Initiation à l'analyse de données Oxford Nanopore
un focus sur la reconstruction de virus

Adapted from ONT training SG 2021
Prepared by Aurore Comte , Julie Orjuela and Denis Filloux

Ouagadougou, Septembre 2022

Let's discover Jupyter through the IFB cloud

Working environment

What is jupyter book ?

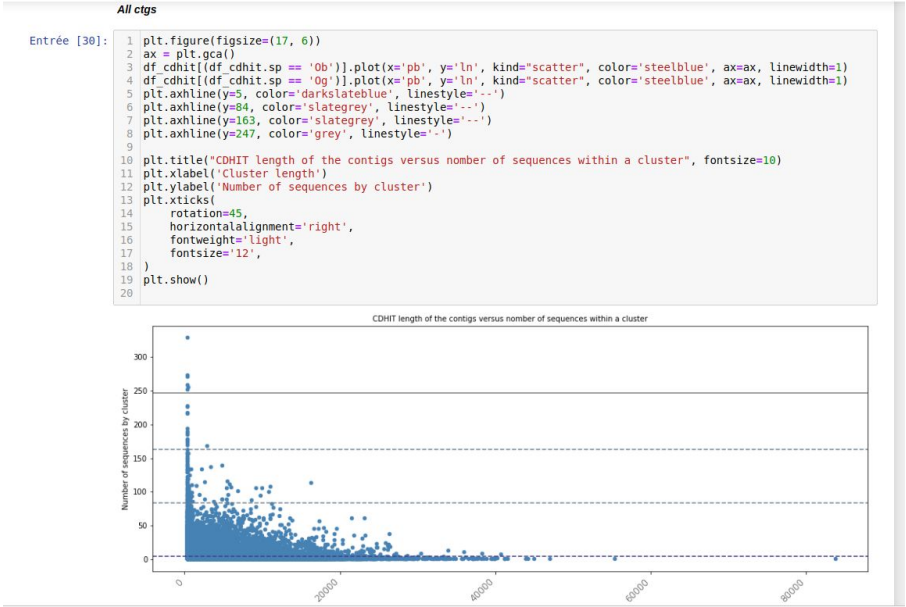
- One of the most popular tool among data scientists to perform data analysis
- Provides a complete environment in which numerous programming languages can be use through a simple web browser

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 :

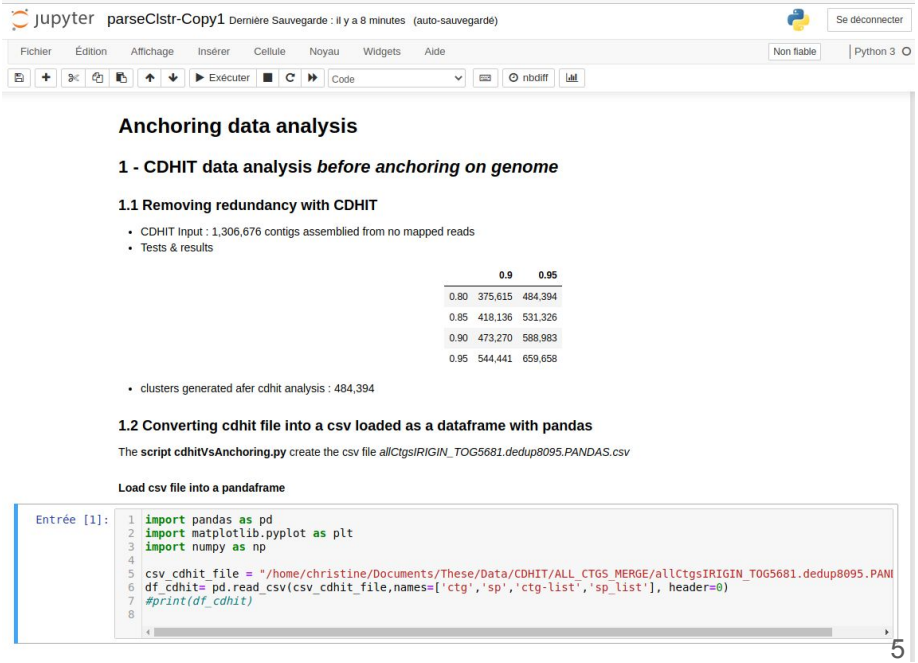
- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook



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



Lab notebook for science data ?

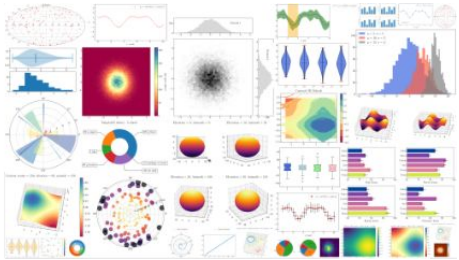
- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

How to become a super datascientist ?

- facilement importer des fichiers tabulés dans des dataframes, similaires aux dataframes sous R.
  (et exporter)

- manipuler ces tableaux de données / DataFrames

- facilement tracer des graphes à partir de ces DataFrames grâce à matplotlib



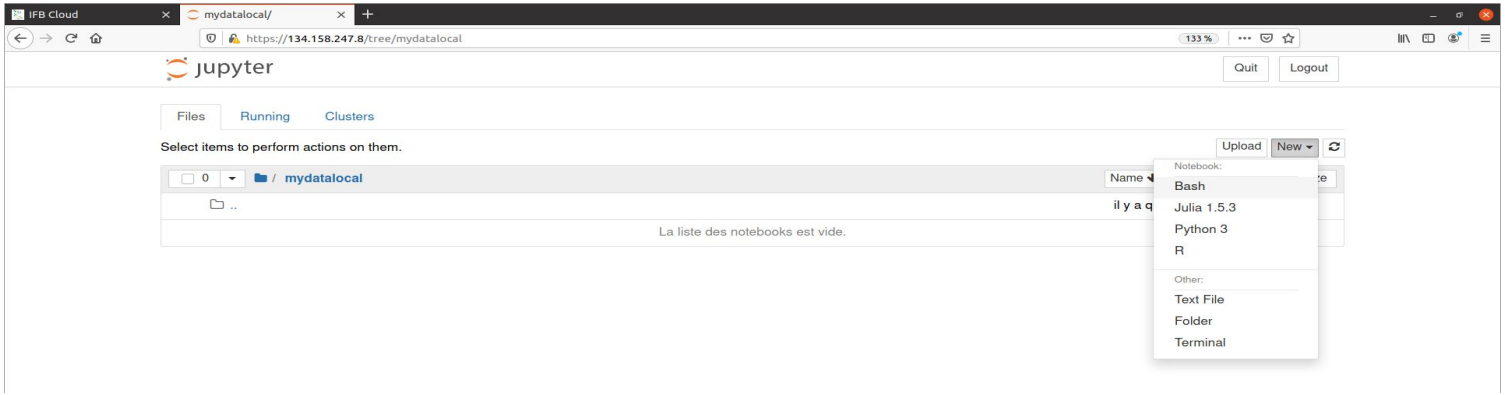
How will you use Jupyter Notebook ?

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”

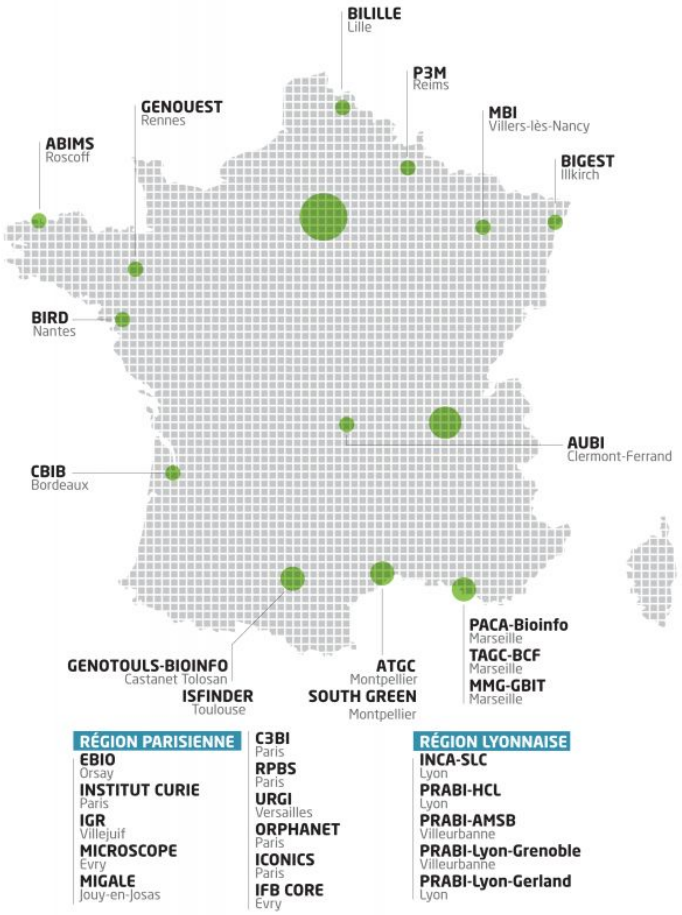
- Through this virtual machine, we will create jupyter books and execute all our analysis



IFB ?

