioinformatics #longreads Traduire le Tweet 5:25 PM · 28 avr. 2021 · Twitter Web App

16 Retweets 57 J'aime

Replies 12 Likes 12 Shares 1

Retweetez Répondre

harish @harishkt19... · 30 avr. ... En réponse à @GiessenSingh @kazumachack et @nanopore Do you have any comparisons as to how Bonito basecalled and HiFi reads behave? 1 1 1 1

Harmeet Singh @... · 30 avr. ... Not yet!! but I will have that in some time 1 2 1 1 Voir les réponses

Nick Verecke @m... · 29 avr. ... 1 1 1 1

<https://hal.inria.fr/hal-03123133/document>

summary_file.txt

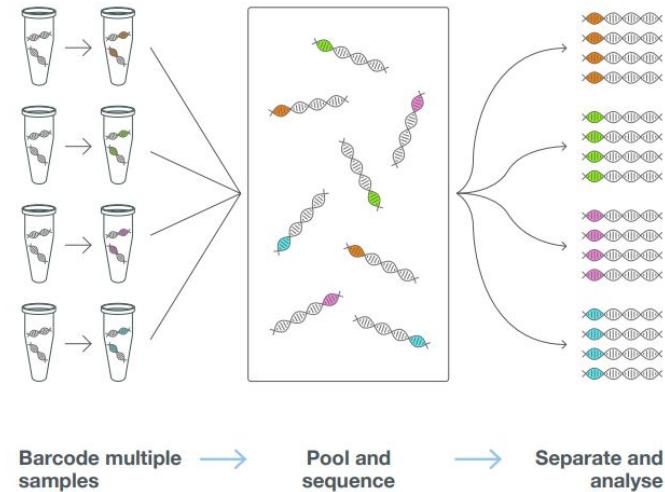
filename	FAK47038_aa36ef836fd50817477a5770772dffc63bfed2eb_30
read_id	188e2a0b-780c-440d-9223-61d8979dd002
run_id	aa36ef836fd50817477a5770772dffc63bfed2eb
batch_id	0
channel	70
mux	3
start_time	9688.985500
duration	1.610500
num_events	1288
passes_filtering	TRUE
template_start	9689.318000
num_events_template	1022
template_duration	1.278000
sequence_length_template	545
mean_qscore_template	11.462492
strand_score_template	3.165753
median_template	79.270927
mad_template	9.512511
scaling_median_template	79.270927
scaling_mad_template	9.512511

ONT demultiplexing

Deepbinner: Demultiplexing barcoded ONT reads with deep convolutional neural networks (CNN). The network is trained to classify barcodes based on the raw nanopore signal.

Guppy

In contrast to Deepbinner, guppy barcoding requires basecalling of all reads and detects barcodes in the sequence



ONT Read calling, cleaning and filtering

Sequencer ONT : raw fast5 files

- Transform fast5 signal in fastq standard format *Guppy, Bonito*
- Optional Demultiplexing and removing adapters *Guppy options*
- Optional Find and remove adapters from reads *Porechop*
- Optional Quality filtering using the *sequencing_summary.txt* information : *Guppy options, filtlong, nanofilt*

Guppy is a neural network based basecaller that in addition to basecalling also performs filtering of low quality reads, clipping of Oxford Nanopore adapters and estimation of methylation probabilities per base

