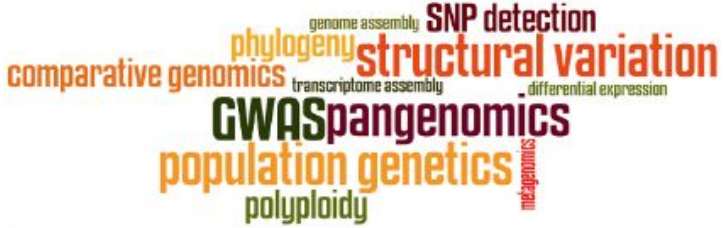
Initiation à l'analyse de données Oxford Nanopore/Assemblage

Alliance

South Green bioinformatics platform

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

Alliance Bioversity & CIAT

IRD Institut de Recherche pour le Développement FRANCE

cirad

INRAE

4 institutes

diade

PHIM

agAP

3 research units

25+

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Storage and computing resources

Tools

Trainings

400+

Collaborative development of tools

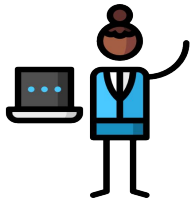
Genomics

- Pangenomic
- Gene families
- Comparative
- Phylogeny
- Assemblies
- Annotation
- Data mining

Diversity exploration

- genotype manipulation
- mosaic manipulation

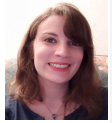
Metagenomic



+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 Dennu				

Bioinformatics resources

On va travailler sous Linux !

- 2 façons d'utiliser linux :

en ***mode graphique***

En mode terminal

- 2 façons d’utiliser linux :

  en ***mode console***

avant tout !

Bases de Linux

https://github.com/SouthGreenPlatform/training_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 ?

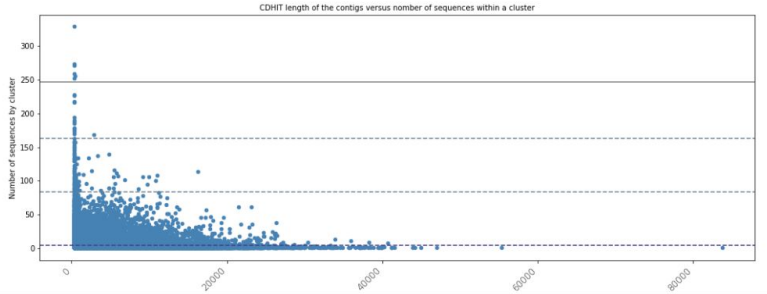
An unique interface/file where text,code and output codes can be mixed :

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

All ctgs

Entrée [30]:

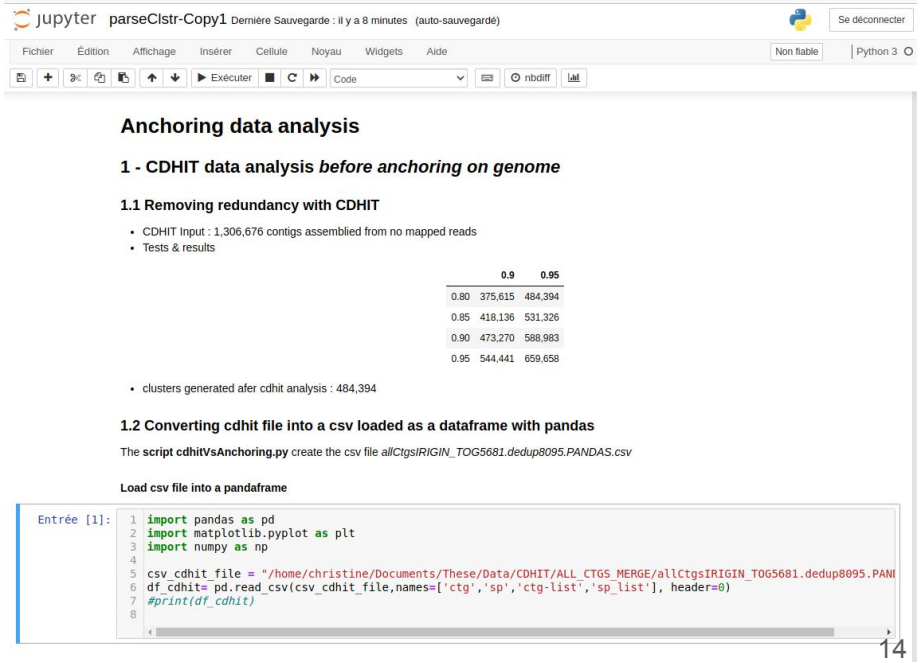
```
plt.figure(figsize=(17, 6))
ax = plt.gca()
df_cdhit[(df_cdhit.sp == 'Ob')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
df_cdhit[(df_cdhit.sp == 'Og')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
plt.axhline(y=5, color='darkslateblue', linestyle='--')
plt.axhline(y=84, color='slategrey', linestyle='--')
plt.axhline(y=163, color='slategrey', linestyle='--')
plt.axhline(y=247, color='grey', linestyle='-')

plt.title("CDHIT length of the contigs versus nomber of sequences within a cluster", fontsize=10)
plt.xlabel('Cluster length')
plt.ylabel('Number of sequences by cluster')
plt.xticks(
    rotation=45,
    horizontalalignment='right',
    fontweight='light',
    fontsize='12',
)
plt.show()
```



