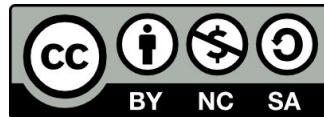


# Plant Genomic Sequencing with Oxford Nanopore technology

Quito, Ecuador  
9-15th November 2023  
PUCE University



# *Bioinformatics resources*

# We will work under LINUX!!

- 2 ways to use it:

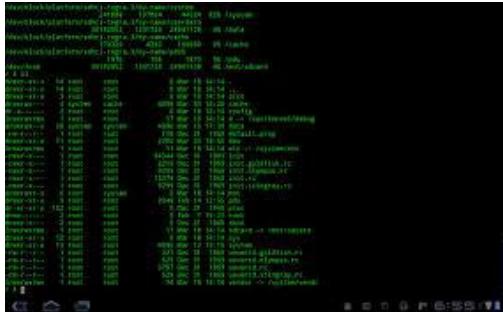
in *graphical mode*



# Terminal mode

- 2 ways to use it:

in *console mode*



A screenshot of a terminal window displaying a large amount of command-line text. The text includes various file paths, file names, and numerical values, likely representing a log or a dump of system data. The terminal has a dark background with white text.



Before everything else !

# Bases for Linux

[https://github.com/SouthGreenPlatform/training\\_ONT\\_teaching/b  
lob/ONT\\_LA/slides/intro-2-linux\\_slides.pdf](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/ONT_LA/slides/intro-2-linux_slides.pdf)

# In *jupyter book mode*

- A third way to us Linux

in *jupyter book mode*



On the IFB Cloud !



*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

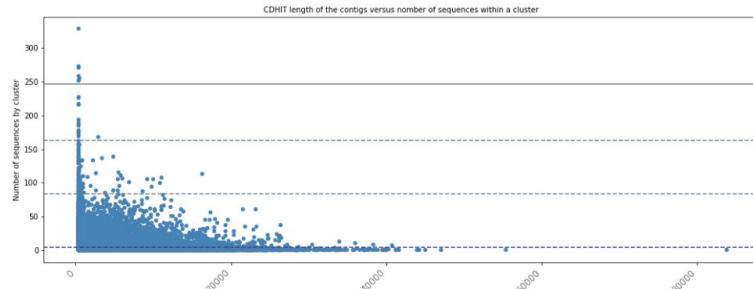


# Why use jupyter book ?

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

```
All ctgs
Entrée [30]: 1 plt.figure(figsize=(17, 6))
2 ax = plt.gca()
3 df_cdhit[(df_cdhit.sp == '0b')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
4 df_cdhit[(df_cdhit.sp == '0g')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
5 plt.axhline(y=5, color='darkslateblue', linestyle='--')
6 plt.axhline(y=84, color='slategrey', linestyle='--')
7 plt.axhline(y=163, color='slategrey', linestyle='--')
8 plt.axhline(y=247, color='grey', linestyle='--')
9
10 plt.title("CDHIT length of the contigs versus number of sequences within a cluster", fontsize=10)
11 plt.xlabel('Cluster length')
12 plt.ylabel('Number of sequences by cluster')
13 plt.xticks(
14     rotation=45,
15     horizontalalignment='right',
16     fontweight='light',
17     fontsize=12,
18 )
19 plt.show()
20
```



# 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:**

**Anchoring data analysis**

**1 - CDHIT data analysis before anchoring on genome**

**1.1 Removing redundancy with CDHIT**

  - CDHIT Input : 1,306,676 contigs assembled from no mapped reads
  - Tests & results

	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

**1.2 Converting cdhit file into a csv loaded as a dataframe with pandas**

The script cdhitVsAnchoring.py creates the csv file allCtgtsIRIGIN\_TOG5681.dedup8095.PANDAS.csv

Load csv file into a pandasframe

**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 ?

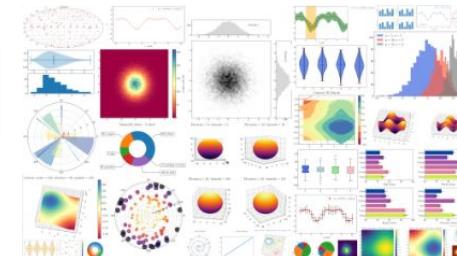
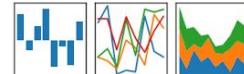


- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

# How to become a super datascientist ?

- easily import/export tabular files into/from dataframes (similar to R dataframe).
- manipulate these data tables / DataFrames
- easily draw beautiful graphs from these DataFrames with 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

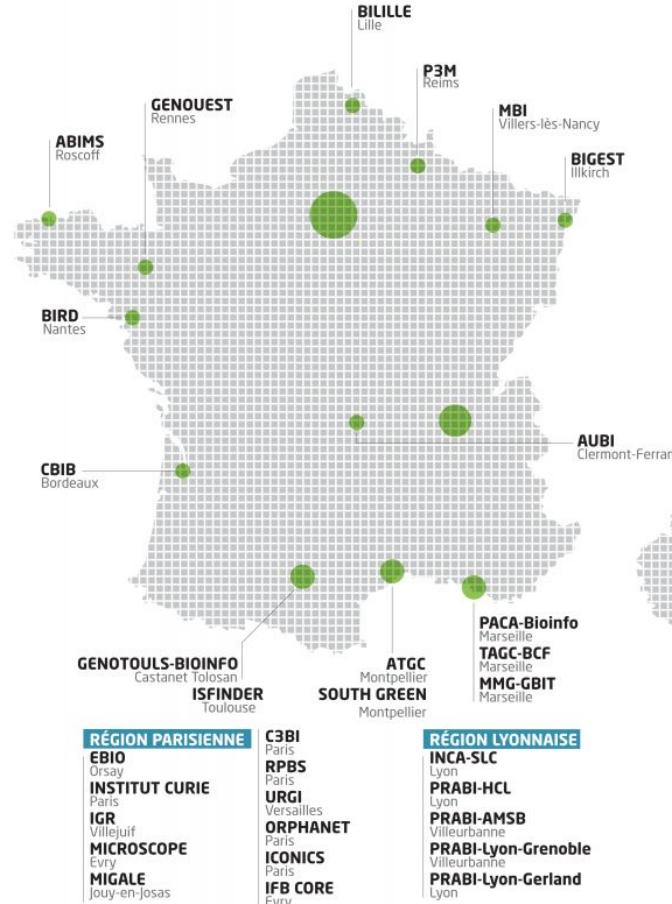


A screenshot of a web browser window titled "IFB Cloud". The address bar shows "mydatalocal/" and the URL "https://134.158.247.8/tree/mydatalocal". The main content area displays a Jupyter interface with tabs for "Files", "Running", and "Clusters". A message at the top says "Select items to perform actions on them." Below this, there is a file tree showing a folder named "mydatalocal". A message in the center says "La liste des notebooks est vide." To the right, there is a "Notebook" dropdown menu with options: "Bash", "Julia 1.5.3", "Python 3", "R", "Text File", "Folder", and "Terminal". A "New" button is also visible in this menu. At the bottom right of the interface, there are "Upload" and "Logout" buttons.

# IFB ?



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

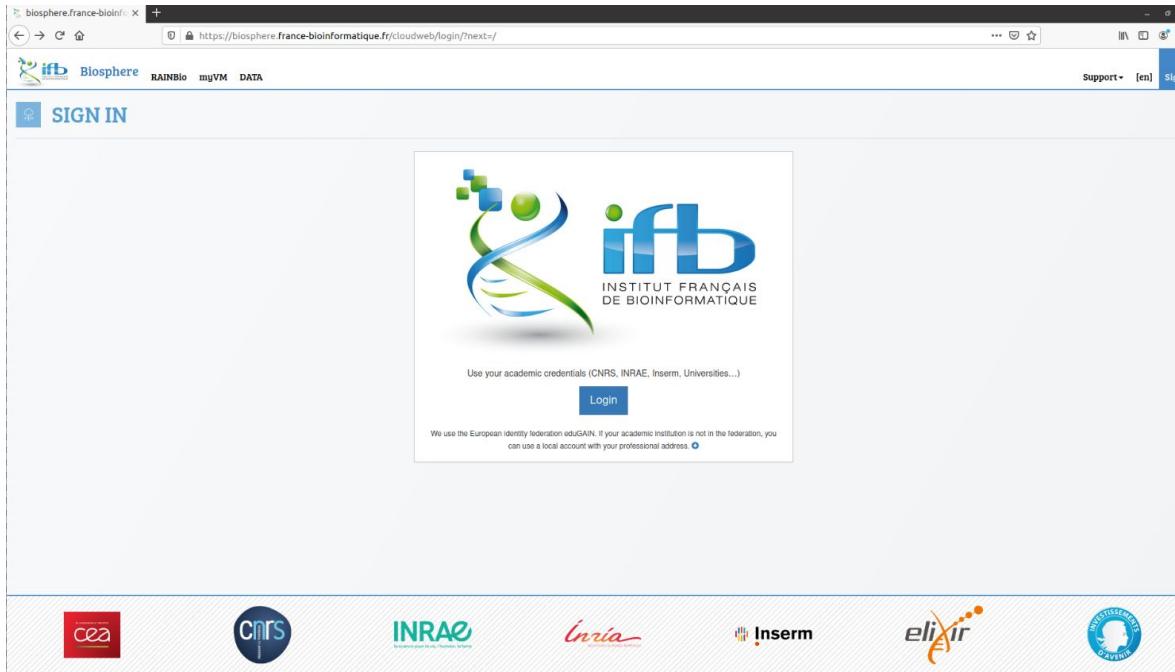


# Biosphere, IFB CLOUD FOR LIFE SCIENCES

- 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 resources, 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 header includes the IFB logo, navigation links for Biosphere, RAINBio, mgVM, and DATA, and a "Support" section. A "Pubkey missing" message is visible in the top right. The main content area is titled "BIOSPHERE FEATURES" and lists various services and resources 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, Fiiii, 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 ....   ⚙️

<b>AnalysesSV</b> ★ bcftools, BEDTools, BWA, Jupyter, Matplotlib, pandas, SAMtools ⚡ DNA polymorphism, Genetic variation, Genotyping experiment, GWAS study	<b>CoursAnalysesNanoporeSG</b> ★ bandage, Jupyter ⚡ Data architecture, analysis and design, Mathematics, Statistics and probability 🔧	<b>virus_ONT</b> ★ Jupyter ⚡ Data architecture, analysis and design, Mathematics, Statistics and probability 🔧	<b>ANF MetaBioDiv</b> ★ DESeq2, ggplot2, phyloseq, RStudio ⚡ Transcriptomics, Microbiology, Metagenomics, Sequence analysis
---	--	---	---



# Let's run your vm through the cloud

The screenshot shows a web-based interface for managing virtual machines. At the top, there is a navigation bar with the IFB Biosphere logo, followed by links to RAINBio, myVM, and DATA. On the right side, there is a user profile section with an email address (julie.orjuela@ird.fr) and a support link.

The main content area displays a virtual machine configuration for "CoursAnalysesNanoporeSG". The configuration includes:

- 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 (with a green dot) and Sequence analysis (with a green dot).
- Outils:** Jupyter
- OS:** Debian 11
- Recette de l'app (git):** [https://github.com/SouthGreenPlatform/training\\_ONT\\_VM/tree/2022](https://github.com/SouthGreenPlatform/training_ONT_VM/tree/2022)
- App de base:** Jupyter

On the far right, there is a vertical sidebar with buttons for "EDITER", "LANCER", and "DÉPLOIEMENT AVANCÉ". A large black arrow points from the text "Let's run your vm through the cloud" towards the "LANCER" button.

# 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 is set to "Julie\_ONT". The "Groupe à utiliser" dropdown shows "virus\_ont (Initiation à l'analyse de la séquençage de virus)" and "tagé nome viraux) 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.

Configurer le déploiement d'une appliance

Déploiement de l'appliance "virus\_ONT"

Name: Julie\_ONT

Groupe à utiliser: virus\_ont (Initiation à l'analyse de la séquençage de virus) tagé nome viraux) 828.01

Cloud: ifb-core-cloudbis

Gabarit d'image cloud:

- 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.m4.4xlarge (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)

Computational biology

Annuler

VM/tree/2022

Support julie.orjuela@ird.fr (eduGAIN)

EDITER LANCER DÉPLOIEMENT AVANCÉ

# 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 eduGAIN.
- Left Sidebar:** CLOUD
- Deployment List:** A table titled "Déploiements" showing two entries:
  - ID: 19804, Name: virus\_ONT (1.0), Status: testontvirus, Started: Sep 05 2022, 17h00, Groups: virus\_ont, Specification: Broker (8 cores, 32GB RAM), Cloud: da98, Access: ifb-core-cloudbis.
  - ID: 19759, Name: virus\_ONT (1.0), Status: ● (red dot), Started: Sep 05 2022, 10h25, Groups: DIADE, Specification: Broker (1 core, 4GB RAM), Cloud: b680, Access: (empty).
- Buttons:** Arrêter les déploiements (Stop deployments) and Tout voir (6) (View all 6).

A red arrow points to the "Cloud" column header in the deployment table, specifically highlighting the "da98" entry.

# 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	8 32 200	da98	ifb-core-cloudbis	https Params 134.158.248.119

Arrêter les déploiements

Tout voir (4)

Cloud

Déploiements

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	https Params 134.158.248.119

Arrêter les déploiements

Tout voir (4)

# Let's run your vm through the cloud

get the url... link "https"

The screenshot shows a web-based interface for managing cloud deployments. At the top, there is a header with a cloud icon and the word 'CLOUD'. Below the header, a navigation bar includes 'Déploiements' and a search icon. The main area displays a table of deployed VMs. The columns are labeled 'ID', 'Nom', 'Début', 'Groupes', 'Spécification', 'Broker', 'Cloud', and 'Accès'. A single row is visible, representing a deployment named 'virus\_ONT (1.0)' with ID 19804. The 'Début' column shows 'Sep 05 2022, 17h00'. The 'Groupes' column contains 'virus\_ont'. The 'Spécification' column shows memory settings: '8' (32, 200). The 'Cloud' column lists 'da98' and 'ifb-core-cloudbis'. The 'Accès' column shows 'https Params' and the IP '134.158.248.119'. A red trash can icon is present in this column. A large black arrow points from the bottom right towards the 'Accès' column. At the bottom left, there is a red button labeled 'Arrêter les déploiements'.