FASTQ FORMAT

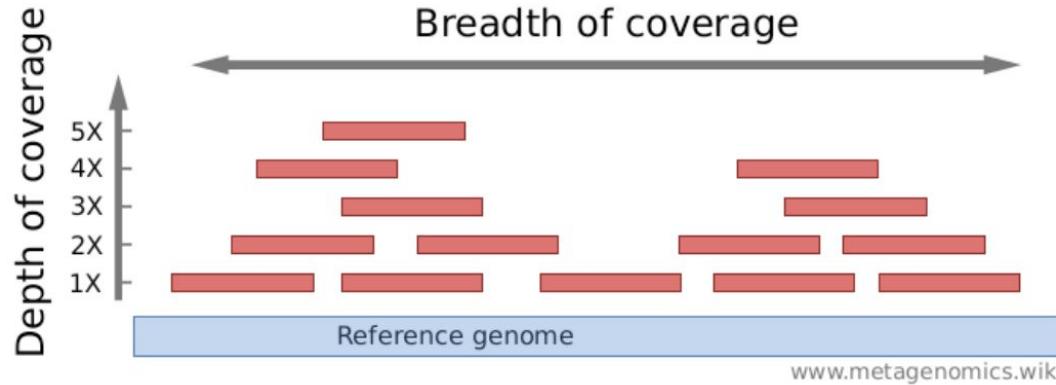
1 séquence = 4 lignes

```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTTCTTAGTATTTTGATGTCAATTCCGTGTTGGTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTGTT
+
@@@DDDD>FFFFAFBEABB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTCAGTAACGACCTATAAATTAAAGTAGC
+
@CFFFFFFGGHHHHJJJJIEE<HHHIJJIGBHGGEEIJJEIEIJIHHJFIIJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTCGGAAAAGTTGGGTATGGCTCTAGTAGCTTTGTCTTAT
+
@C@FFFFFFGGHHDHGIIEEHIII<CGHIJJIJJ:FC9DGAFGHII?DGBFIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAAGCTAAAAGAACACAGTTGCTGAAGCAGCAAACACAAGAAC
+
B@@@DFFFFFGHHHHJIIIIJJJIIGJCHHEIII>GHIG@GHIDHGJIIFHIIJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGAAGAGTTAAAAACAAATTATGAGCAACTGAGTTTC
+
@@@CFFFFFFHFFFHFGIJJIHIIJJIIHJJECGHJJCHGICDGHHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTGTCGGGGTTGGAAATTGTCACTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```



- @identifiant de la séquence
- Séquence
- + (id séquence).
- Qualité de la séquence = un caractère ASCII pour chaque base

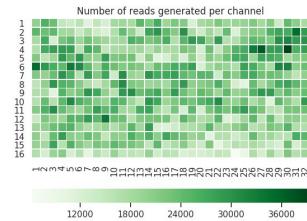
Calculate depth of coverage



depth of coverage estimation :

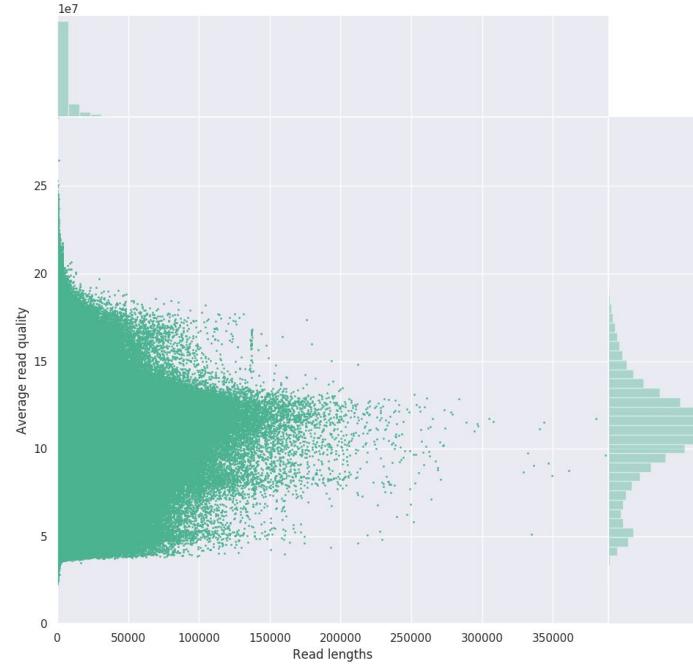
- Count how much base pairs in all sequenced reads? *total_pb*
- What is the expected genome size? *genome_size*

$\text{depth_of_coverage} = \text{total_pb}/\text{genome_size}$



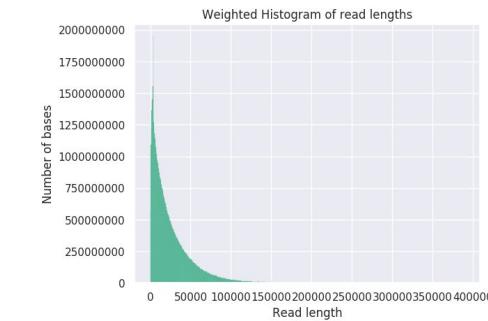
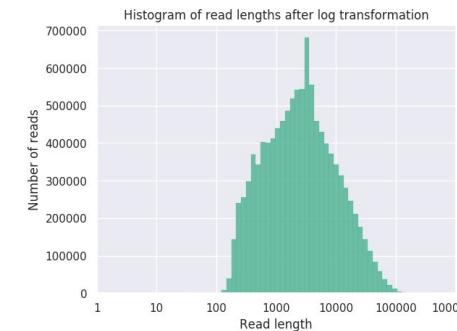
Reads Quality control : *NanoPlot*

Read lengths vs Average read quality plot

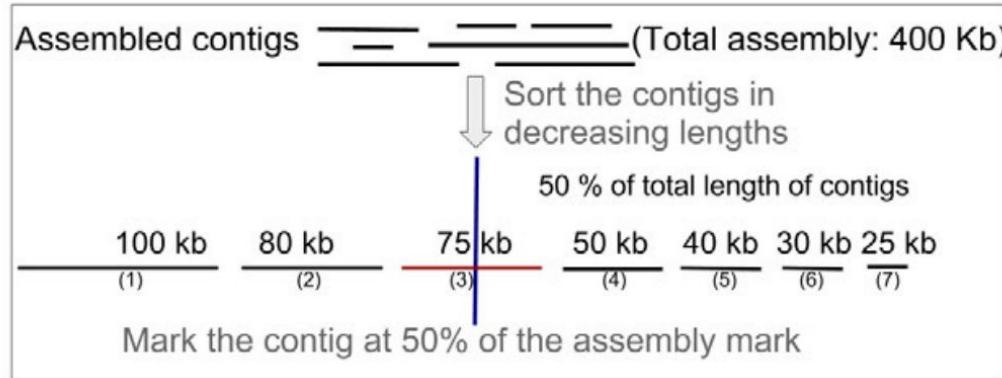


Summary statistics

General summary	
Active channels	512.0
Mean read length	6,315.6
Mean read quality	10.9
Median read length	2,517.0
Median read quality	11.1
Number of reads	10,847,854.0
Read length N50	16,816.0
Total bases	68,510,227,164.0



What is N50 and L50?



- N50, length of the contig at 50% assembly: 75 kb
→ L50, number of contigs until 50% assembly: 3

Reads Quality control

NanoPlot : <https://github.com/wdecoster/NanoPlot>

NanoComp : <https://github.com/wdecoster/nanocomp>

mini_qc : https://github.com/roblanf/minion_qc

Conclusion : check reads N50, reads length distribution, and calculate coverage !