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

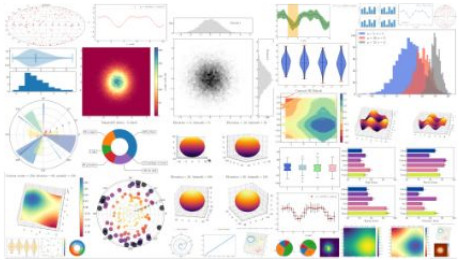


Lab notebook for science data ?

- 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 ?

- 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



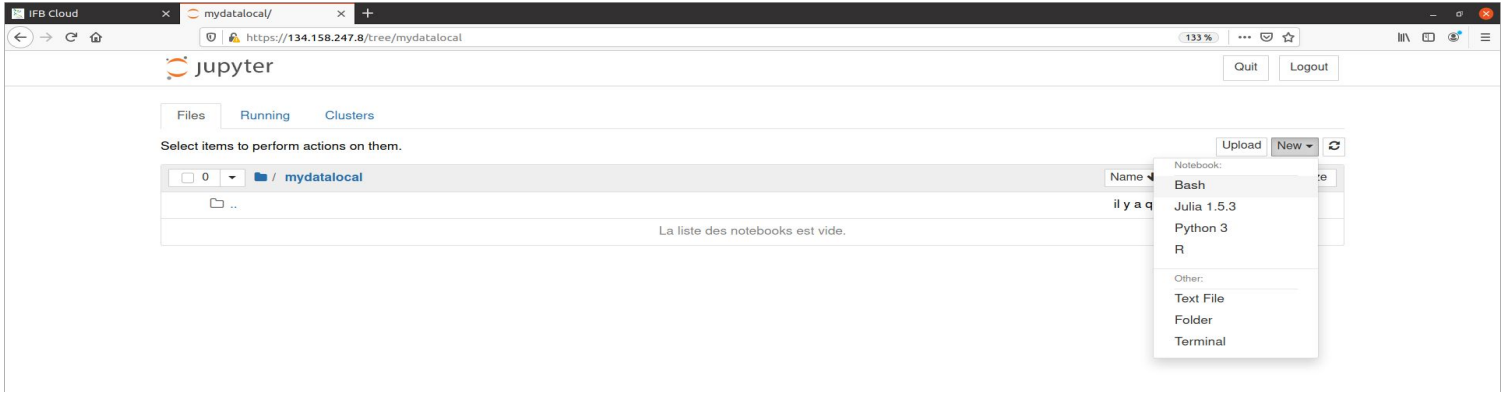
How will you use Jupyter Notebook ?

- 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”