# 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" in the foreground. The dialog lists a single parameter:

nom	valeur
JUPYTER_TOKEN	28f9a32ae92eaecbc816880489c9217e3263f9fd4614352

In the background, a table displays a job named "virus". The "Accès" column for this job contains the URL "https://134.248.119.134" with a yellow arrow pointing to it. The table also includes columns for Début, Groupes, Spécification, Broker, Cloud, and Accès.

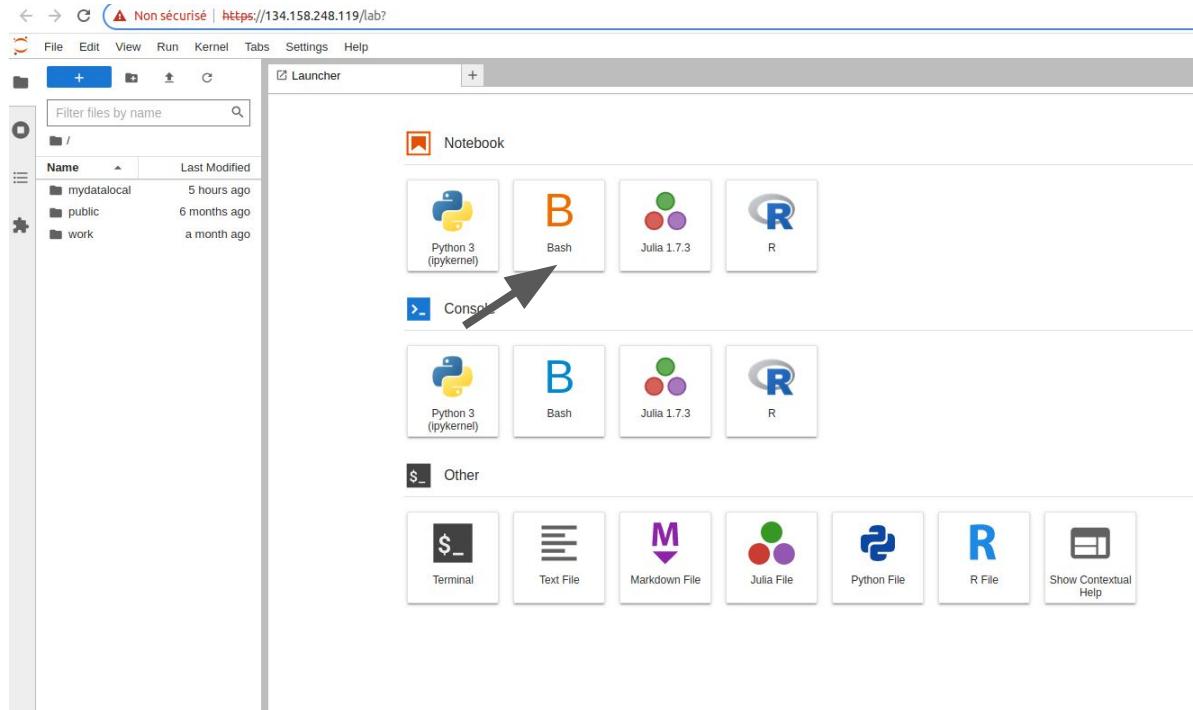
# Let's run our vm through the cloud

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

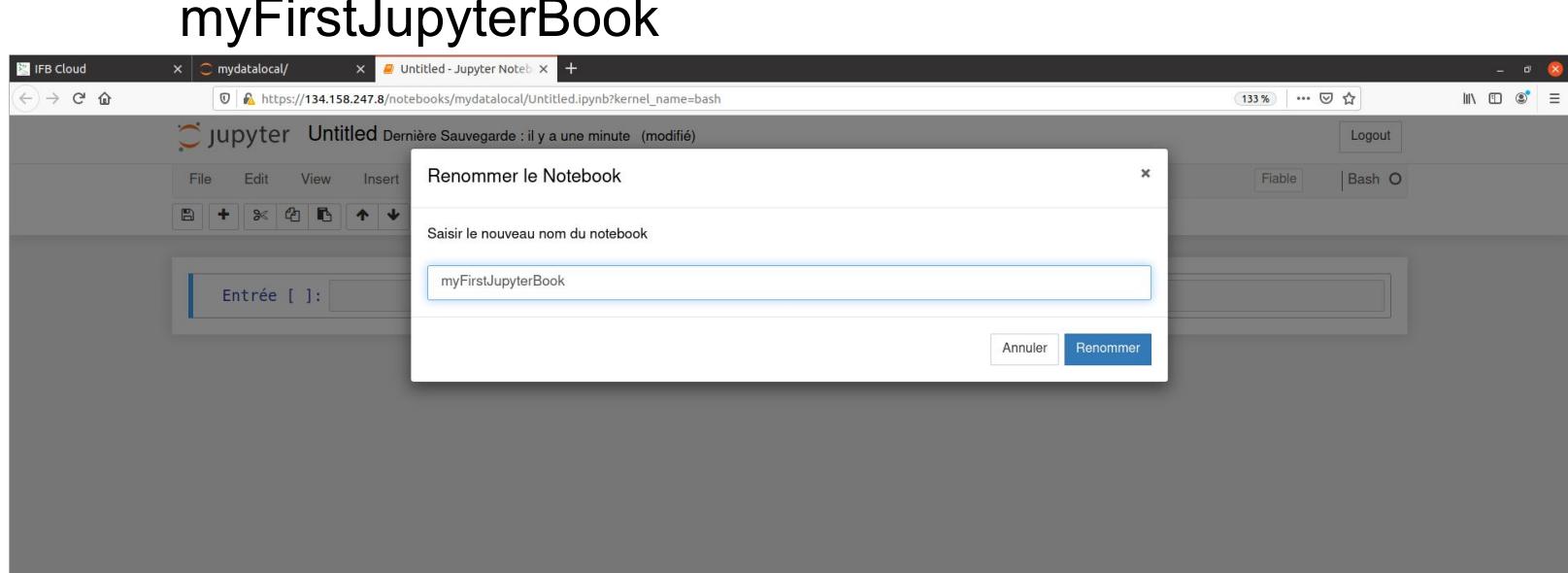


# Create your first jupyter book

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/tree/ONT\\_LA](https://github.com/SouthGreenPlatform/training_ONT_teaching/tree/ONT_LA)

## Run your first bash command - *git clone*

Download all the jupyter books with the following *git* commands  
(all in oneline)

`git clone --branch ONT_LA`

`https://github.com/SouthGreenPlatform/training\_ONT\_teaching.git`

# Run your first bash command - *git clone*

The screenshot shows a Jupyter Notebook interface with a terminal window open. The terminal window is titled 'Bash' and contains the following text:

```
[4]: pwd  
/home/jovyan/work
```

Below the terminal, there is explanatory text and a URL:

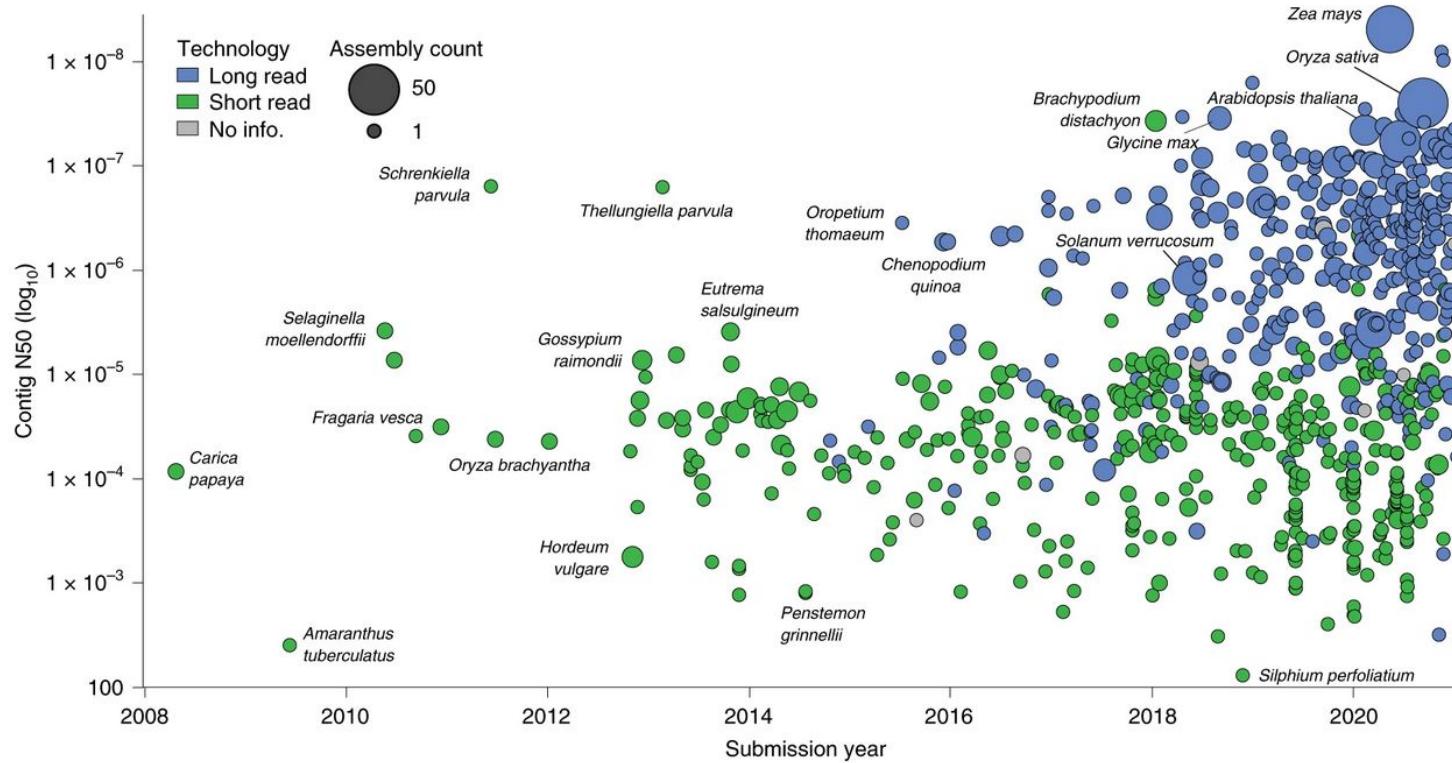
My first Juptyer book - Training SG SV  
My first linux command - `pwd`

Download all jupyter book we will use for this week - `git clone`

url [https://github.com/SouthGreenPlatform/training\\_SV\\_teaching/tree/2022](https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022)

```
[3]: 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.
```

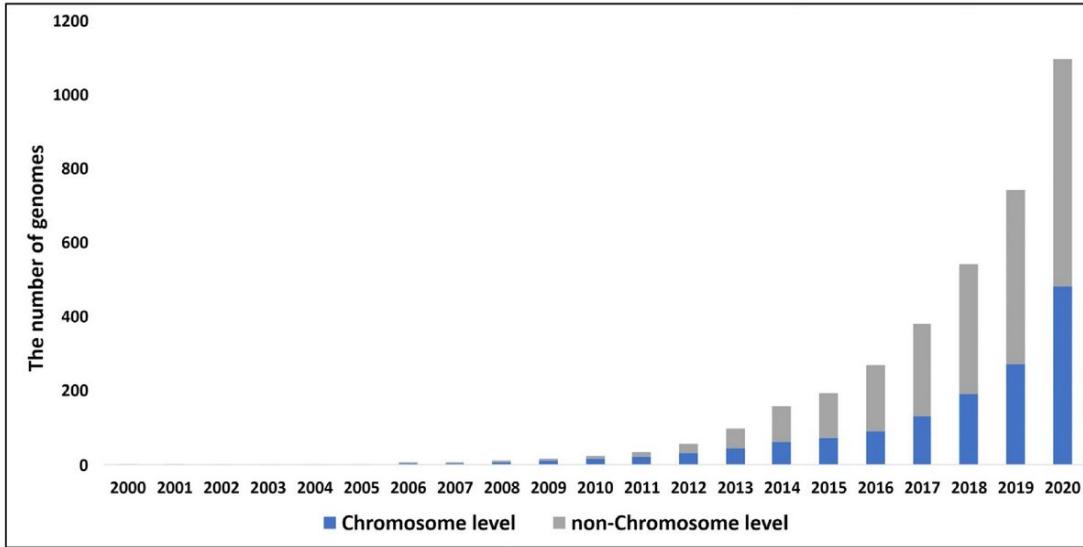
*Let's start !*



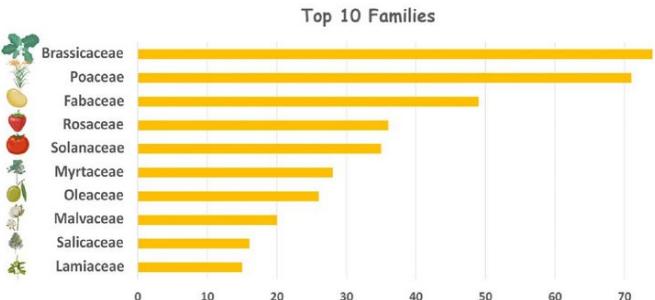
Assembly contiguity by submission date for 798 land plant species with publicly available genome assemblies. Points are coloured by the type of sequencing technology used and scaled by the number of assemblies available for that species. There is an improvement in contiguity associated with the advent of long-read sequencing technology, and a noticeable increase in the number of genome assemblies generated annually. All assemblies generated before 2008 have since been updated and are therefore not included.