TP1. Reads Quality Control

- TP1

[https://github.com/SouthGreenPlatform/training_ONT_teaching/
blob/main/1.raw_quality_control.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/main/1.raw_quality_control.ipynb)

Chapitre 2.

Assemblies

What do you want to do with these long reads?



Research areas

- Microbiology
- Microbiome
- Environmental
- Plant
- Animal
- Infectious disease
- Human genomics
- Clinical research
- Cancer
- Transcriptome
- Populations genomics
- COVID-19

Investigations

- Structural variation
- SNVs and phasing
- Gene expression
- Identification
- Splice variation
- Assembly
- Fusion transcripts
- Epigenetics
- Single cell
- Chromatin conformation

Techniques

- Whole genome
- Targeted
- Whole transcriptome
- Metagenomics
- Short Fragment Mode

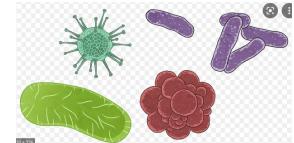
Type	Reference	Application	
Aligners/Alignment-based classifiers			
BLAST, MEGABLAST	[58,59]	Targeted; Shotgun	
minimap2	[33]	Targeted; Shotgun	
Alignment-free classifiers			
Kraken, Kraken2	[35,64]	Targeted; Shotgun	
KrakenUniq	[65]	Shotgun	
Bracken	[66]	Targeted; Shotgun	
Metamaps	[69]	Shotgun	
Centrifuge	[34]	Targeted; Shotgun	
Mash	[72]	Targeted; Shotgun	
Long-read assemblers			
Canu	[90]	Shotgun	
miniasm	[73]	Shotgun	
wtdbg2	[91]	Shotgun	
OPERA-MS	[95]	Shotgun	
MetaFlye	[96]	Shotgun	
MetaSPAdes	[74]	Shotgun	https://doi.org/10.1016/j.csbj.2021.02.020
Sequence correction and polishing tools			
Nanopolish		https://github.com/jts/nanopolish	Targeted; Shotgun
Medaka		https://github.com/nanoporetech/medaka	Targeted; Shotgun
Metagenomic analysis pipelines			
MEGAN-LR	[60]		Shotgun
NanoCLUST	[25]		Targeted
Reticulatus		https://github.com/SamStudio8/reticulatus	Shotgun
MUFFIN	[70]		Shotgun
NanoSPC	[71]		Shotgun
BusyBee		https://ccb-microbe.cs.uni-saarland.de/busybee/	Shotgun

Type	Reference	Application	
Aligners/Alignment-based classifiers			
BLAST, MEGABLAST	[58,59]	Targeted; Shotgun	
minimap2	[33]	Targeted; Shotgun	
Alignment-free classifiers			
Kraken, Kraken2	[35,64]	Targeted; Shotgun	
KrakenUniq	[65]	Shotgun	
Bracken	[66]	Targeted; Shotgun	
Metamaps	[69]	Shotgun	
Centrifuge	[34]	Targeted; Shotgun	
Mash	[72]	Targeted; Shotgun	
Long-read assemblers			
Canu	[90]	Shotgun	
miniasm	[73]	Shotgun	
wtdbg2	[91]	Shotgun	