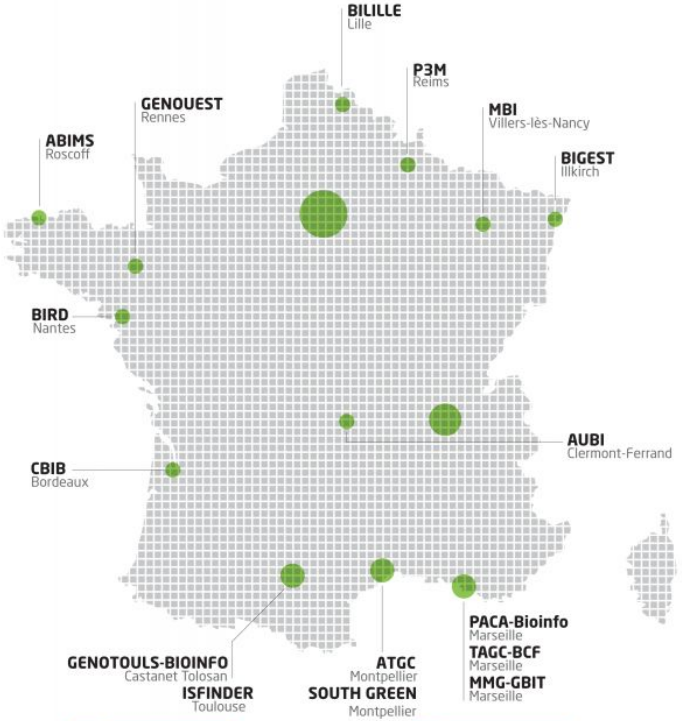
- Through this virtual machine, we will create jupyter books and execute all our analysis



IFB ?

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

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

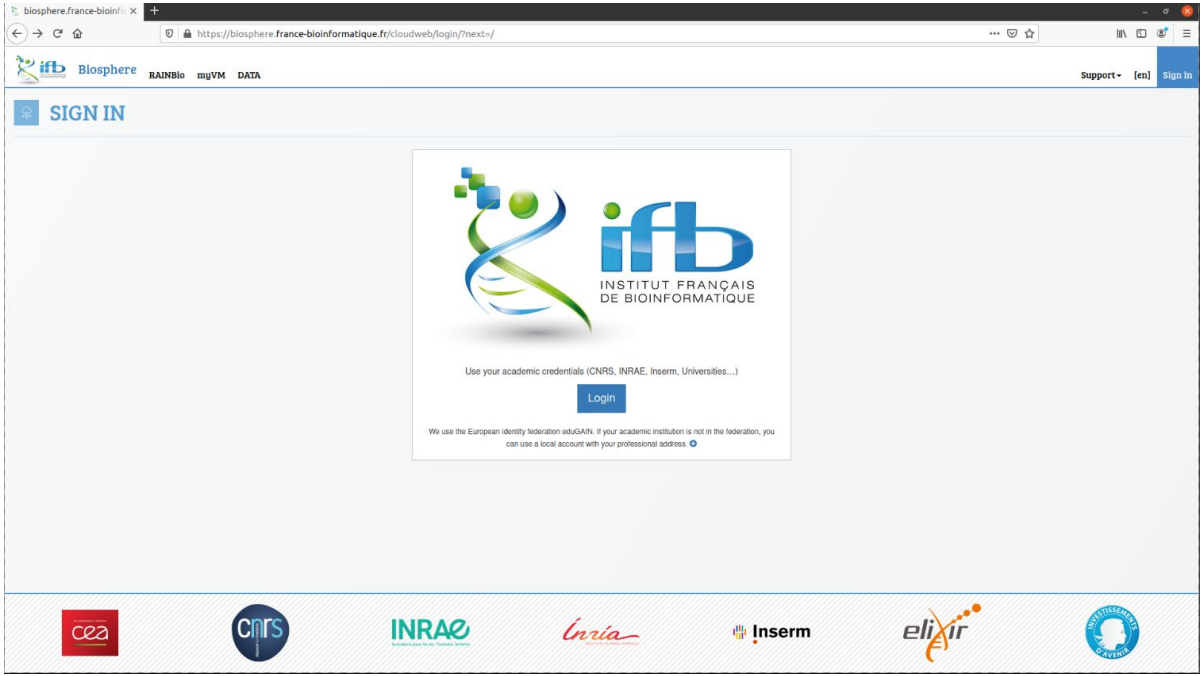


Biosphere, IFB CLOUD FOR LIFE SCIENCES

- 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 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.