# Published plant genomes from 2000

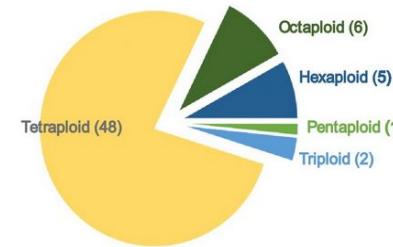
(A)



(B)



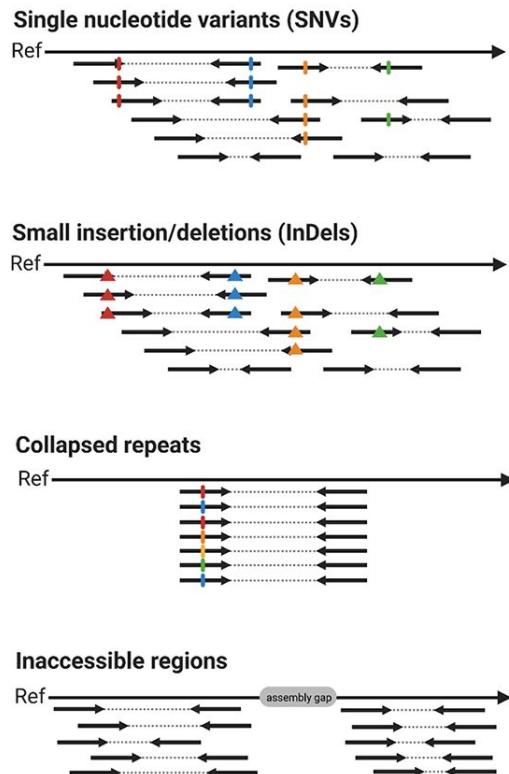
(C)



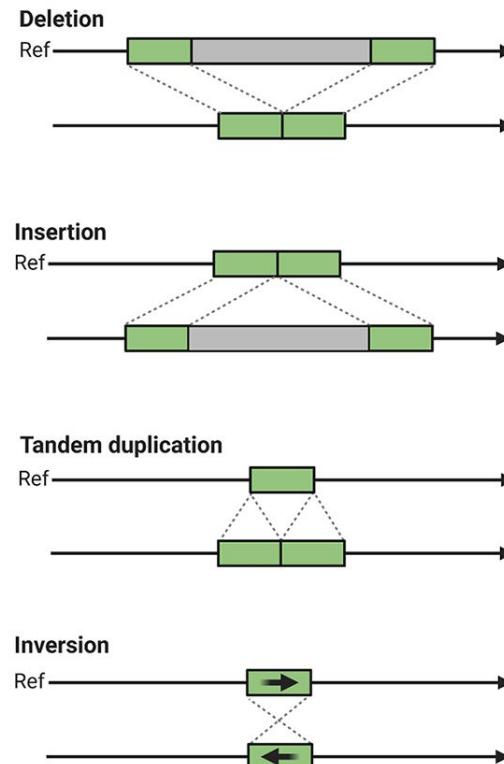
Sun et al, 2022

# Long Reads - What you can do with it ?

## A NGS variant calling

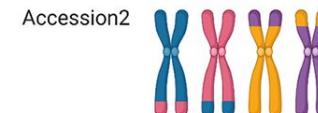
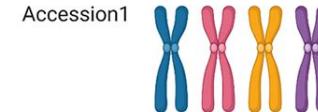


## B long read variant calling

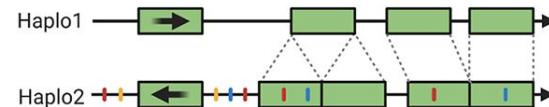


## C *de novo* assembly

### Chromosomal rearrangements



### Separated haplophases

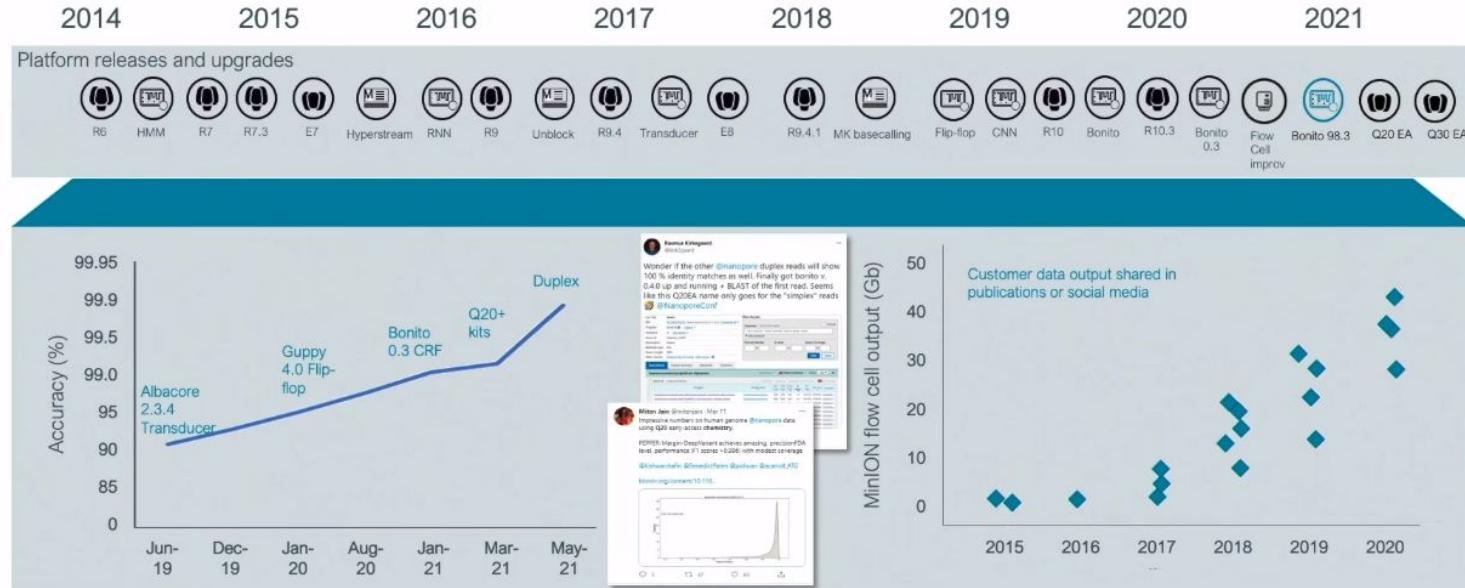


# (Plant) genome project workflow from DNA extraction over ONT sequencing to data submission

	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		

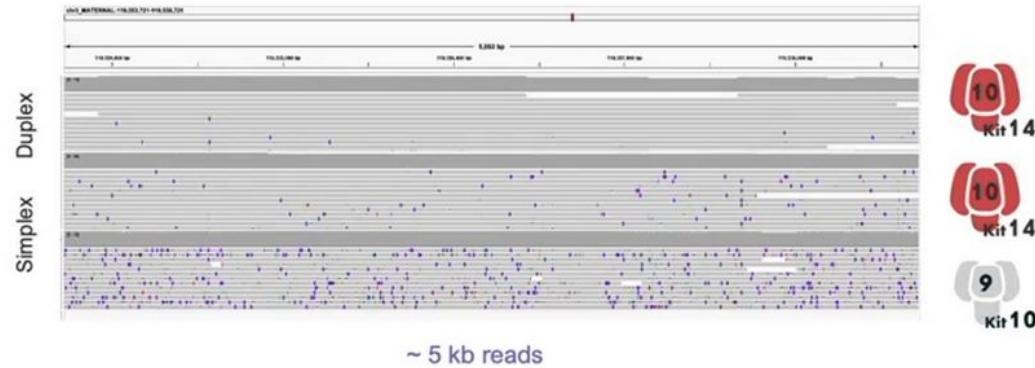
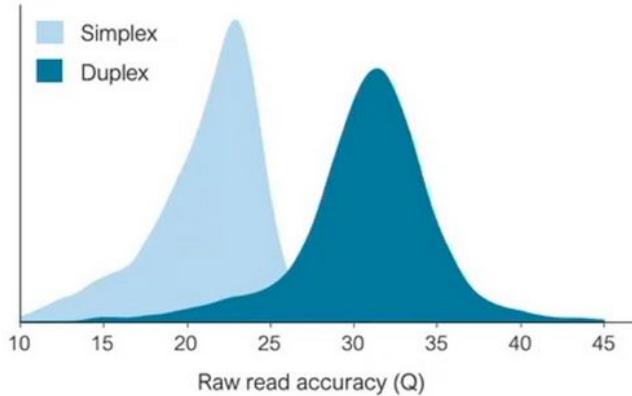
## Upgrades drive performance enhancements

...and core ones ship in consumables and software



# Duplex Accuracy

## Duplex accuracy



- Accuracy of Duplex raw reads ~ Q32 @ 5 kHz
- Runtime of new Stereo base caller similar to Simplex calls
- Duplex accuracy independent of read length
- Clearly visible improvements to errors



Alexander Wittenberg @AW\_NGS · 30 Nov

@KeyGeneInfo received novel High duplex PromethION flow cells from @nanopore as Developer access. This pore is optimised to generate significant higher duplex data. Very high accuracy & long reads! Evaluation next week on first crops 🌱 #nanoporeconf



# A lot of data !



MinION



MinION Mk1C



MinION Mk1D



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	1	5		2	2	24	48
Maximum number of channels per flow cell	512	512	512	512		2,675	2,675	2,675	2,675
Run time	Up to 72 hours		Up to 72 hours	Up to 72 hours	Up to 72 hours	Up to 72 hours			
Device TMO <sup>†</sup>	50 Gb	50 Gb	50 Gb	250 Gb		580 Gb	580 Gb	~7 Tb	~14 Tb
Maximum number of flow cells per year*	104	104	104	520		208	208	2,596	4,992
Offer sequencing as a service	No	No	No	Yes		Yes	Yes	Yes	Yes

1	Platform	Run time max: (d)	Yield low: (Gb)	Yield high: (Gb)	Rate max: (Gb/d)	Reagents max: (\$)	Price per Gbp max: (\$)	Price per Gbp min: (\$)	hg-30x min instr amm (\$)	hg-30x min plus 5yr amm 10% serv 90% util (\$)	Machine: (\$ K)	Install base	Availability	Max total theoretical output per day (Gb) from current day (Gb)	Max total output per day (Gb) from installbase	
32	ONT P2 Solo 1fcell R10.4HD Q30+ data with Kit14	3.00	NA	60	20	1530-816	15.00	12.42	1242	1247	1271	10.5	430	worldwide	20	17,200
33	ONT P2 1fcell R10.4HD Q30+ data with Kit14	3.00	NA	60	20	1530-816	15.00	12.42	1242	1275	1410	60		worldwide	20	0
34	ONT P24 24fcell R10.4HD Q30+ data with Kit14	3.00	NA	1440	480	1530-655	15.00	10.00	1000	1005	1024	225	15	worldwide	535	8,025
35	ONT P48 48fcell R10.4HD Q30+ data with Kit14	3.00	NA	2880	960	1530-655	15.00	10.00	1000	1004	1018	310	60	worldwide	960	57,600

all prices of all technologies at may 2023

[https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/edit?hl=en\\_GB&hl=en\\_GB#gid=1569422585](https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/edit?hl=en_GB&hl=en_GB#gid=1569422585)



Albert Vilella  
@AlbertVilella

...

Oxford @nanopore has updated the pricing for the PromethION flowcells and so did I in the NGS specs spreadsheet. Numbers in the lined squares are for Q30+ Duplex data, based on the 60Gb per R10.4HD flowcell as per announced at #NanoporeConf



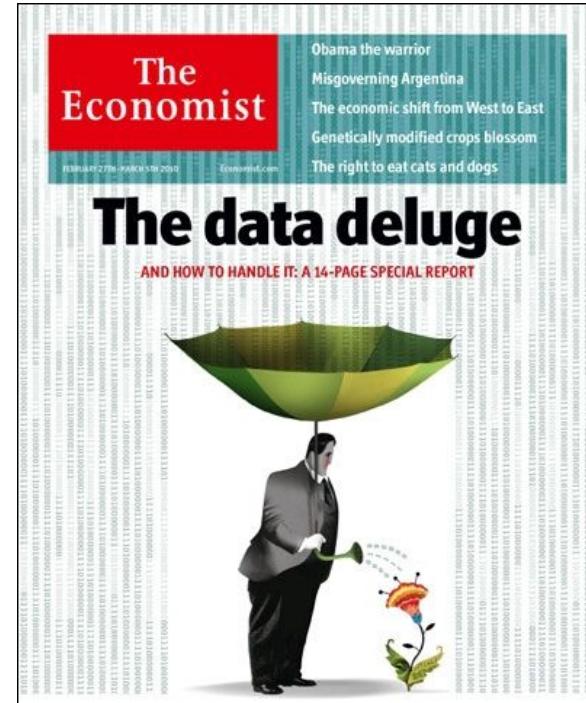
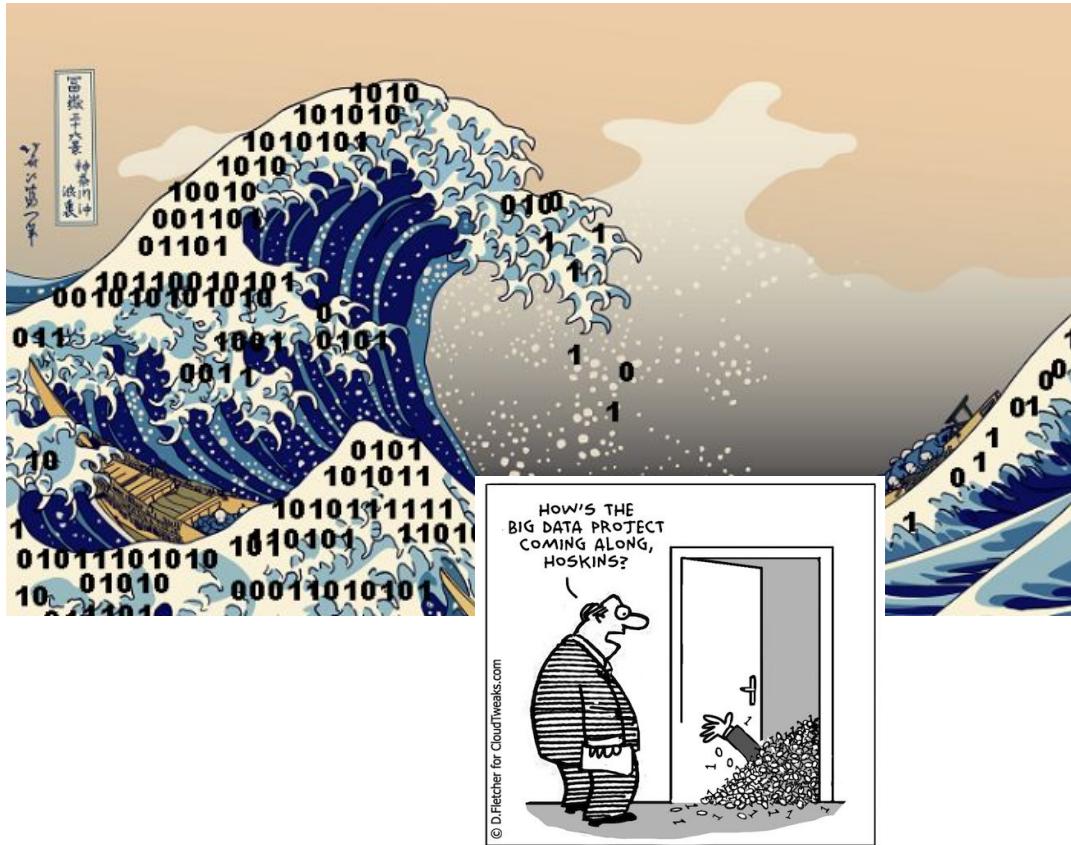
Matthew Miller - ↗ ... @b... · May 18

