

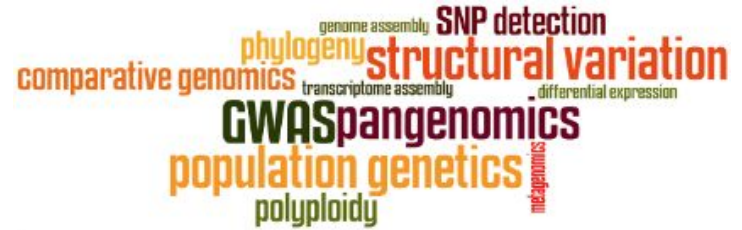
Initiation à l'analyse de données Oxford Nanopore/Assemblage



bioinformatics platform dedicated to the genetics and genomics of tropical and Mediterranean plants and their pathogens



Mutualisation



Cacao

Banana

Coffee

Rice

Palm

Cassava

Pseudocercospora

Magnaporthe

South Green

bioinformatics platform



4 institutes



3 research units



25+



Storage and computing
resources

Tools

Trainings



400+

Collaborative development of tools

Genomics

Pangenomic

Gene families

Comparative

Phylogeny

Assemblies

Annotation

Data mining

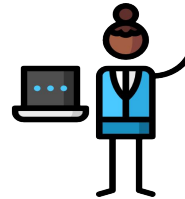
Diversity

genotype manipulation

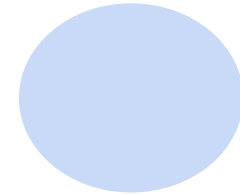
exploration

mosaic manipulation

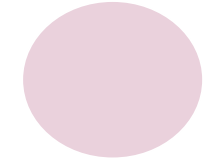
Metagenomic



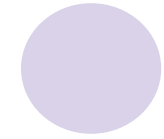
+20
tools



web applications (16)



visualisation (8)






workflows(5)



packages (4)



<https://github.com/SouthGreenPlatform/>

	Conception	Formation		
	ONT	ONT		
		2021	2022	2023
 Julie ORJUELA				
 Aurore COMTE				
 François SABOT				
Louis Denuu				



Bioinformatics resources

On va travailler sous Linux !

- 2 façons d'utiliser linux :

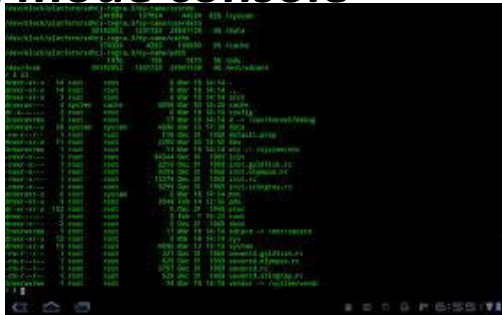
en *mode graphique*



En mode terminal

- 2 façons d'utiliser linux :

en *mode console*



```
Linux 3.10.0-112.el7.x86_64 #1 SMP Tue Aug 12 2014 22:03:12 PDT
root@localhost:~# cat /etc/passwd
root:x:0:0:root:/root:/bin/bash
bin:x:1:1:bin:/bin:/usr/sbin/nologin
daemon:x:2:2:daemon:/sbin:/usr/sbin/nologin
adm:x:3:4:adm:/var/adm:/usr/sbin/nologin
lp:x:7:7:lp:/var/spool/lpd:/usr/sbin/nologin
postfix:x:8:8:postfix:/var/spool/postfix:/usr/sbin/nologin
xfs:x:13:13:xfs:/etc:/usr/sbin/nologin
nfsnobody:x:65534:65534:nfsnobody:/var/lib/nfs:/usr/sbin/nologin
root@localhost:~#
```



avant tout !

Bases de Linux

https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/main/slides/GuideDeSurvieLinux-french2022.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 use through a simple web browser

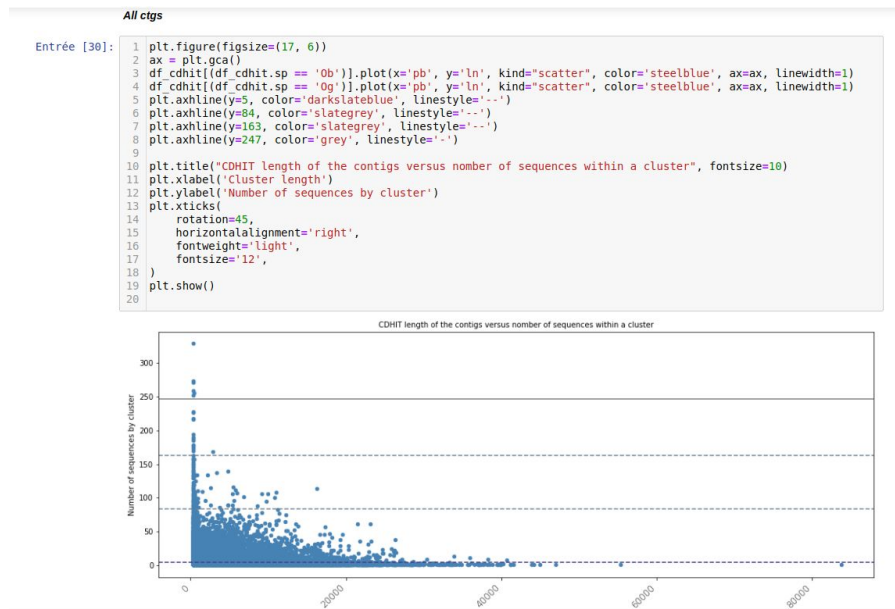
ex : Bash (Linux), Python, Java, R, Julia,
Matlab, Octave, Scheme, Processing,
Scala



Why use jupyter book ?

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

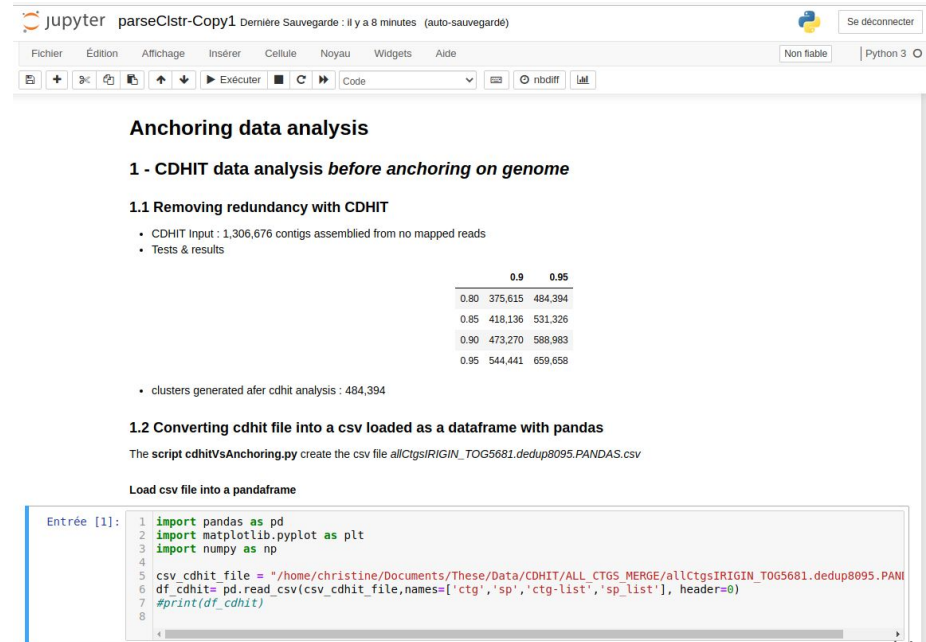
- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook



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 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	375,615	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` create the csv file `allCtgsIRIGIN_TOG5681.dedup8095.PANDAS.csv`