OPERA-MS	[95]	Shotgun	
MetaFlye	[96]	Shotgun	
MetaSPAdes	[74]	Shotgun	https://doi.org/10.1016/j.csbj.2021.02.020
 Sequence correction and polishing tools			
Nanopolish		https://github.com/jts/nanopolish	Targeted; Shotgun
Medaka		https://github.com/nanoporetech/medaka	Targeted; Shotgun
Metagenomic analysis pipelines			
MEGAN-LR	[60]		Shotgun
NanoCLUST	[25]		Targeted
Reticulatus		https://github.com/SamStudio8/reticulatus	Shotgun
MUFFIN	[70]		Shotgun
NanoSPC	[71]		Shotgun
BusyBee		https://ccb-microbe.cs.uni-saarland.de/busybee/	Shotgun

Which assembler to use over my favorite organism?

Long reads simplify genome assembly, with the ability to span repeat-rich sequences (characteristic of antimicrobial resistance genes) and structural variants. Nanopore sequencing also shows a lack of bias in GC-rich regions, in contrast to other sequencing platforms. To perform microbial genome assembly, we suggest using the third-party de novo assembly tool Flye. We also recommend one round of polishing with Medaka.

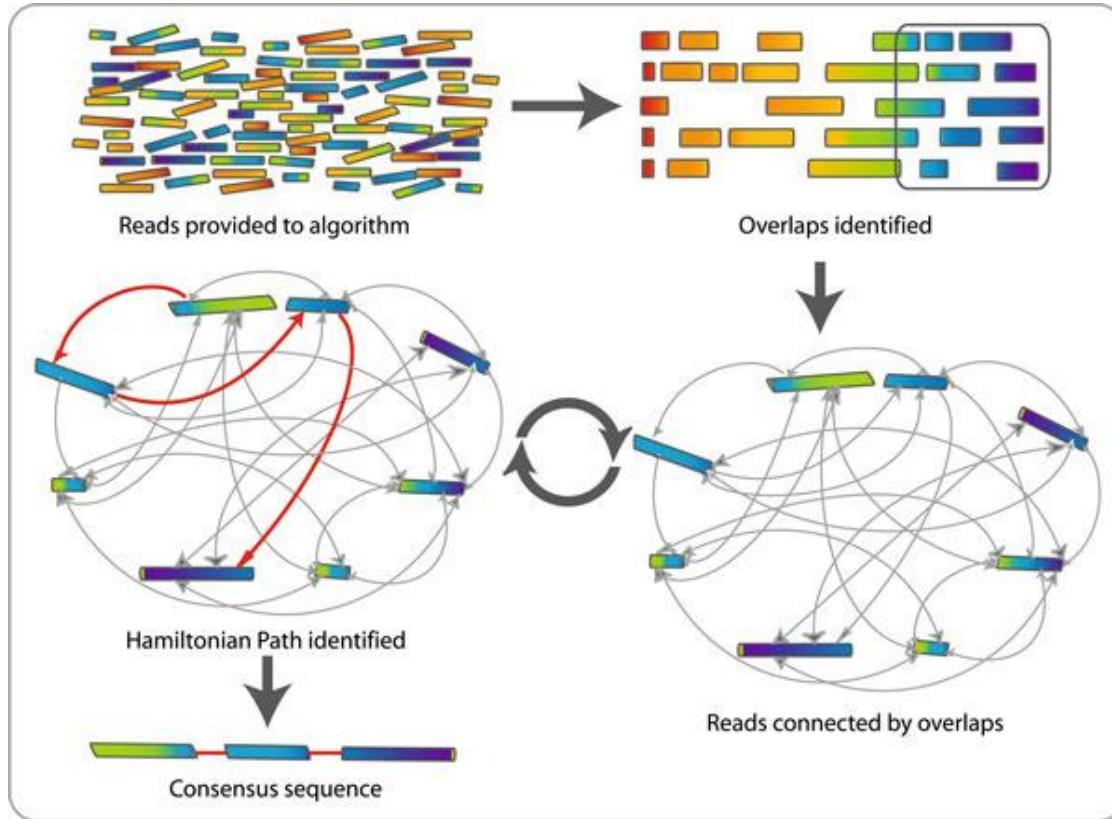
<https://nanoporetech.com/sites/default/files/s3/literature/microbial-genome-assembly-workflow.pdf>



For assembly, ONT recommend sequencing a human genome to a minimum depth of 30x of 25–35 kb reads. However, sequencing to a depth of 60x is advisable to obtain the best assembly metrics. We also recommend basecalling in high accuracy mode. Greatest contig N50 is usually obtained with Shasta and Flye. Polishing/Correction is also recommended (Racon and Medaka).

<https://nanoporetech.com/sites/default/files/s3/literature/human-genome-assembly-workflow.pdf>

Overlap–layout–consensus genome assembly algorithm (OLC)



[Canu](#), [Flye](#), [Miniasm](#), [Raven](#), [Smartdenovo](#), [Shasta](#)

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3055744/>

Polishing / Correction

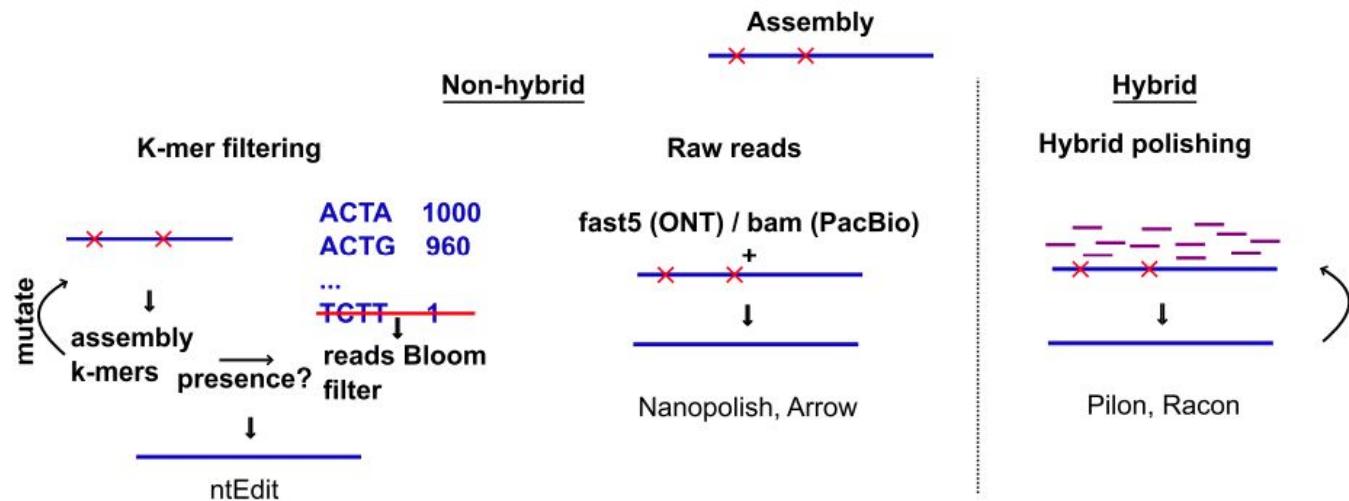
[Racon](#) correct raw contigs generated by rapid assembly methods which do not include a consensus step. It can polish with either Illumina data or data produced by third generation of sequencing. (recursive use)

[Medaka](#) and [Nanopolish](#) create a consensus sequence of nanopore sequencing data. (mapping + consensus)

- + Medaka uses neural networks where Nanopolish uses HMMs.