Wow, a bit of an early @nanopore surprise. Promethion flow cells now are just 900 USD, even at the four pack. As expected, this is about a 30% reduction in the cost to get a Promethion worth of Nanopore data. Huge news! #NanoporeConf

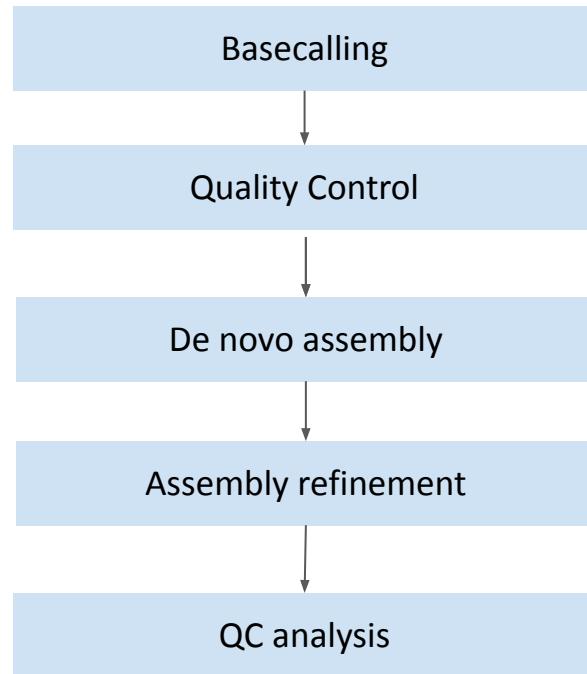
The data that these platforms produce differ qualitatively from second-generation sequencing, thus necessitating tailored analysis tools



# From data rarity to data deluge

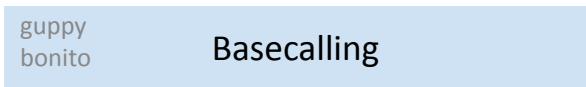


# Typical long-read analysis pipelines for ONT data



# Typical long-read analysis pipelines for ONT data

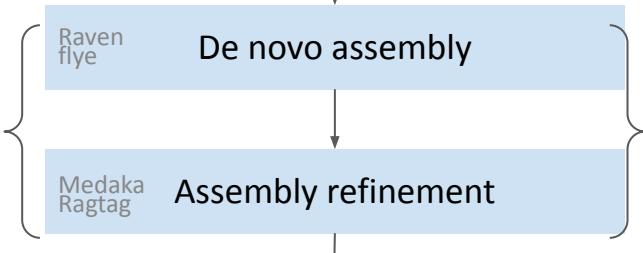
Demo



Practical 1

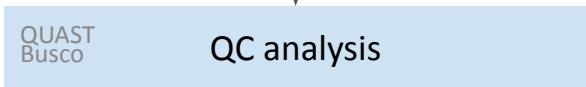


Practical 2



[Optional]

Practical 3



*The Data !*

# The Pan-Genome of the cosmopolitan picophytoplankton *Bathycoccus prasinus*

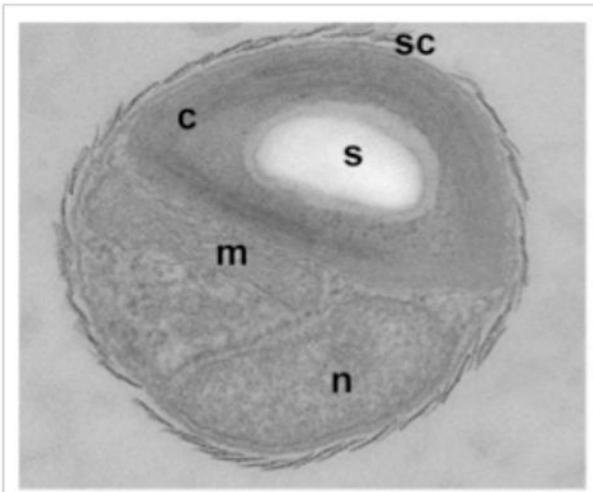
A first step towards understanding adaptation to latitude and seasons

Louis Denu, LOMIC

supervised by

François-Yves Bouget, LOMIC - Martine Devic, LOMIC - François Sabot, IRD Montpellier

## *Bathycoccus prasinus*

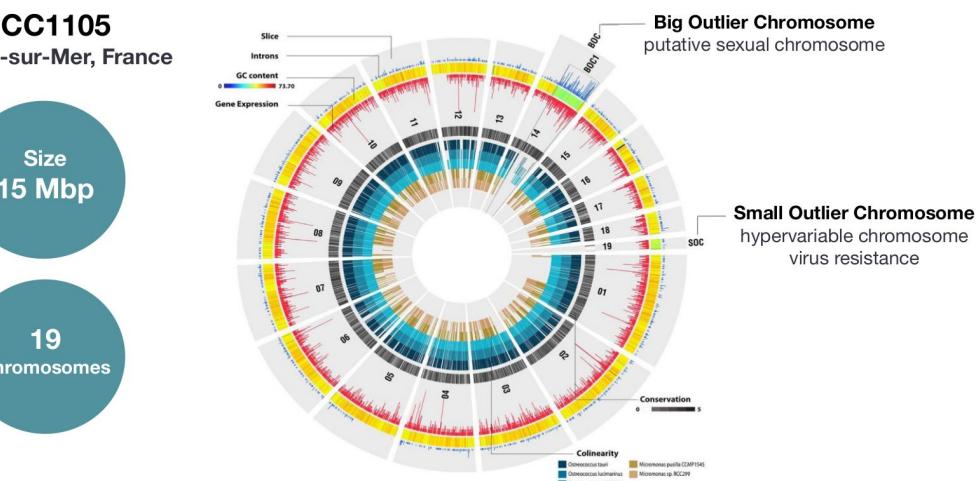


RCC1105  
Banyuls-sur-Mer, France

Size  
15 Mbp

19  
Chromosomes

Moreau et al., *Genome Biology* 2012



# *Chapter 1*

# *Reads Quality Control*

# MinION Mk1B (MN44570) Final report



30 Oct 23, 17:20 — 02 Nov 23, 17:25 · AfricaRice · WAB0001642 · MN44570  
Protocol run ID: ceed113-3fe5-4d22-a41e-2dc48162bb6e

[Run summary](#) | [Run configuration](#) | [Sequence output](#) | [Run health](#) | [Run log](#)

## ▲ Run summary

### DATA OUTPUT

Estimated bases	6.97 Gb
Reads generated	87.49 GB
Estimated N50	726.39 k
Reads generated	20.45 kb

### RUN DURATION

Run time	
Elapsed time	
Run limit	
Run status	<b>FINISHED</b> · Target runtime has been reached

### BASECALLING

Reads called	100%
Bases called (min Q score: 8)	4.22 Gb
Pass	2.29 Gb

[View unit abbreviations used in this report](#)

## ▲ Run configuration

### ▲ RUN SETUP

Flow cell type	FLO-MIN114
Flow cell type alias	FLO-MIN114
Flow cell ID	FAW45275
Kit type	SQK-LSK114

### ▲ RUN SETTINGS

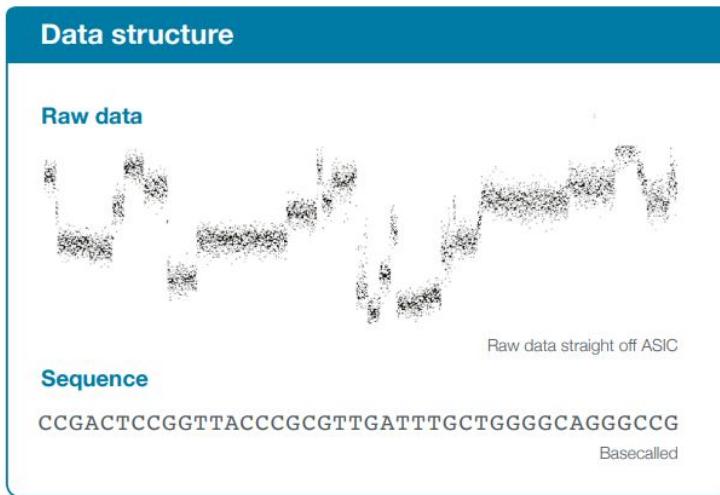
Run limit

72h

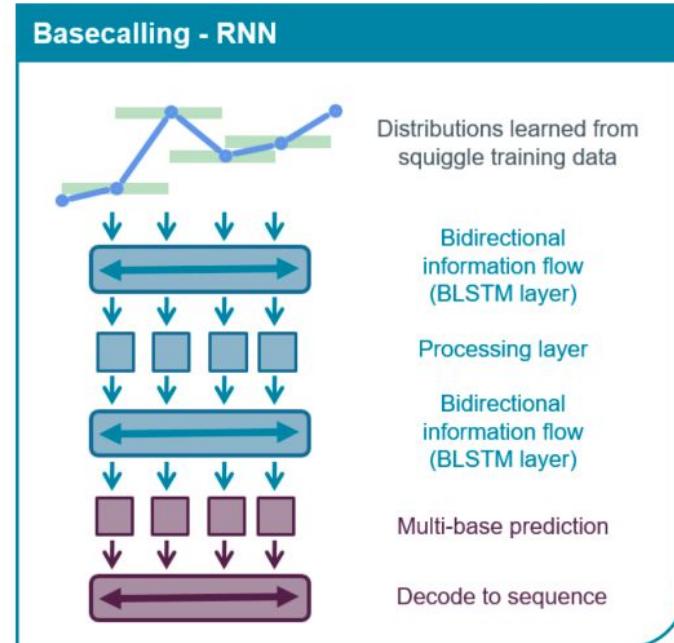
### ▲ DATA OUTPUT SETTINGS

FAST5 output	vbz_compress
FAST5 reads per file	4000
FASTQ output	gzip_compress
FASTQ reads per file	4000
BAM output	Off
Bulk file output	Off
Data location	/var/lib/minknow/data/.

## 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



Chris Seymour ✅ @iiSeymour · May 18

Huge dorado release 🎉

...

- 1.4x simplex perf HAC/SUP on A100
- 5khz v4.2 models
- All context 5mC, 6mA models
- Significantly faster duplex calling
- Duplex pairing handled automatically
- BAM output
- Aligned output via minimap2

#nanoporeconf



New Release v0.3.0

# v0.3.0

[0.3.0] (18 May 2023)

This is a major release of Dorado which introduces: Duplex pairing and splitting for directly going from POD5 to duplex reads, major...



github.com

Release v0.3.0 · nanoporetech/dorado

[0.3.0] (18 May 2023) This is a major release of Dorado which introduces: Duplex pairing and splitting for directly going from POD...

4

65

156

23.2K



Chris Seymour ✅ @iiSeymour · 9 oct.

Dorado v0.4.0 is live!

...

- barcode demux & trimming
- simplex read splitting
- improved 6mA, 5mC/5hmC GpG models
- new 5mC/5hmC all context model
- increasing duplex pairing rates

Plus a whole bunch more..

[Traduire avec DeepL](#) 🇫🇷

New Release v0.4.0

# v0.4.0

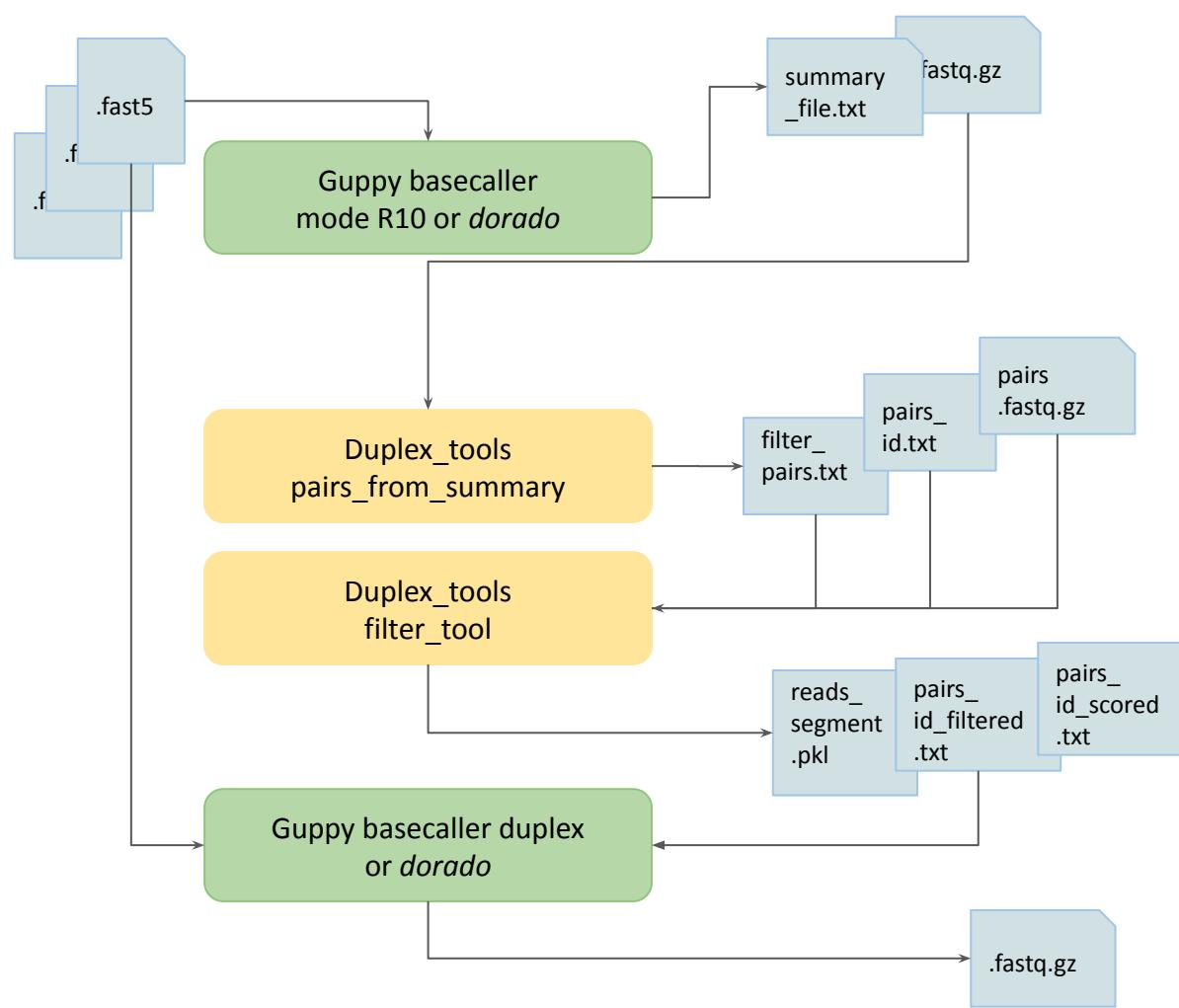
[0.4.0] (9 Oct 2023)