Load csv file into a dataframe

```
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_CTGS_MERGE/allCtgsIRIGIN_TOG5681.dedup8095.PAN
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 ?



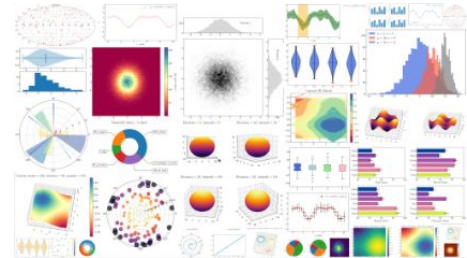
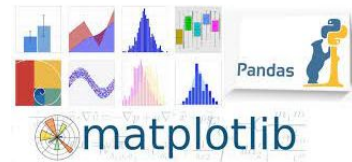
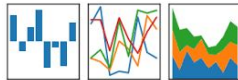
- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

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”

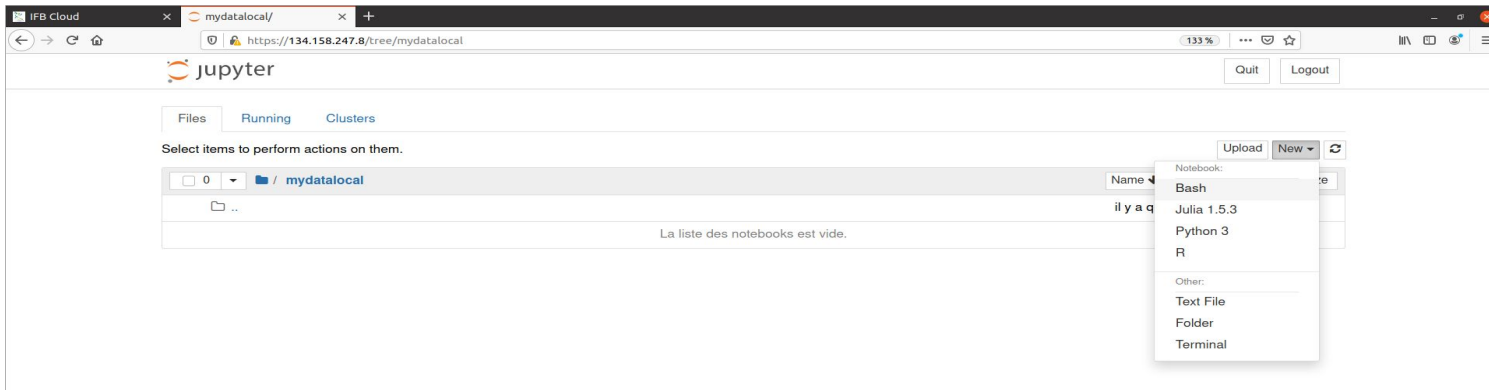


How will you use Jupyter Notebook ?

- 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

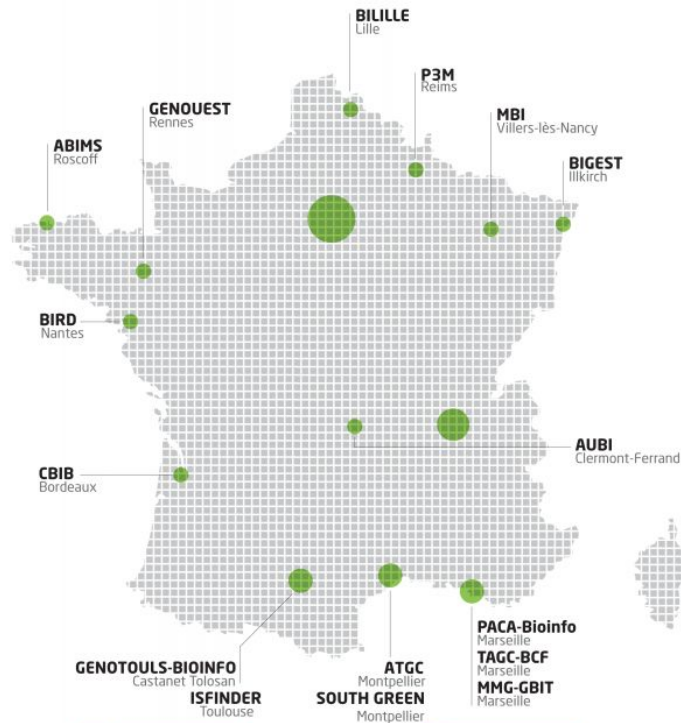


IFB ?



INSTITUT FRANÇAIS DE BIOINFORMATIQUE

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



RÉGION PARISIENNE

EBIO
Orsay
INSTITUT CURIE
Paris
IGR
Villejuif
MICROSCOPE
Evry
MIGALE
Jouy-en-Josas

C3BI
Paris
RPBS
Paris
URGI
Versailles
ORPHANET
Paris
ICONICS
Paris
IFB CORE
Evry

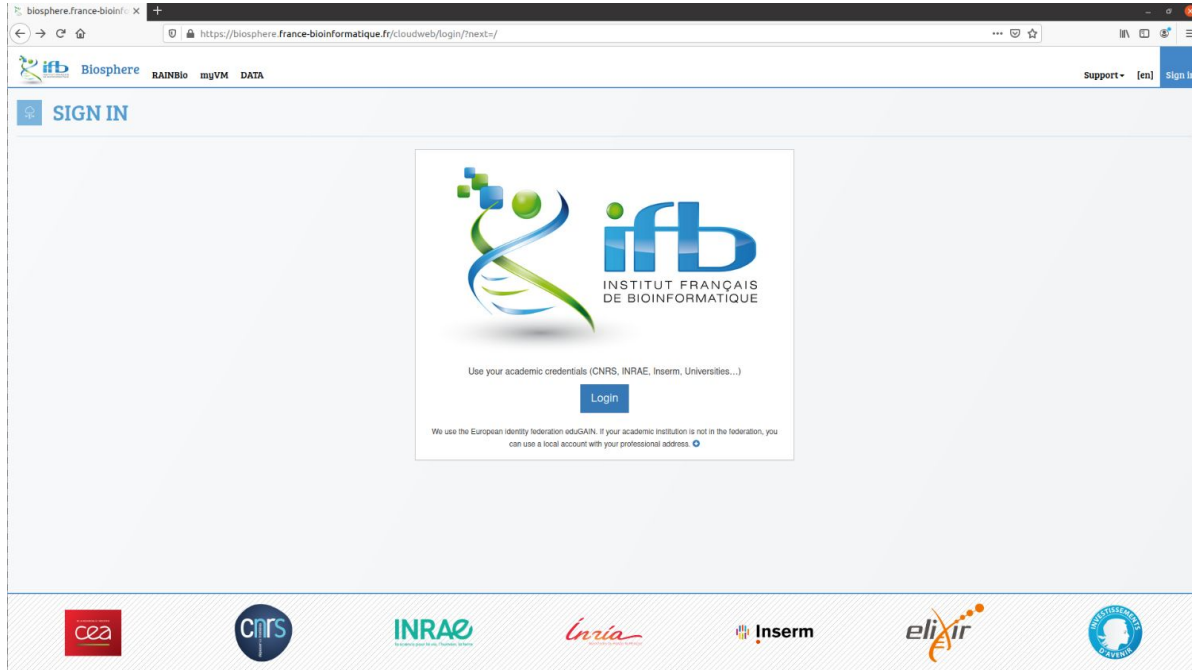
RÉGION LYONNAISE

INCA-SLC
Lyon
PRABI-HCL
Lyon
PRABI-AMSB
Villeurbanne
PRABI-Lyon-Grenoble
Villeurbanne
PRABI-Lyon-Gerland
Lyon

- 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



Let's start with biosphere

- Select a specific group

The screenshot shows the IFB Biosphere dashboard. At the top, there are navigation tabs for RAINBio, myVM, and DATA. The main header includes the IFB Biosphere logo and the text 'Support'. Below the header, there is a 'CLOUD' section with a 'Déploiements' tab. A table is visible with columns for 'ID', 'Nom', 'Début', 'Groupes', and 'Spécification'. A red arrow points to the 'Groupes' menu item in the user profile dropdown menu. The dropdown menu also includes options for 'Langues', 'Paramètres', 'Quota', and 'Se déconnecter'.

IFB Biosphere RAINBio myVM DATA Support

Connecté.e en tant que francois.sabot@ird.fr

Langues
[en] English
[fr] Français

Paramètres
Groupes
Quota

Se déconnecter

Cloud

Déploiements


Arrêter les déploiements

Appiances et déploiements favoris Déploiements récemment terminés Quota

ID	Broker	Nom	Der. dém.	Paramétrage
----	--------	-----	-----------	-------------

Let's start with biosphere































- Ask for joining *M2UMASM* (Master2 bioinfo UM Assemblage)

 **IFB Biosphere** RAINBio myVM DATA

LISTE DES GROUPES


Mes groupes Rejoindre un groupe Créer un groupe

Afficher 10 éléments Recherche

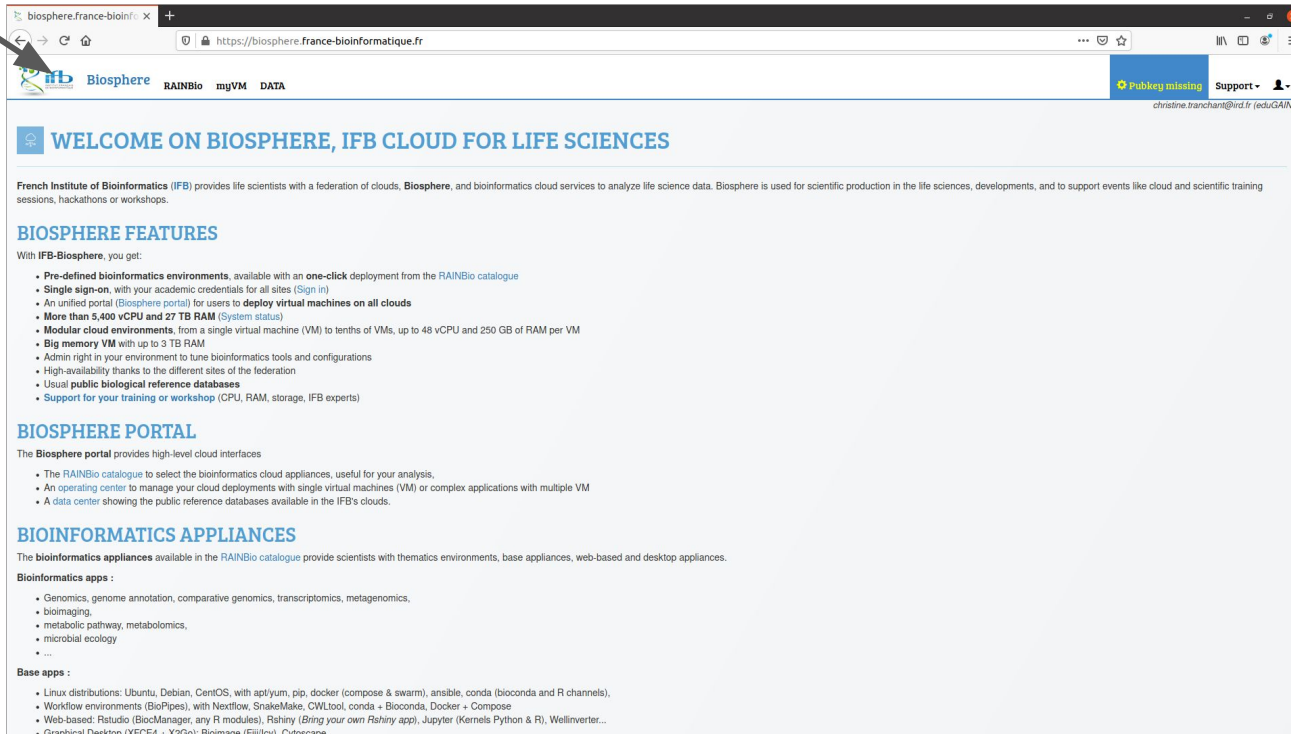
	Nom	Site web	Actions
	2AD (Acquisition et Analyse de Données pour l'Histoire Naturelle), UAR2700		
	AGAP (Amélioration génétique et adaptation des plantes méditerranéennes et tropicales), 1098		
	AgroImpact (UR1158 AgroImpact Agroressources et Impacts environnementaux), UR1158		
	AMAP (botAnique et Modélisation de l'Architecture des Plantes et des végétations), UMR5120		
	ANF_Metabodiv#2 (Exploration de la Diversité Taxonomique des Ecosystèmes par Metabarcoding)		
	ANSES - Sophia-Antipolis (ANSES - Laboratoire de Sophia-Antipolis), 13001202400043		
	ANSES AVB (ANSES Antibiorésistance et Virulence Bactériennes), -		
	ANSES Fougères (ANSES Laboratoire de Fougères), -		
	ANSES Ploufragan (ANSES Ploufragan-Plouzané), 13001202400118		
	ANSES-MYCO (ANSES - Mycoplasmoses animales), na		

Affichage de l'élément 1 à 10 sur 178 éléments

Précédent 1 2 3 4 5 ... 18 Suivant