- + Medaka uses basecalled reads, not the raw signal.
- + Medaka propose the ability to train one's own basecalling model

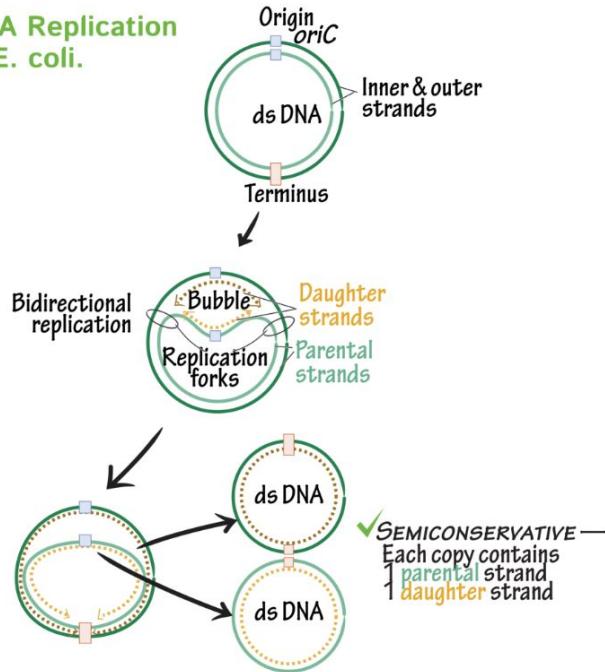
[Pilon](#) correct assemblies using illumina reads. (recursive use)

Autres : [NeuralPolish](#) , [ntEdit](#)



Circularisation ?

DNA Replication
in E. coli.



Some assemblers give you information about circularisation of assembled molecules (flye, canu).

Circularisation can be found also on GFA files generated by assemblers. (miniasm, raven, shasta)

You can try to circularise assembled molecules using tools as [circlator](#)

it could be interesting tagging and rotation of circular molecule before each polishing step.

As well as, fixing (*dnaA* gene) the start position on circular genome. This is efficient when multiple genome alignments are envisaged.

TP2. Assemblies

- TP2

[https://github.com/SouthGreenPlatform/training_ONT_teaching/
blob/main/2.assemblies.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/main/2.assemblies.ipynb)

Chapitre 3.

Contigs Quality

QUAST

Quality Assessment Tool for Genome Assemblies by [CAB](#)

26 March 2021, Friday, 07:37:40

[View in Icarus contig browser](#)

All statistics are based on contigs of size ≥ 3000 bp, unless otherwise noted (e.g., "# contigs (≥ 0 bp)" and "Total length (≥ 0 bp)" include all contigs).

Aligned to "TIGRv7_ok" | 375 096 285 bp | 16 fragments | 43.57 % G+C

Worst	Median	Best	<input type="checkbox"/> Show heatmap
Genome statistics			
Genome fraction (%)	65.801	65.916	65.417
Duplication ratio	1.036	1.041	1.041
Largest alignment	2 503 013	2 501 477	1 739 590
Total aligned length	255 403 246	257 194 821	255 339 839
NGA50	48 559	48 062	42 714
LGA50	1338	1333	1404
Misassemblies			
# misassemblies	9633	9923	7666
Misassembled contigs length	373 371 138	373 825 172	335 007 830
Mismatches			
# mismatches per 100 kbp	2776.55	2831.25	2669.89
# indels per 100 kbp	321.69	301.83	330.99
# N's per 100 kbp	0	0.23	0
Statistics without reference			
# contigs	181	250	250
Largest contig	43 938 576	43 971 118	14 121 367
Total length	383 158 522	384 147 370	387 291 200
Total length (≥ 1000 bp)	383 173 133	384 197 574	387 291 200
Total length (≥ 10000 bp)	382 901 616	383 618 037	387 291 200