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

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

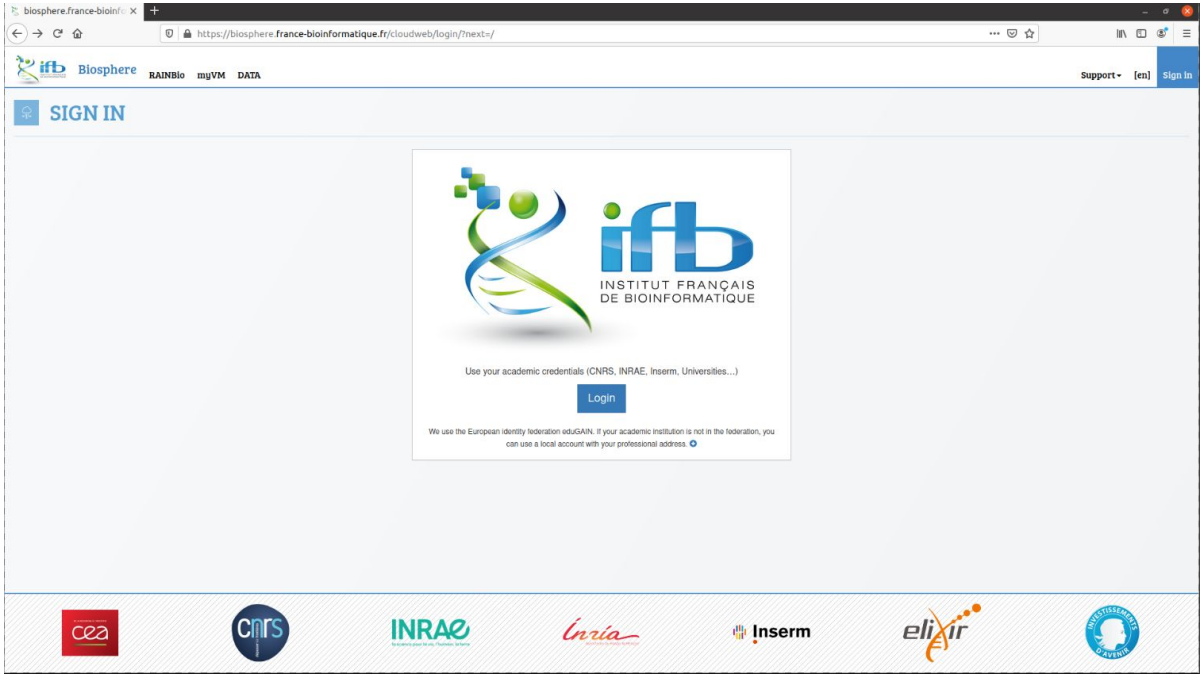


Biosphere, IFB CLOUD FOR LIFE SCIENCES

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

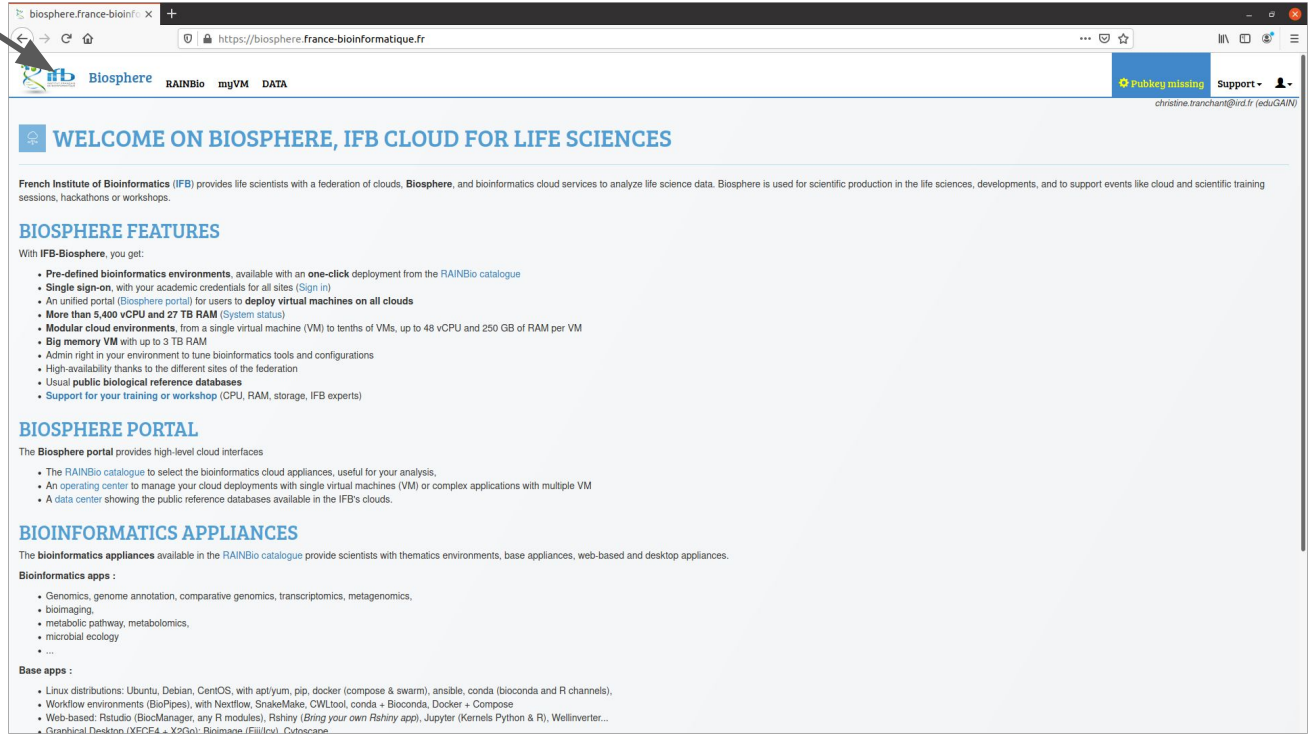
Let's start with biosphere

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



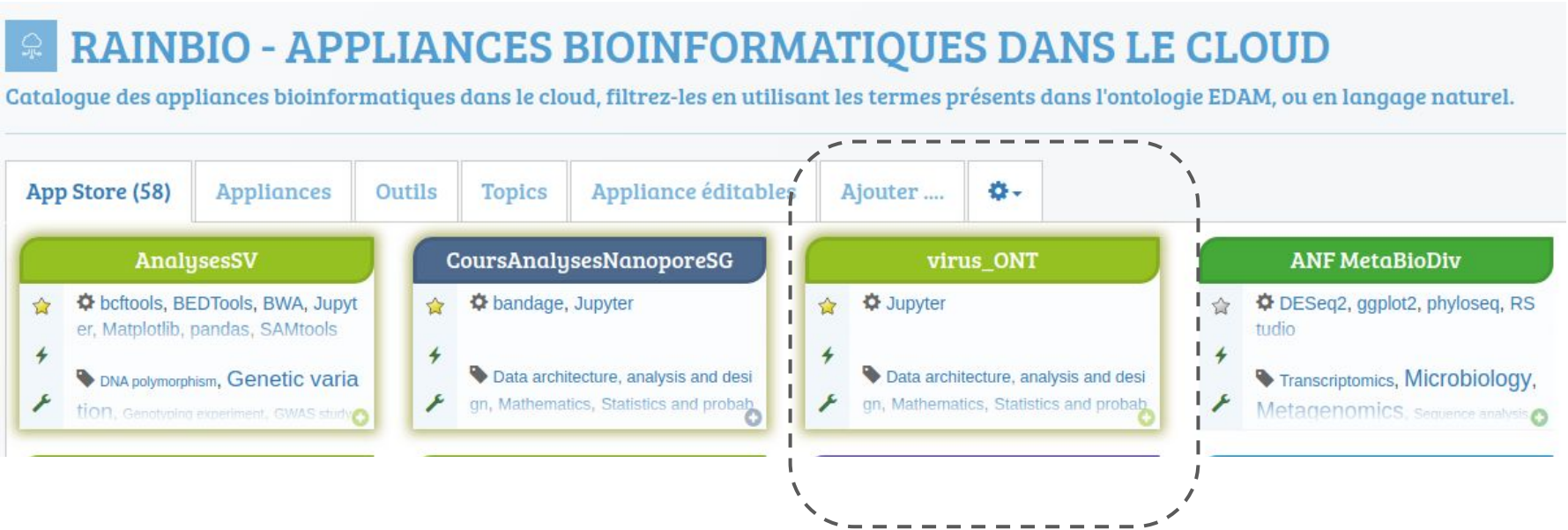
Connected / here we are

RAINBIO catalog to access our Virtual Machine (VM)

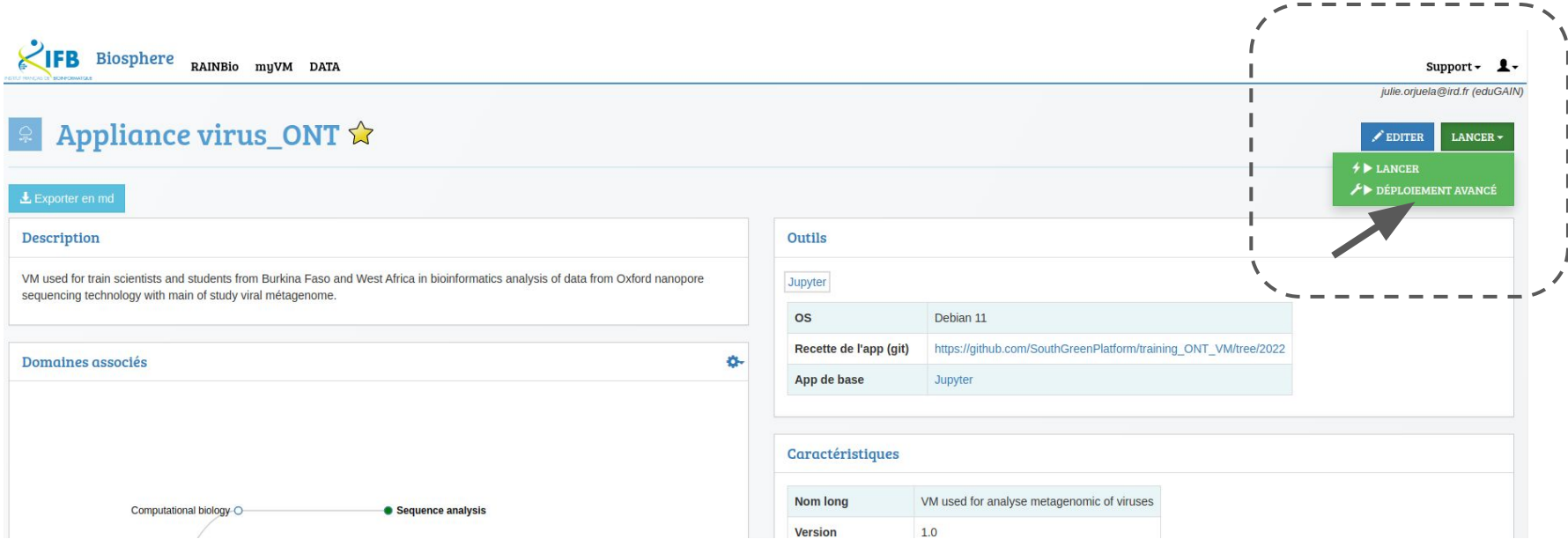


Searching for the vm we will use

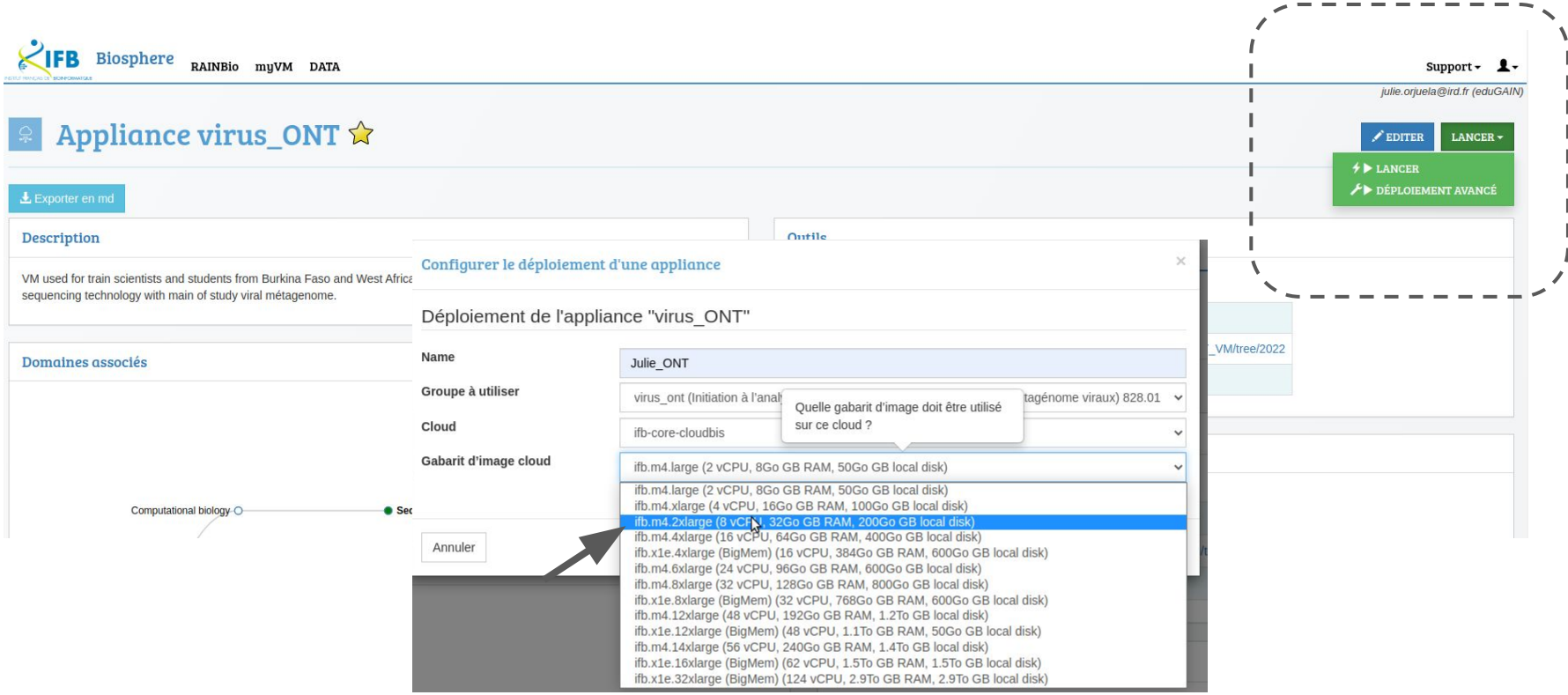
vm's name : **virus_ONT**



Let's run your vm through the cloud

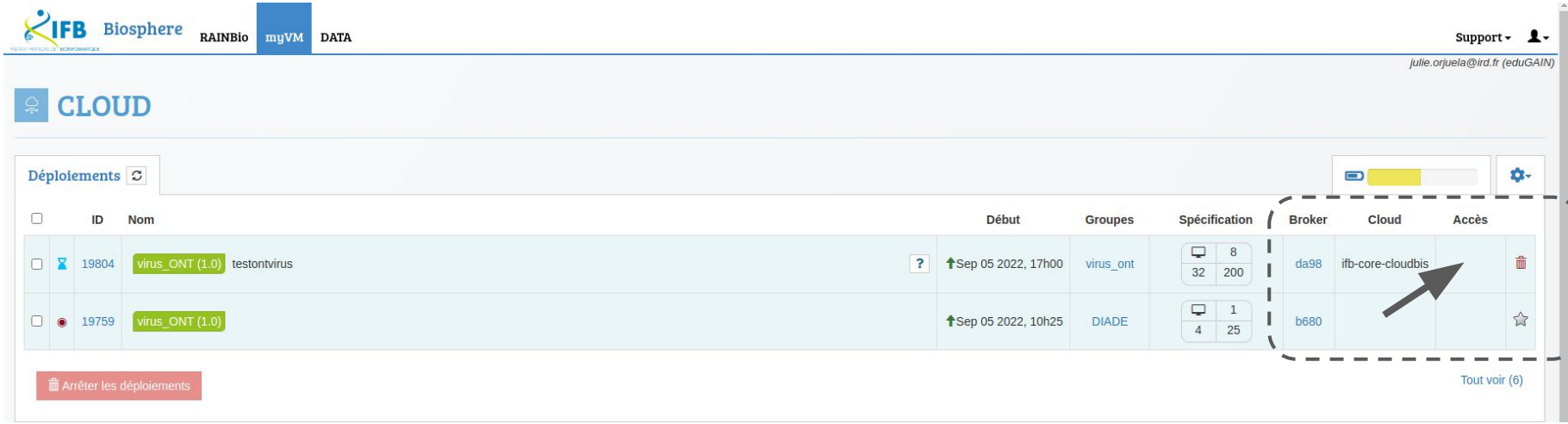


Let's run your vm through the cloud



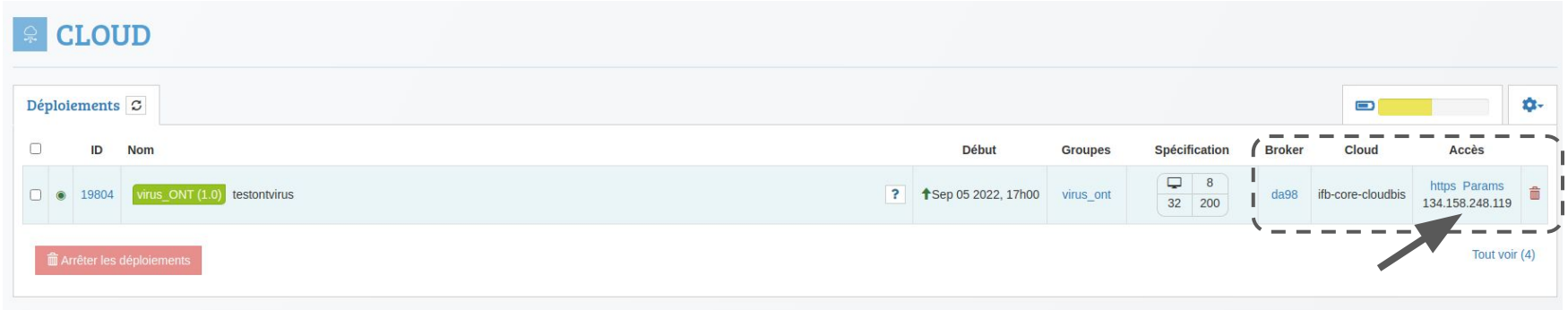
Let's run your vm through the cloud

Loading…



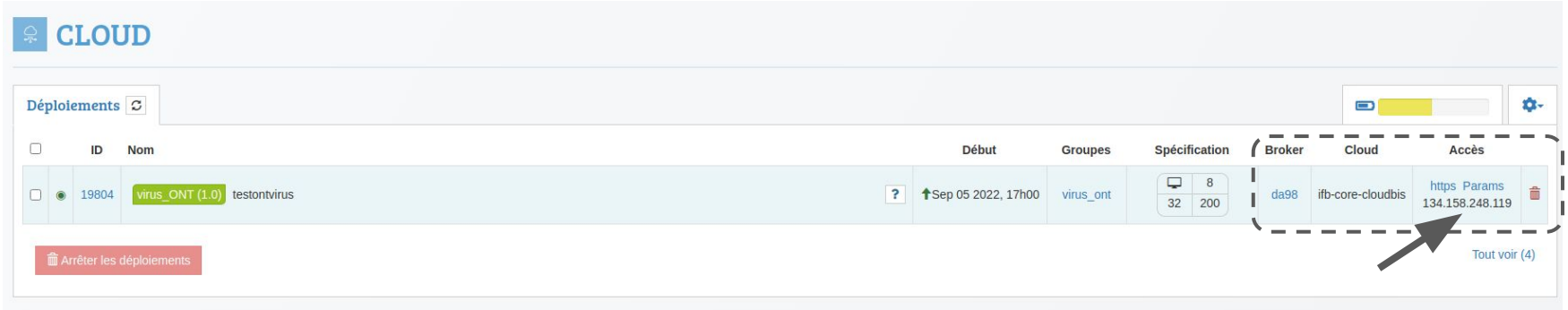
Let’s run your vm through the cloud

ready !