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 header includes the IFB Biosphere logo and navigation links for RAINBio, myVM, and DATA. A user profile for christine.tranchant@rd.fr (edu.GAIN) is visible in the top right corner.

WELCOME ON BIOSPHERE, IFB CLOUD FOR LIFE SCIENCES

French Institute of Bioinformatics (IFB) provides life scientists with a federation of clouds, **Biosphere**, and bioinformatics cloud services to analyze life science data. Biosphere is used for scientific production in the life sciences, developments, and to support events like cloud and scientific training sessions, hackathons or workshops.

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,
- Bioimaging
- 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 (BioC/Manager, any R module), Rshiny (*Bring your own Rshiny app*), Jupyter (Kernels Python & R), Wellinverter...
- Graphical Desktop: XCFE4_X2Go: Bioimane_Ellilcov_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 éditables** | Ajouter ... | ⚙️

Appliance	Tools	Topics
AnalysesSV	bcftools, BEDTools, BWA, Jupyter, Matplotlib, pandas, SAMtools	DNA polymorphism, Genetic variation, Genotyping experiment, GWAS study
CoursAnalysesNanoporeSG	bandage, Jupyter	Data architecture, analysis and design, Mathematics, Statistics and probability
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

CoursAnalysesNanoporeSG

↓ Exporter en md

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

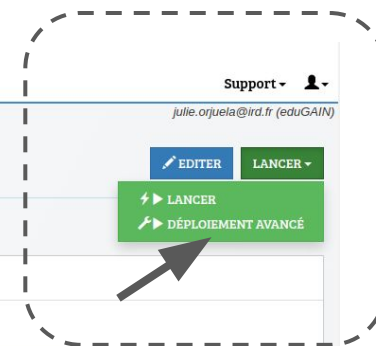
Outils

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



Let's run your vm through the cloud

CoursAnalysesNanoporeSG

Exporter en md

Description

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

Domaines associés

Computational biology Sec

Configurer le déploiement d'une appliance

Déploiement de l'appliance "virus_ONT"

Name	Julie_ONT
Groupe à utiliser	virus_ont (Initiation à l'analyse du tagénome viraux) 828.01
Cloud	ifb-core-cloudbis
Gabarit d'image cloud	<ul style="list-style-type: none">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.m4.4xlarge (16 vCPU, 64Go GB RAM, 400Go GB local disk)ifb.x1e.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.x1e.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.x1e.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.x1e.16xlarge (BigMem) (62 vCPU, 1.5To GB RAM, 1.5To GB local disk)ifb.x1e.32xlarge (BigMem) (124 vCPU, 2.9To GB RAM, 2.9To GB local disk)

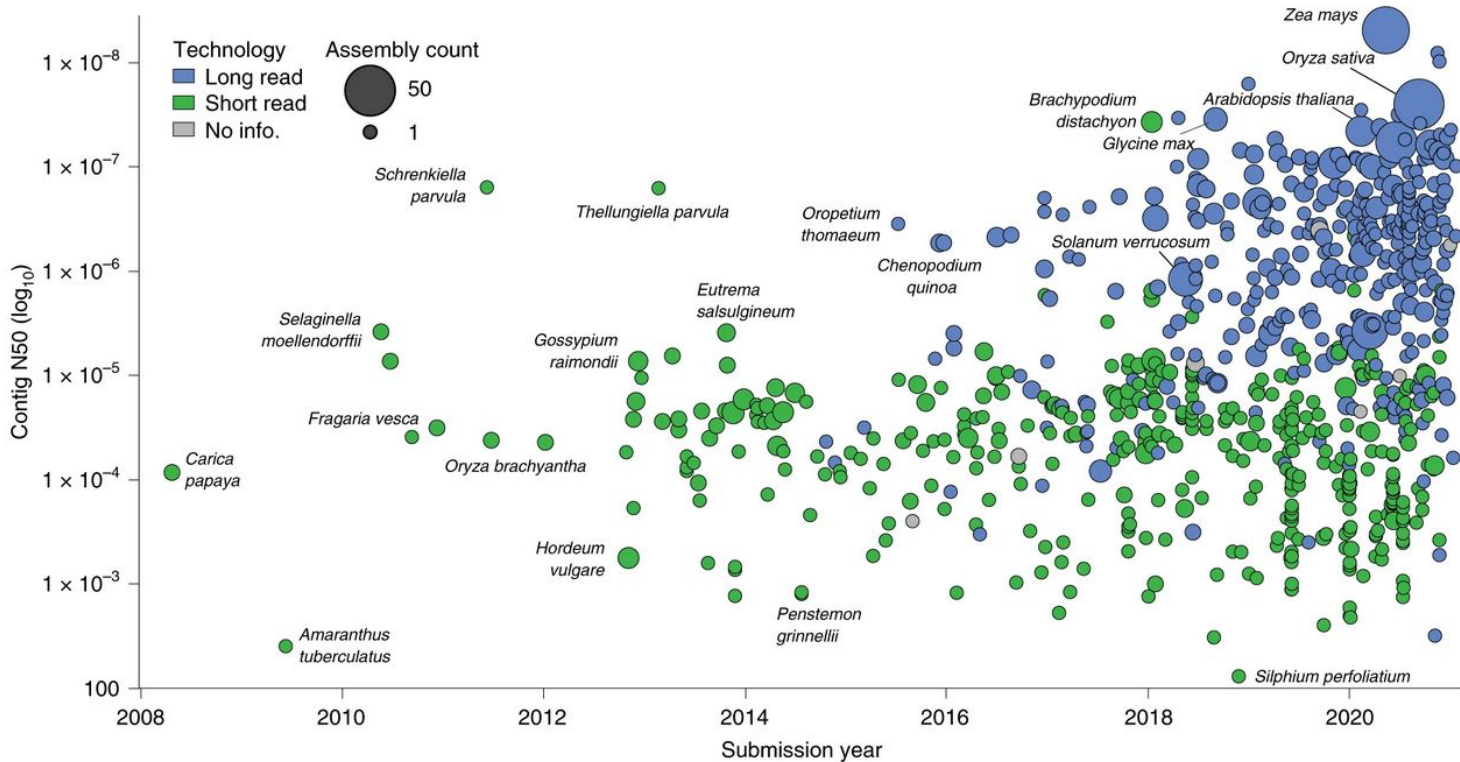
Annuler

Quelle gabarit d'image doit être utilisé sur ce cloud ?

EDITER LANCER

LANCER DÉPLOIEMENT AVANCÉ

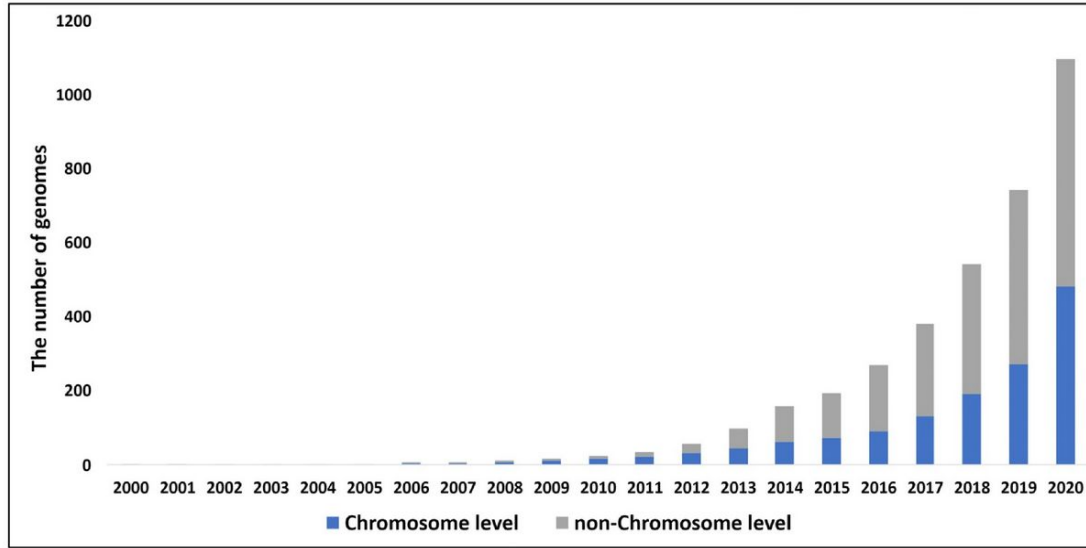
Let's start!



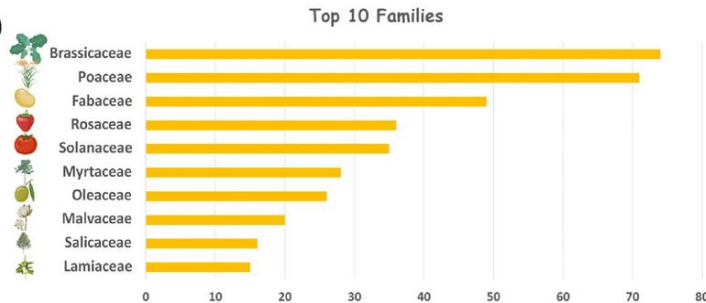
Assembly contiguity by submission date for 798 land plant species with publicly available genome assemblies. Points are coloured by the type of sequencing technology used and scaled by the number of assemblies available for that species. There is an improvement in contiguity associated with the advent 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

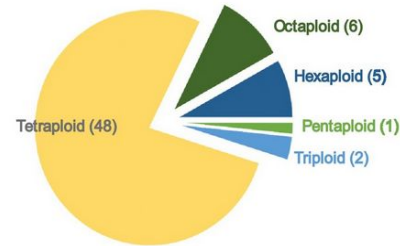
(A)



(B)










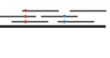




(C)



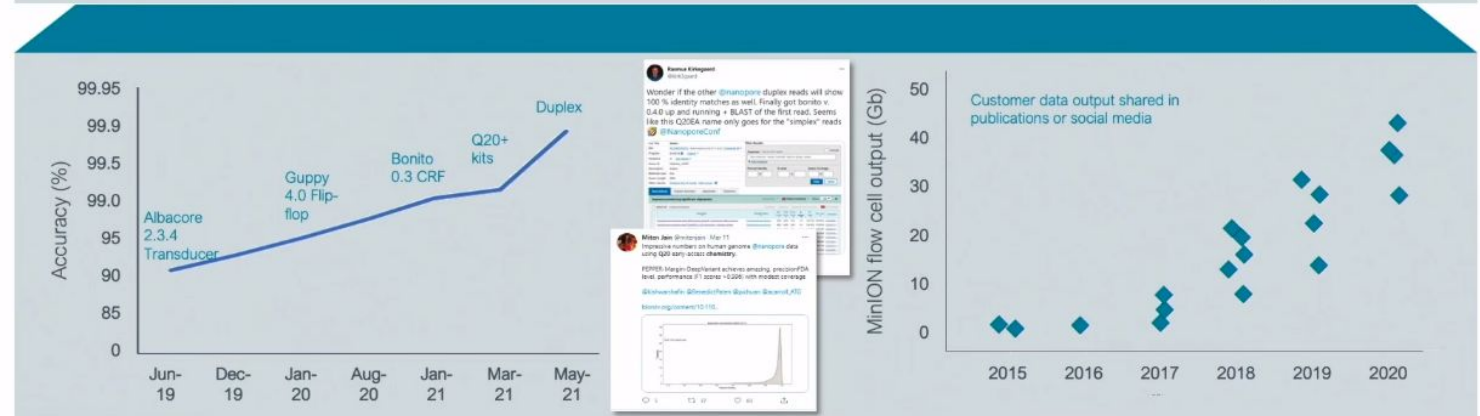
Sun et al, 2022

(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



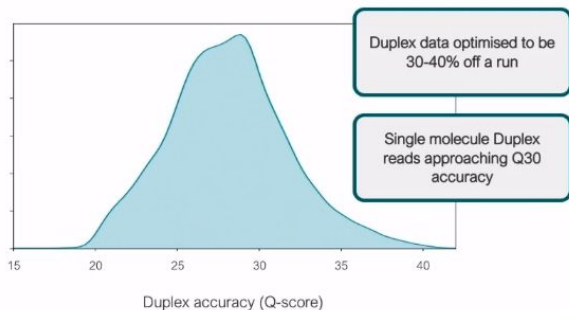