This release of Dorado introduces barcode demultiplexing, barcode trimming, simplex read splitting, and updated models for calling 6m...



github.com

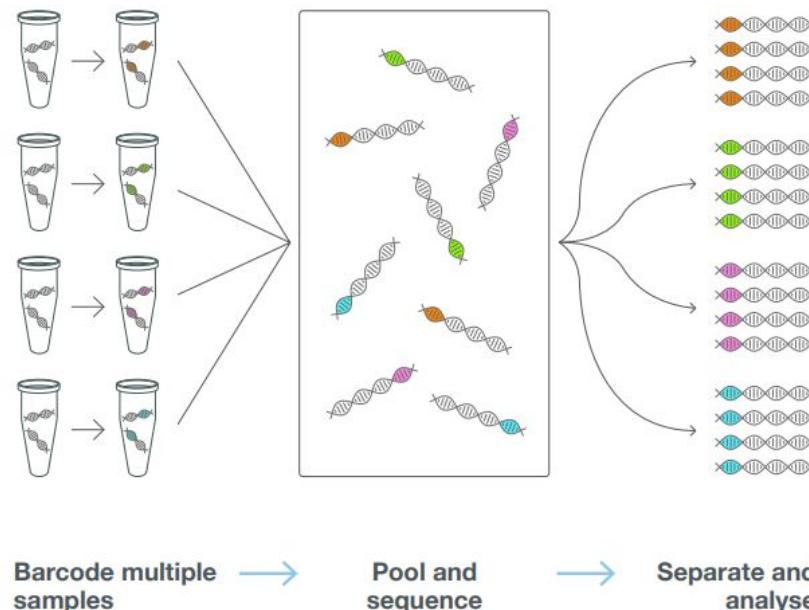




# ONT demultiplexing

## Guppy

```
guppy_basecaller \  
--input_path <folder containing .fast5 files> \  
--save_path <output folder> \  
--config dna_r9.4.1_450bps_fast.cfg \  
--barcode_kits SQK-RBK004
```



## ONT Read calling, cleaning and filtering

Sequencer ONT : raw fast5 files (or POD5)