Total length (≥ 50000 bp)	381 421 486	381 880 053	387 291 200
250	13 998 410	383 785 534	369 892 751
729	6 500 937	383 785 534	369 966 935
854	6 543 040	383 785 534	373 136 825
		368 865 072	373 406 571
		365 953 108	371 578 702
			368 382 574

[Extended report](#)

plus petit nb de contigs : flye+racon puis raven+racon
plus long contigs : flye+racon

<https://github.com/ablab/quast>

Genome statistics	FLYE_STEP_POLISHING_RACon	FLYE_STEP_ASSEMBLY	RAVEN_STEP_POLISHING_RACon	RAVEN_STEP_ASSEMBLY	SHASTA_STEP_POLISHING_RACon	SHASTA_STEP_ASSEMBLY
Statistics without reference						
# contigs	181	250	250	250	729	854
# contigs (>= 0 bp)	194	285	250	250	767	1149
# contigs (>= 1000 bp)	188	274	250	250	763	1000
# contigs (>= 5000 bp)	168	207	250	250	674	746
# contigs (>= 10000 bp)	139	156	250	250	564	587
# contigs (>= 25000 bp)	97	99	250	250	487	488
# contigs (>= 50000 bp)	74	75	250	250	444	445
Largest contig	43 938 576	43 971 118	14 121 367	13 998 410	6 500 937	6 543 040
Total length	383 158 522	384 147 370	387 291 200	383 785 534	369 892 751	373 136 825
Total length (>= 0 bp)	383 176 103	384 204 105	387 291 200	383 785 534	369 969 110	373 471 297
Total length (>= 1000 bp)	383 173 133	384 197 574	387 291 200	383 785 534	369 966 935	373 406 571
Total length (>= 5000 bp)	383 108 497	383 977 711	387 291 200	383 785 534	369 668 739	372 705 755
Total length (>= 10000 bp)	382 901 616	383 618 037	387 291 200	383 785 534	368 865 072	371 578 702
Total length (>= 25000 bp)	382 215 424	382 691 571	387 291 200	383 785 534	367 717 125	370 136 458
Total length (>= 50000 bp)	381 421 486	381 880 053	387 291 200	383 785 534	365 953 108	368 382 574
N50	14 538 350	14 555 248	3 455 235	3 425 125	1 355 467	1 360 886
N75	10 163 758	10 173 888	1 497 559	1 483 567	738 018	741 772
L50	10	10	28	28	79	80
L75	17	17	68	68	173	174
GC (%)	43.56	43.61	43.59	42.81	43.43	43.36
Similarity statistics						
# similar correct contigs	260	247	263	0	255	60
# similar misassembled blocks	1251	1178	1257	0	1245	499

less contigs : flye+racon puis raven+racon

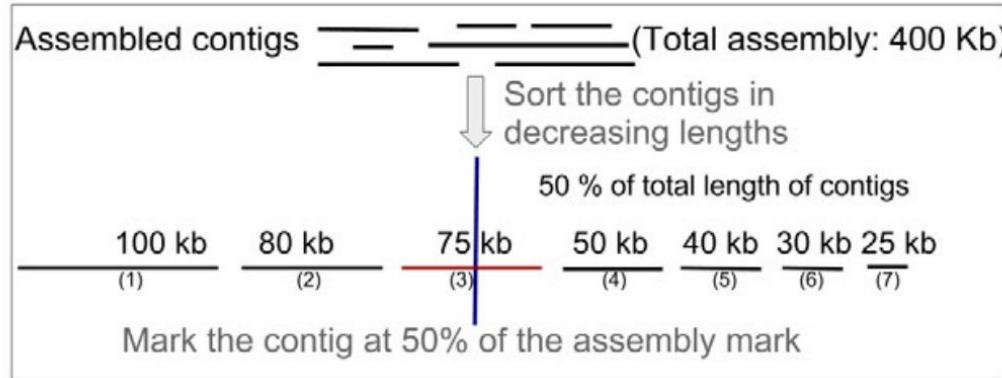
largest contig : flye+racon

largest N50 : flye

largest L50 : flye

what is N50 and L50?

What is N50 and L50?



- N50, length of the contig at 50% assembly: 75 kb