Last upgrades !

Nanopore accuracy

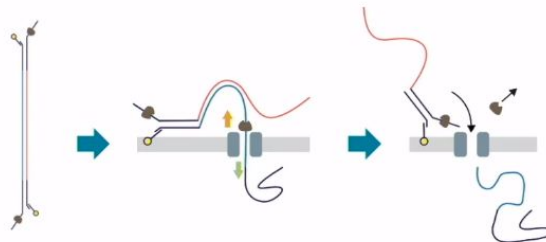
When we last spoke...

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/3CEIj19

Raw read modal 99.3%, >Q20

Raw read accuracy (%)

8:30 AM · Sep 23, 2021 · HubSpot

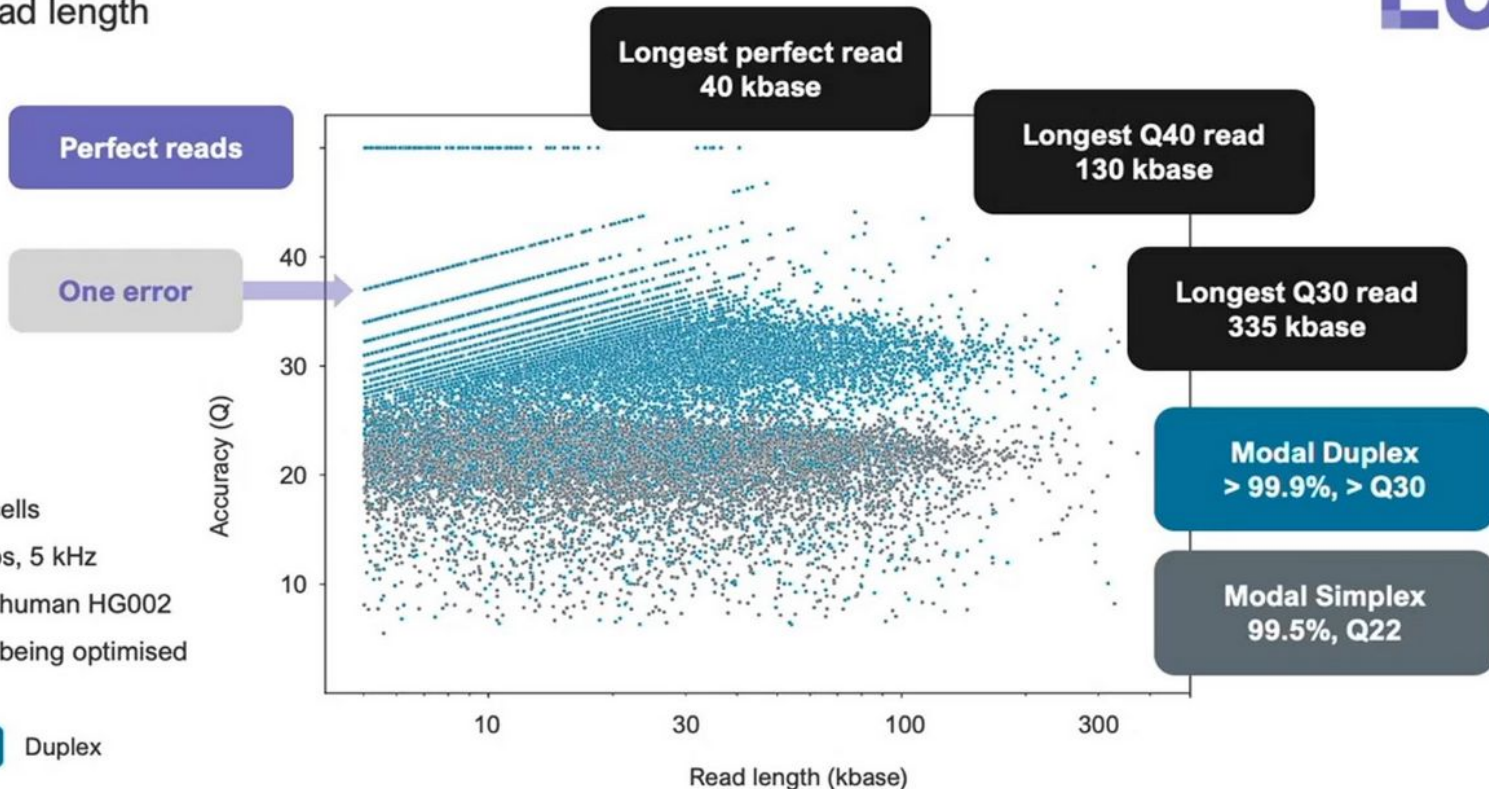
33 Retweets 1 Quote Tweet 62 Likes

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

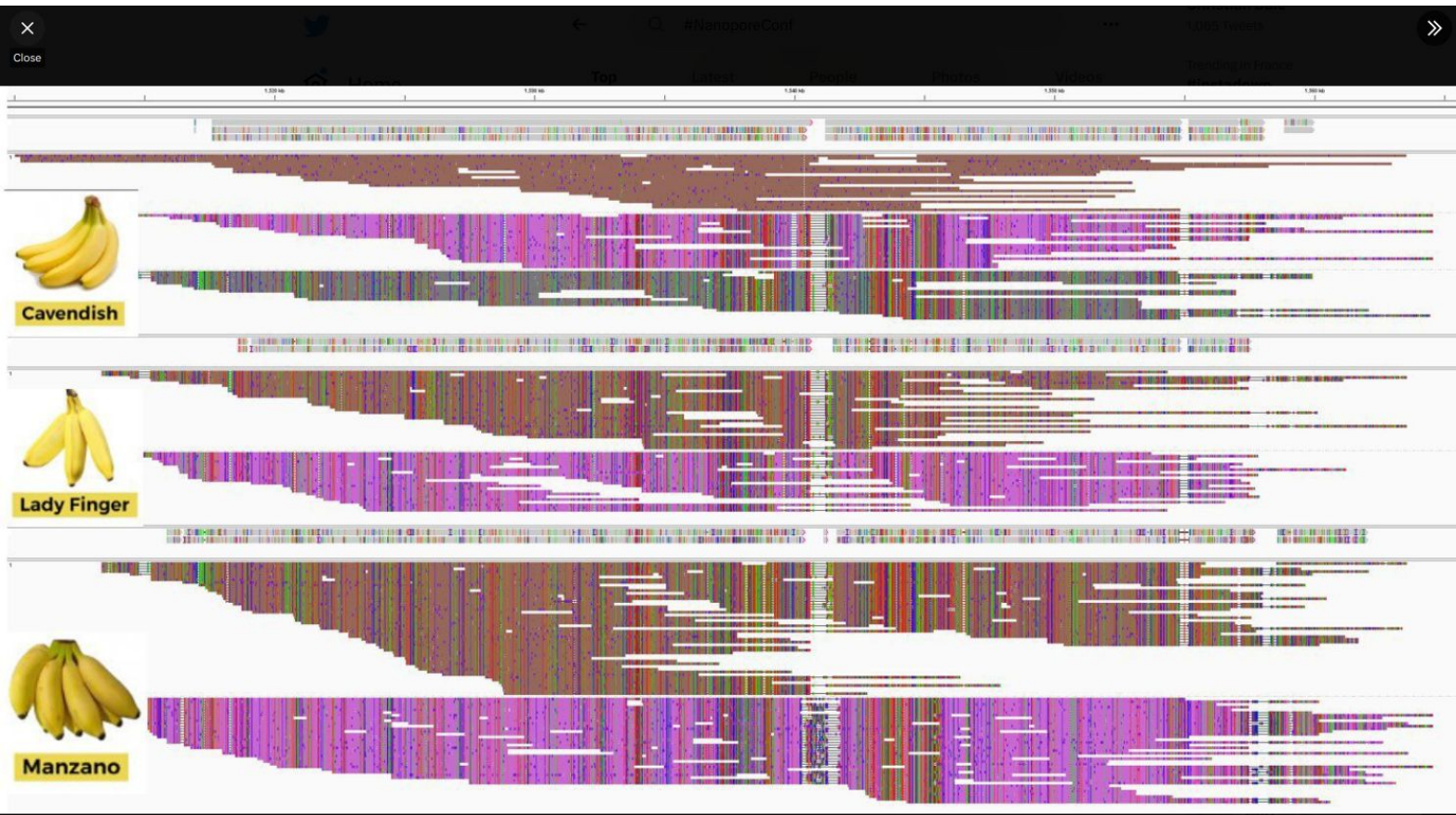
Duplex

Accuracy and read length



Ultra-long Duplex:

- PromethION flowcells
- ULK114 @ 400 bps, 5 kHz
- Unsheared native human HG002
- ULK Duplex rates being optimised



Alexander Wittenberg
@AW_NGS

[@keygene](#) Unlocking the Power of Haplotype Based Molecular Breeding using Long Read [@nanopore](#) adaptive sequencing screening resistance gene regions in banana 🍌 [#nanoporeconf](#) Poster#53 and online #114

10:17 AM · May 18, 2023 · 271 Views

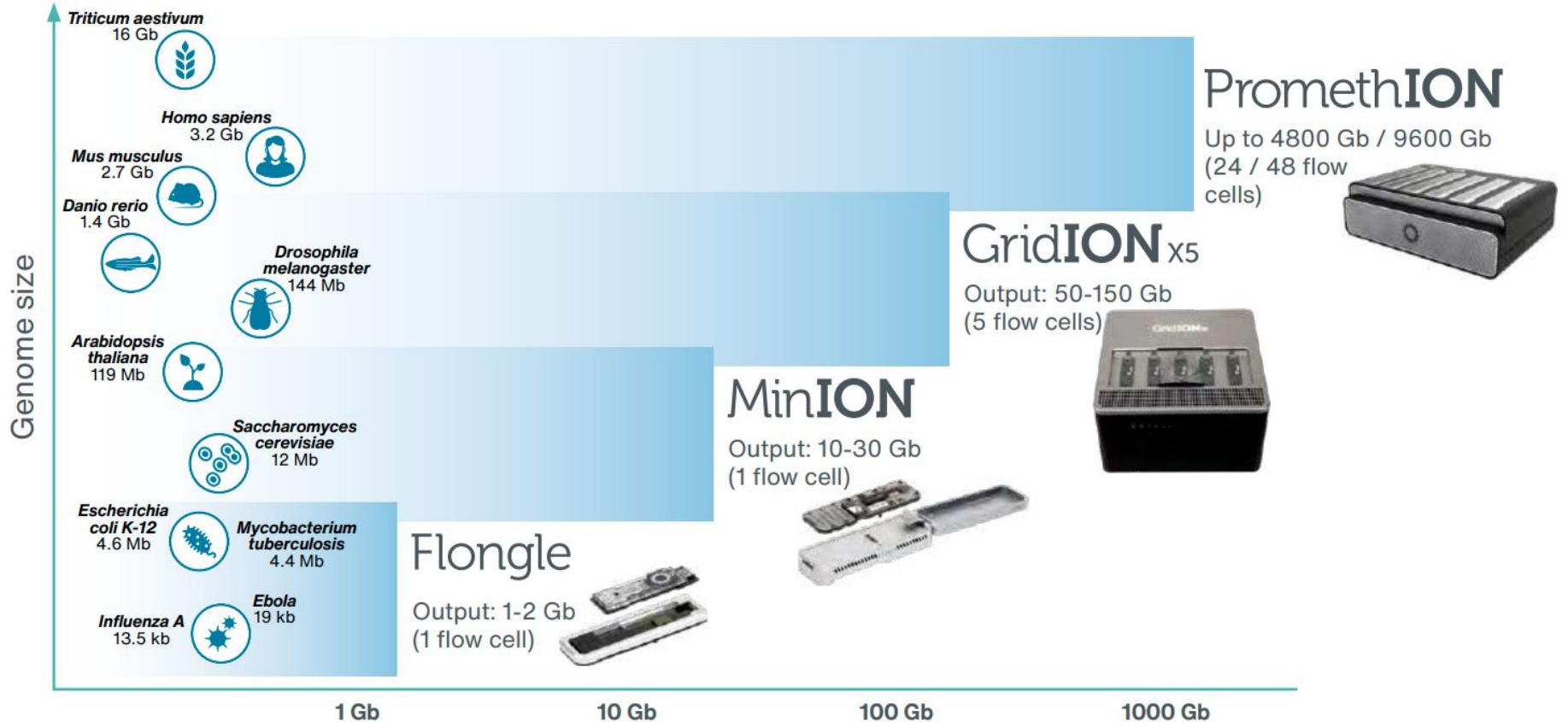
2 Retweets 6 Likes



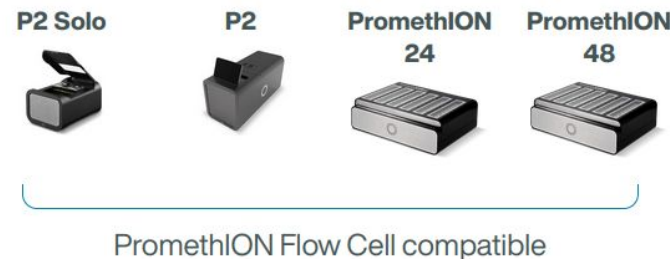
Tweet your reply!

Reply

A lot of data !



A lot of data !

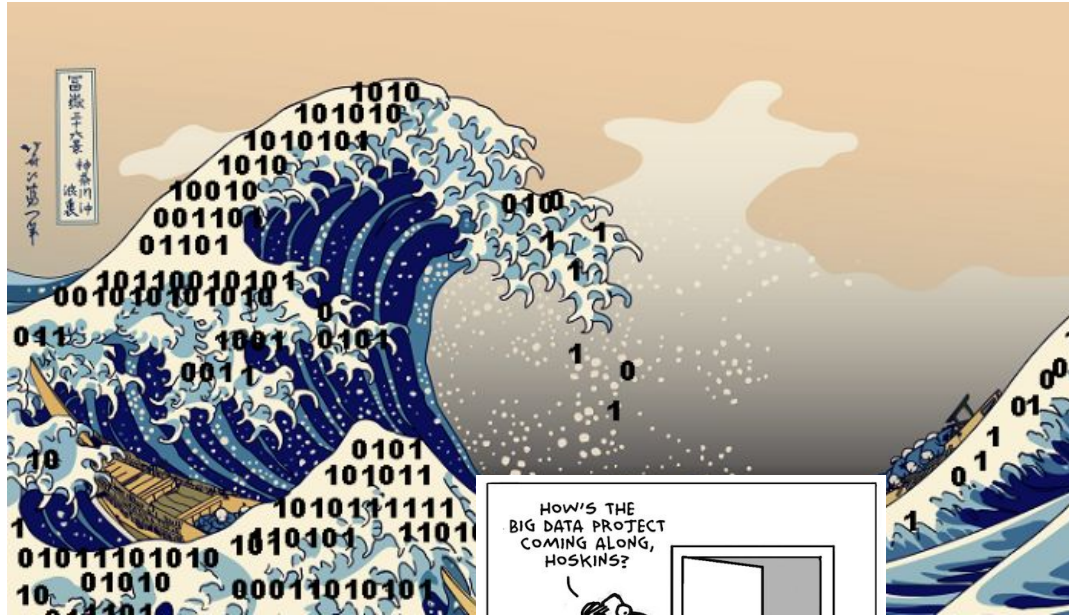


Configuration	Platform			Techniques		Tech specifications	
Number of flow cells per device	1	1	5	2	2	24	48
Maximum number of channels per flow cell	512	512	512	2,675	2,675	2,675	2,675
Run time	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours
Device TMO [†]	50 Gb	50 Gb	250 Gb	580 Gb	580 Gb	~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

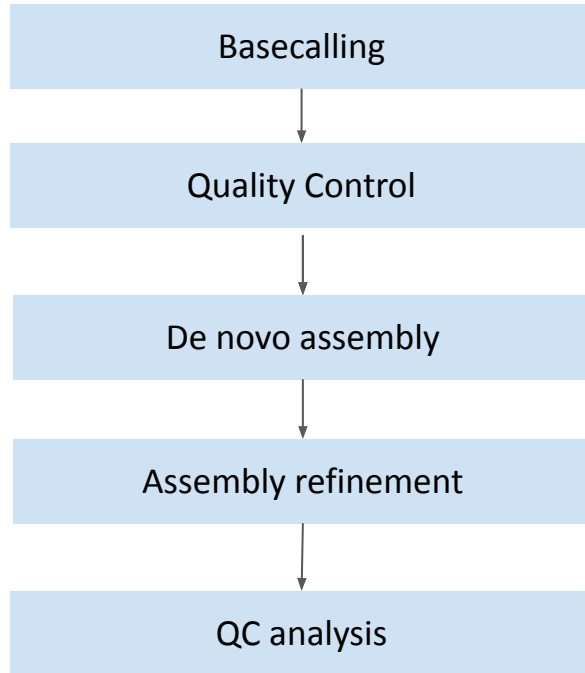
The data that these platforms produce differ qualitatively from second-generation sequencing, thus necessitating tailored analysis tools



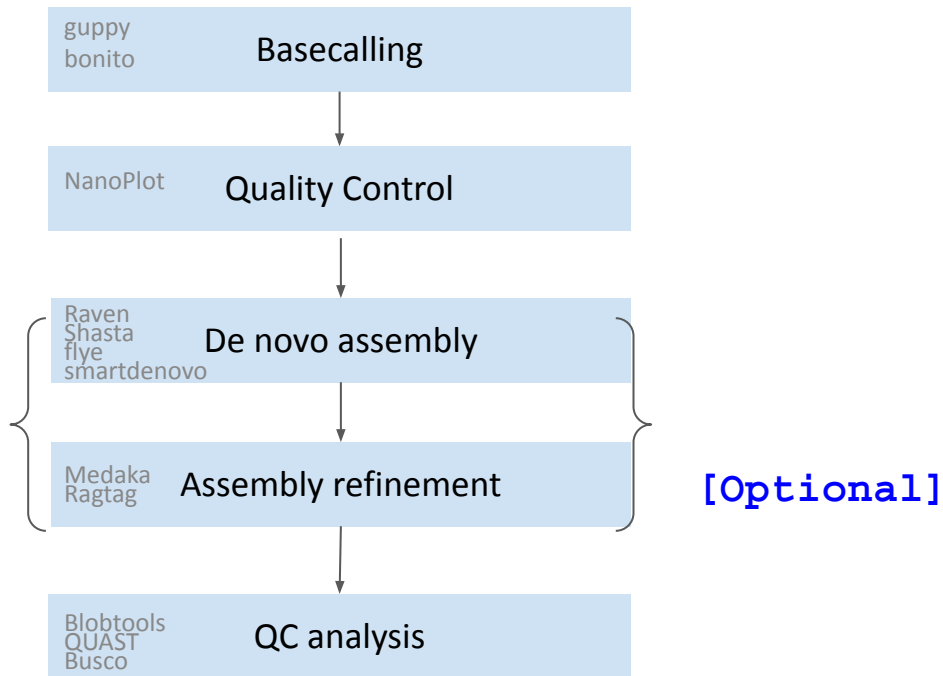
From data rarity to data deluge



Typical long-read analysis pipelines for ONT data



Typical long-read analysis pipelines for ONT data



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

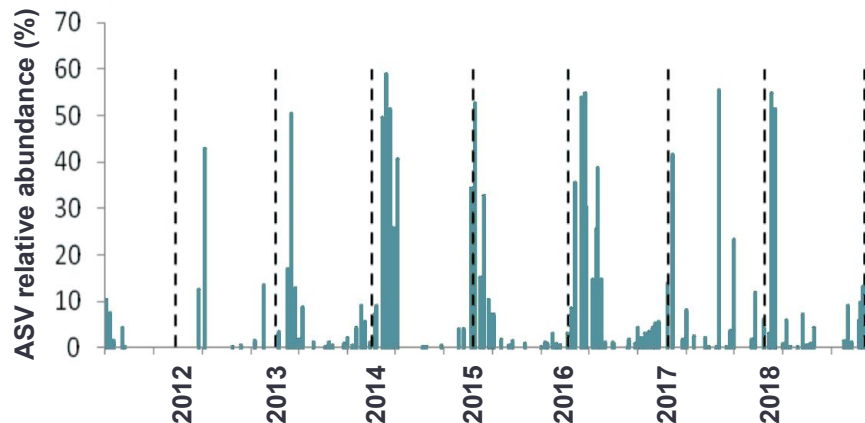
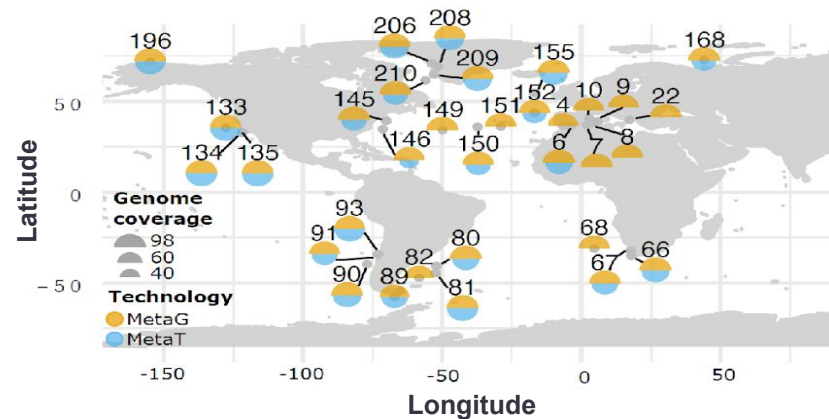
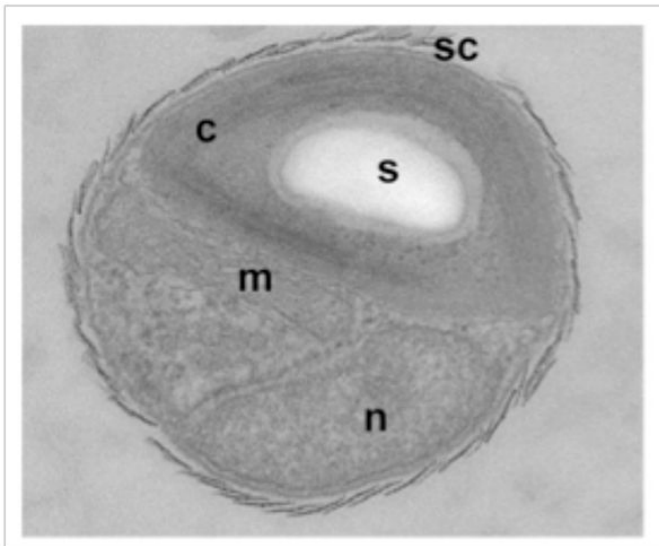
Observatoire Océanologique de Banyuls-sur-mer
Laboratoire d'Océanographie Microbienne



Bathycoccus prasinus

A cosmopolitan model to study adaptation to latitude and seasons

Bathycoccus prasinus

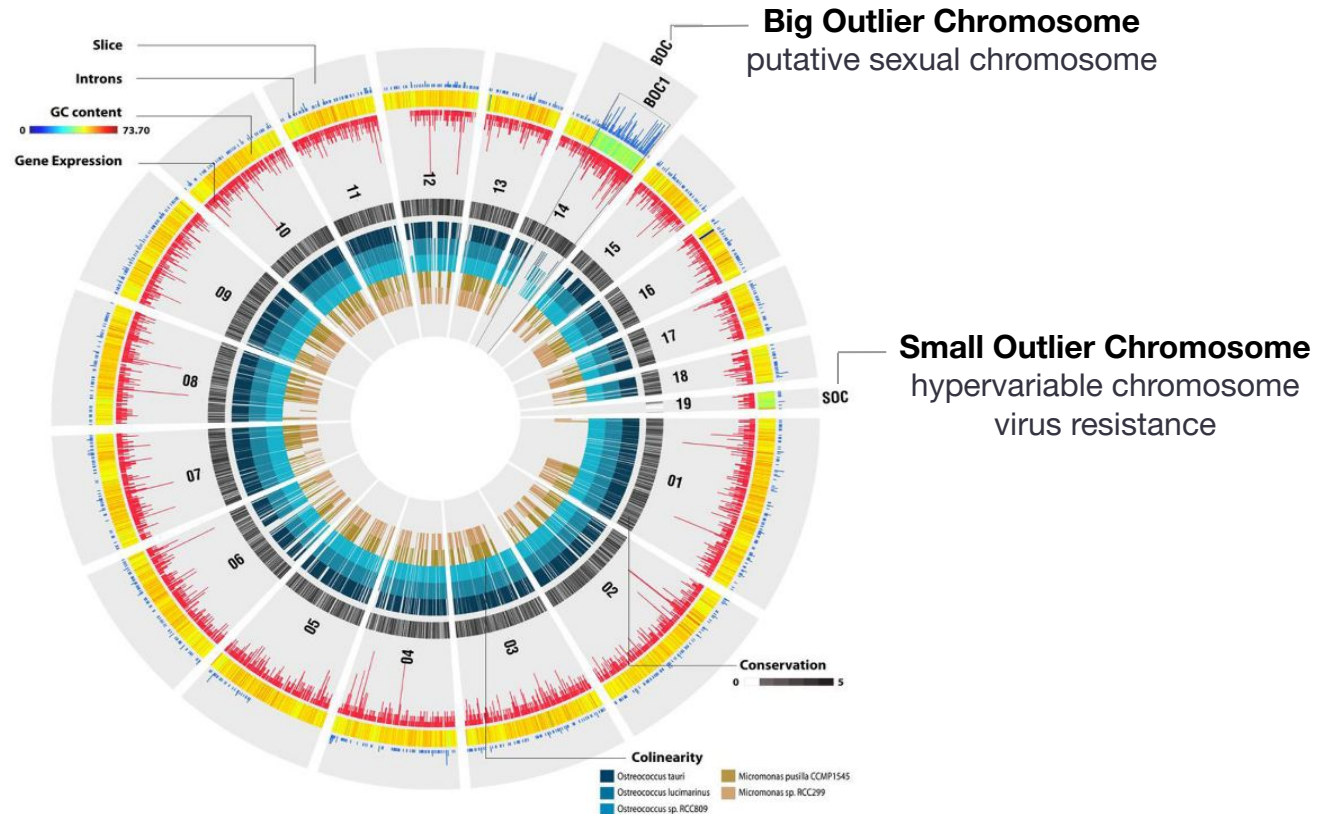


Bathycoccus prasinus single reference genome

RCC1105
Banyuls-sur-Mer, France

Size
15 Mbp


19
Chromosomes




Back to Machines!






Let's run your vm through the cloud


Loading...

IFB Biosphere RAINBio myVM DATA Support 
julie.orjuela@ird.fr (eduGAIN)

CLOUD

Déploiements 

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

 Arrêter les déploiements

Tout voir (6)

Let's run your vm through the cloud

ready !

CLOUD

Déploiements

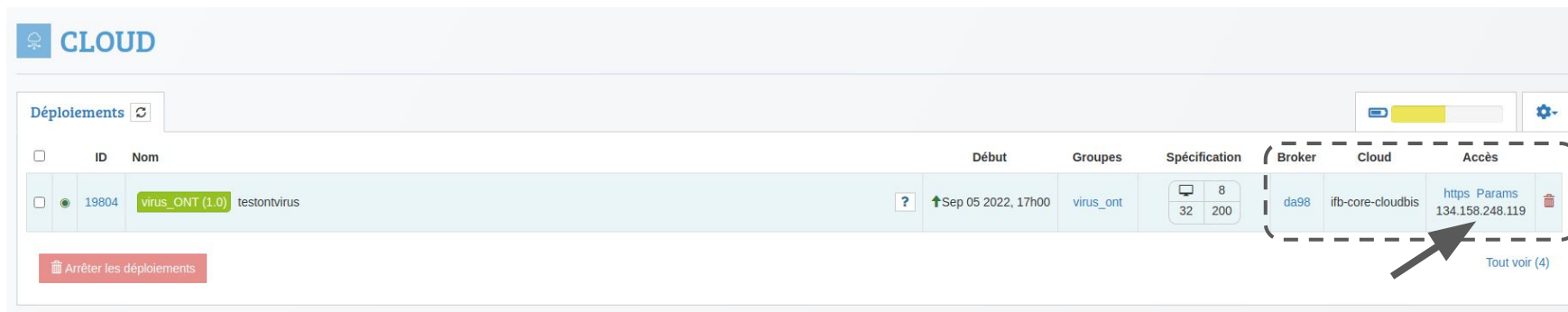
ID	Nom	Début	Groupes	Spécification	Broker	Cloud	Accès
19804	virus_ONT (1.0) testontvirus	↑ Sep 05 2022, 17h00	virus_ont	8 32 200	da98	ifb-core-cloudbis	https Params

Arrêter les déploiements

Tout voir (4)

Let's run your vm through the cloud

get the url... link "https"



The screenshot shows a cloud management interface with a table of deployments. The table has columns for ID, Nom, Début, Groupes, Spécification, Broker, Cloud, and Accès. A red dashed box highlights the 'Accès' column for the first deployment, which contains the text 'https Params 134.158.248.119'. An arrow points to the 'https' link. Below the table is a red button labeled 'Arrêter les 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 our vm through the cloud