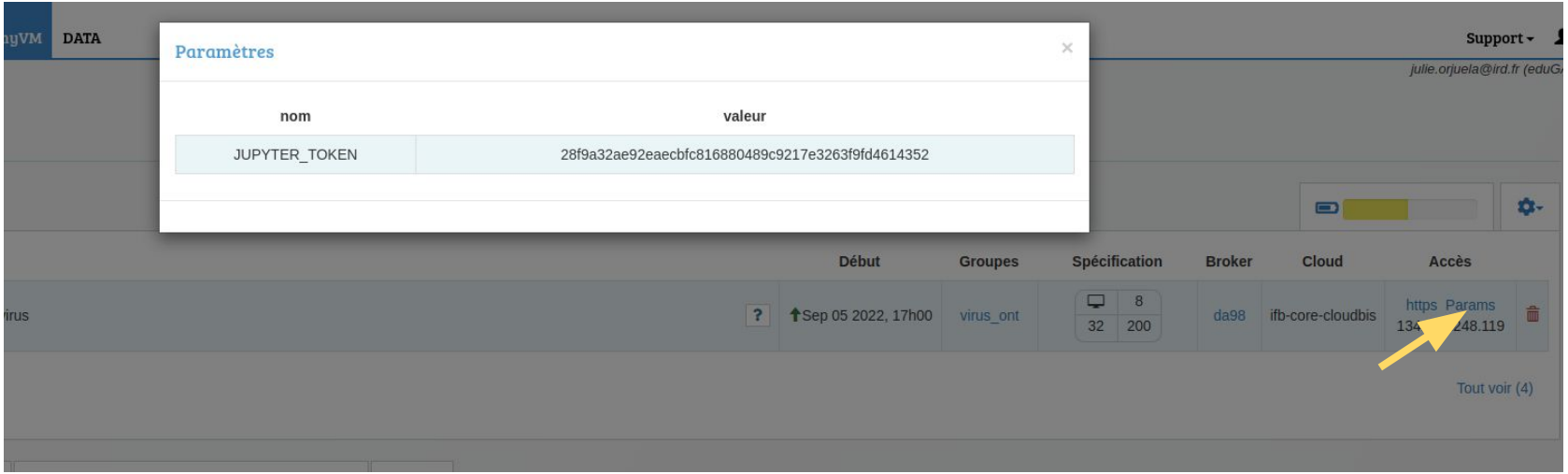
Let’s run your vm through the cloud

get the url… link “https”



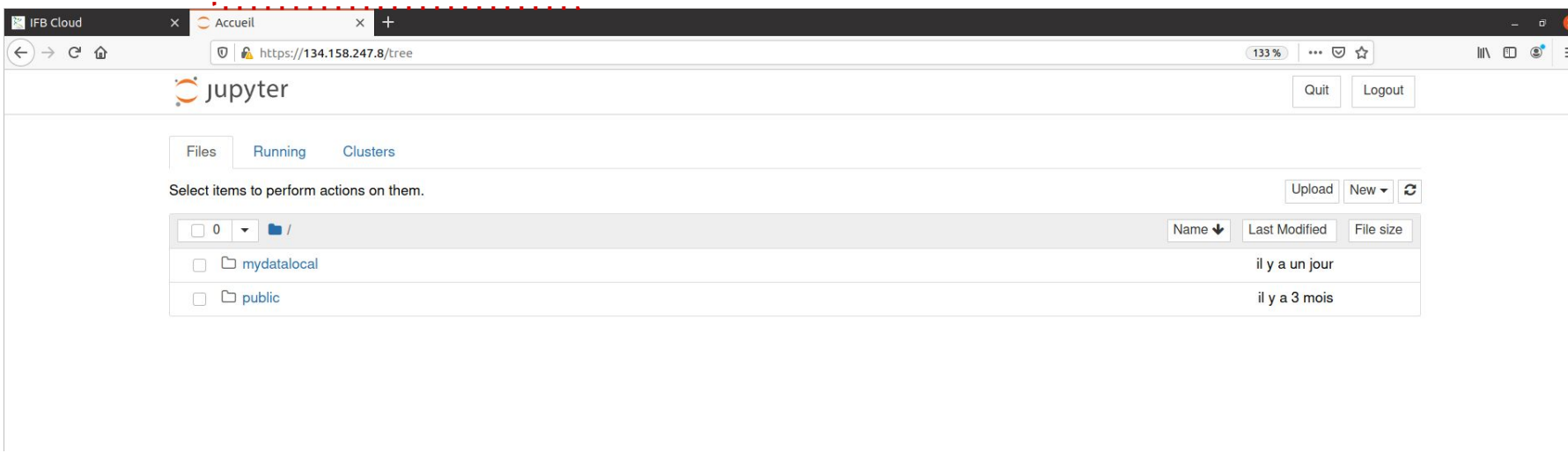
Let’s run our vm through the cloud

Get the token identifiant… link “Params”



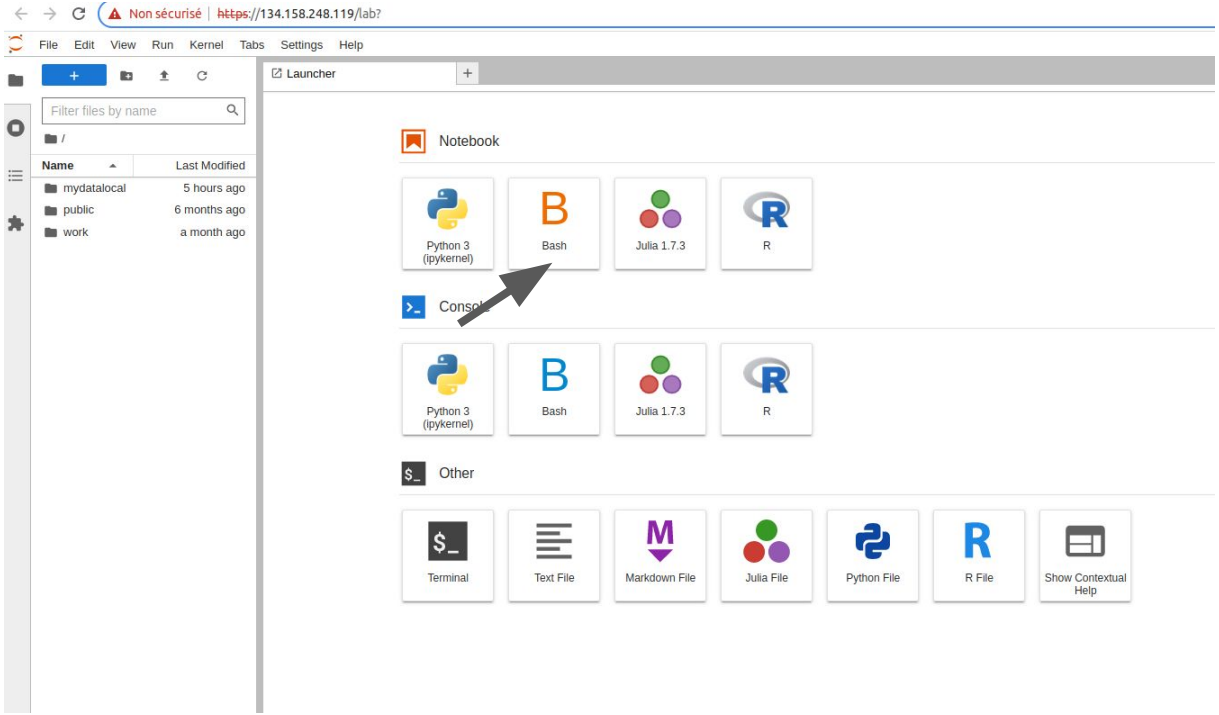
Let's run our vm through the cloud

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



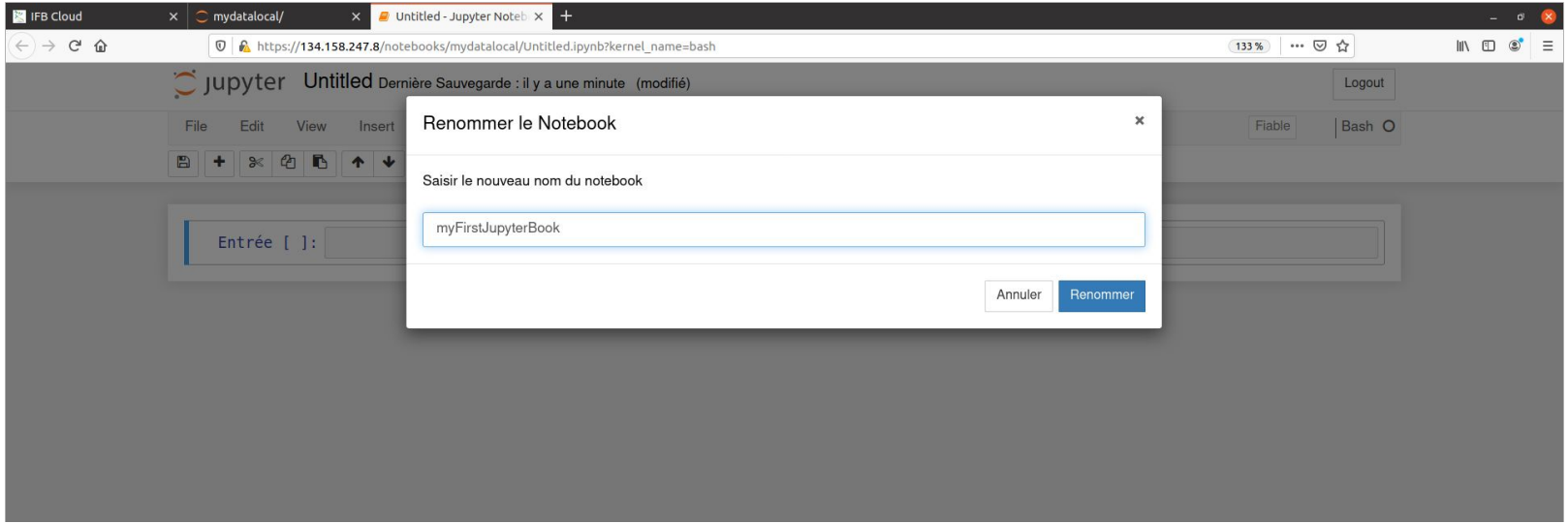
Create your first jupyter book

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



Rename your first jupyter book

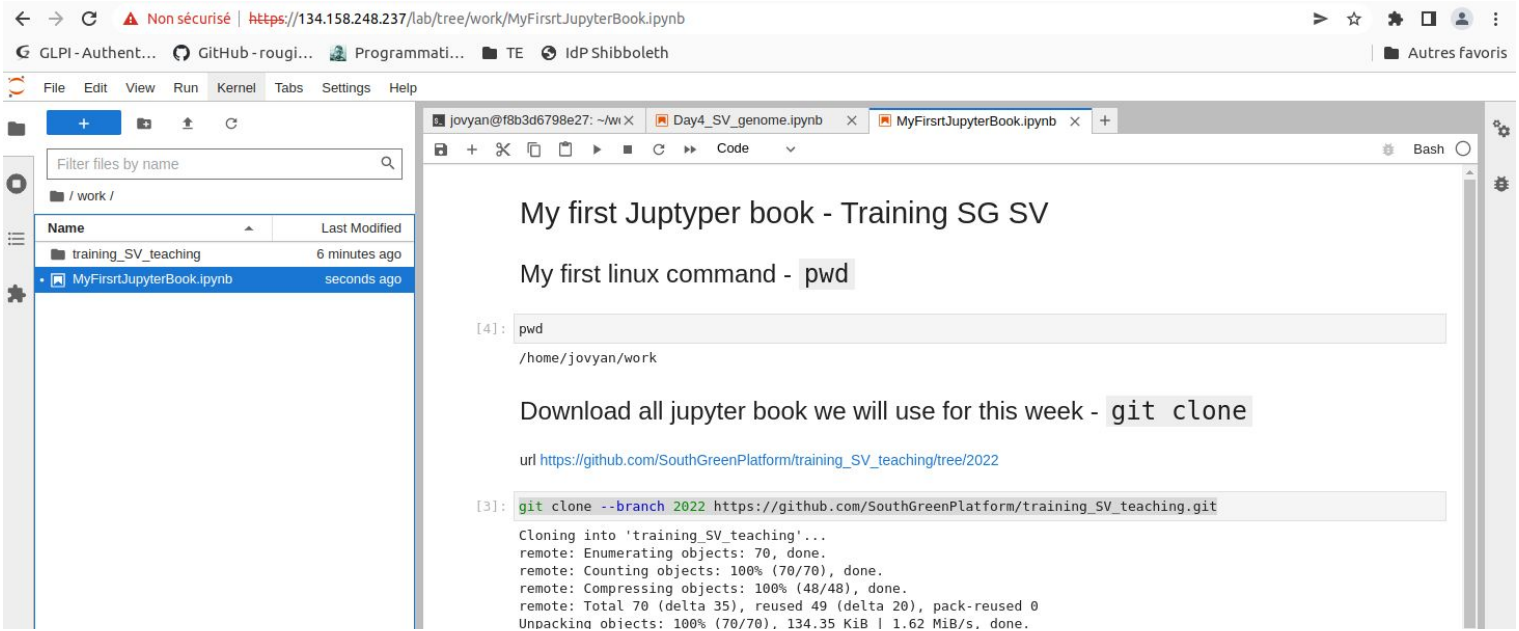
myFirstJupyterBook



Run your first bash command - *git clone*

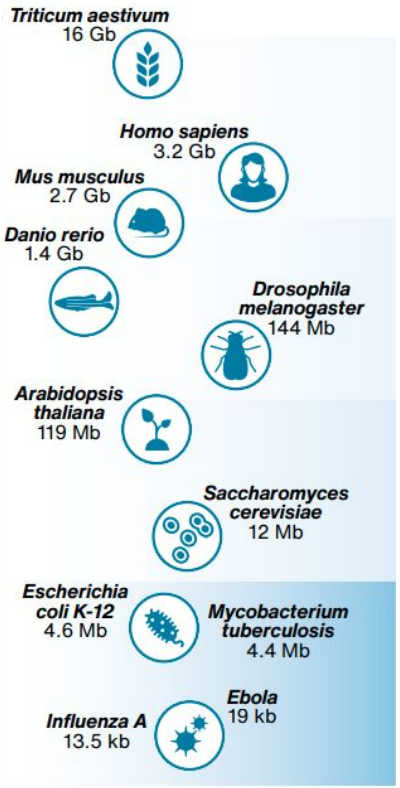
All jupyterbook used for practice are here :
https://github.com/SouthGreenPlatform/training_ONT_teaching/tree/2022