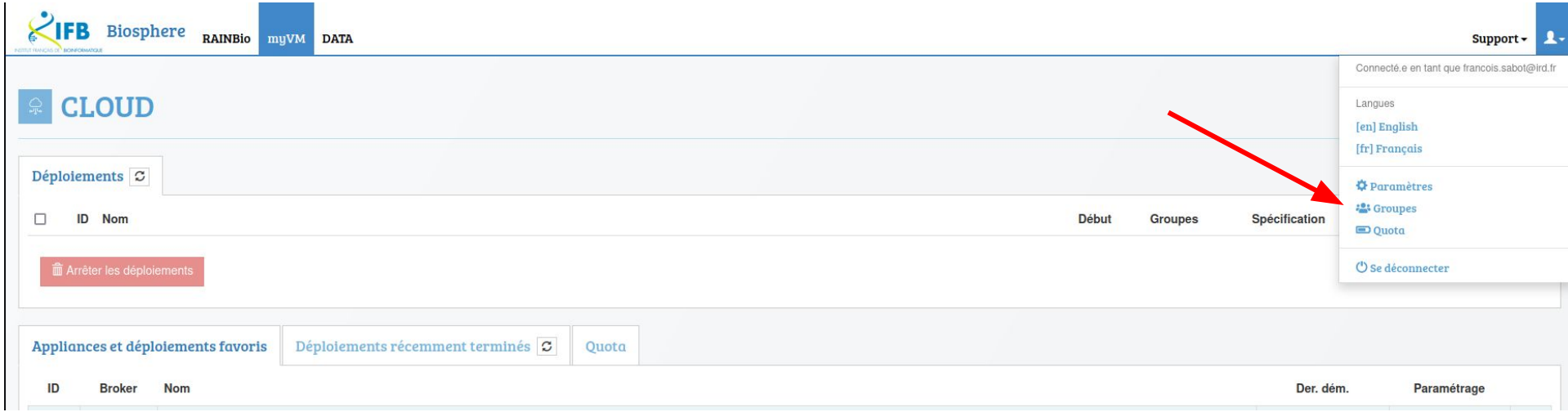
Let's start with biosphere

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



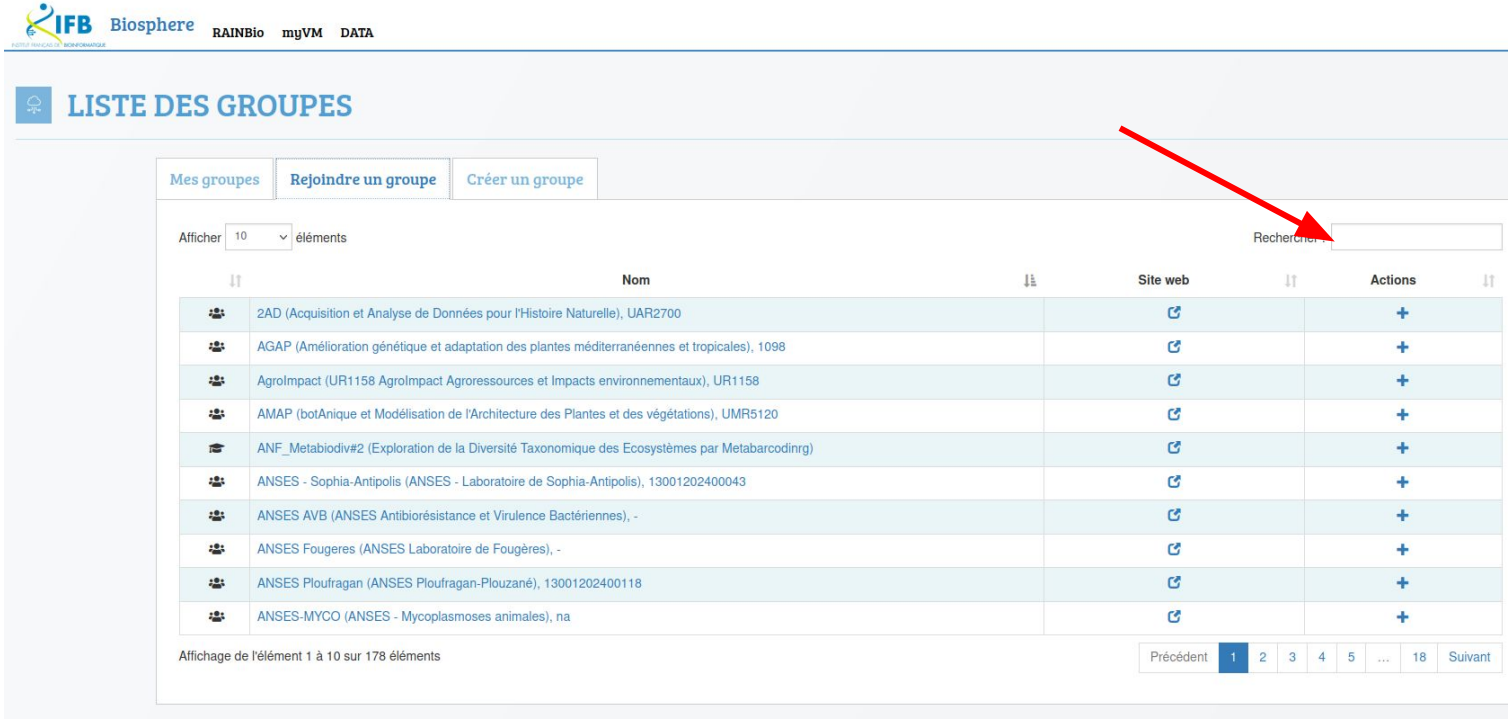