Get the token identifiant... link "Params"

The screenshot shows a web interface with a 'Paramètres' dialog box open. The dialog box has a table with two columns: 'nom' and 'valeur'. The first row contains 'JUPYTER_TOKEN' and '28f9a32ae92eaecbfc816880489c9217e3263f9fd4614352'. Below the dialog box, there is a table with columns: 'Début', 'Groupes', 'Spécification', 'Broker', 'Cloud', and 'Accès'. The 'Accès' column contains a link 'https://134.248.119' with the text 'Params' next to it. A yellow arrow points to this link. The background interface includes a 'Support' dropdown and a user email 'julie.orjuela@ird.fr (eduG'.

nom	valeur
JUPYTER_TOKEN	28f9a32ae92eaecbfc816880489c9217e3263f9fd4614352

Début	Groupes	Spécification	Broker	Cloud	Accès
↑ Sep 05 2022, 17h00	virus_ont	8 32 200	da98	ifb-core-cloudbis	https://134.248.119 Params

Let's run our vm through the cloud

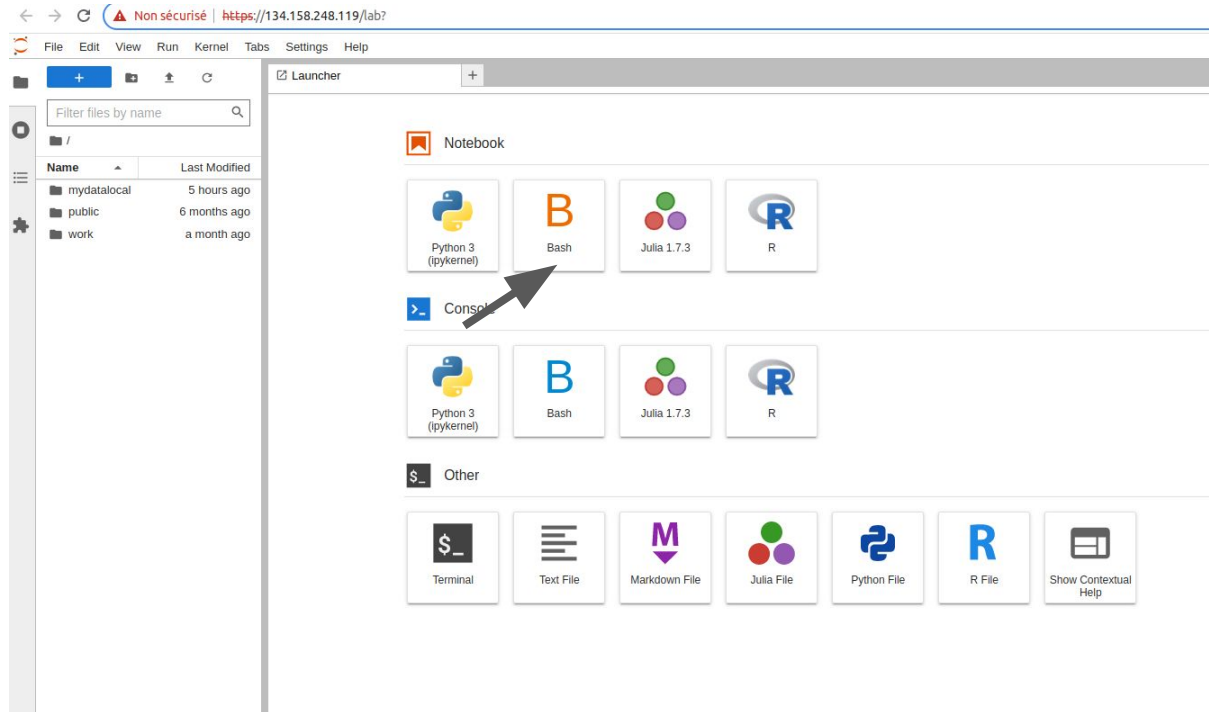
Open your vm (https link) to access to your own jupyter lab

The screenshot shows a web browser window with a single tab titled "Accueil". The address bar displays the URL "https://134.158.247.8/tree". The page content includes the Jupyter logo on the left and "Quit" and "Logout" buttons on the right. Below the navigation tabs, there is a section for file management with "Upload", "New", and refresh buttons. A table lists files and folders:

	Name	Last Modified	File size
<input type="checkbox"/>	0 /		
<input type="checkbox"/>	mydatalocal	il y a un jour	
<input type="checkbox"/>	public	il y a 3 mois	

Create your first jupyter book

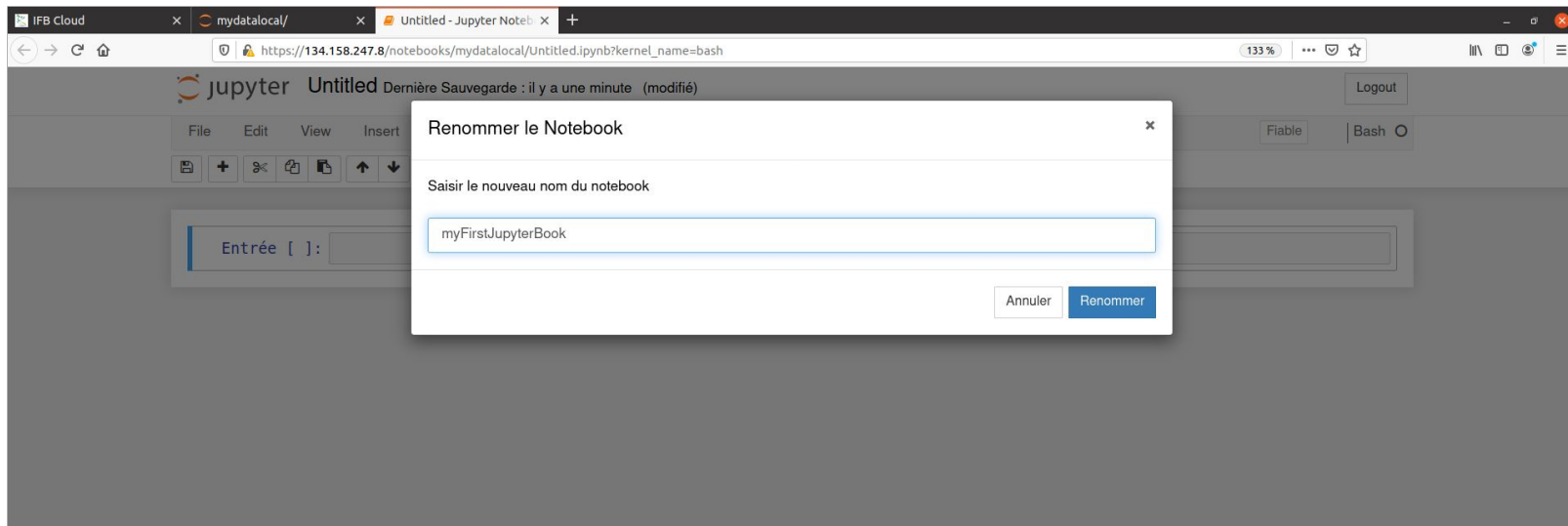
Go into the directory “work” and create a new jupyter book
-> kernel : bash



The screenshot shows the JupyterLab interface in a web browser. The address bar displays "Non sécurisé | https://134.158.248.119/lab?". The interface includes a menu bar (File, Edit, View, Run, Kernel, Tabs, Settings, Help) and a sidebar on the left with a file browser. The file browser shows a directory structure with folders: mydatalocal (5 hours ago), public (6 months ago), and work (a month ago). The main area is the "Launcher" view, which displays a grid of kernel options. The "Notebook" section contains four options: Python 3 (ipykernel), Bash, Julia 1.7.3, and R. An arrow points to the "Bash" option. Below the "Notebook" section is the "Console" section, which also contains the same four kernel options. At the bottom is the "Other" section, which includes Terminal, Text File, Markdown File, Julia File, Python File, R File, and Show Contextual Help.

Rename your first jupyter book

myFirstJupyterBook



Run your first bash command - *git clone*

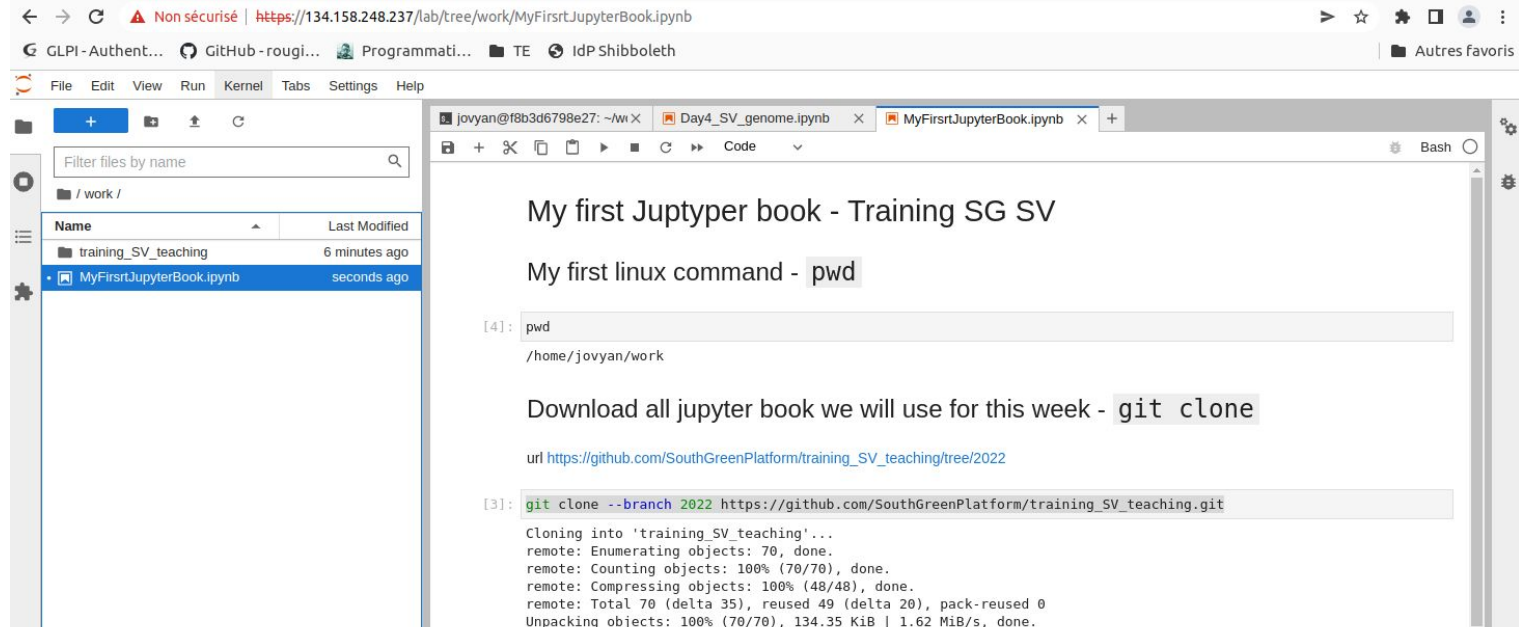
All jupyterbook used for practice are here :

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`

`git checkout 2023_MTP`



The screenshot shows a JupyterLab environment. On the left, a file browser displays the directory structure: `/ work /` with subdirectories `training_SV_teaching` (modified 6 minutes ago) and `MyFirstJupyterBook.ipynb` (modified seconds ago). The main area shows a terminal window with the following content:

```
My first Juptyper book - Training SG SV

My first linux command - pwd

[4]: pwd

/home/jovyan/work

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

url https://github.com/SouthGreenPlatform/training\_SV\_teaching/tree/2022

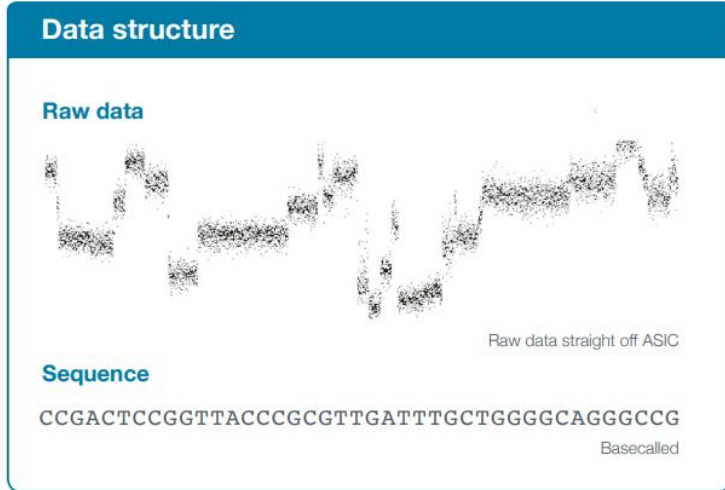
[3]: git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git

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.
```

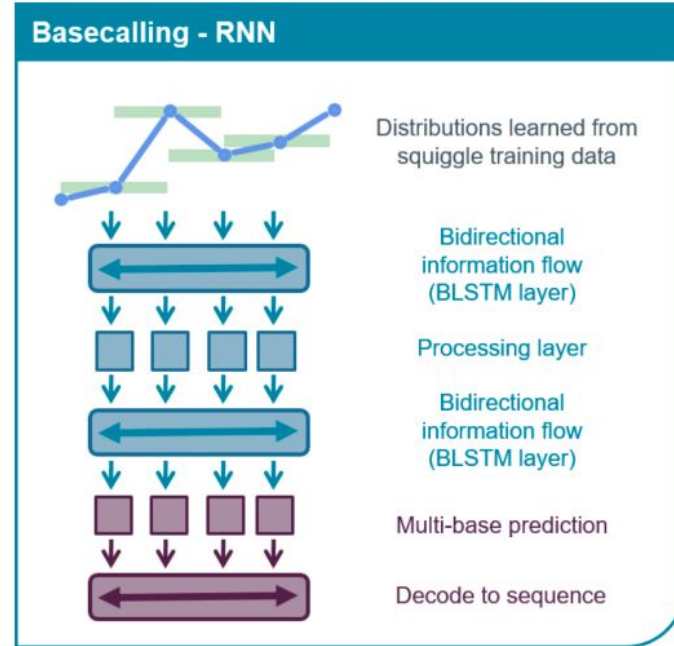
Chapitre 1

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

FASTQ FORMAT

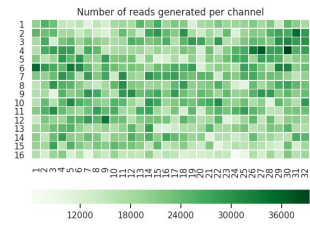
1 séquence = 4 lignes

```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTTCTTAGTATTTTTGTAGTCATCCGTGTTGGTTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTTGGT
+
@@@DDDD>FFFAFBEABB4C+3?:CBB@<<A?E4A??9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAAGTGCAGTAACGACCTATAAATTTTAAAGTAGC
+
@CFFFFFFGGHHHHIJJJIEE<HHHIJJIGBHGGEEIJJJEIJIHHJFIIJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTTCGGGTATGGCTCTAGTAGCTTTTGTCTTAT
+
@C@FFFFFFGGHHDHGIIEEHIII<CGHIJJIIJ:?FC9DGAFGHII?DGBFIIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@DFFFFGHHHHJJIIIIJJIGJCHHEIII>GHIG@GHIDHGJIIFHIIJJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGTAAGAGTTAAAAACAATTAATGAGCAACTGAGTTC
+
@@CFFFFFFHFFHFGIJJIIHHIIJJIIHIIJJJECGHIJJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTTGTGGGGGTTGGGAAATTGTCACCTTCTGCTACAATGCCG
+
@<@DDDDHFFFDIIBDFGHHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```

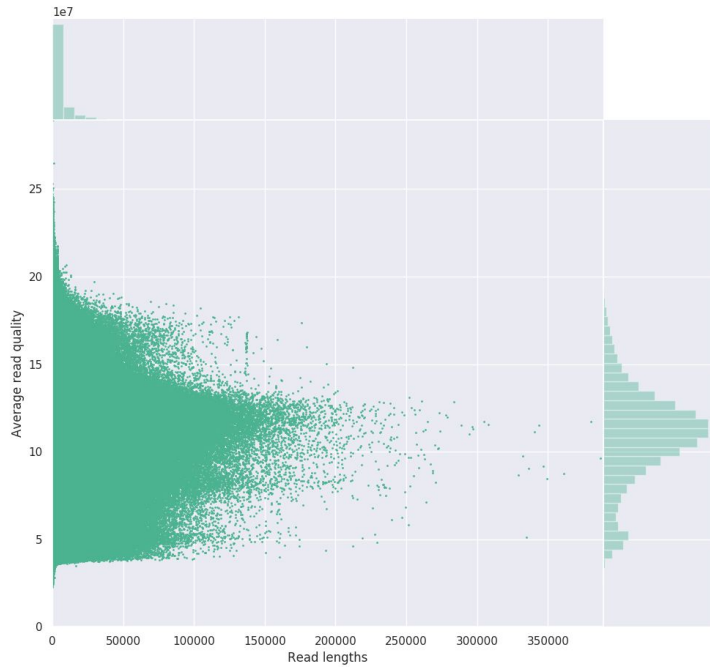


- @identifiant de la séquence
- Séquence
- + (id séquence).
- Qualité de la séquence = un caractère ASCII pour chaque base