Download all the jupyter books with the command *git clone*
git clone --branch 2022 https://github.com/SouthGreenPlatform/training_ONT_teaching.git



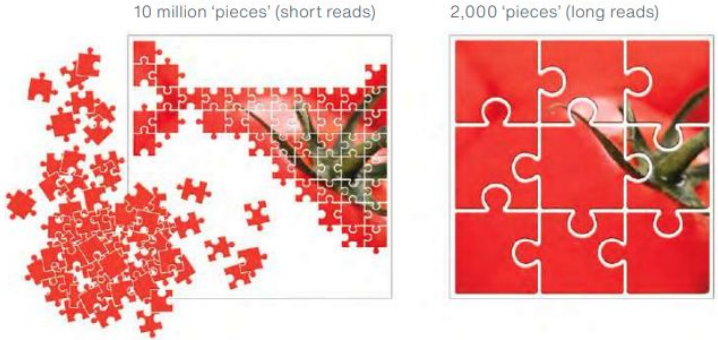
let’s start !

Why use Long reads ?

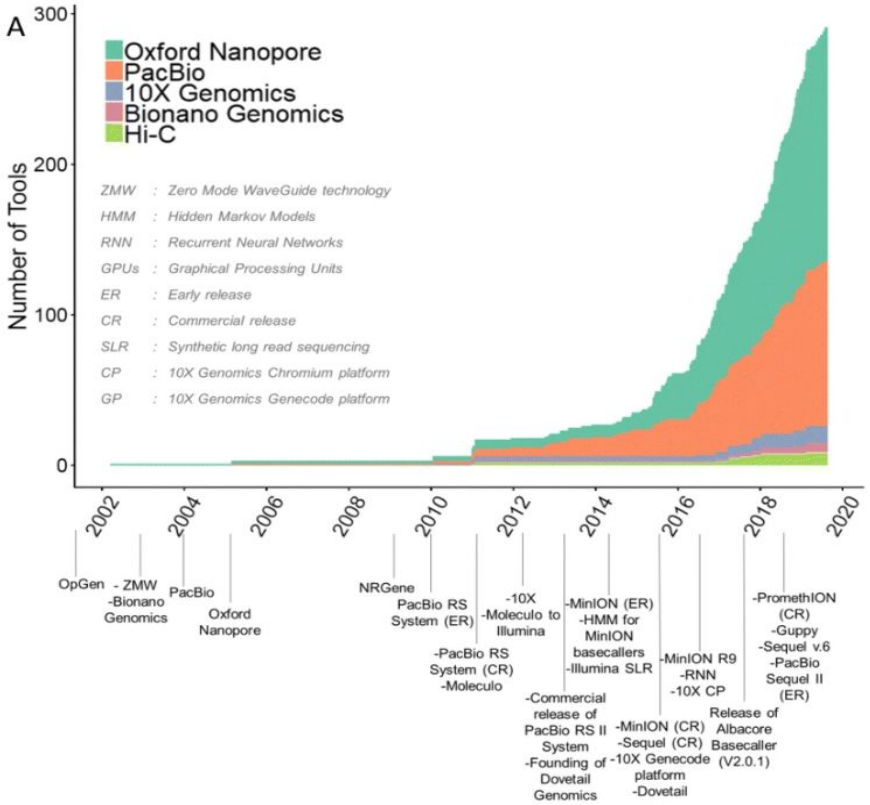


Microbial genomes | Human genomes | Animal genomes | Plant genomes

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

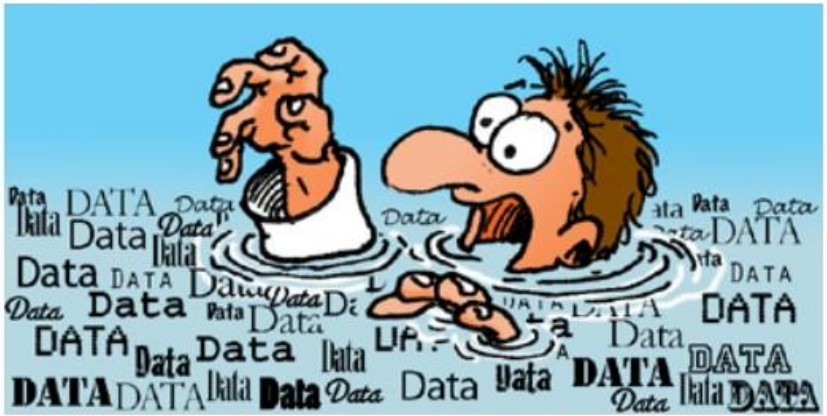


A lot of tools are being developed and upgraded frequently !