Let's start with biosphere

- Select a specific group



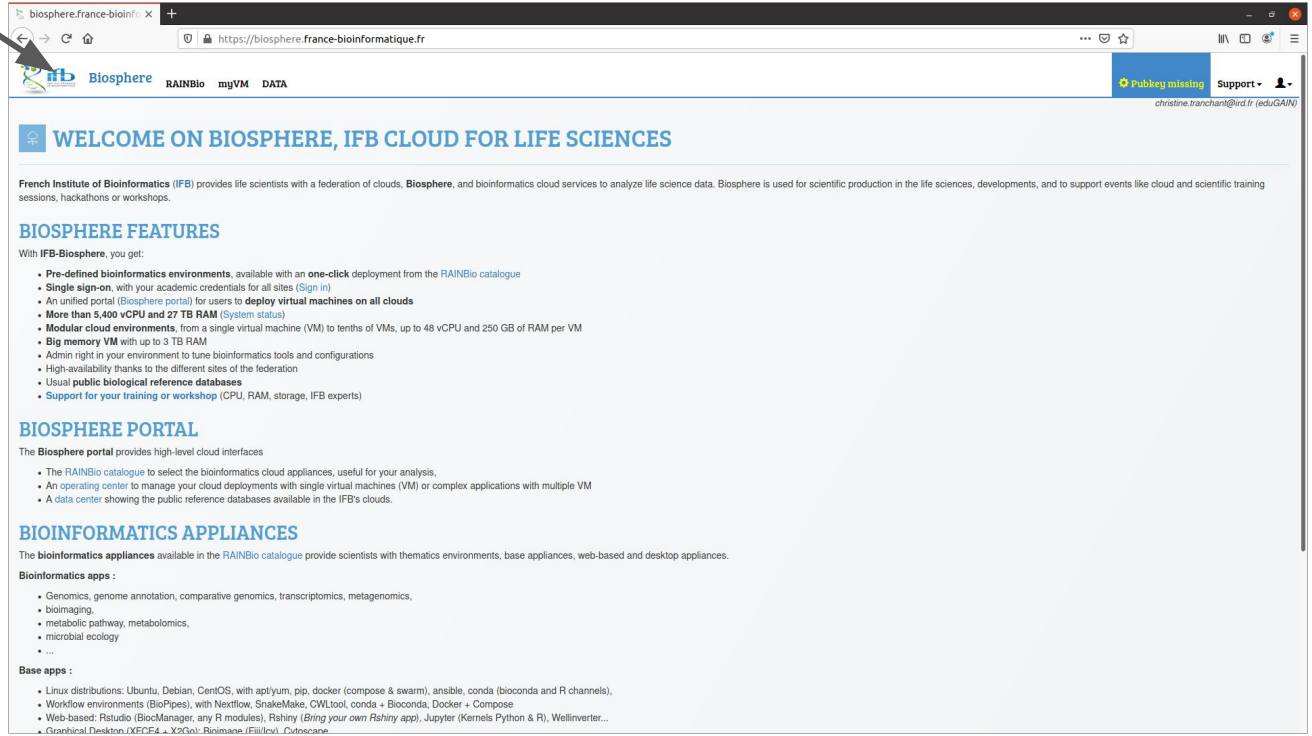
Let's start with biosphere