Reads Quality control : *NanoPlot*

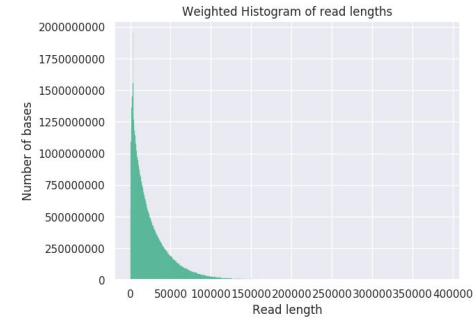


Read lengths vs Average read quality plot

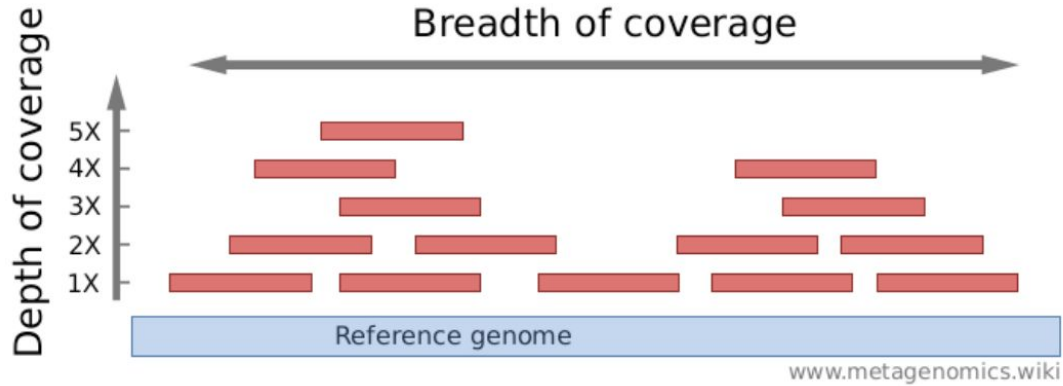


Summary statistics

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



Calculate depth of coverage

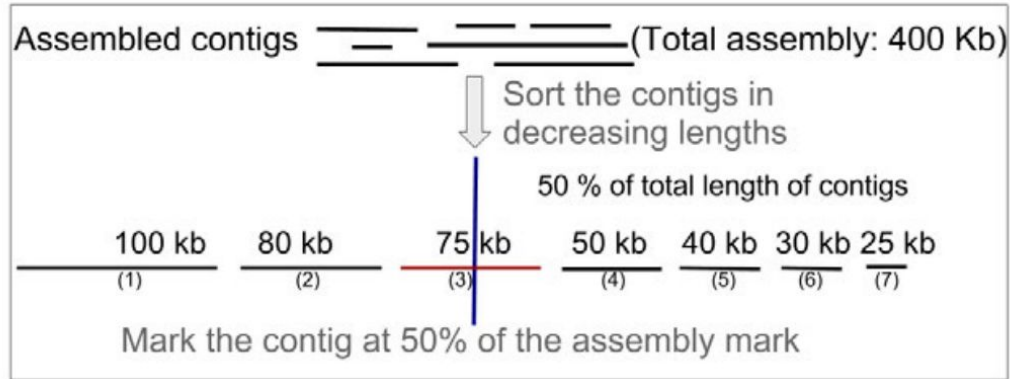


depth of coverage estimation :

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

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

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 !

Chapitre 2

Assemblies

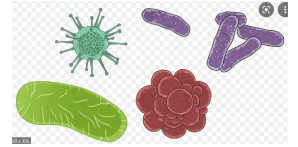
Which assembler to use over my favorite organism?

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.

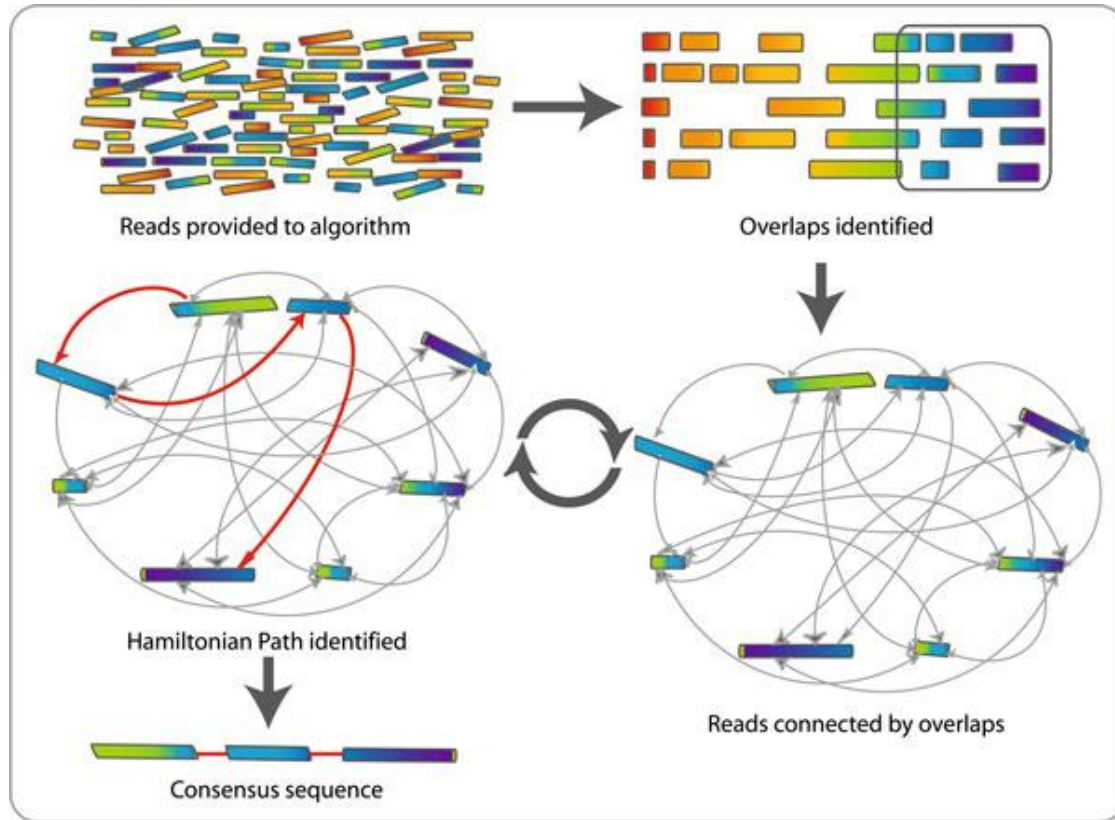
<https://nanoporetech.com/sites/default/files/s3/literature/microbial-genome-assembly-workflow.pdf>

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)



Polishing / Correction

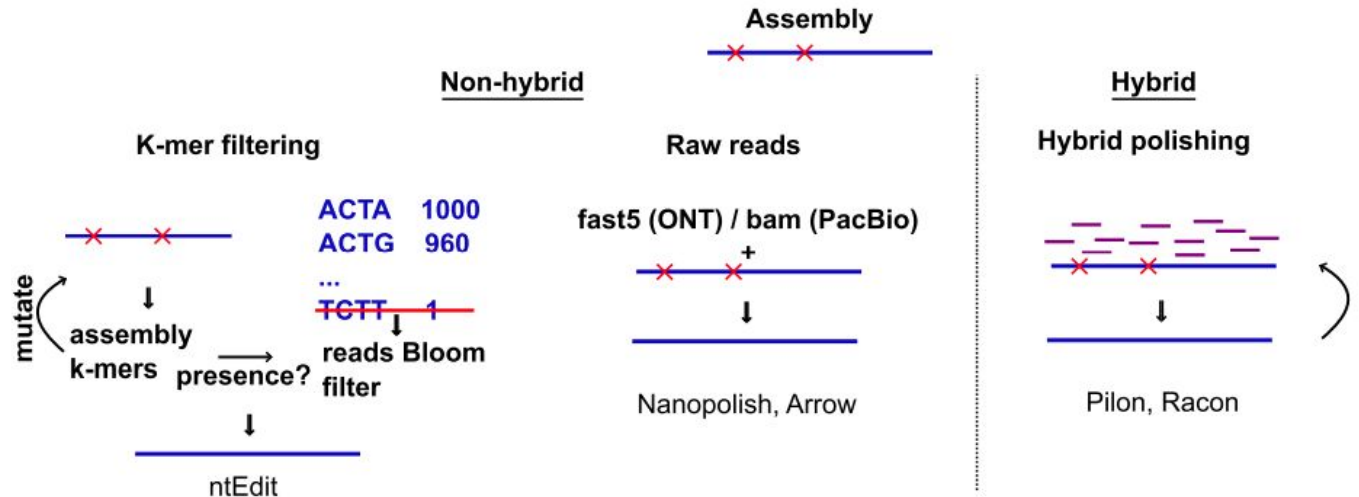
Racon correct raw contigs generated by rapid assembly methods which do not include a consensus step. It can polish with either Illumina data or data produced by third generation of sequencing. (recursive use)

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

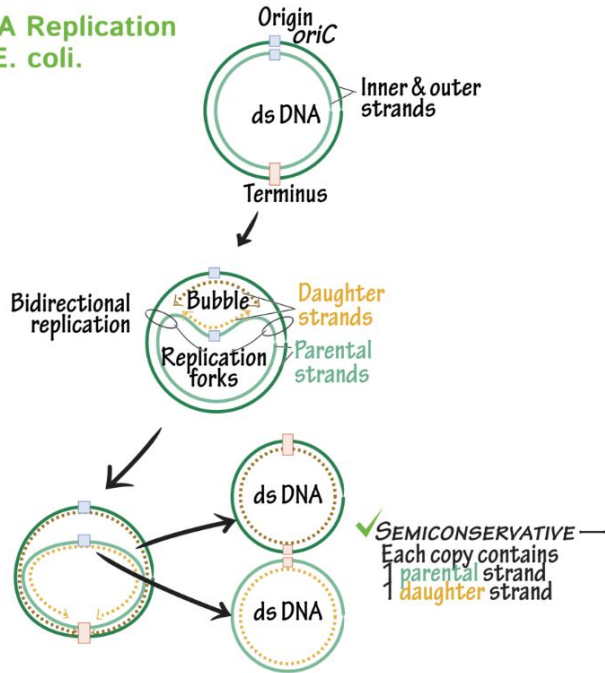
Pilon correct assemblies using illumina reads. (recursive use)

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.

Chapitre 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 ≥ 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

 Show heatmap

Genome statistics	FLYE_STEP_POLISHING_RACON	FLYE_STEP_ASSEMBLY	RAVEN_STEP_POLISHING_RACON	RAVEN_STEP_ASSEMBLY	SHASTA_STEP_POLISHING_RACON	SHASTA_STEP_ASSEMBLY
Genome fraction (%)	65.801	65.916	65.417	15.444	65.191	59.206
Duplication ratio	1.036	1.041	1.041	1.001	1.027	1.02
Largest alignment	2 503 013	2 501 477	1 739 590	52 085	1 894 135	1 779 077
Total aligned length	255 403 246	257 194 821	255 339 839	57 942 144	251 070 979	226 440 774
NGA50	48 559	48 062	42 714	-	42 891	14 680
LGA50	1338	1333	1404	-	1432	2186
Misassemblies						
# misassemblies	9633	9923	7666	11	6928	4350
Misassembled contigs length	373 371 138	373 825 172	335 007 830	226 513	318 941 237	283 654 397
Mismatches						
# mismatches per 100 kbp	2776.55	2831.25	2669.89	1271.33	2668.91	2675.87
# indels per 100 kbp	321.69	301.83	330.99	1344.38	318.53	437.86
# N's per 100 kbp	0	0.23	0	0	0	0
Statistics without reference						
# contigs	181	250	250	250	729	854
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 (≥ 1000 bp)	383 173 133	384 197 574	387 291 200	383 785 534	369 966 935	373 406 571
Total length (≥ 10000 bp)	382 901 616	383 618 037	387 291 200	383 785 534	368 865 072	371 578 702
Total length (≥ 50000 bp)	381 421 486	381 880 053	387 291 200	383 785 534	365 953 108	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