→ L50, number of contigs until 50% assembly: 3

QUAST

Quality Assessment Tool for Genome Assemblies by [CAB](#)

26 March 2021, Friday, 07:37:40

[View in Icarus contig browser](#)

All statistics are based on contigs of size ≥ 3000 bp, unless otherwise noted (e.g., "# contigs (≥ 0 bp)" and "Total length (≥ 0 bp)" include all contigs).

Aligned to "TIGRv7_ok" | 375 096 285 bp | 16 fragments | 43.57 % G+C

Worst	Median	Best	<input type="checkbox"/> Show heatmap
Genome statistics			
Genome fraction (%)	65.801	65.916	65.417
Duplication ratio	1.036	1.041	1.041
Largest alignment	2 503 013	2 501 477	1 739 590
Total aligned length	255 403 246	257 194 821	255 339 839
NGA50	48 559	48 062	42 714
LGA50	1338	1333	1404
Misassemblies			
# misassemblies	9633	9923	7666
Misassembled contigs length	373 371 138	373 825 172	335 007 830
Mismatches			
# mismatches per 100 kbp	2776.55	2831.25	2669.89
# indels per 100 kbp	321.69	301.83	330.99
# N's per 100 kbp	0	0.23	0
Statistics without reference			
# contigs	181	250	250
Largest contig	43 938 576	43 971 118	14 121 367
Total length	383 158 522	384 147 370	387 291 200
Total length (≥ 1000 bp)	383 173 133	384 197 574	387 291 200
Total length (≥ 10000 bp)	382 901 616	383 618 037	387 291 200
Total length (≥ 50000 bp)	381 421 486	381 880 053	387 291 200

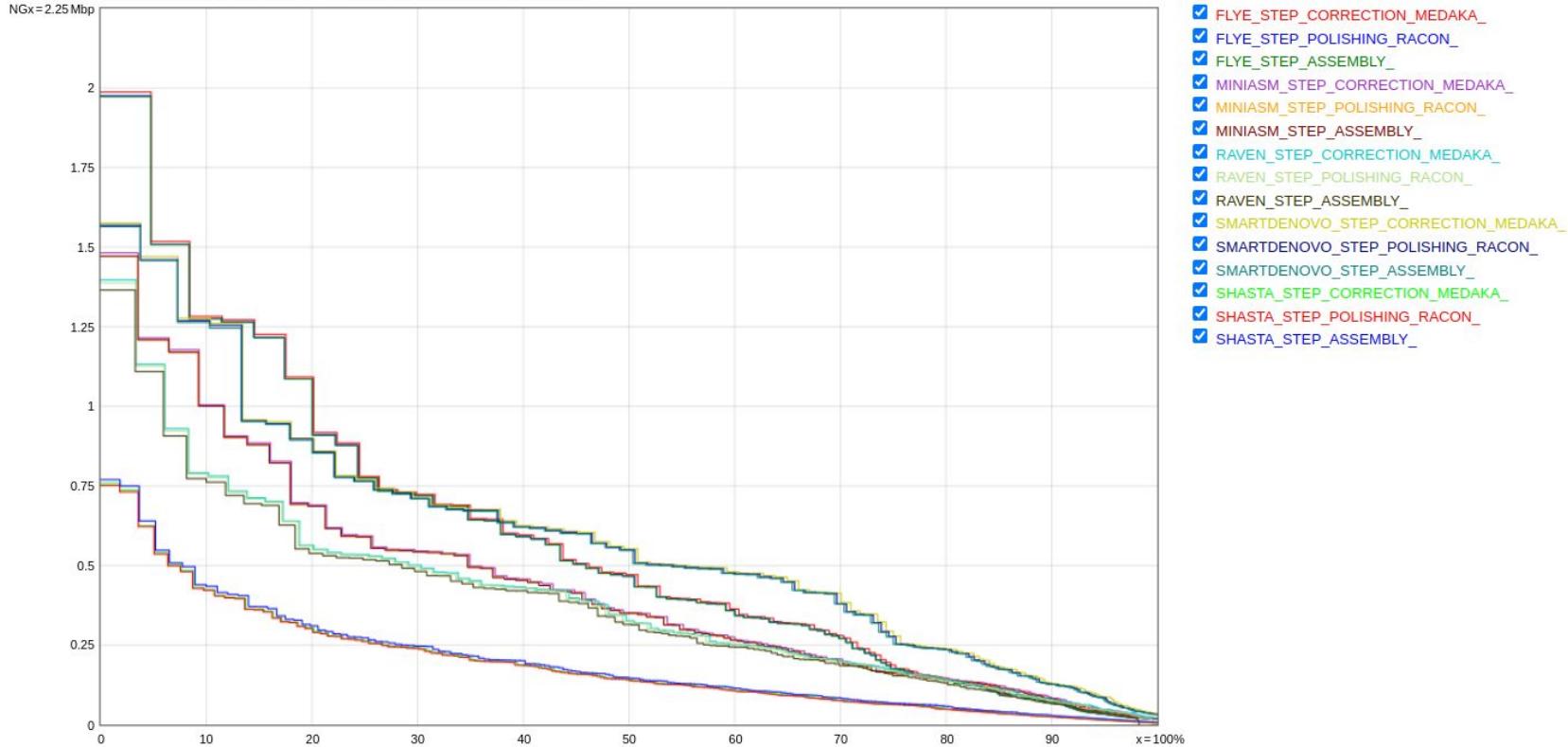
[Extended report](#)

Check misassemblies and N percentage.

BE CAREFUL! A misassembly for QUAST can be a structural variation!

Nx graph

Plots: Cumulative length Nx NAx NGx NGAx Misassemblies GC content



The greater the area under the curve AUC, the better is the assembly.
Nx represent N50 but also N10 to N100

BUSCO

from QC to gene prediction and phylogenomics

BUSCO v5.2.2 is the current stable version!

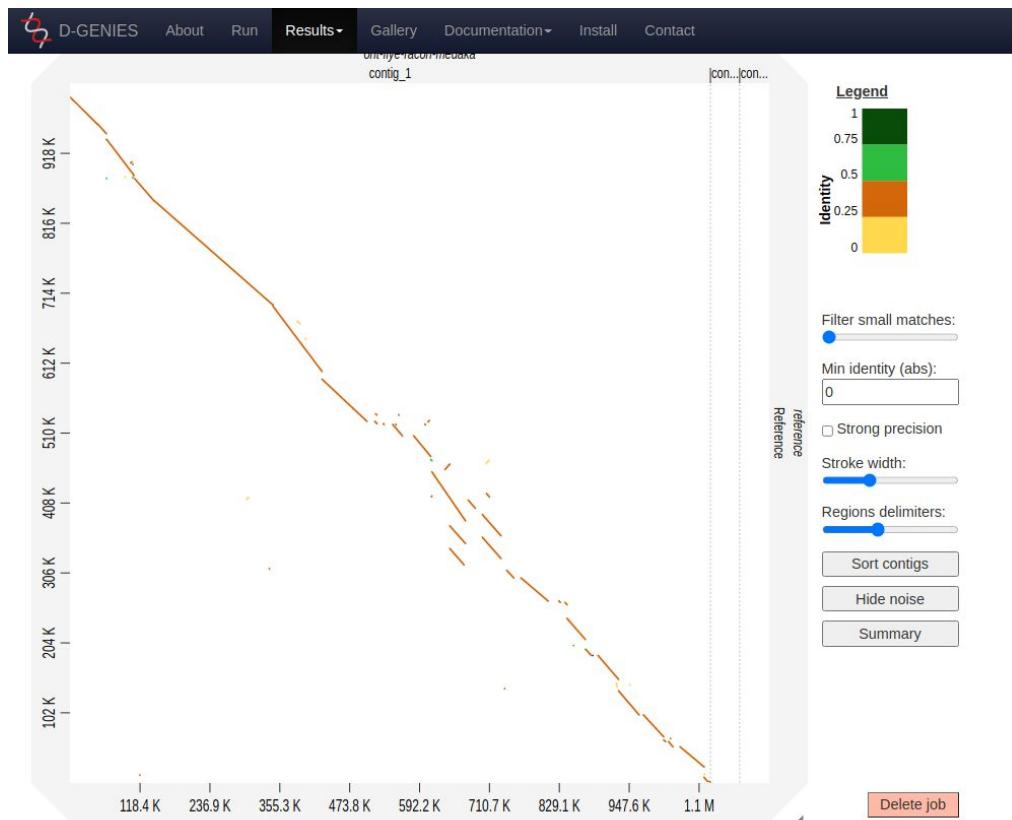
Gitlab [🔗](#), a Conda package [🔗](#) and Docker container [🔗](#) are also available.

Based on evolutionarily-informed expectations of gene content of near-universal single-copy orthologs, BUSCO metric is complementary to technical metrics like N50.

Helps to check if you have a good assembly, by searching the expected single-copy lineage-conserved orthologs in any newly-sequenced genome from an appropriate phylogenetic clade.