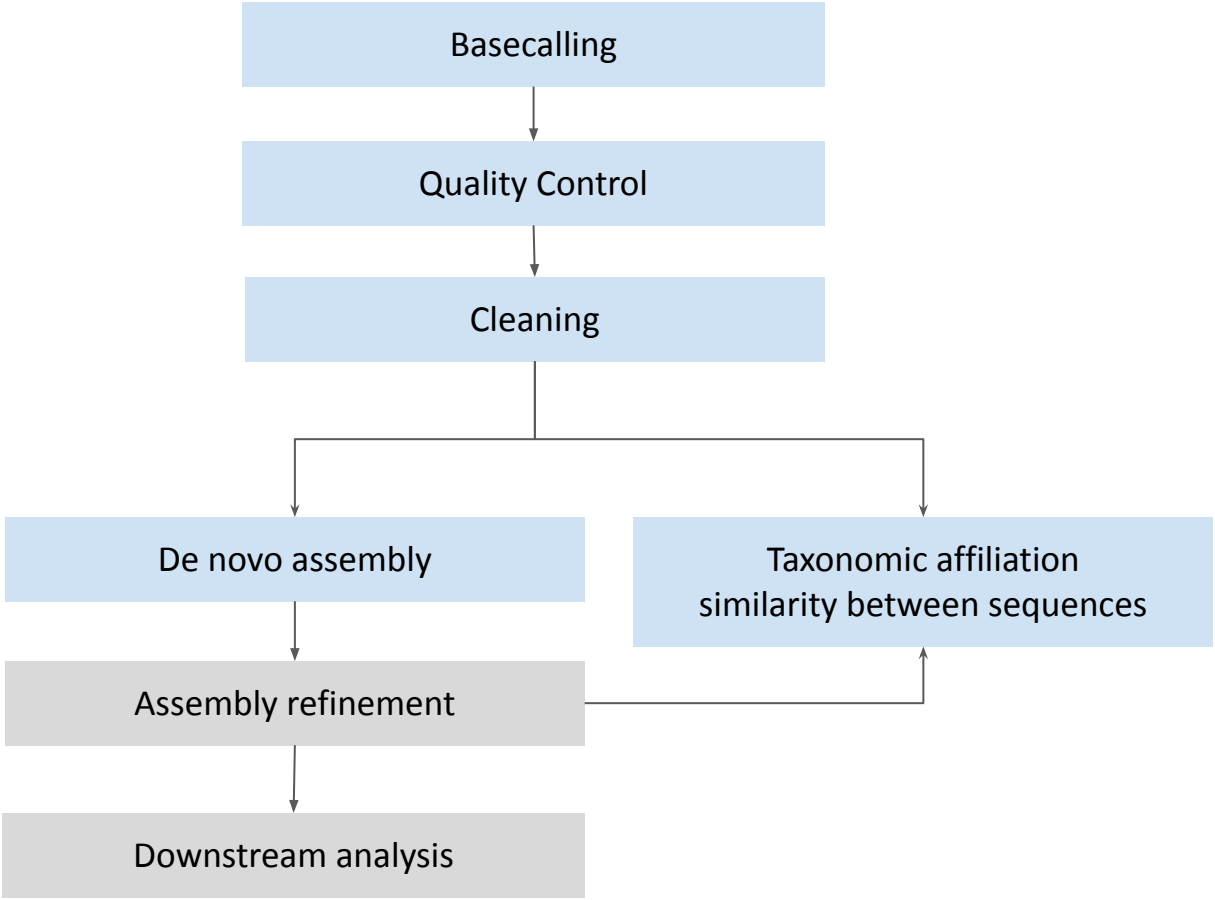


https://doi.org/10.1186/s13059-020-1935-5


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com/

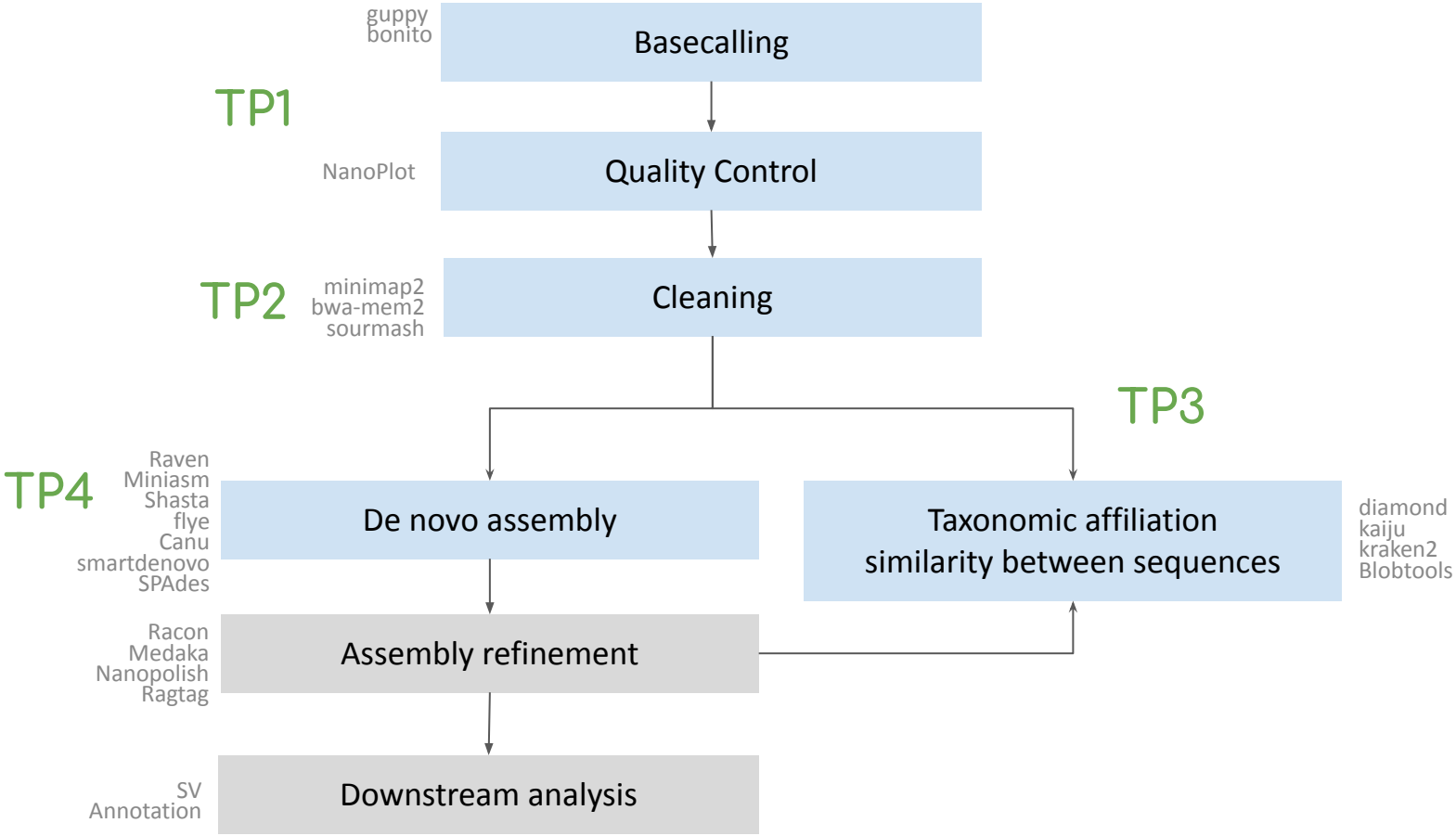

Data
DATA

Typical pipeline for virus reconstruction analysis



adapted from https://doi.org/10.1186/s13059-020-1935-5

Typical pipeline for virus reconstruction analysis



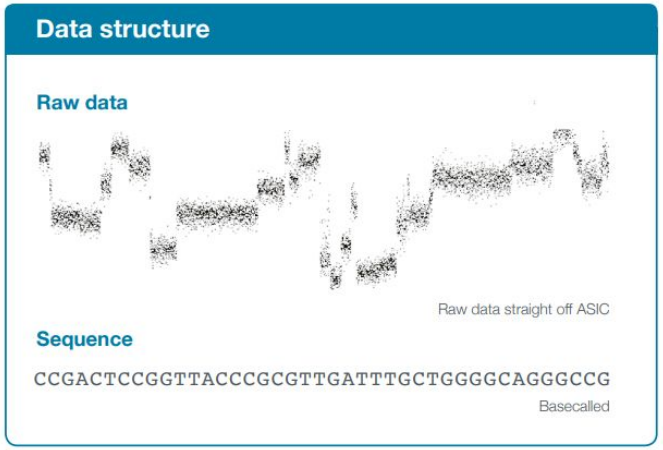
Les données

@denis

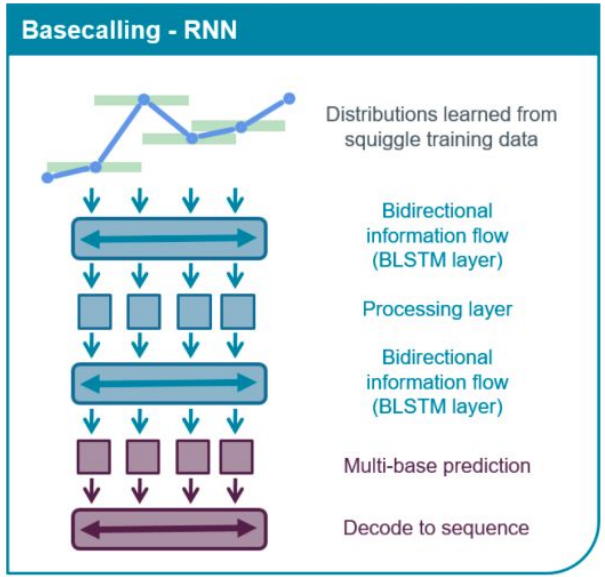
Chapitre 1

Reads Quality Control

ONT Read calling



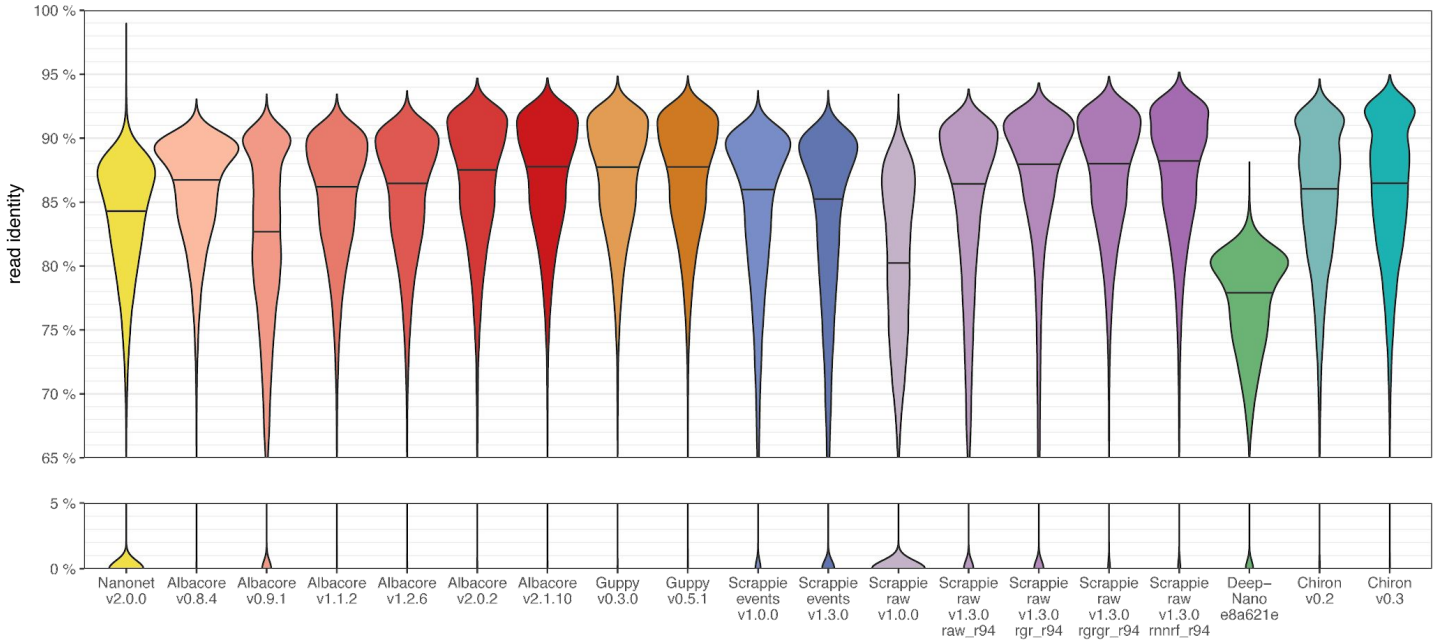
Reccurent Neural Network (RNN) – works like your brain! It can learn on the previous data and improve its performance on new data



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

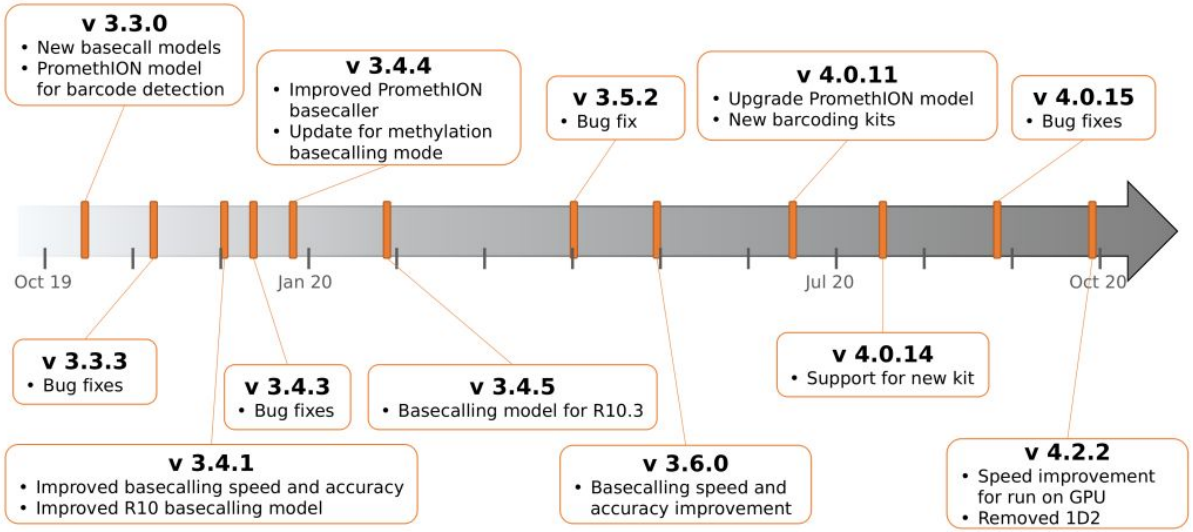
https://bioinformatics.uni-muenster.de/graid/education/presentations/DataAnalysis_Sapporo2019.pdf

ONT Read calling



https://github.com/rrwick/Basecalling-comparison

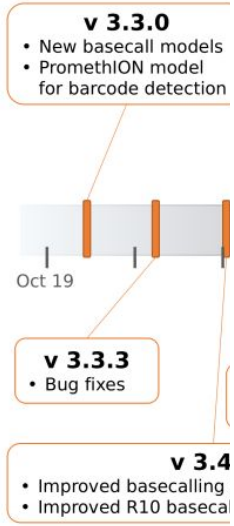
Guppy basecaller releases



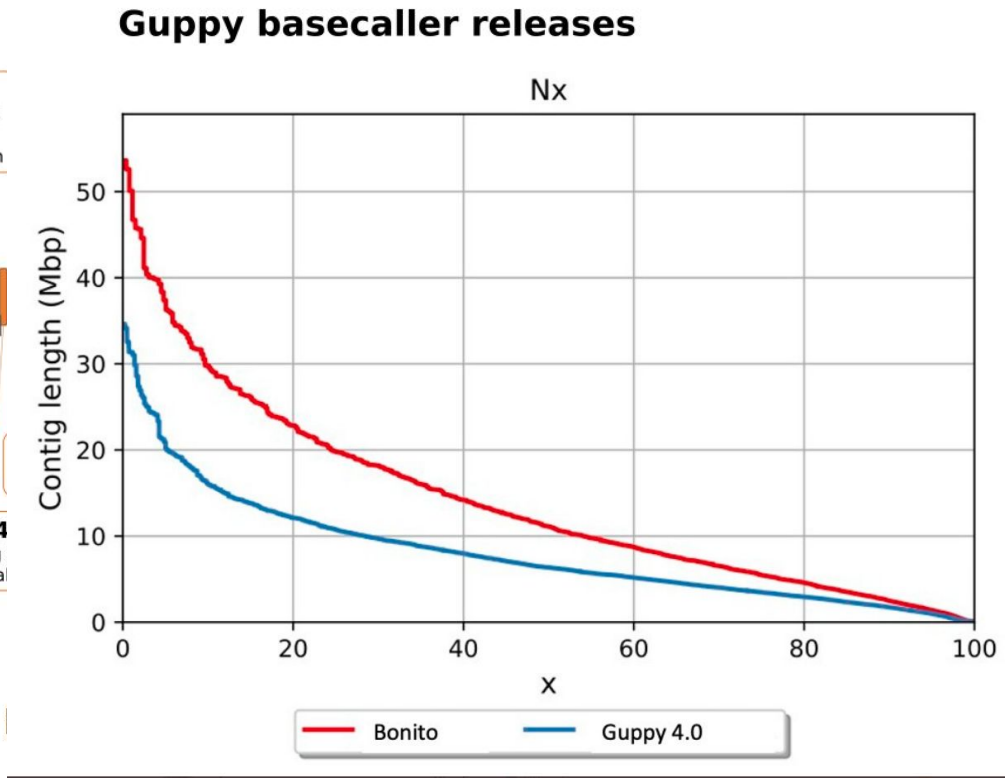
(+ Many other basecallers prior to Guppy [1] and to come.)

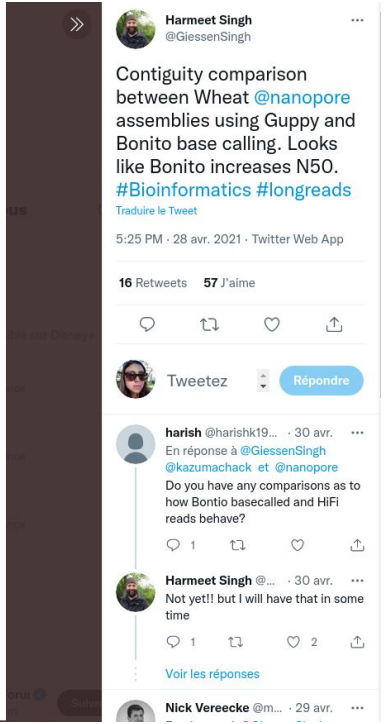
https://hal.inria.fr/hal-03123133/document

bonito



(+ Many other





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

summary_file.txt

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

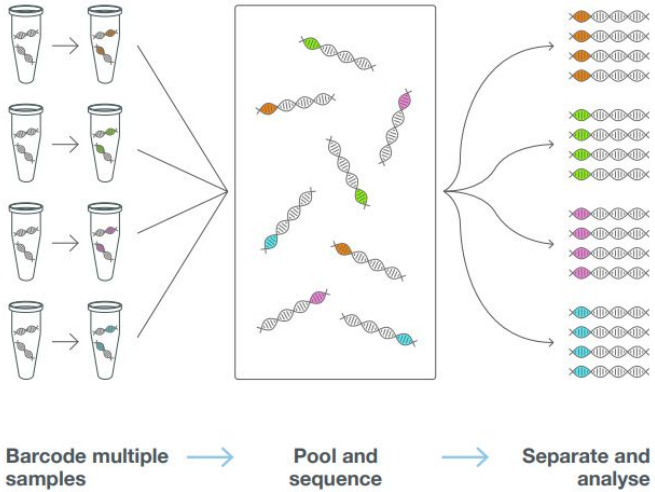
TP1a. Basecalling

https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2022/TP1.Basecalling_QC.ipynb