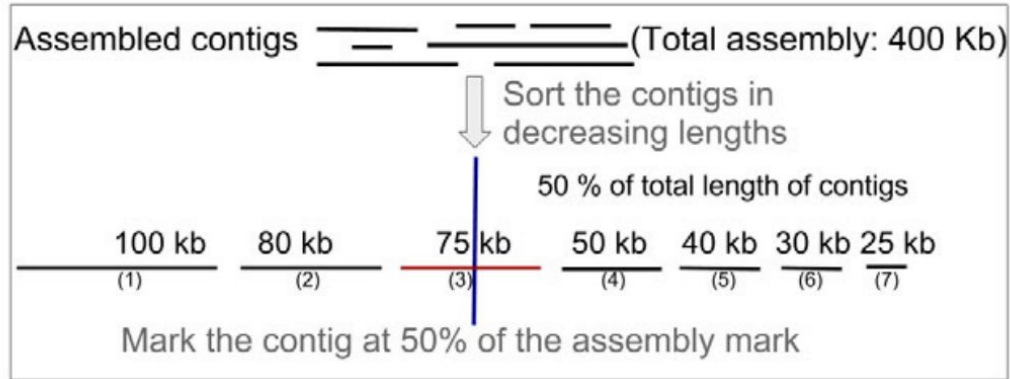
largest contig : flye+racon

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

 Show heatmap

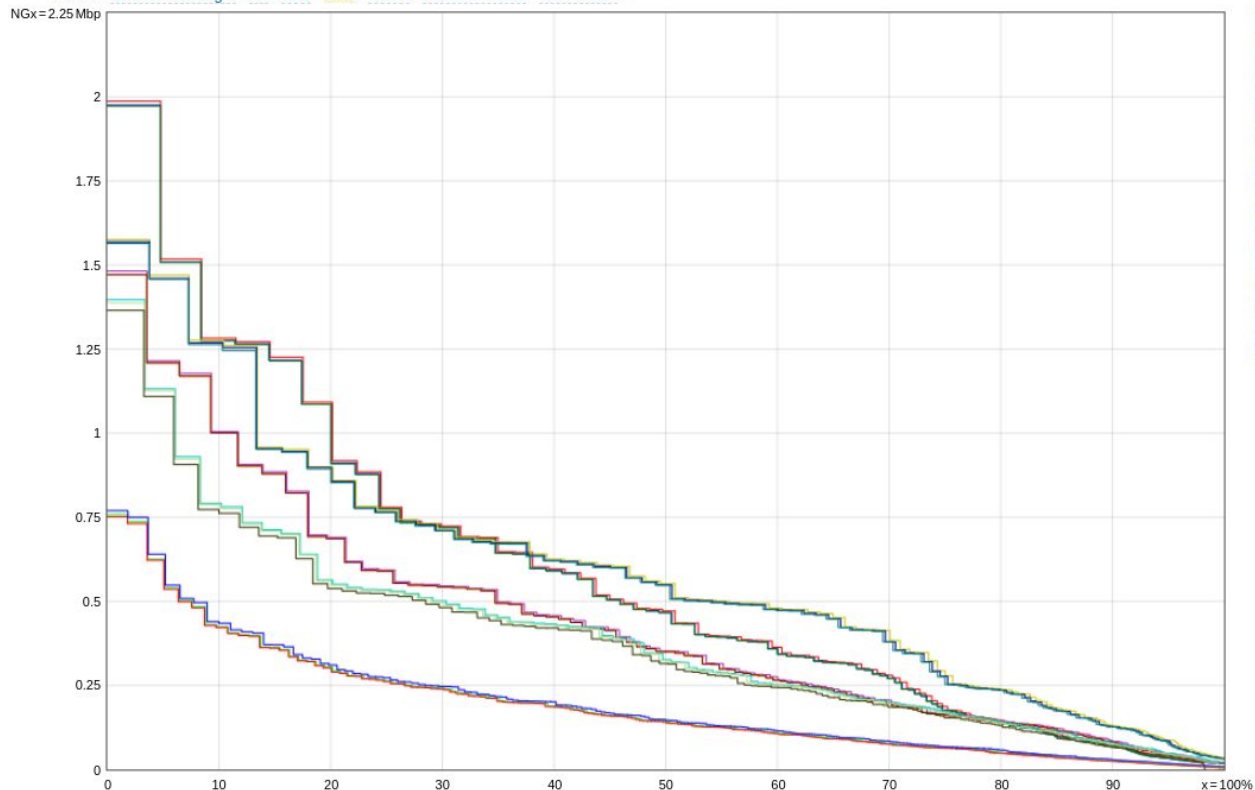
Genome statistics	FLYE_STEP_POLISHING_RACON	FLYE_STEP_ASSEMBLY	RAVEN_STEP_POLISHING_RACON	RAVEN_STEP_ASSEMBLY	SHASTA_STEP_POLISHING_RACON	SHASTA_STEP_ASSEMBLY
Genome fraction (%)	65.801	65.916	65.417	15.444	65.191	59.206
Duplication ratio	1.036	1.041	1.041	1.001	1.027	1.02
Largest alignment	2 503 013	2 501 477	1 739 590	52 085	1 894 135	1 779 077
Total aligned length	255 403 246	257 194 821	255 339 839	57 942 144	251 070 979	226 440 774
NGA50	48 559	48 062	42 714	-	42 891	14 680
LGA50	1338	1333	1404	-	1432	2186
Misassemblies						
# misassemblies	9633	9923	7666	11	6928	4350
Misassembled contigs length	373 371 138	373 825 172	335 007 830	226 513	318 941 237	283 654 397
Mismatches						
# mismatches per 100 kbp	2776.55	2831.25	2669.89	1271.33	2668.91	2675.87
# indels per 100 kbp	321.69	301.83	330.99	1344.38	318.53	437.86
# N's per 100 kbp	0	0.23	0	0	0	0
Statistics without reference						
# contigs	181	250	250	250	729	854
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 (≥ 1000 bp)	383 173 133	384 197 574	387 291 200	383 785 534	369 966 935	373 406 571
Total length (≥ 10000 bp)	382 901 616	383 618 037	387 291 200	383 785 534	368 865 072	371 578 702
Total length (≥ 50000 bp)	381 421 486	381 880 053	387 291 200	383 785 534	365 953 108	368 382 574

[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!

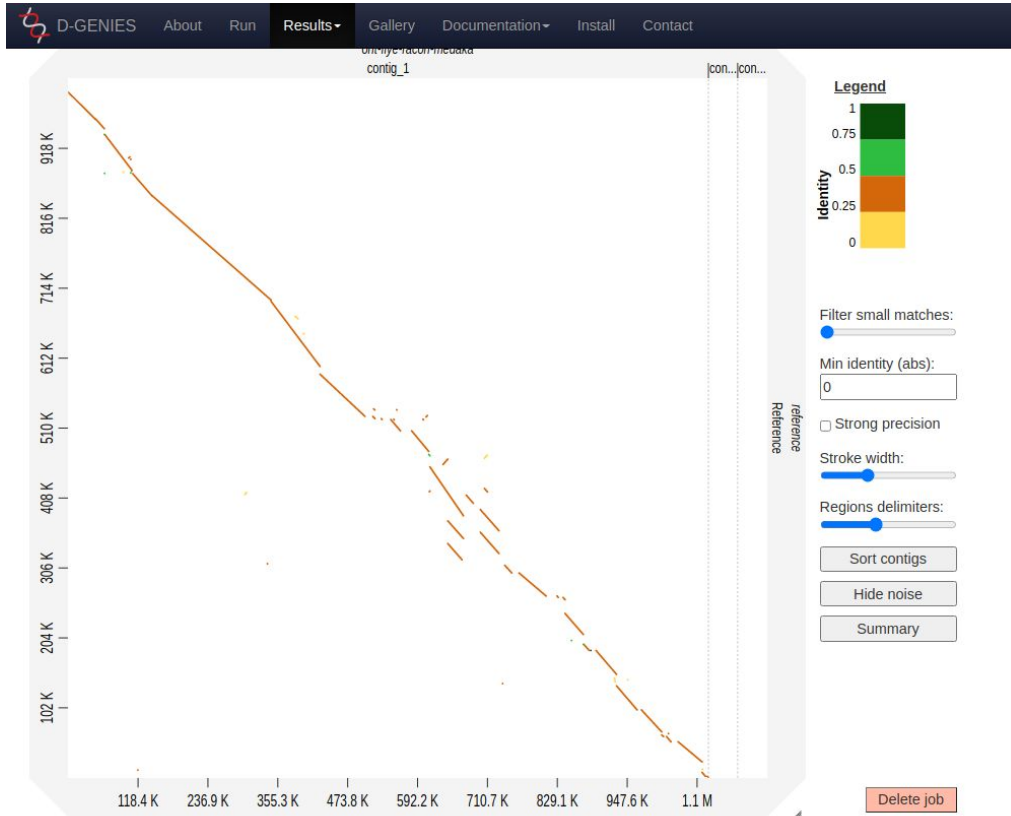
[Gitlab](#), a [Conda package](#) and [Docker container](#) are also available.

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

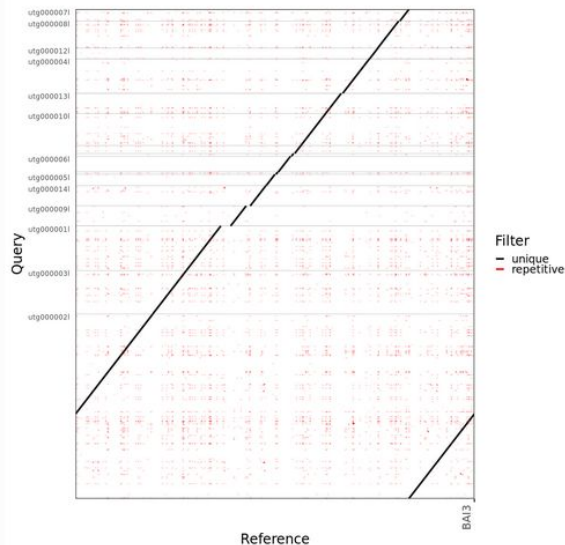
RAVEN

SMARTDENOVO

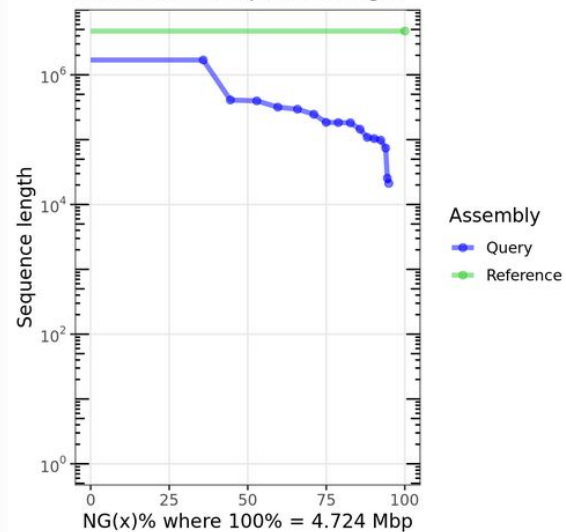
SHASTA

STEP_CORRECTION_NANOPOLISH_STARTFIXED

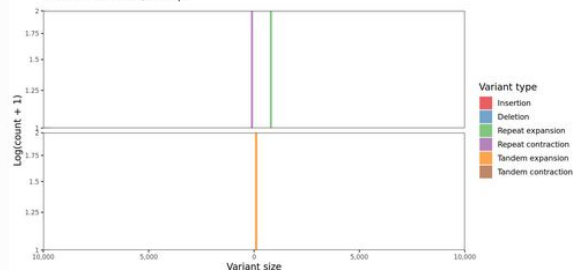
Dot plot of Assemblytics filtered alignments



Cumulative sequence length

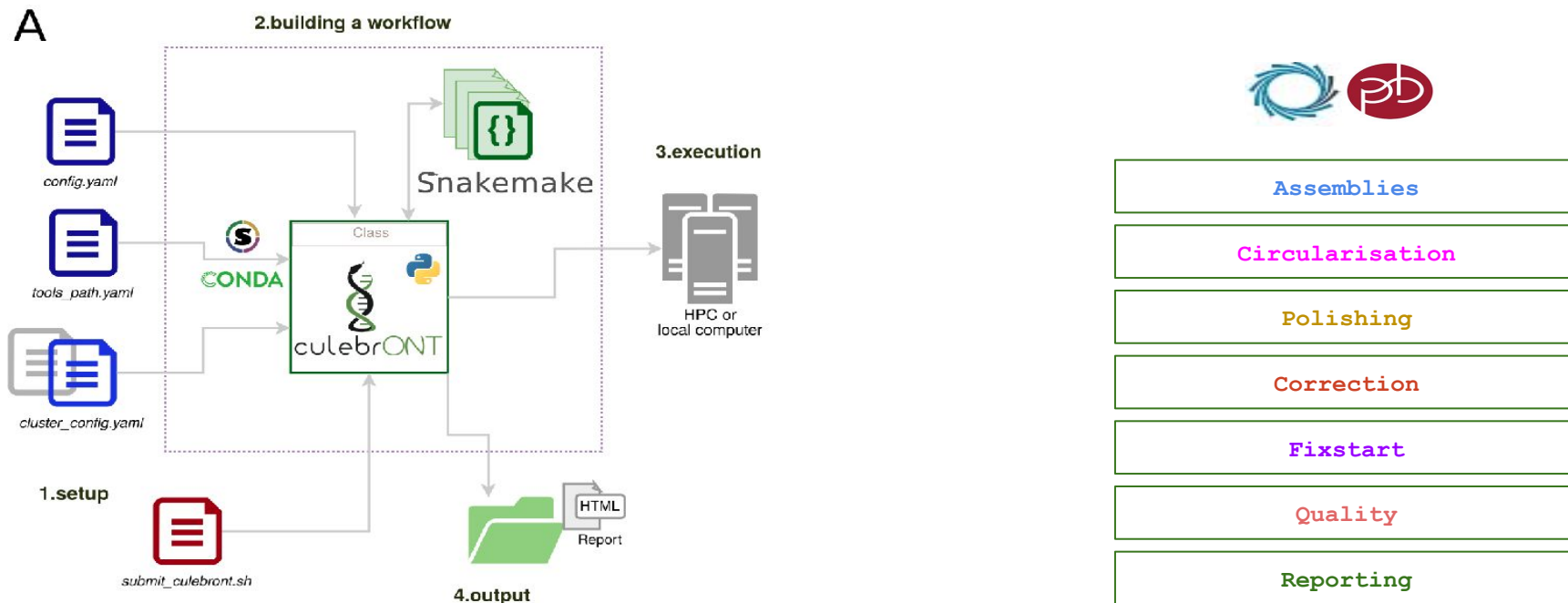


Variants 50 to 10,000 bp



A flexible and reproducible pipeline for LR assembly and evaluation

```
pip install culebrONT
```



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

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~~ now polishing is not mandatory.
- There are still genomes very difficult to assemble



Merci pour votre attention !



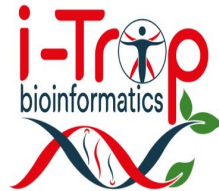
Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International:

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

SUIVEZ NOUS SUR TWITTER !



South Green : [@green_bioinfo](https://twitter.com/green_bioinfo)



I-Trop : [@ltropBioinfo](https://twitter.com/ltropBioinfo)