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



Connected / here we are

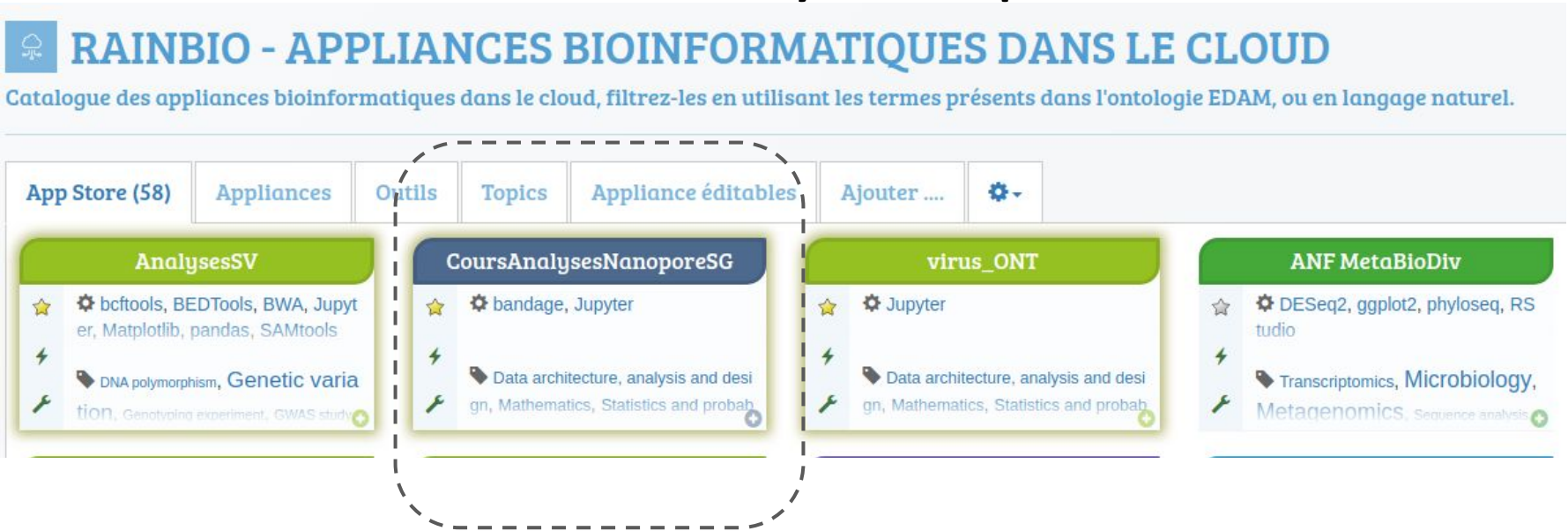
RAINBIO catalog to access our Virtual Machine (VM)