ONT demultiplexing

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

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



ONT Read calling, cleaning and filtering

Sequencer ONT : raw fast5 files

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

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

FASTQ FORMAT

```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTTCTTAGTATTTTTGTAGTCATTCCGTGTTGGTTTAGTTGCAAGGT
+
@@@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
CATGTGCTATTACCAAAAGTGCAGTAACGACCTATAAATTTTAAAGTAGC
+
@CFFFFFGGHHHHIJJJIEE<HHHIJJIGBHGGEEIIJJEIEIJIHHJFIIJJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTTCGGGTATGGCTCTAGTAGCTTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIEEHIII<CGHIJIJIIJ:?FC9DGAFGHII?DGBFIIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@DFFFFGHHHHJIIIIIJJJIIGJCHHEIII>GHIG@GHIDHGJIIFIFHIJJIJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGTAAGAGTTAAAAAACAATTAATGAGCAACTGAGTTC
+
@@CFFFFFHFFHFGIJJIHHIIJJIIIHIJJJECGHIJJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTTGTCGGGGGTTGGGAAATTGTCACTTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```

1 séquence = 4 lignes

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

Quality in reads, is it similar to illumina phred score ?

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

Score indicates probability **P** of a wrong base:

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

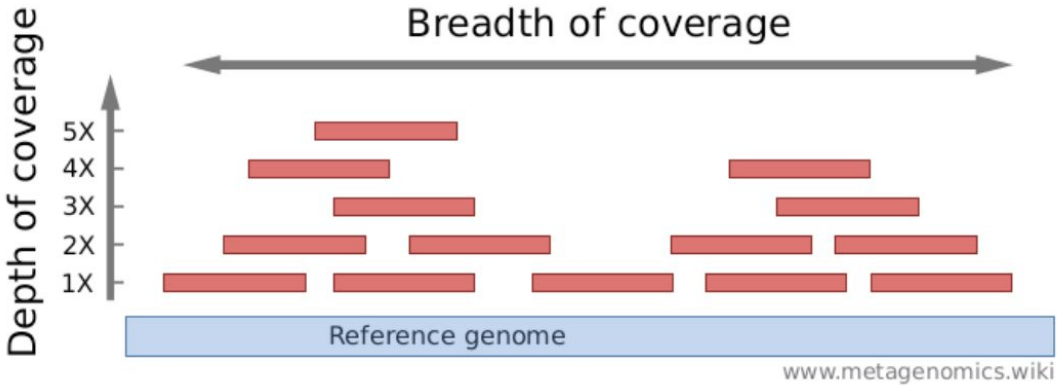
Phred score of 10 ↔ 10% error rate ; score of 20 ↔ 1% error rate

Nanopore quality score (Q) does not follow Phred scores
Yet enables to estimate error rate (E) (locally and at read level)

- HAC (High-Accuracy models) mode reduces error rate by 2%
- HAC mode basecalls homopolymers up to twice better than FAST (but also library R10 instead of R9)
- FAST mode is only about 2 times faster now

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

Calculate depth of coverage

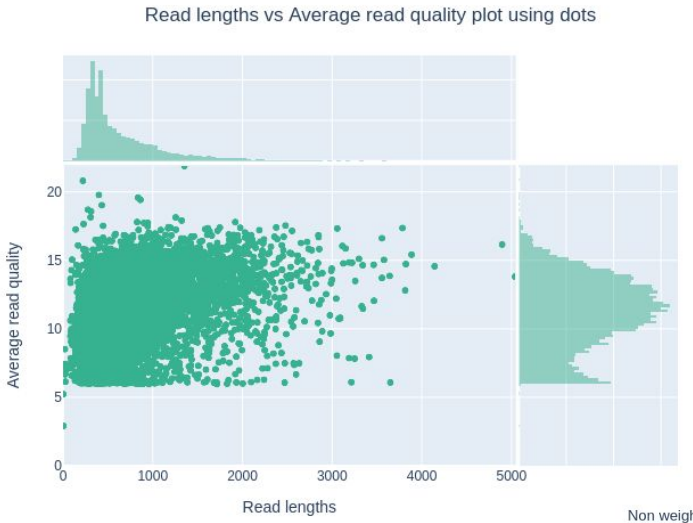


depth of coverage estimation :

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

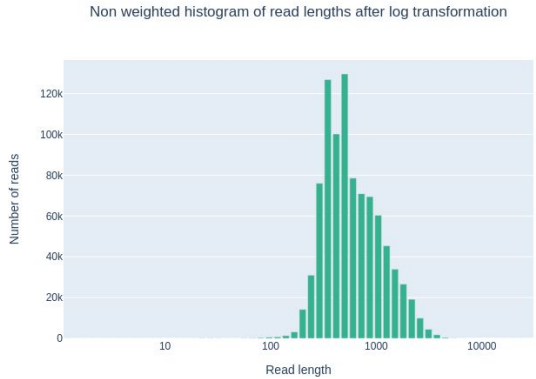
depth_of_coverage = total_pb/genome_size

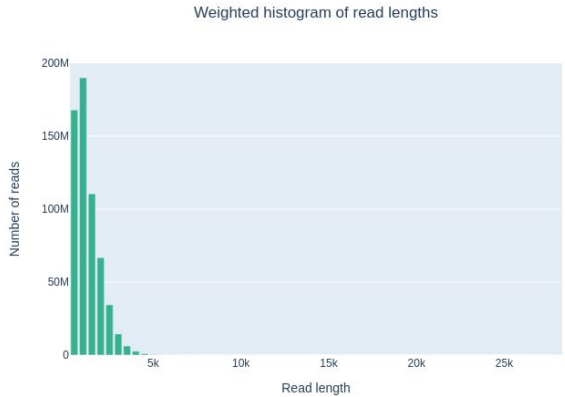
Reads Quality control : *NanoPlot*



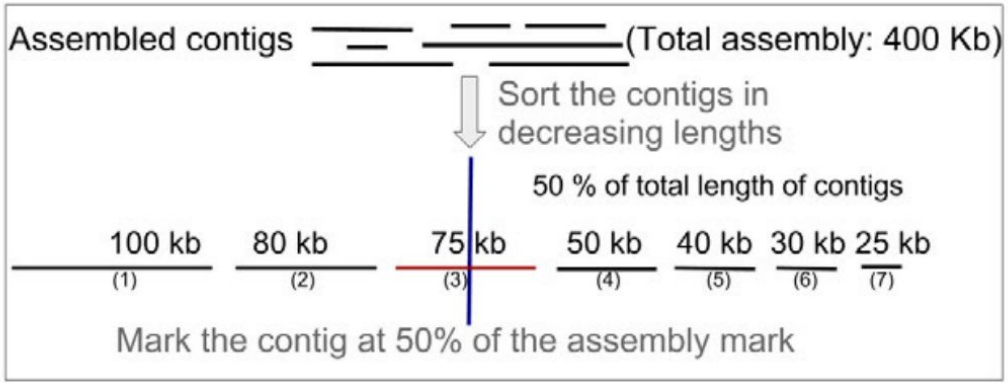
Summary statistics

General summary	
Mean read length	656.0
Mean read quality	11.2
Median read length	463.0
Median read quality	11.4
Number of reads	906,090.0
Read length N50	823.0
STDEV read length	488.3
Total bases	594,378,530.0





What is N50 and L50?



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

https://core.ac.uk/download/pdf/157813812.pdf

Reads Quality control

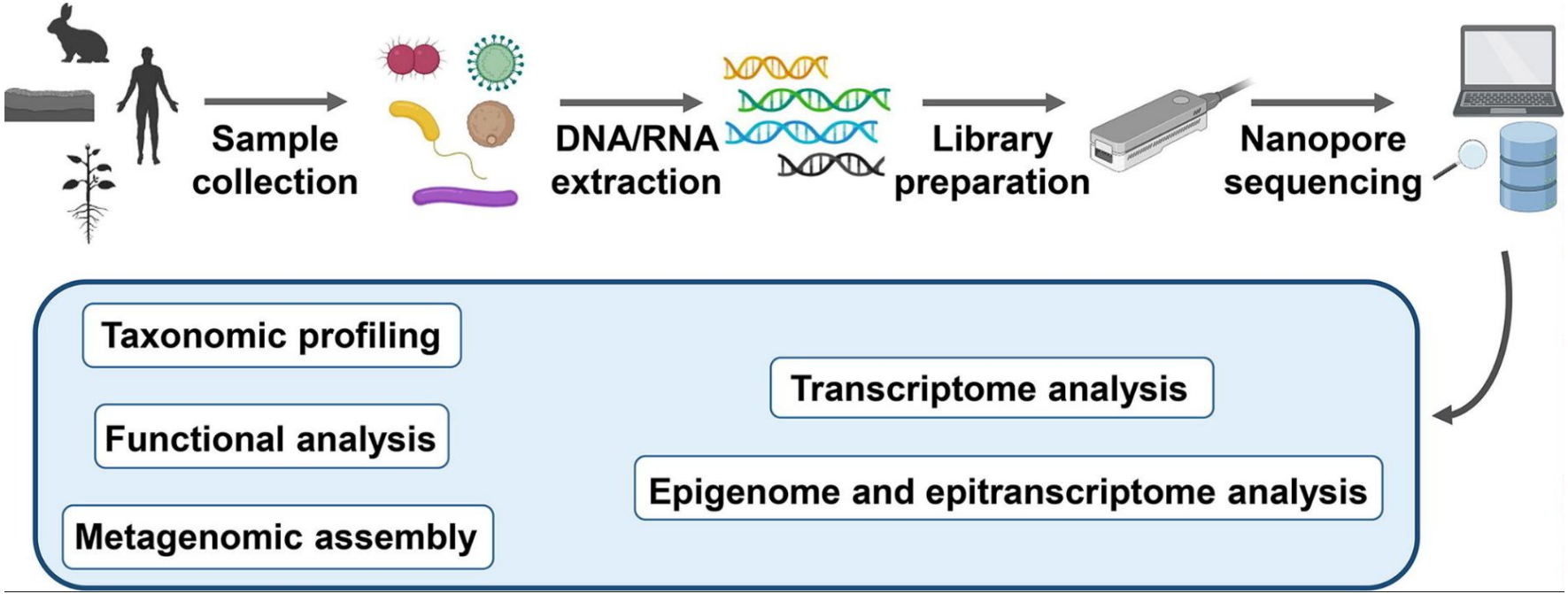
NanoPlot : https://github.com/wdecoster/NanoPlot
NanoComp : https://github.com/wdecoster/nanocomp
mini_qc : https://github.com/roblanf/minion_qc

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

TP1b. Quality Control

https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2022/TP1.Basecalling_QC.ipynb

What do you want to do with these long reads?



https://doi.org/10.1016/j.csbj.2021.02.020

Problem

We are interested in the composition of the metavirome of the pineapple.

1) What are the viruses present in our dataset?

Chapitre 2

- mapping for cleaning and searching
- affiliation taxonomique

2) Can we identify and assemble new viruses in this metavirome?

Chapitre 3

- Assemblies

Chapitre 2

1. Cleaning by Mapping

Remove unnecessary reads from the dataset

Sequencing with cDNA-PCR Barcoding kit => All RNA with a poly(A) tail sequenced:

- Host (Pineapple)
- Viruses
- Fungi
- Bacteriae
- Other Eukaryote (human…)

Taxonomic assignation and **assembly of long reads** are long processes which need a lot of resources.

→ we need to remove the maximum of unnecessary reads

Prepare the “contamination” library

RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.
https://ftp.ncbi.nlm.nih.gov/refseq/release/

Index of /refseq/release

Name	Last modified	Size
Parent Directory		-
announcements/	2022-07-14 10:59	-
archaea/	2022-07-14 17:27	-
bacteria/	2022-07-14 16:16	-
complete/	2022-07-15 03:52	-
fungi/	2022-07-15 04:11	-
invertebrate/	2022-07-14 18:08	-
mitochondrion/	2022-07-14 16:19	-
other/	2022-07-15 03:55	-
plant/	2022-07-14 11:24	-
plasmid/	2022-07-14 11:28	-
plastid/	2022-07-15 03:58	-
protozoa/	2022-07-14 17:33	-
release-catalog/	2022-07-15 04:12	-
release-error-notice/	2022-07-14 10:58	-
release-notes/	2022-07-15 09:04	-
release-statistics/	2022-07-15 04:12	-
vertebrate_mammalian/	2022-07-14 17:23	-
vertebrate_other/	2022-07-14 12:22	-
viral/	2022-07-14 16:19	-
README	2022-07-15 04:12	4.6K
RELEASE_NUMBER	2022-07-15 04:12	4

fungi.1.1.genomic.fna.gz	2022-07-15 04:03	31M
fungi.1.genomic.gbff.gz	2022-07-15 04:03	13M
fungi.1.protein.faa.gz	2022-07-15 04:03	10M
fungi.1.protein.gpff.gz	2022-07-15 04:03	22M
fungi.1.rna.fna.gz	2022-07-15 04:03	17M
fungi.1.rna.gbff.gz	2022-07-15 04:03	48M

wget pour télécharger:

wget ftp://ftp.ncbi.nlm.nih.gov/refseq/release/fungi/fungi.1.protein.faa.gz

Mapping des reads sur la library

K. Sahlin et al. 2022

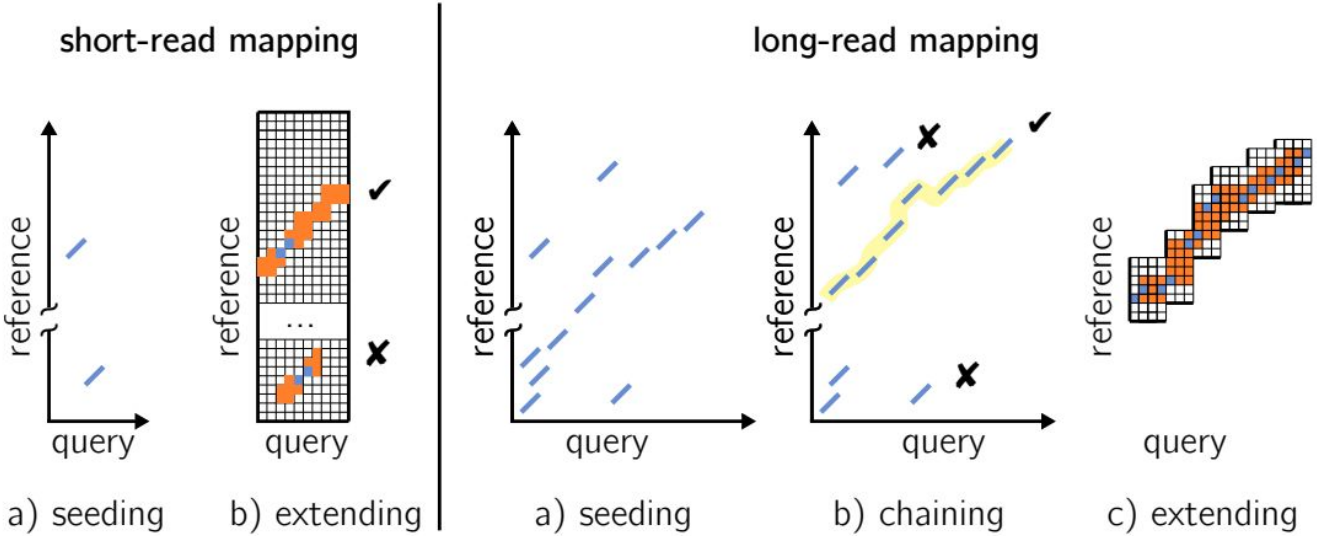


Figure 1 Differences in the main steps between short-read mapping (left) and long-read mapping (right). *Query* denotes the read and *reference* denotes a genome region. Mainly, short-read approaches extend (orange parts) from a single anchor (in blue) on the whole read length while long-read approaches gather multiple anchors, and chain (yellow line) them in for a candidate extending procedure that is done between pairs of anchors.

BWA

minimap2

SAM format

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *    NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC
r001  83 ref 37 30 9M         =  7 -39 CAGCGCC
```

Header

- Ligne commençant par @
- Metadonnees sous forme de tag

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
SQ - Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG - read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
PG - Program	ID*	Program name
	VN	Program version
	CL	Command line
CO - comment		One-line text comments

SAM format

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

alignement

Format tabulé

SAM format : http://samtools.sourceforge.net/samtools.shtml

SAM format

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *    NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAG
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT
```

alignement

Format tabulé

SAM format : http://samtools.source

Col	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string (operations: MIDNSHP)
7	NRNM	Mate Reference NaMe (`=' if same as RNAME)
8	MPOS	1-based leftmost Mate POSition
9	ISIZE	inferred Insert SIZE
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33=Phred base quality)

TP2. Cleaning

https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2022/TP2.Cleaning_data.ipynb

Chapitre 3

Taxonomic affiliation

Taxonomic Assignation

Taxonomic assignment is the process of assigning an Operational Taxonomic Unit (OTUs) to sequences, that can be reads or contigs.

To assign an OTU to a sequence it is compared against a database.

There are many programs for doing taxonomic mapping, we will see 2 strategies:

1. **BLAST or DIAMOND**, these mappers search for the most likely hit for each sequence within a database of genomes.

2. **K-mers (KRAKEN)**: The algorithm breaks the query sequence (reads, contigs) into pieces of length k, look for where these are placed within the tree and make the classification with the most probable position.

Pairwise alignment: BLAST / DIAMOND

If you have two or more sequences, you may want to know :
– How similar are they?
– Which residues correspond to each other?
– Is there a pattern to the conservation/variability of the sequences?
– What are the evolutionary relationships of these sequences?