- Transform fast5 signal in fastq standard format *Guppy Bonito/Dorado/Rerio*
- 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 sequence = 4 lines

```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTTCTTAGTATTTTGATGTCAATTCCGTGTTGGTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTGTT
+
@@@DDDD>FFFAFBEBB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTCAGTAACGACCTATAAATTAAAGTAGC
+
@CFFFFFGGHHHHIIJJJIEE<HHHIJJIGBHGGEEIJJEIEIJHHJFIIJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTCGAAAAGTCGGGTATGGCTCTAGTAGCTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIEEHIII<CGHIJIIJ:?:FC9DGAFGHII?DGBFIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAAGCTAAAAGAACACAGTTGCTGAAGCAGCAAACACAAGAAC
+
B@@@DFFFFFGHHHHJIIIIJJJIIGJCHHEIII>GHIG@GHIDHGJIIFHIIJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGAAGAGTTAAAAACAATTATGAGCAACTGAGTTTC
+
@@@CFFFFFHFFFHFGIJJIIHHIIJJIIHJJECGHJJCHGICDGHHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTGTCGGGGTTGGAAATTGTCACTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```



- @sequence ID
- Sequence in IUPAC code
- +
- Sequence quality in ASCII values

## ***Reads quality***

**Phred quality score:** confidence score for each sequenced base

Ranging from 0 to 93 (the higher the better)

Base	T	G	A	T	A	G	T	T	A	T	G
Score	32	40	41	35	29	23	26	32	36	32	14
ASCII	A	I	J	D	>	8	;	A	E	A	/

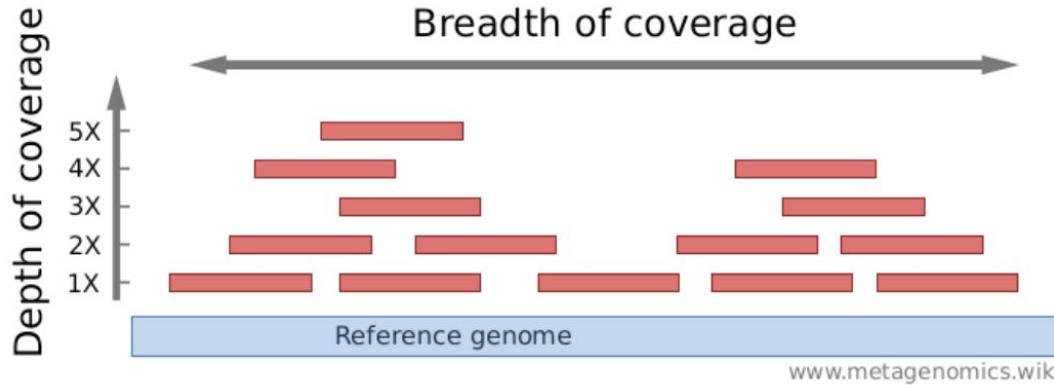
In FASTQ files scores are encoded in ASCII characters (33 to 126)

Score indicates probability  $P$  of a wrong base:

$$P = 10^{\frac{-Q}{10}}$$

Phred score of 10  $\leftrightarrow$  10% error rate ; score of 20  $\leftrightarrow$  1% error rate

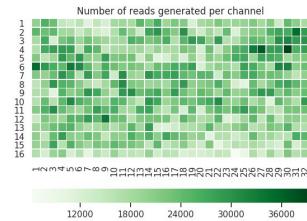
## Computing depth of coverage



depth of coverage estimation :

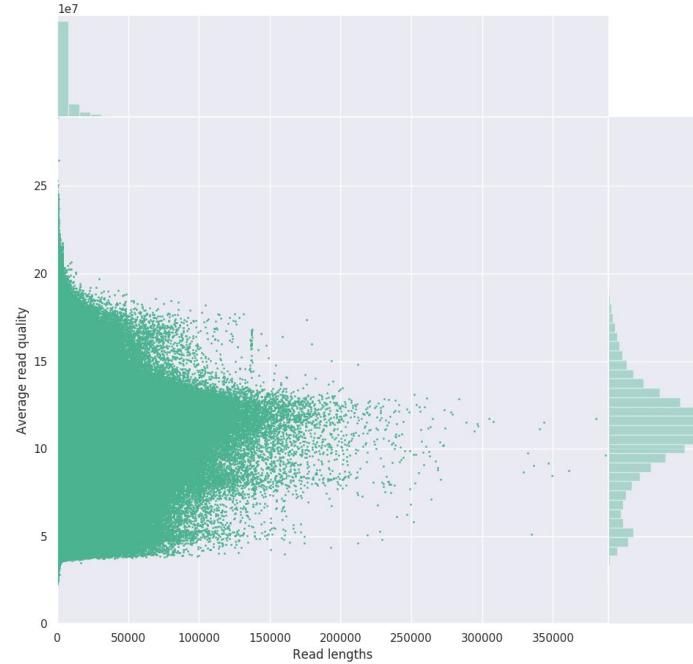
- Count how much base pairs in all sequenced reads? *total\_pb*
- What is the expected genome size? *genome\_size*

$\text{depth\_of\_coverage} = \text{total\_pb}/\text{genome\_size}$



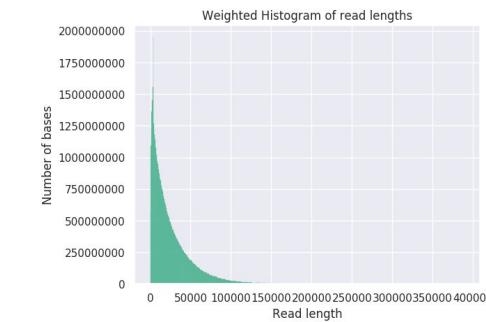
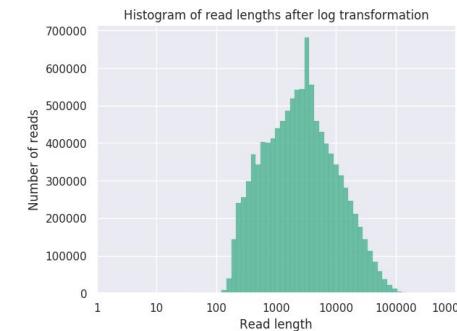
# Reads Quality control : *NanoPlot*

Read lengths vs Average read quality plot

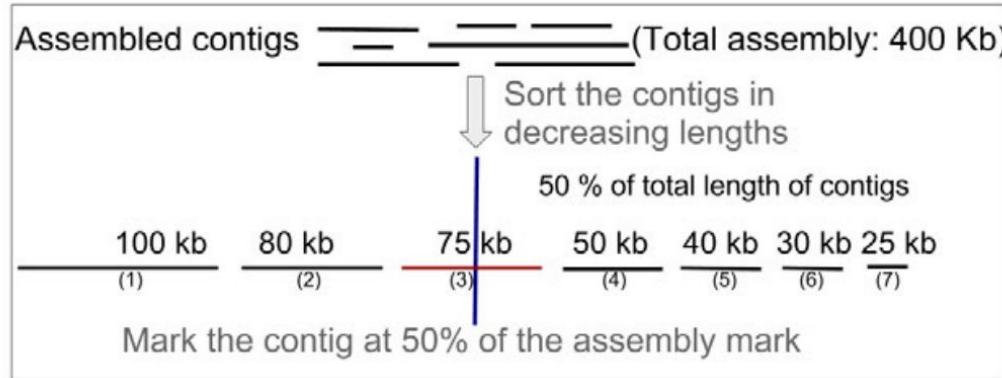


## Summary statistics

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](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/ONT\\_LA/1.light\\_raw\\_quality\\_control.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/ONT_LA/1.light_raw_quality_control.ipynb)

# *Chapter 2*

# *Assemblies*

# What do you want to do with these long reads?



## Research areas

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

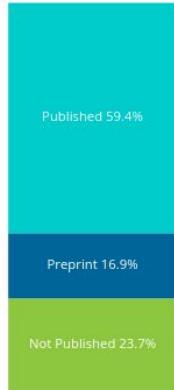
## Investigations

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

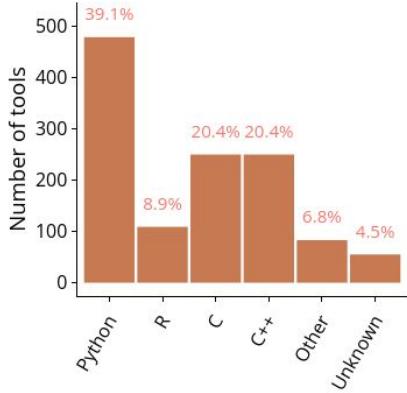
## Techniques

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

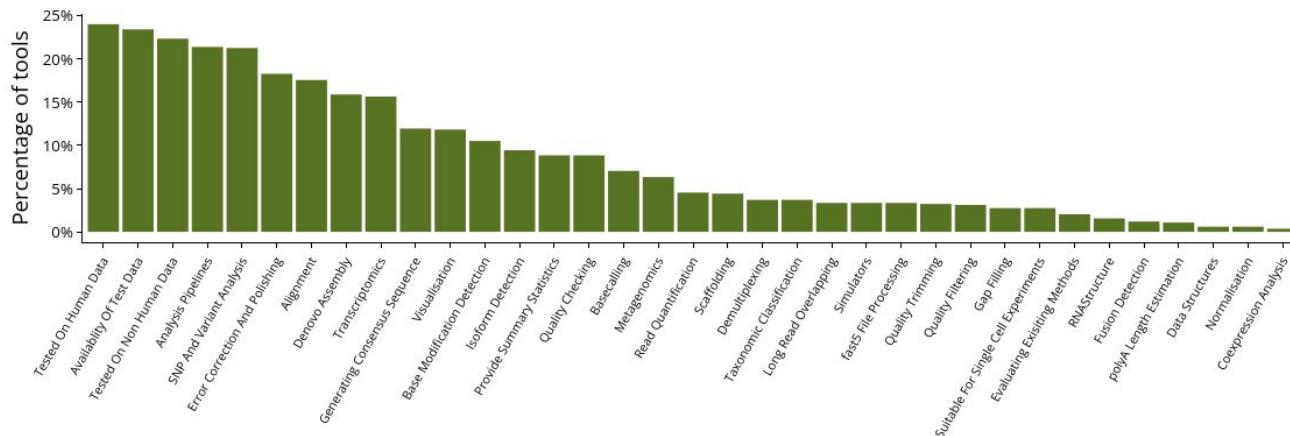
## Publication status



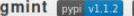
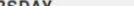
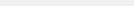
## Platforms



## Categories



# Most cited assembly tools for ONT

SPADEs	Canu
Unicycler	Flye
ALLPATHS-LG	MaSuRCA
Miniasm	MHAP
PBcR	StringTie2
Wtdbg2	Circlator   
Purge Haplotigs	RaGOO
HiCanu	Shasta   
MECAT	DBG2OLC
ABruijn	OPERA-MS
LINKS	SDA
DALIGNER	FinisherSC
TULIP	NpScarf
Platanus-Allee	Tigmint   
MdBG	PoreSeq
Hinge	ONTrack
Darwin	RagTag   
ASA3P	Verkko
MicroBIE   	l5sdav

<https://long-read-tools.org/>

Pilon

Racon

Zika Bioinformatics Pipeline

Wtdbg2

SQANTI3

NextPolish

MECAT

SQANTI2

NaS

INC-Seq

SiCeLoRe

Tigmint   

Hapo-G

Homopolish

CoLoRMap

TallyNN

TranscriptClean

MicroPIPE   

Flye

Nanocorrect

PBcR

Nanopolish

NextPolish

Nanocorr

SQANTI

Quickmerge

Longread-UMI-Pipeline

FinisherSC

FMLRC

LoRMA

NtEdit

CrossStitch

ASA3P

Daccord

Verkko

LRSDAY

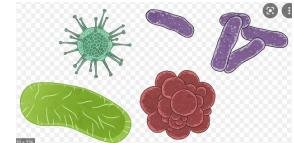
Most cited polishing tools for  
ONT

<https://long-read-tools.org/>

# 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.

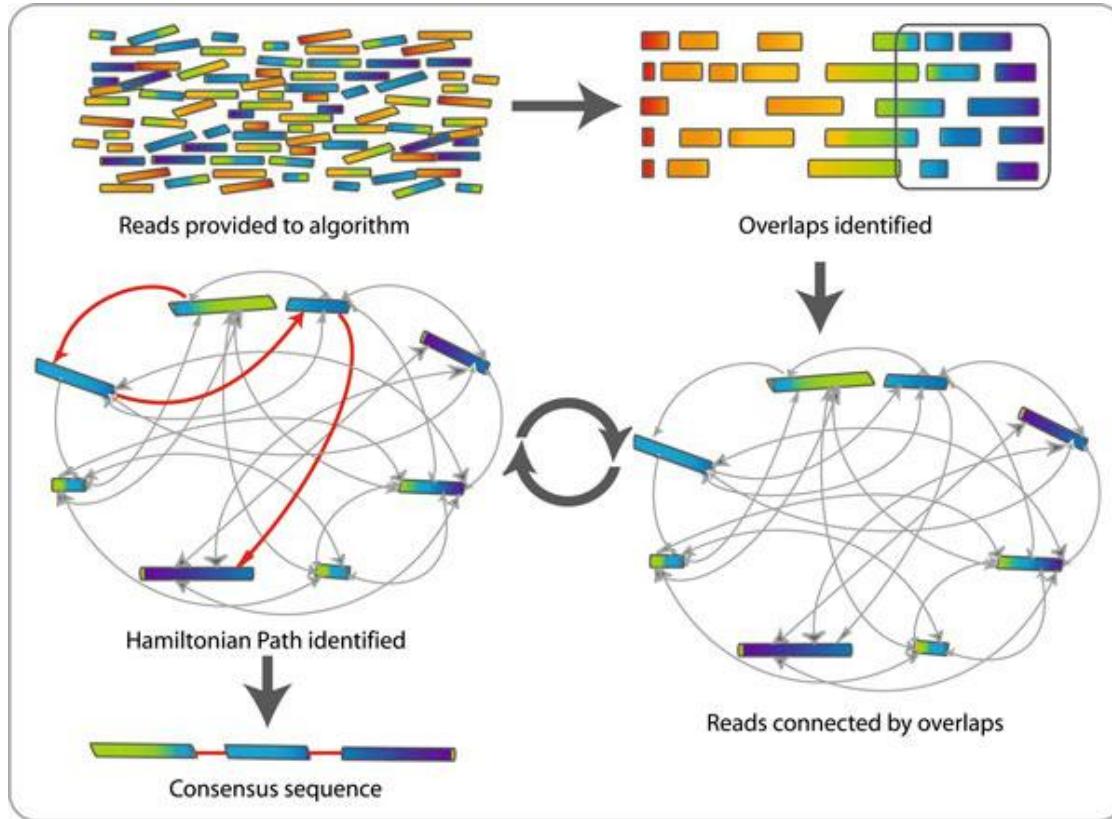
<https://nanoporetech.com/sites/default/files/s3/literature/microbial-genome-assembly-workflow.pdf>



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>

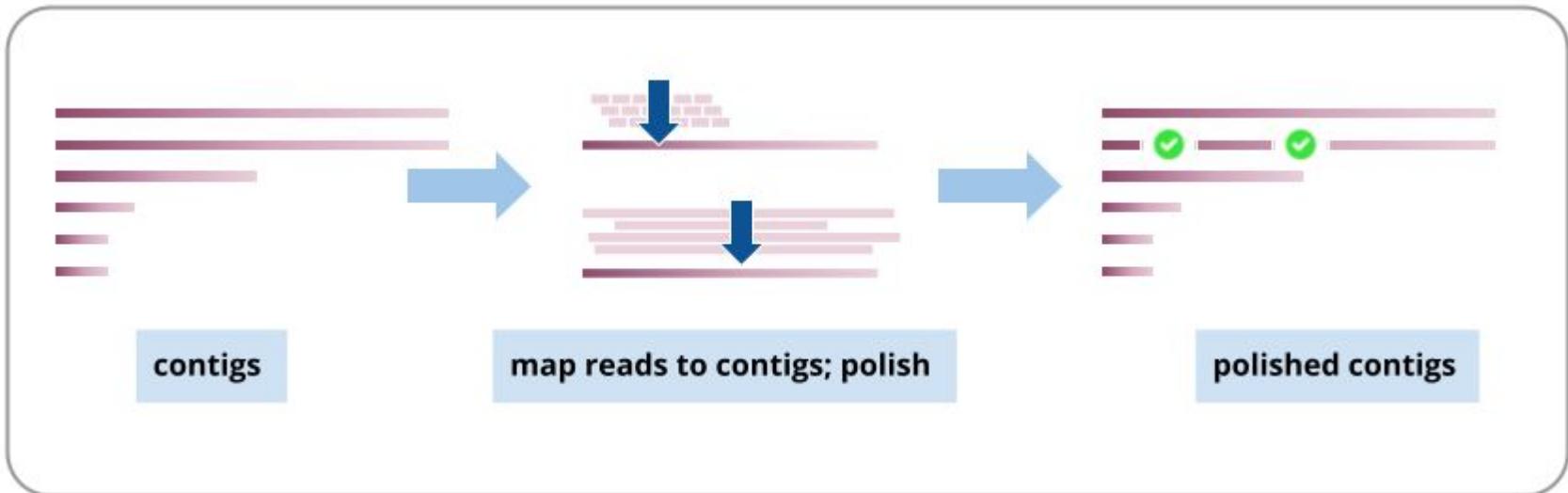
## Overlap–layout–consensus genome assembly algorithm (OLC)



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

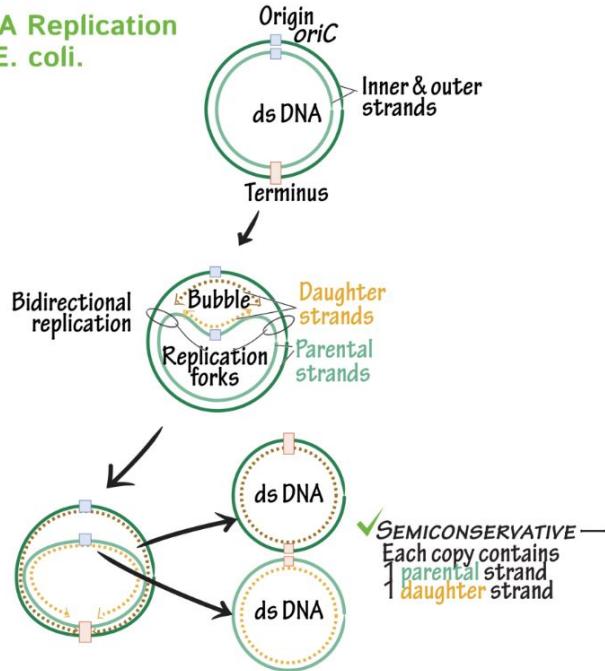
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3055744/>

# Polishing / Correction



# 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/ONT\\_LA/2.light\\_assemblies.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/ONT_LA/2.light_assemblies.ipynb)

# *Chapter 3*

# *Contigs Quality Control*

# 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  $\geq 3000$  bp, unless otherwise noted (e.g., "# contigs ( $\geq 0$  bp)" and "Total length ( $\geq 0$  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
<b>Genome statistics</b>			
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
<b>Misassemblies</b>			
# misassemblies	9633	9923	7666
Misassembled contigs length	373 371 138	373 825 172	335 007 830
<b>Mismatches</b>			
# 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
<b>Statistics without reference</b>			
# 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 ( $\geq 1000$ bp)	383 173 133	384 197 574	387 291 200
Total length ( $\geq 10000$ bp)	382 901 616	383 618 037	387 291 200
Total length ( $\geq 50000$ bp)	381 421 486	381 880 053	387 291 200
250	13 998 410	383 785 534	369 892 751
729	6 500 937	369 785 534	369 966 935
854	6 543 040	368 865 072	371 578 702
373 136 825	368 382 574	365 953 108	373 406 571

[Extended report](#)

smaller nb of contigs : flye+racon puis raven+racon  
longer 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
<b>Statistics without reference</b>						
# contigs	181	250	250	250	729	854
# contigs ( $\geq 0$ bp)	194	285	250	250	767	1149
# contigs ( $\geq 1000$ bp)	188	274	250	250	763	1000
# contigs ( $\geq 5000$ bp)	168	207	250	250	674	746
# contigs ( $\geq 10000$ bp)	139	156	250	250	564	587
# contigs ( $\geq 25000$ bp)	97	99	250	250	487	488
# contigs ( $\geq 50000$ 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 ( $\geq 0$ bp)	383 176 103	384 204 105	387 291 200	383 785 534	369 969 110	373 471 297
Total length ( $\geq 1000$ bp)	383 173 133	384 197 574	387 291 200	383 785 534	369 966 935	373 406 571
Total length ( $\geq 5000$ bp)	383 108 497	383 977 711	387 291 200	383 785 534	369 668 739	372 705 755
Total length ( $\geq 10000$ bp)	382 901 616	383 618 037	387 291 200	383 785 534	368 865 072	371 578 702
Total length ( $\geq 25000$ bp)	382 215 424	382 691 571	387 291 200	383 785 534	367 717 125	370 136 458
Total length ( $\geq 50000$ 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
<b>Similarity statistics</b>						
# 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

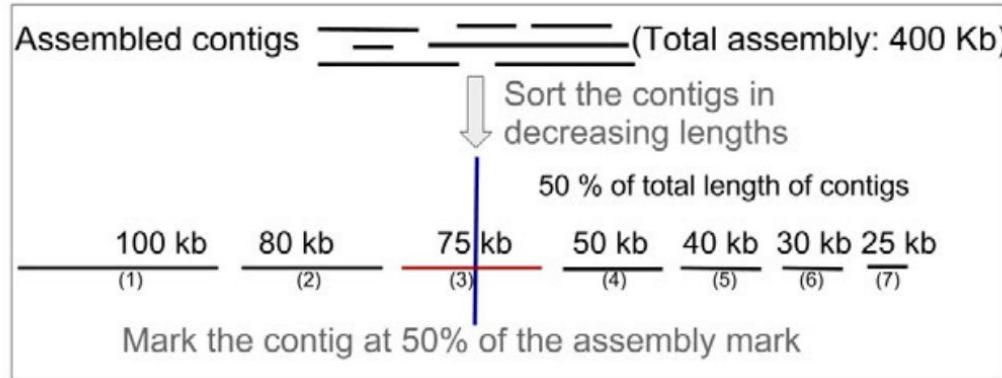
largest contig : flye+racon

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  $\geq 3000$  bp, unless otherwise noted (e.g., "# contigs ( $\geq 0$  bp)" and "Total length ( $\geq 0$  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
<b>Genome statistics</b>			
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
<b>Misassemblies</b>			
# misassemblies	9633	9923	7666
Misassembled contigs length	373 371 138	373 825 172	335 007 830
<b>Mismatches</b>			
# 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
<b>Statistics without reference</b>			
# 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 ( $\geq 1000$ bp)	383 173 133	384 197 574	387 291 200
Total length ( $\geq 10000$ bp)	382 901 616	383 618 037	387 291 200
Total length ( $\geq 50000$ bp)	381 421 486	381 880 053	387 291 200

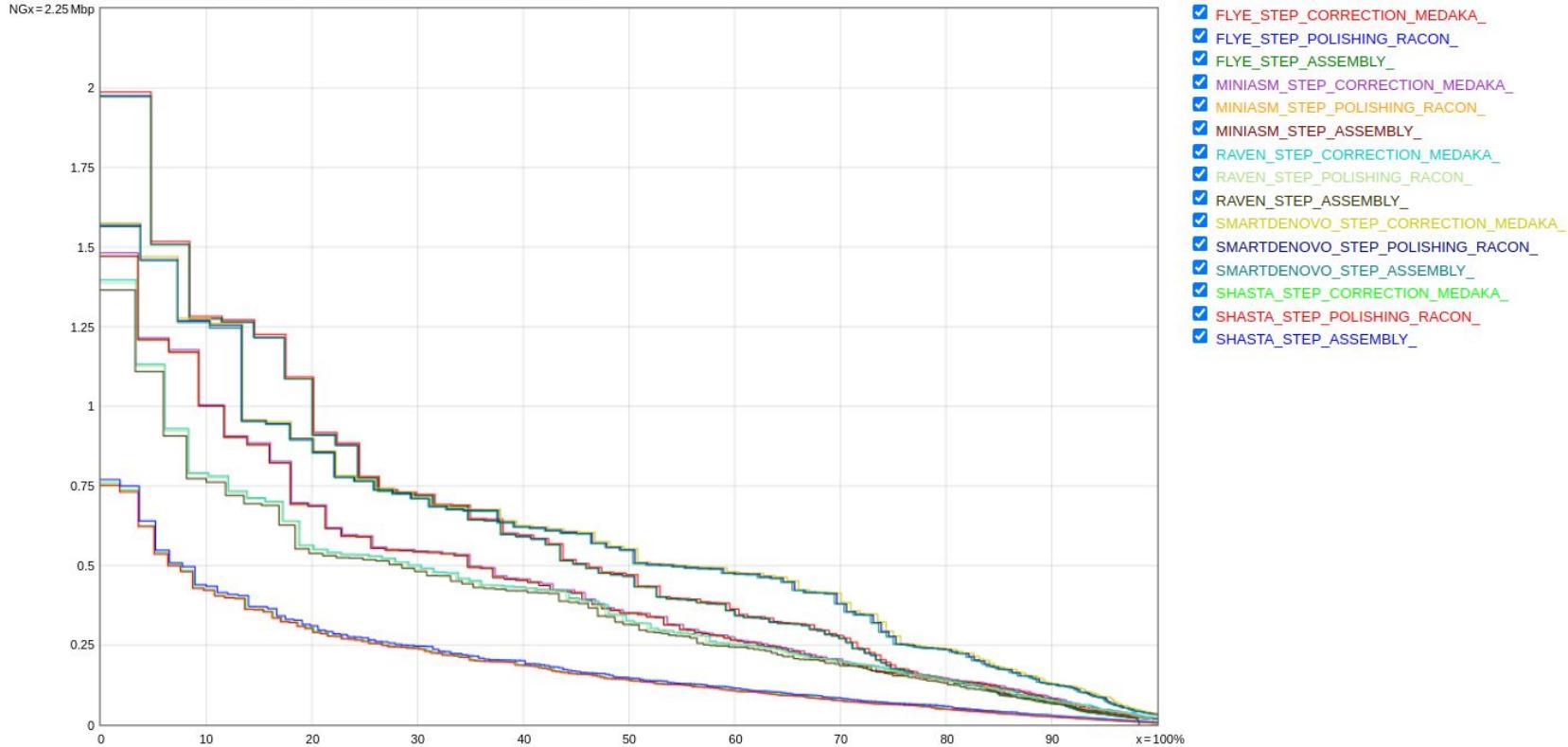
[Extended report](#)

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/
```