```
INFO Results:  
INFO C:95.6%[S:73.6%,D:22.0%],F:1.4%,M:3.0%,n:1759  
INFO 1682 Complete BUSCOs (C)  
INFO 1295 Complete and single-copy BUSCOs (S)  
INFO 387 Complete and duplicated BUSCOs (D)  
INFO 25 Fragmented BUSCOs (F)  
INFO 52 Missing BUSCOs (M)  
INFO 1759 Total BUSCO groups searched  
INFO BUSCO analysis done. Total running time: 621.2351775169373 seconds  
INFO Results written in /tmp/orjuela/BUSCO/run_trinity_busco/
```

Comparison with a reference genome



- NUCMER : Aligns a set of draft sequence contigs to a finished sequence
<http://mummer.sourceforge.net/>
- D-Genies : Online tool to compare two genomes by dot plot method
<http://dgenies.toulouse.inra.fr/>
- autre: *Gepard*

CANU

FLYE

MINIASM

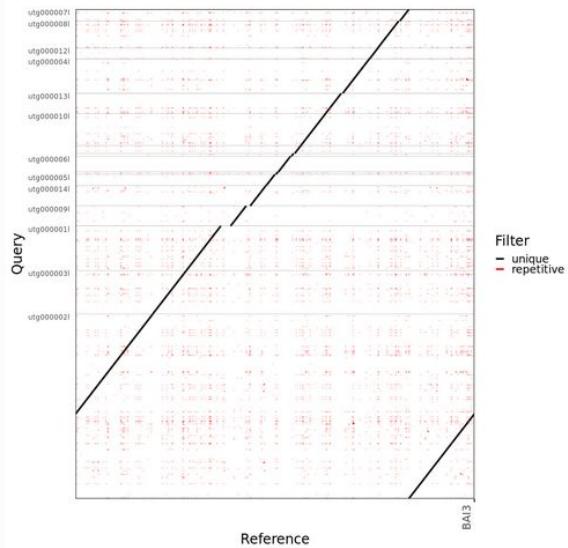
RAVEN

SMARTDENOVO

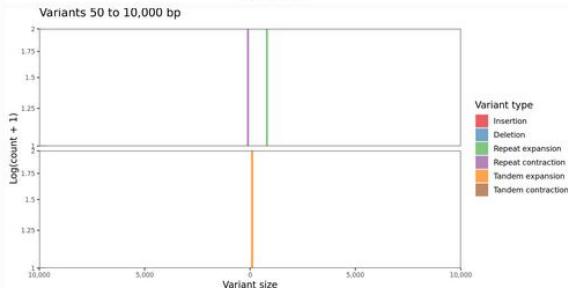
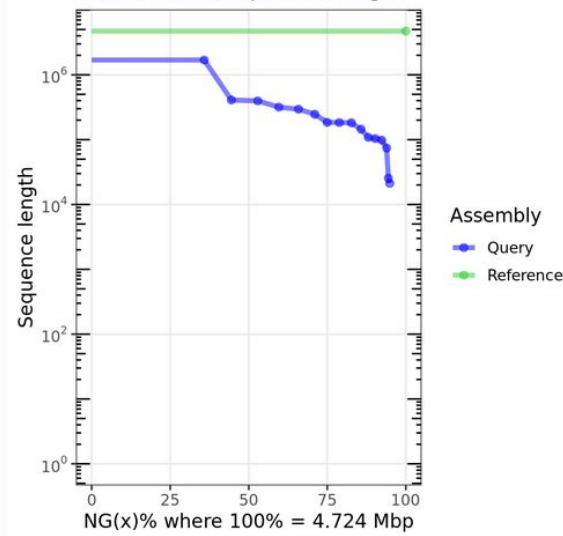
SHASTA

STEP_CORRECTION_NANOPOLISH_STARTFIXED

Dot plot of Assemblytics filtered alignments



Cumulative sequence length



TP3. Contigs Quality

- TP3

[https://github.com/SouthGreenPlatform/training_ONT_teaching/
blob/main/3.contigs_quality.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/main/3.contigs_quality.ipynb)

From contigs to chromosomes

Optical mapping : fluorescent marking of restriction sites of very long DNA molecules (up to Mb) to extract signature used to bridge contigs having these signatures.

10x chromium : shallow tagged sequencing of very long DNA fragments with Illumina machines. Read alignments enable scaffolding.

Genetic map : marker assisted contig bridging

HiC : chromosomal interaction sequencing gives the contig order on the chromosomes.

Conclusions

- DNA quality (fragment length) has a direct impact on read length
- We can assemble small to large genomes with Nanopore reads.
- Test a lot of tools to perform assemblies, in any case polishing is mandatory.
- There are still genomes very difficult to assemble

Formateurs

- Julie Orjuela



- François Sabot



- Philippe Cubry



- Ezechiel Tibiri



Merci pour votre attention !



Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International:

<http://creativecommons.org/licenses/by-nc-sa/4.0/>



thank you!