```
Human: ccatcctcagatccgtcttcagaaccaccttcccctcgatccacggctccattttcatcc
       ||||||||||| |||||||||| || |||||   ||| ||| |||| ||| | || |||
Mouse: ccatcctcagacccgtcttcagagccccttc---tcggtccccggccccactgtcttcc
```

BLAST and DIAMOND Compares a QUERY sequence to a DATABASE of sequences

- **Blast**
  - Nucleotide or protein sequences
  - Available as an online web server: http://blast.ncbi.nlm.nih.gov/Blast.cgi

- **Diamond**:
  - For protein and translated DNA searches only
  - High performance analysis of big sequence data
  - Frameshift alignments for long read analysis.
  - 500x-20,000x speed of BLAST

diamond better in our case

What about databases ?

It is important to choose your database wisely.
The database you use will determine the result you get for your data.

Pfam: http://pfam.xfam.org/
db link: ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.fasta.gz
Swiss prot: https://web.expasy.org/docs/swiss-prot_guideline.html
db link: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz
UniProt90: https://www.uniprot.org/help/uniref
db link: ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref90/uniref90.fasta.gz
nr: ftp://ftp.ncbi.nlm.nih.gov/blast//db/FASTA/nr.gz

Diamond, as Blast needs a step of formatage for the database

Diamond output

stitle	qtitle	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore
YP_009664796.1 heat shock protein 70 [Pineapple mealybug wilt-associated virus 2]	c2615778-aa7c-4906-8c53-75cd9fa196f9	95.7	94	4	0	214	495	405	498	3.90e-53	180
NP_813799.1 59 kDa protein [Grapevine leafroll-associated virus 3]	c2615778-aa7c-4906-8c53-75cd9fa196f9	50.6	79	26	2	72	302	358	425	9.91e-15	74.7
YP_008411013.1 heat shock protein 70-like protein [Blackberry vein banding-associated virus]	c2615778-aa7c-4906-8c53-75cd9fa196f9	40.9	93	55	0	217	495	406	498	8.34e-14	72.0
YP_010086802.1 Hsp70 [Pistachio ampelovirus A]	c2615778-aa7c-4906-8c53-75cd9fa196f9	48.4	64	33	0	18	209	338	401	1.75e-12	68.2
YP_004940644.1 HSP70 gene product [Grapevine leafroll-associated virus 1]	c2615778-aa7c-4906-8c53-75cd9fa196f9	49.2	59	30	0	36	212	348	406	1.68e-10	62.4

stitle means Subject Title
qtitle means Query title
pident means Percentage of identical matches
length means Alignment length
mismatch means Number of mismatches
gapopen means Number of gap openings
qstart means Start of alignment in query
qend means End of alignment in query
sstart means Start of alignment in subject
send means End of alignment in subject
evalue means Expect value
bitscore means Bit score

KRAKEN2

Kraken2 is a taxonomic classification system using exact k-mer matches to achieve high accuracy and fast classification speeds.

Like with diamond, you need to choose a database:

- **Minikraken**, a database pre-made by KRAKEN, is a popular database that attempts to conserve its sensitivity despite its small size (Needs 8GB of RAM for the assignment).

Kraken output:

C	799ec77c-6555-4b9f-99a3-e58c9fbc1265	1491	335	0:217 1491:4 0:21 1491:2 0:6 1491:5 0:16 1491:4 2:5 0:20 9606:1
U	37c0c305-d935-4b3b-b336-24b4c4c8021d	0	332	0:56 9606:2 0:240
U	02df7f95-9bbe-4b55-9c8b-78955e3d9210	0	208	0:174
U	b86266e6-4b84-4ed6-abde-302e336f6c24	0	429	0:53 9606:5 0:337
U	b7f946d2-7f1d-492c-b187-3ebc0770a15c	0	292	0:222 131567:2 0:34
U	c2615778-aa7c-4906-8c53-75cd9fa196f9	0	605	0:571
C	b2388cec-c33d-4a6b-948f-4cb151194e5f	1491	417	0:42 9606:1 0:230 1491:1 0:7 1491:2 0:14 1491:5 0:39 1491:1 1239:3 0:38

As we can see, the kraken file is not very readable. So let's look at the report file:

KRAKEN2

Kraken output:

report.txt

0.30	1204	0	7902	432	D	10239	Viruses
0.24	981	0	7065	370	D1	439488	ssRNA viruses
0.24	981	0	7063	369	D2	35278	ssRNA positive-strand viruses
0.24	981	0	7063	369	F	69973	Closteroviridae
0.24	981	0	7063	369	G	217160	Ampelovirus
0.24	981	981	7063	369	S	180903	Pineapple mealybug wilt-associated virus 1
0.05	220	0	806	39	O	2169561	Ortervirales
0.05	220	0	806	39	F	186534	Caulimoviridae
0.05	220	0	806	39	G	10652	Badnavirus
0.05	220	220	800	37	S	2033633	Pineapple bacilliform CO virus

1. Percentage of reads covered by the clade rooted at this taxon
2. Number of reads covered by the clade rooted at this taxon
3. Number of reads assigned directly to this taxon
4. A rank code, indicating (U)nclassified, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies. All other ranks are simply '-'.
5. NCBI taxonomy ID
6. Indented scientific name

KAIJU

Kaiju is a program for sensitive **taxonomic classification** of high-throughput sequencing reads from **metagenomic whole genome sequencing or metatranscriptomics** experiments.

https://kaiju.binf.ku.dk/

Kaiju output:

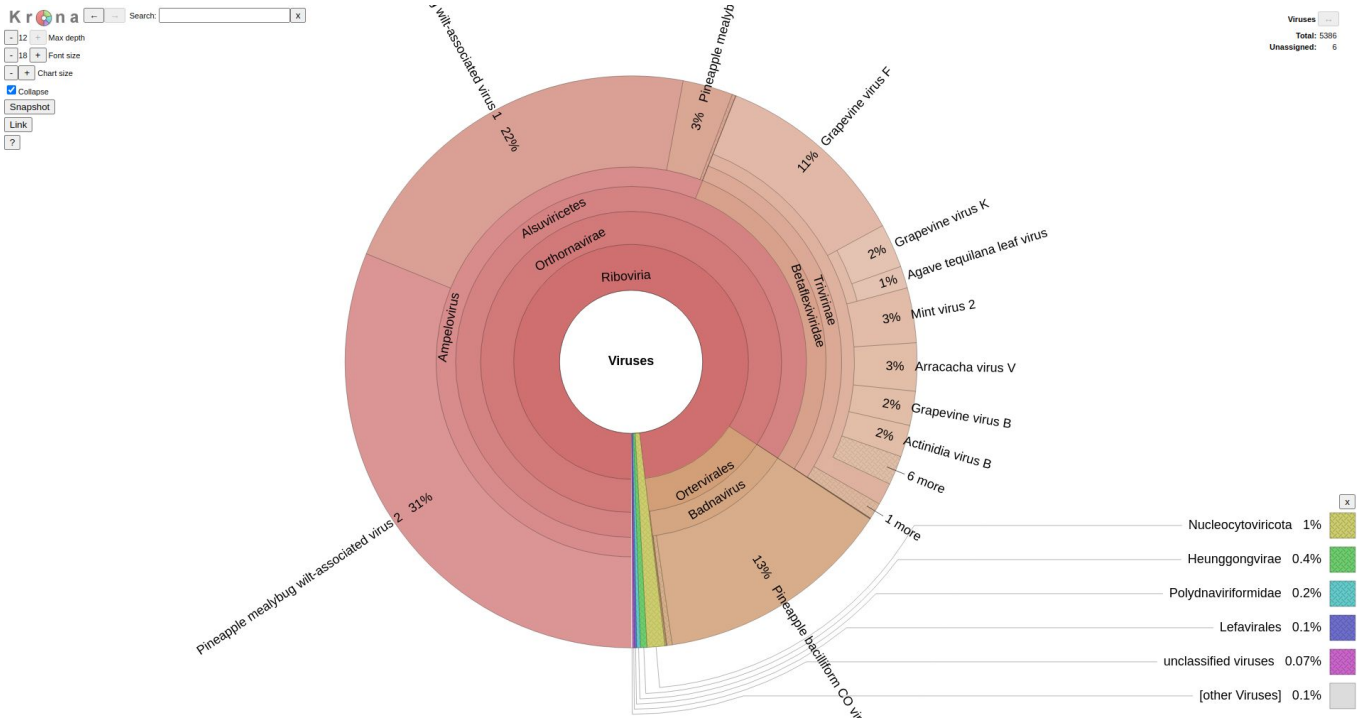
Classified or Unclassified	name of the read	NCBI taxon identifier	score of the best match	taxon identifiers of all database sequences with the best match	accession numbers of all database sequences with the best match	matching fragment sequence	Taxon name