# TP3. Contigs Quality

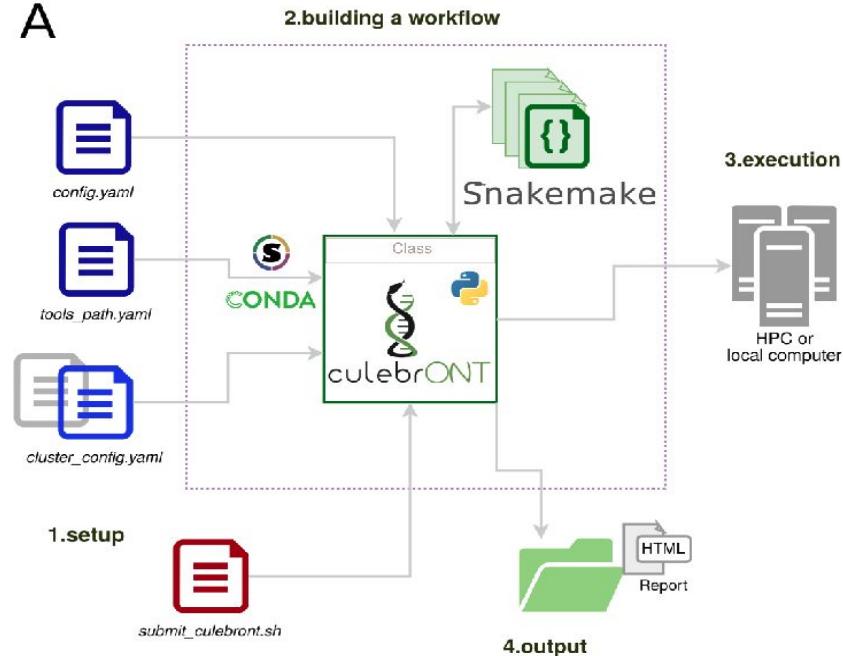
- TP3

[https://github.com/SouthGreenPlatform/training\\_ONT\\_teaching/  
blob/ONT\\_AL/3.light\\_contigs\\_quality.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/ONT_AL/3.light_contigs_quality.ipynb)

# A flexible and reproducible pipeline for LR assembly and evaluation

pip install culebrONT

A



- A recommendation in PCI Genomics <https://genomics.peercommunityin.org/articles/rec?id=158>
- An article in PCJ <DOI:10.24072/pcjournal.153>

## From contigs to chromosomes

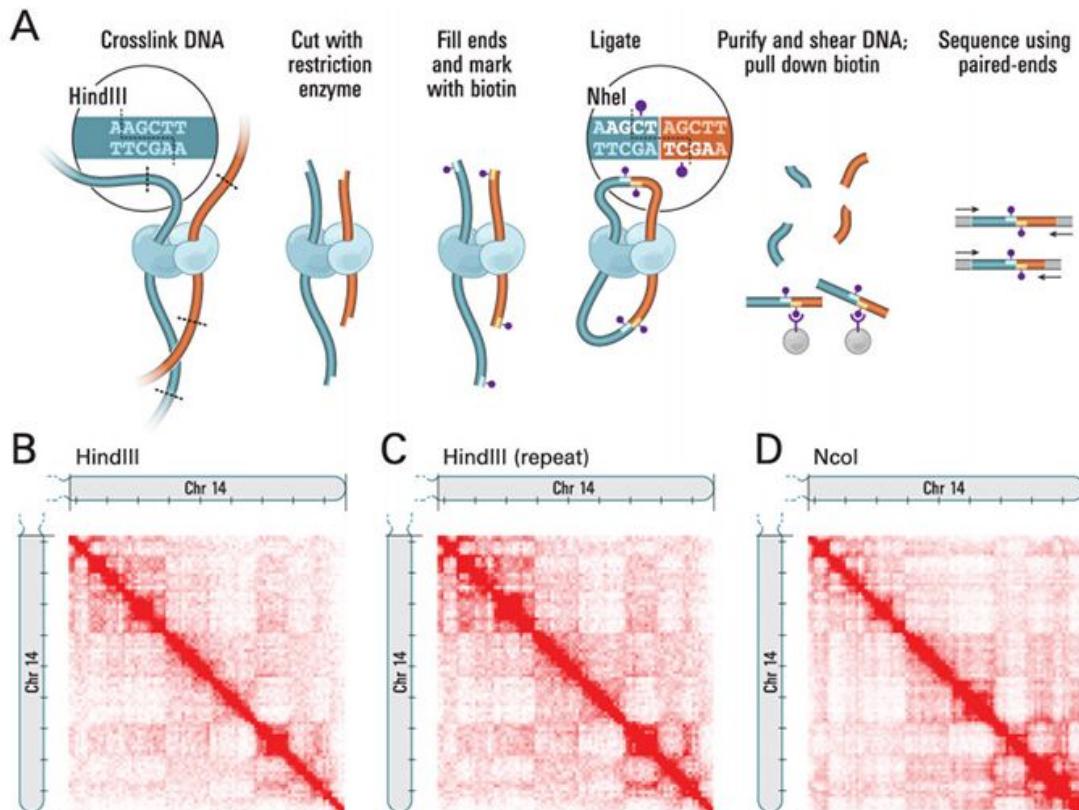
**Optical mapping** : fluorescent marking of restriction sites of very long DNA molecules (up to Mb) to extract signature used to bridge contigs having these signatures.