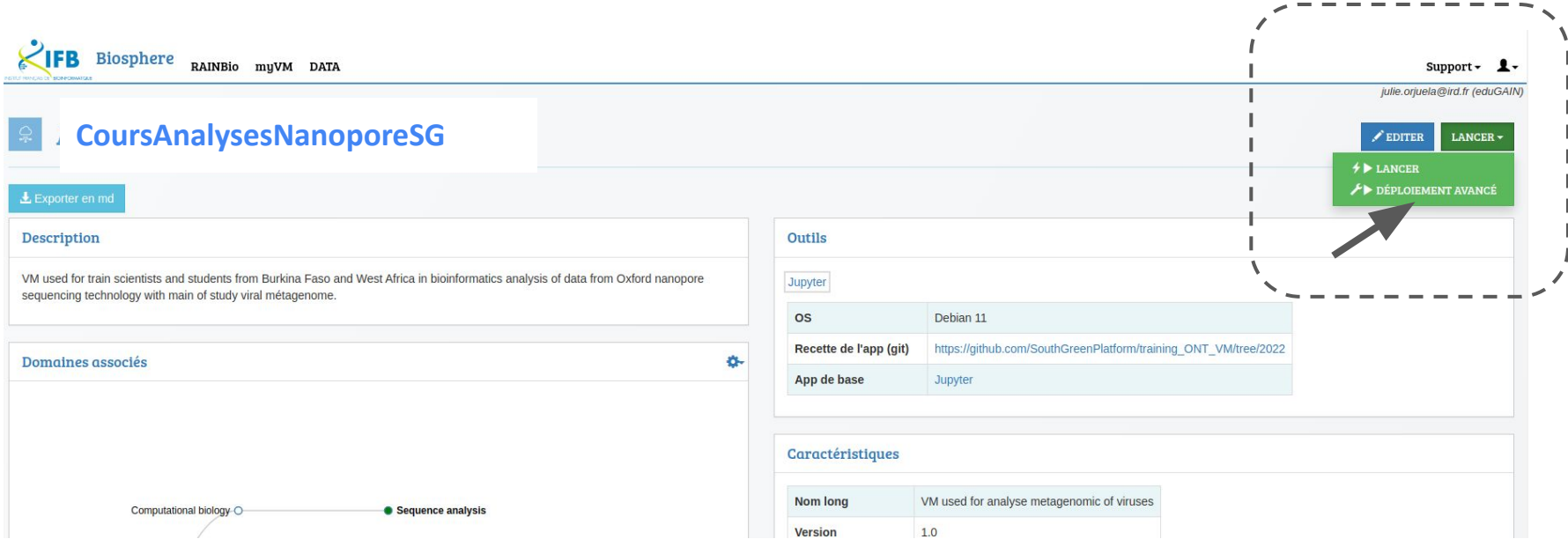
Searching for the vm we will use

vm's name :