U	b7f946d2-7f1d-492c-b187-3ebc0770a15c	0					
C	2615778-aa7c-4906-8c53-75cd9fa196f9	136234	440	136234	YP_009664796.1	RTITFNTGGRKTMYGVYEGEEVRSYLNALTFRGEYISNVEGNRTDSATFSVSSDGILSVSVNGTLLKNDLVPSPPTVFSKNLEYLSNIEK	Pineapple mealybug wilt-associated virus 2

Visualization of taxonomic assignment results

KRONA:

You can transform KRAKEN2 and KAIJU output to visualize them with KRONA.

Krona is a hierarchical data visualization software. Krona allows data to be explored with zooming, multi-layered pie charts and includes support for several bioinformatics tools and raw data formats.



Visualization of taxonomic assignment results

PAVIAN:

Comparison of the composition of several sample.
https://fbreitwieser.shinyapps.io/pavian/

https://carpentries-incubator.github.io/metagenomics/06-taxonomic/index.html

TP3. taxonomic assignation

https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2022/TP3.Assignation_Taxonomique.ipynb

Chapitre 4.
Assemblies

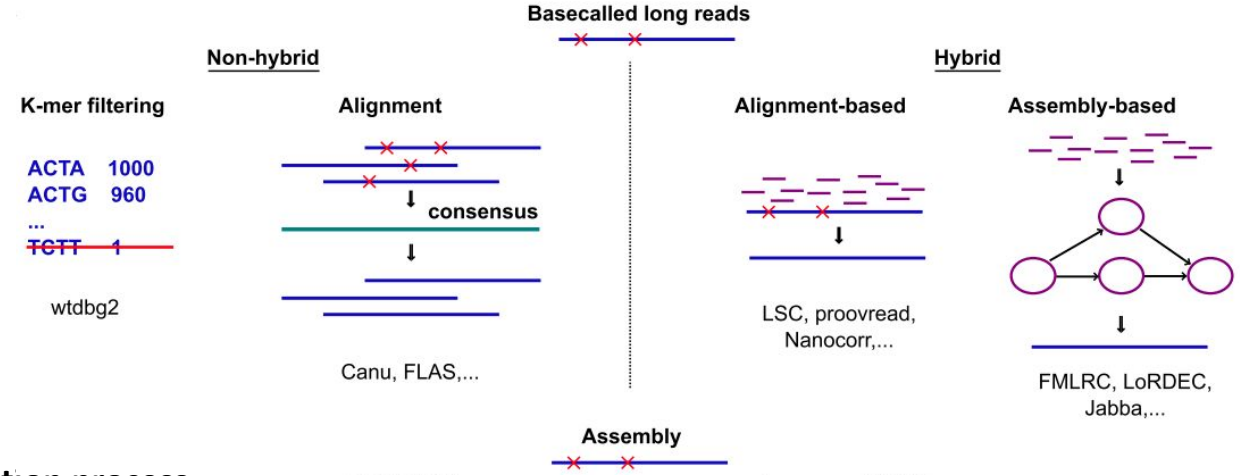
Type	Reference	Application
Aligners/Alignment-based classifiers		
BLAST, MEGABLAST	[58,59]	Targeted; Shotgun
minimap2	[33]	Targeted; Shotgun
Alignment-free classifiers		
Kraken, Kraken2	[35,64]	Targeted; Shotgun
KrakenUniq	[65]	Shotgun
Bracken	[66]	Targeted; Shotgun
Metamaps	[69]	Shotgun
Centrifuge	[34]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

https://doi.org/10.1016/j.csbj.2021.02.020

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

https://doi.org/10.1016/j.csbj.2021.02.020

Reads Correction or not?



Reads Correction process

Correction strategies (*hybrid*)

– External reads : Illumina

– Internal reads : Only long reads or long reads corrected by short ones

Correction pipeline (*non-hybrid*)

– Read alignment

– Consensus calling

Canu module,

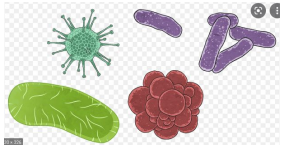
Racon can also be used as a read error-correction tool.

Assembly without reads correction

- Miniasm, Smartdenovo, Flye are members of this “new” family
- Improves speed
- Can work with less read depth.
- Can also assemble corrected reads

https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-020-1935-5.pdf

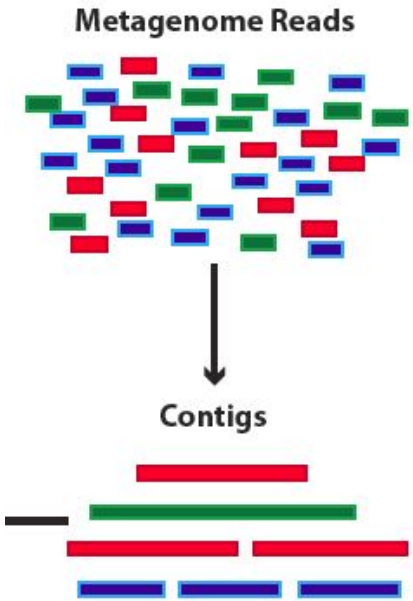
What assembler to use over my favorite organism?



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

assemblage of metagenome

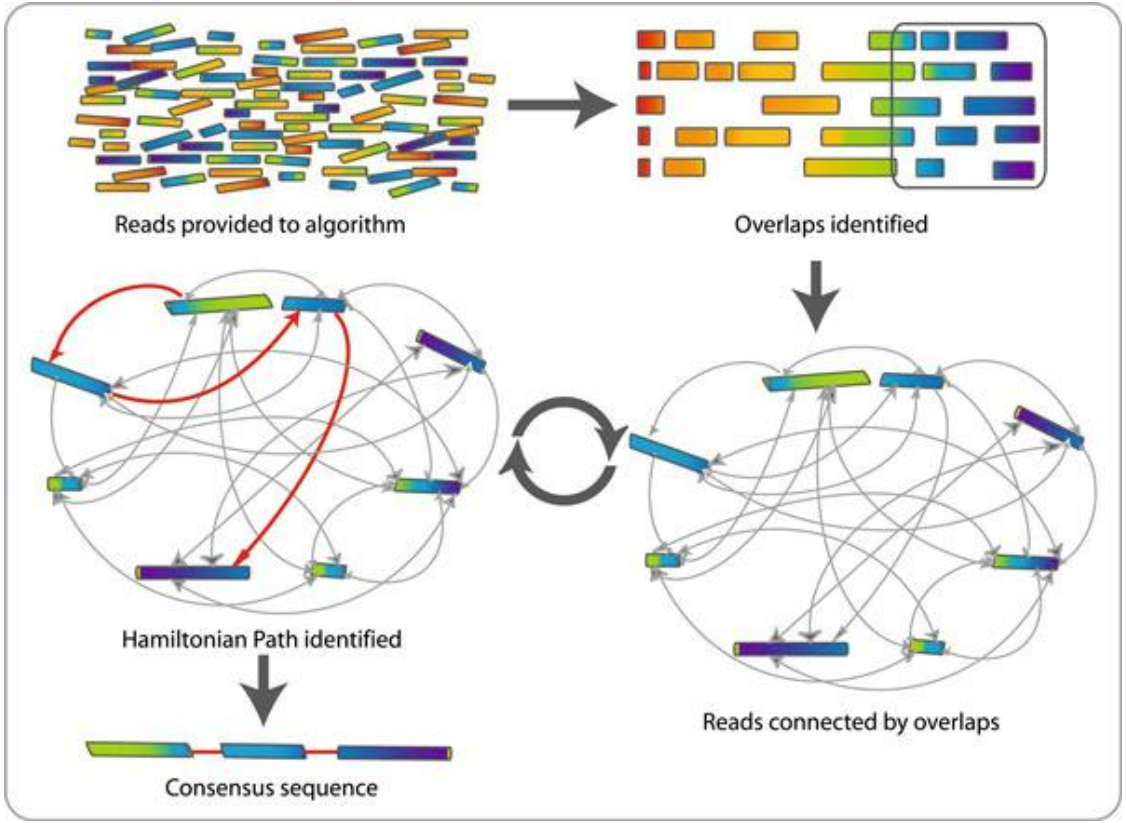


→ alignment and fusion of reads in longer fragments (contigs)

→ objective :

- de novo = reconstruction of new viruses not in database

Overlap–layout–consensus genome assembly algorithm (OLC)



Canu, Flye, Miniasm, Raven, Smartdenovo, Shasta

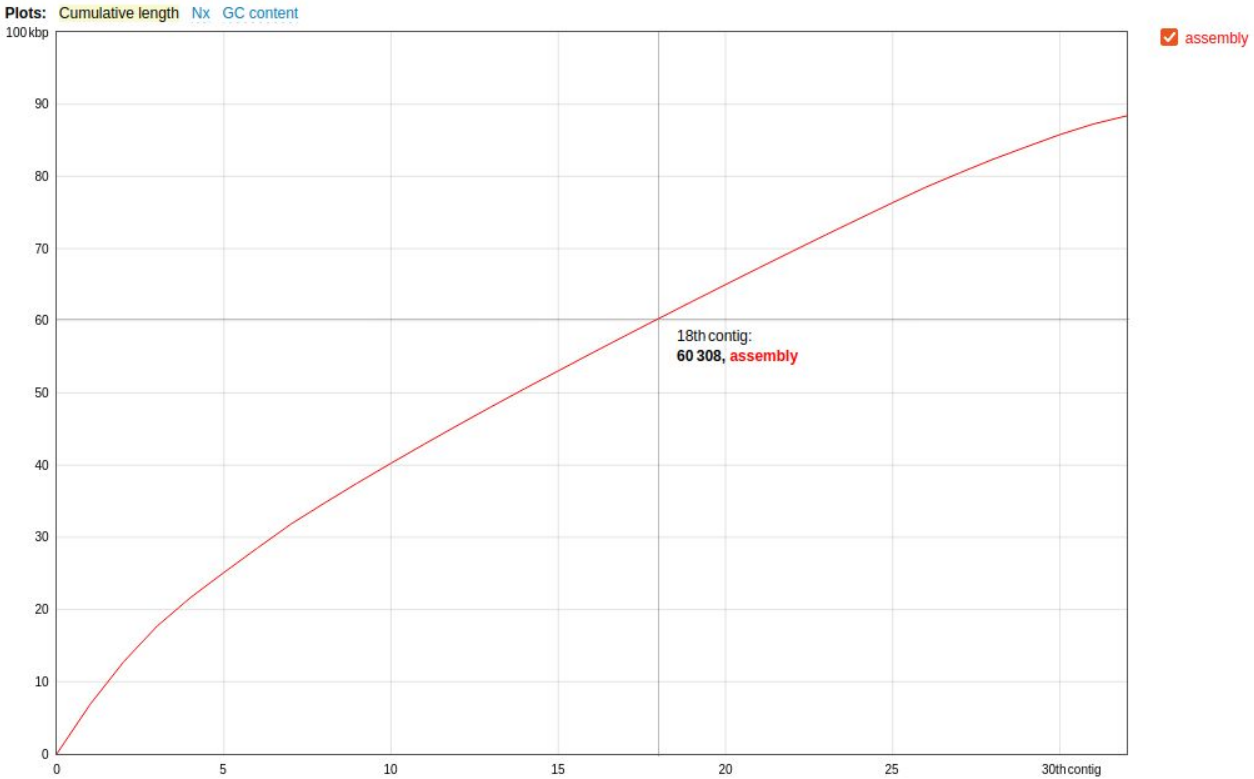
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3055744/

checking assemblies quality

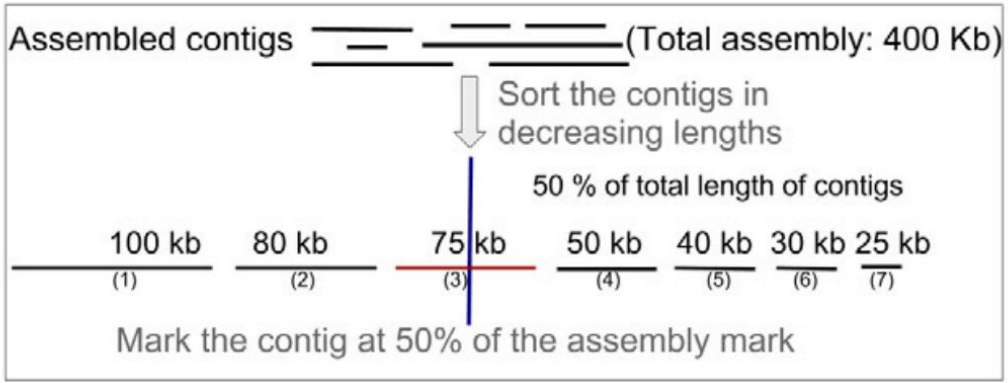
QUAST

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

Statistics without reference	assembly
# contigs	32
# contigs (>= 0 bp)	32
# contigs (>= 1000 bp)	32
# contigs (>= 5000 bp)	2
# contigs (>= 10000 bp)	0
# contigs (>= 25000 bp)	0
# contigs (>= 50000 bp)	0
Largest contig	6877
Total length	88 391
Total length (>= 0 bp)	88 391
Total length (>= 1000 bp)	88 391
Total length (>= 5000 bp)	12 760
Total length (>= 10000 bp)	0
Total length (>= 25000 bp)	0
Total length (>= 50000 bp)	0
N50	2612
N90	1952
auN	3266
L50	12
L90	27
GC (%)	44.93
Mismatches	
# N's per 100 kbp	0
# N's	0



What is N50 and L50?



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

https://core.ac.uk/download/pdf/157813812.pdf

Polishing / Correction

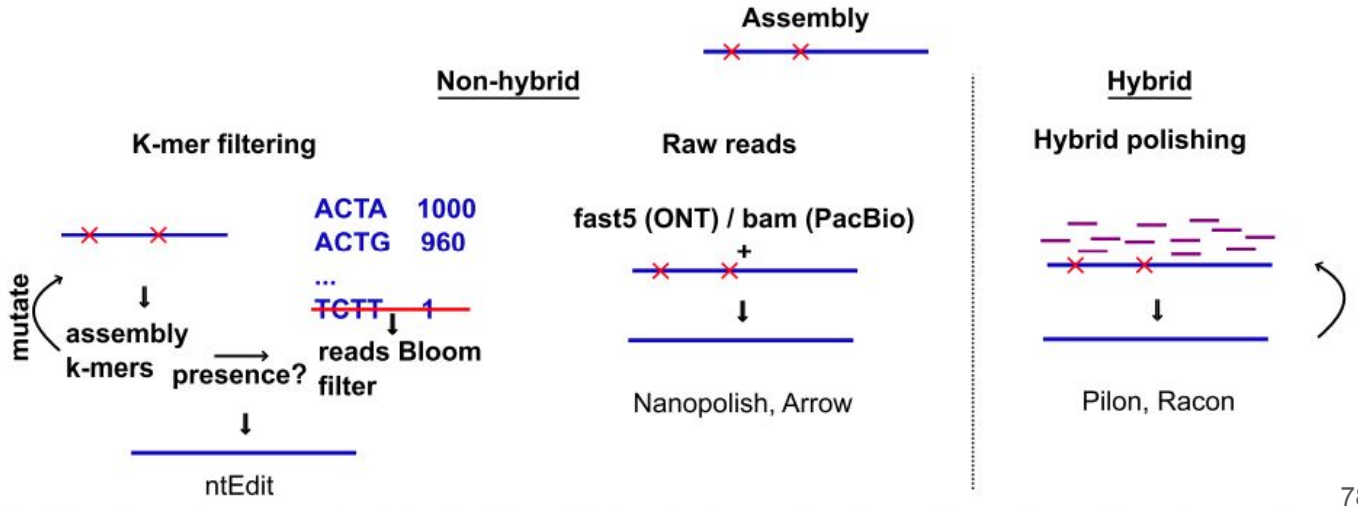
Racon correct raw contigs generated by rapid assembly methods which do not include a consensus step. It can polish with either Illumina data or data produced by third generation of sequencing. (recursive use)

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

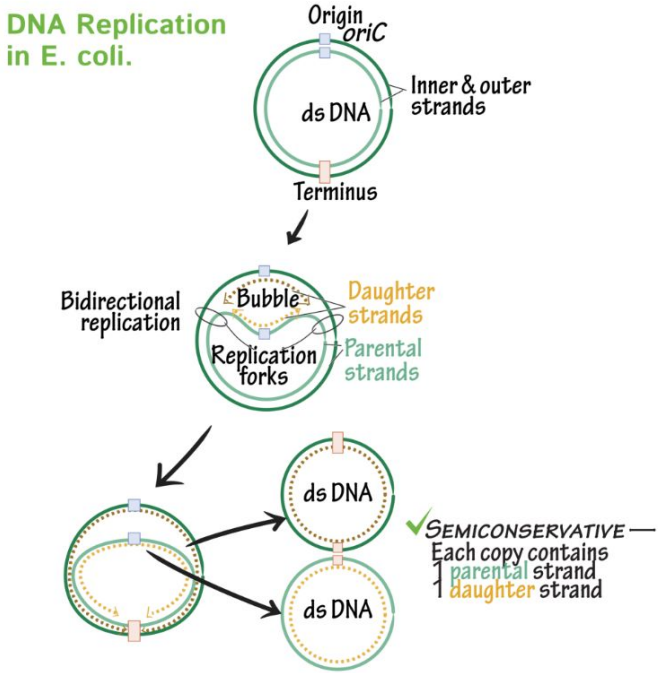
Pilon correct assemblies using illumina reads. (recursive use)

Autres : NeuralPolish , ntEdit



https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-020-1935-5.pdf

Circularisation ?



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.

https://www.drawittoknowit.com/course/immunology/glossary/immunology-microbiology/bacterial-chromosome-replication-e-coli

TP4. assemblies

https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2022/TP4.DeNovoAssembly.ipynb

Formateurs

Aurore Comte IRD

Julie Orjuela IRD

Denis Filloux CIRAD

formateurs de la session 2021 : Francois Sabot, Gautier Sarah et Julie Orjuela

Remerciements à Christine Tranchant pour le matériel pédagogique jupyter notebook.

Merci pour votre attention !

En informatique,
la pensée magique ne fonctionne pas !

Il faut pratiquer … et … *restez calme !*

… à vous de jouer !