**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.

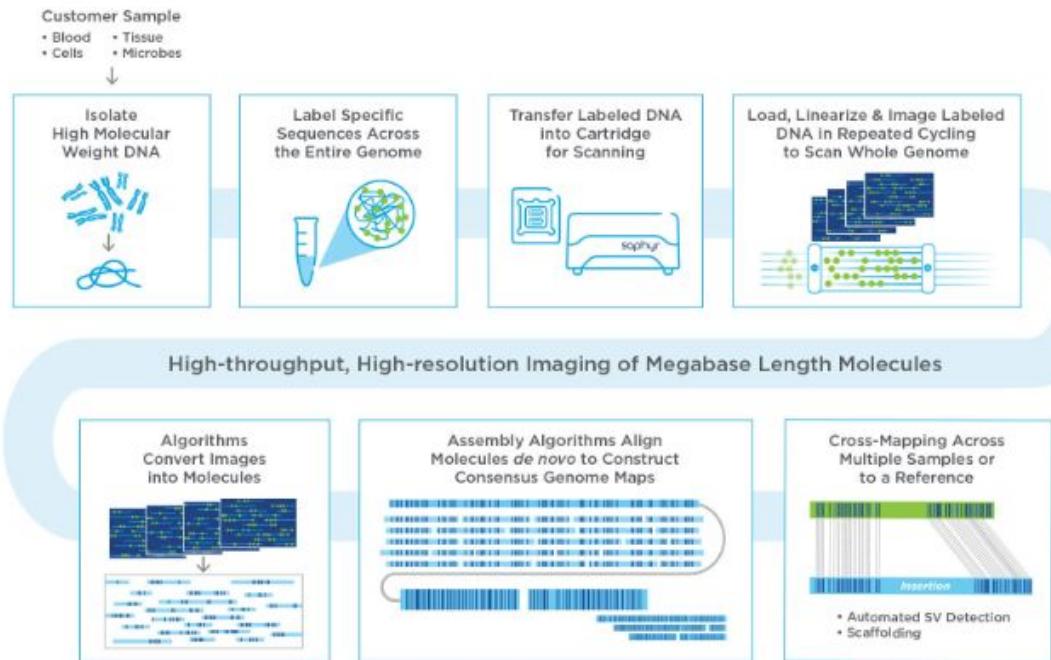
## From contigs to chromosomes: Hi-C



# From contigs to chromosomes: Optical Map

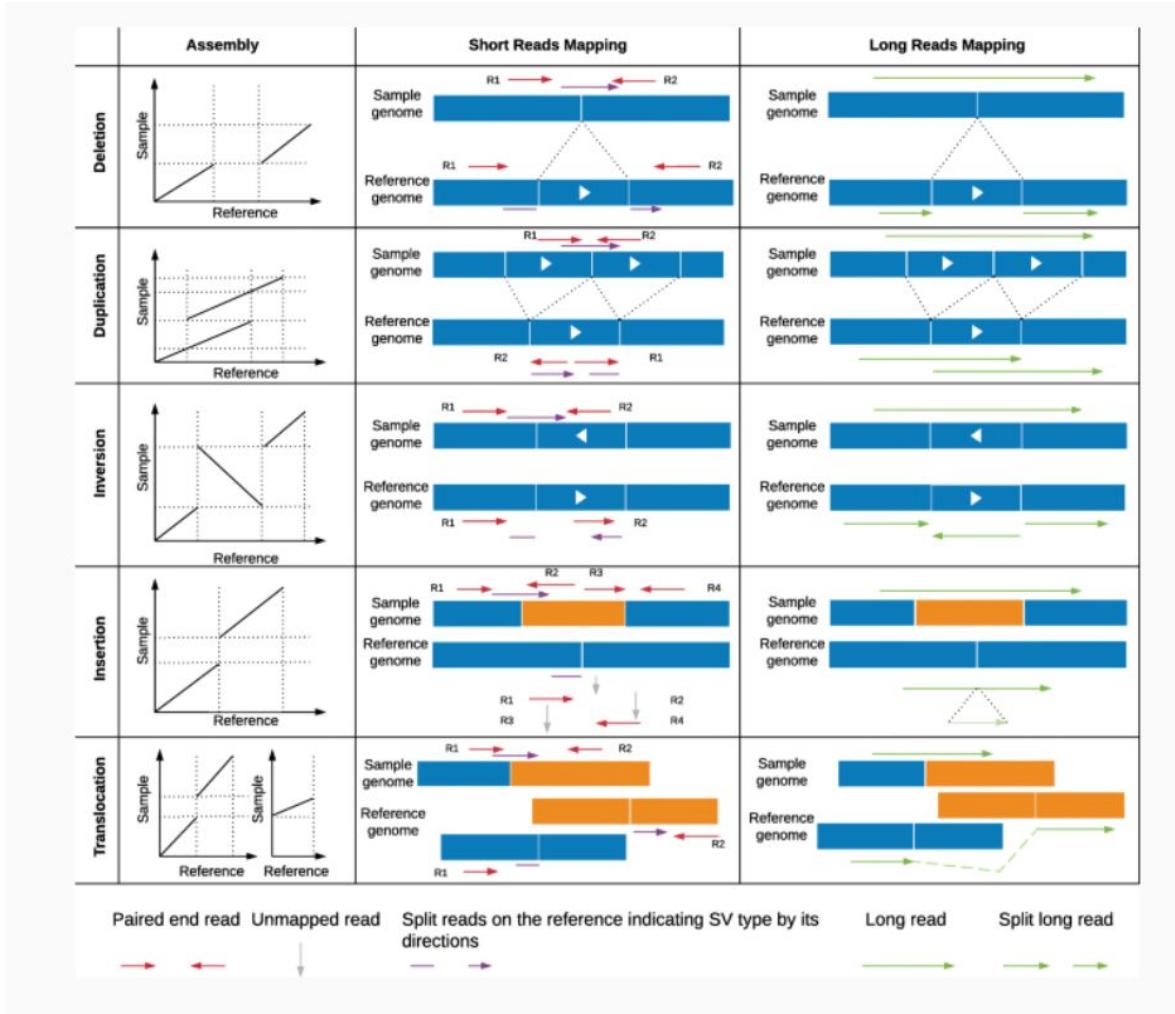


From BioNano website

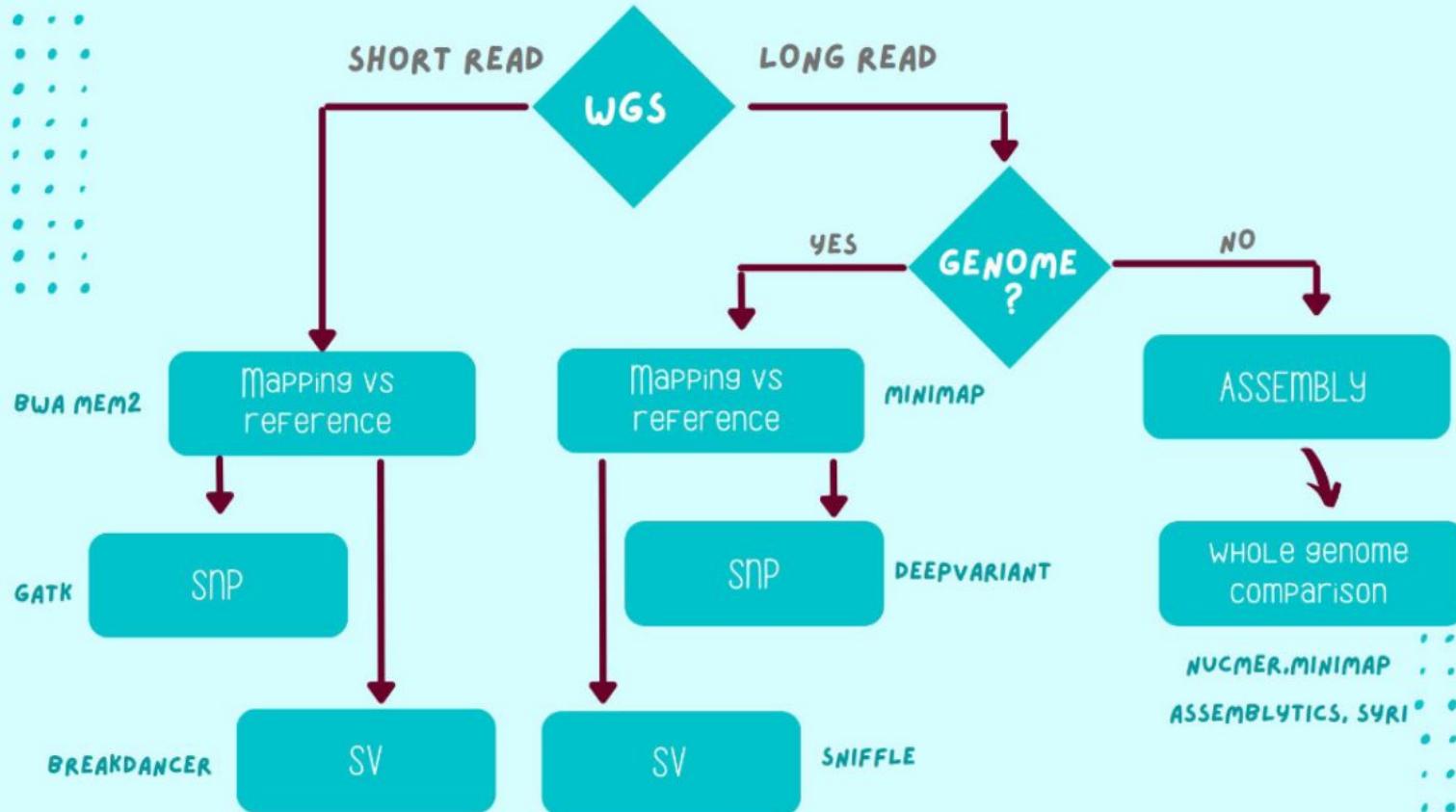


# *Chapter 4*

# *Structural Variants*



# SV DETECTION



# TP4. SV detection

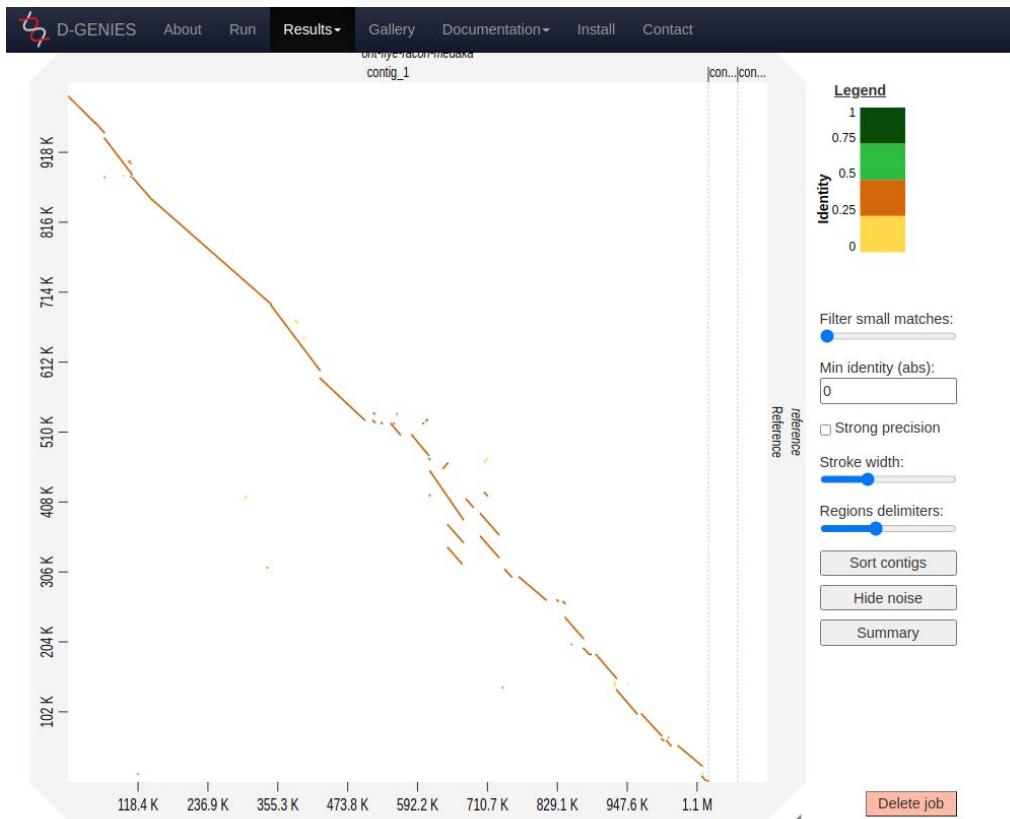
- TP4

[https://github.com/SouthGreenPlatform/training\\_ONT\\_teaching/  
blob/ONT\\_LA/4.light\\_variants\\_detection.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/ONT_LA/4.light_variants_detection.ipynb)

# *Chapter 5*

# *Genome comparison*

# Comparison with another 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/>
- Other: *Gepard*

CANU

FLYE

MINIASM

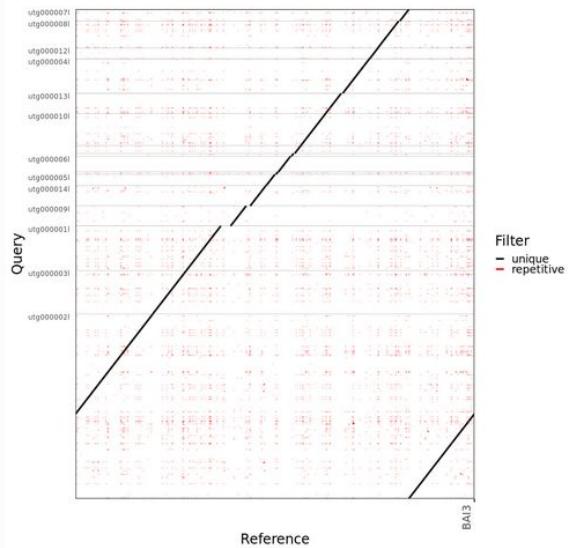
RAVEN

SMARTDENOVO

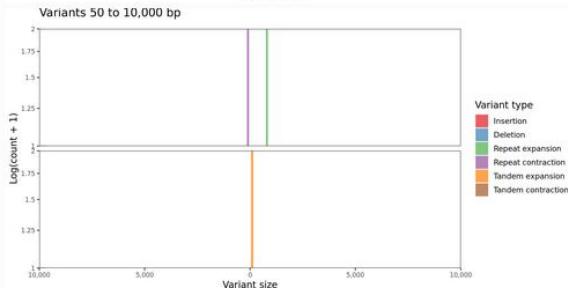
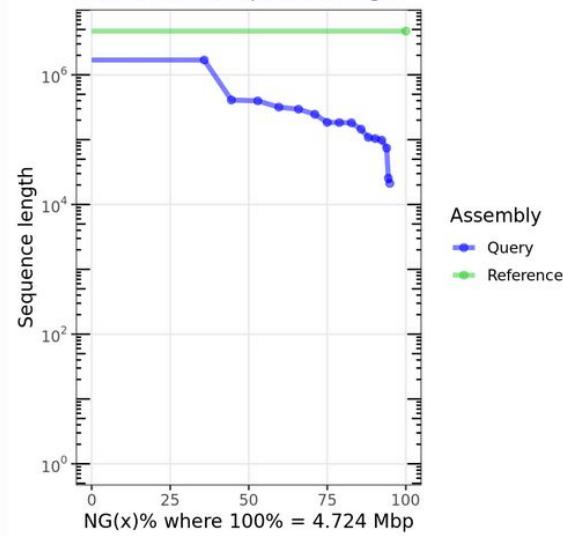
SHASTA

## STEP\_CORRECTION\_NANOPOLISH\_STARTFIXED

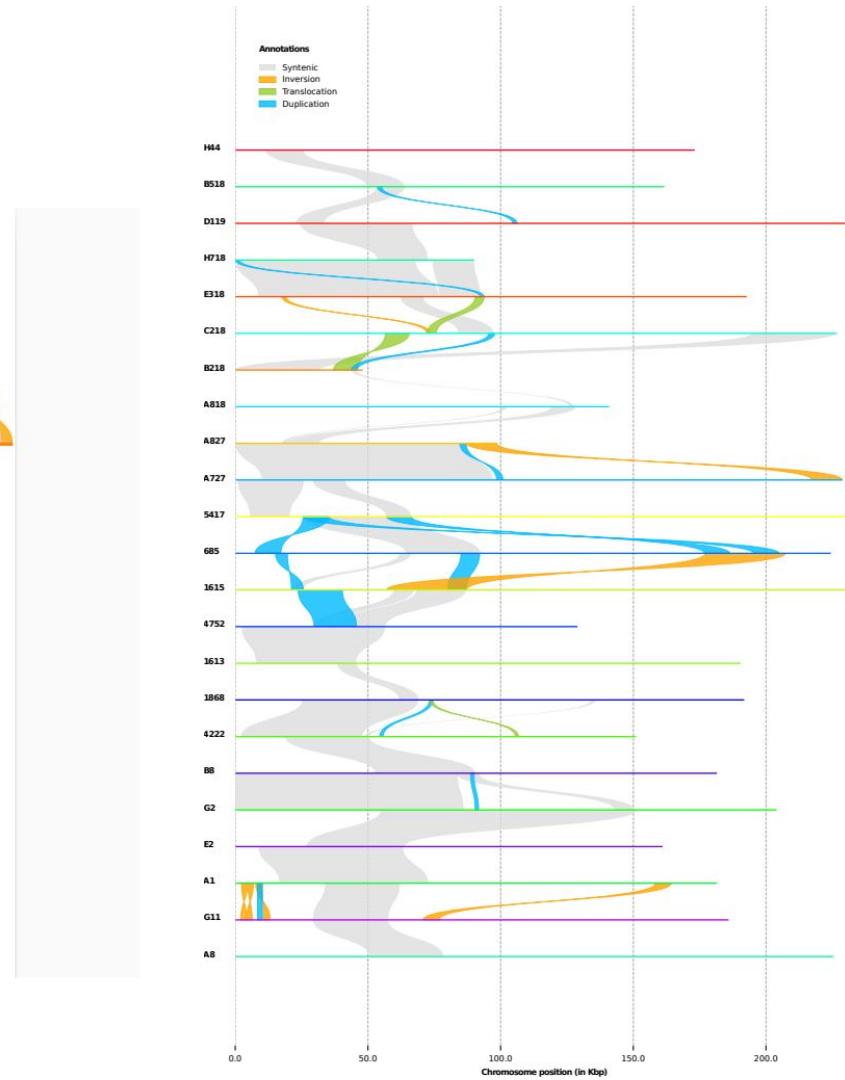
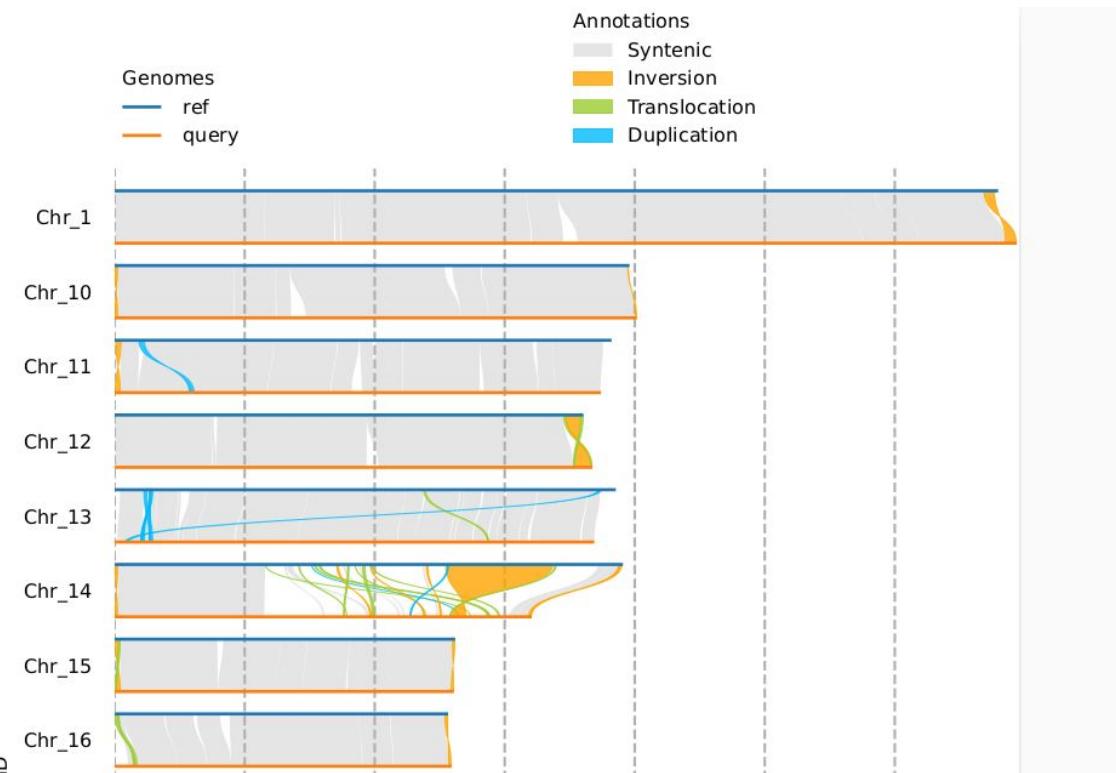
Dot plot of Assemblytics filtered alignments



Cumulative sequence length



# SIRY view of algae genomes



# TP5. Genome comparison

- TP5

[https://github.com/SouthGreenPlatform/training\\_ONT\\_teaching/  
blob/ONT\\_LA/5.light\\_variants\\_detection\\_genome.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/ONT_LA/5.light_variants_detection_genome.ipynb)

# Conclusions

- Be sure of what you sequence
- 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 now~~ polishing is **not** mandatory.
- There are still genomes very difficult to assemble

	Conceptors	Trainers		
		2021	2022	2023
	<b>Julie ORJUELA</b>			
	<b>François SABOT</b>			



# Thanks for your attention!



Pedagogic material used for these teaching is available under the Creative Common Licence CC-BY-NC-SA 4.0 International - No Commercial Use - Same sharing conditions:

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