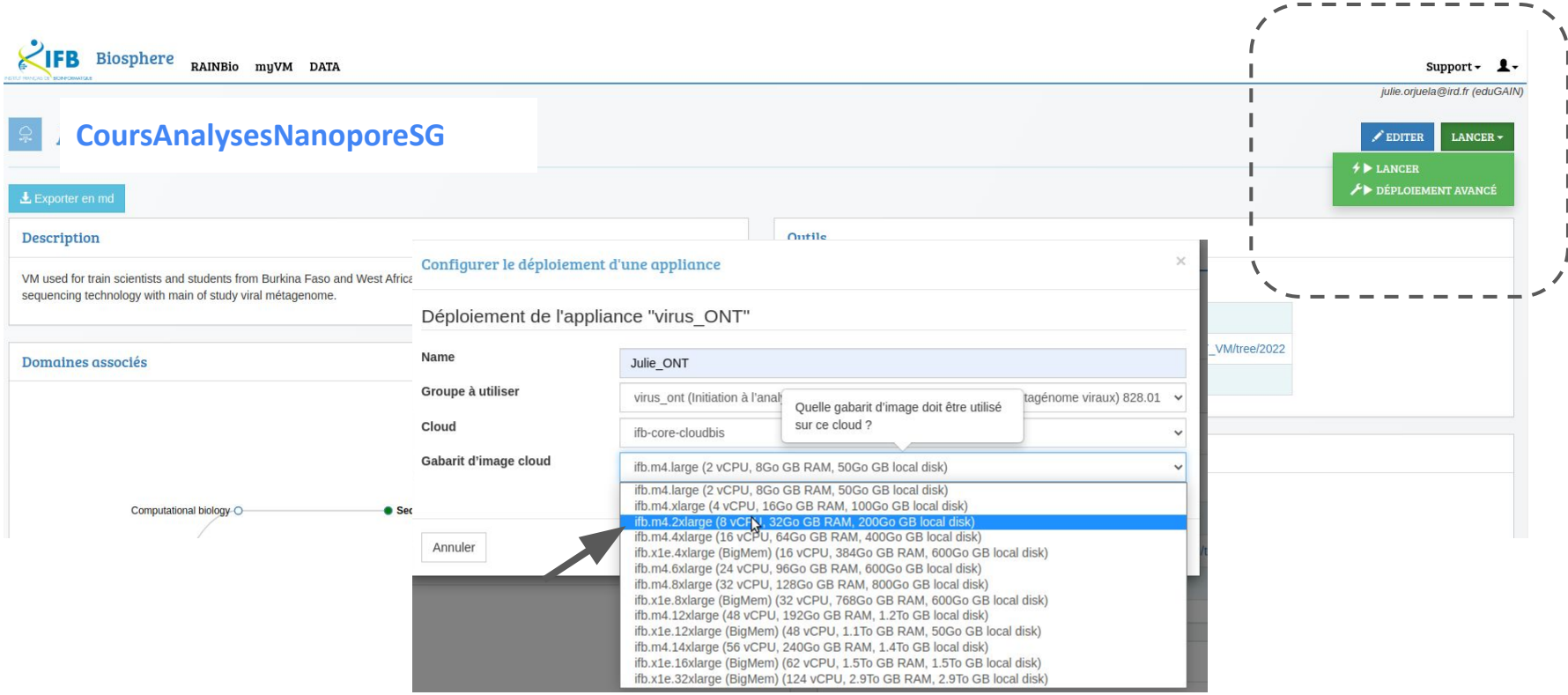
CoursAnalysesNanoporeSG



Let's run your vm through the cloud

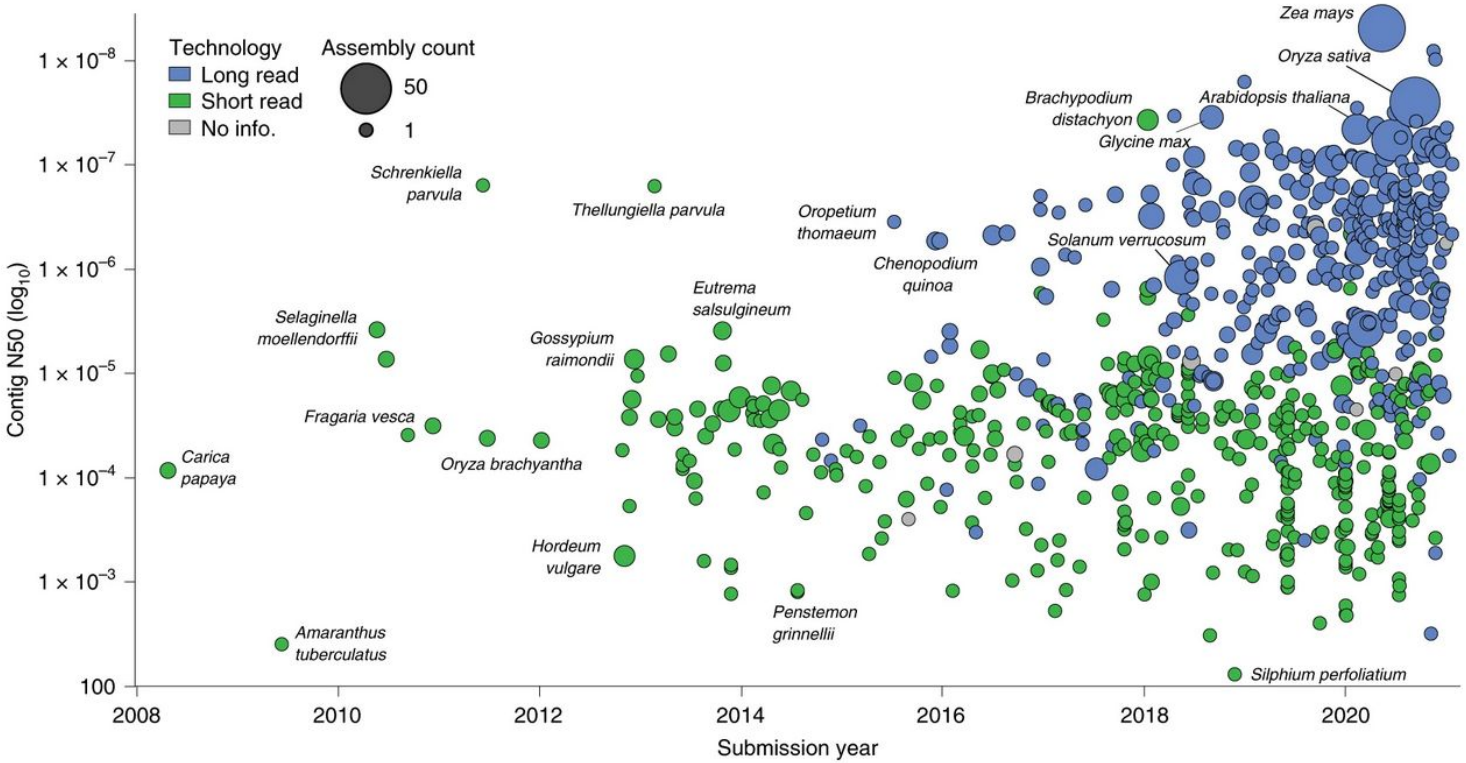


Let's run your vm through the cloud



Let's start !

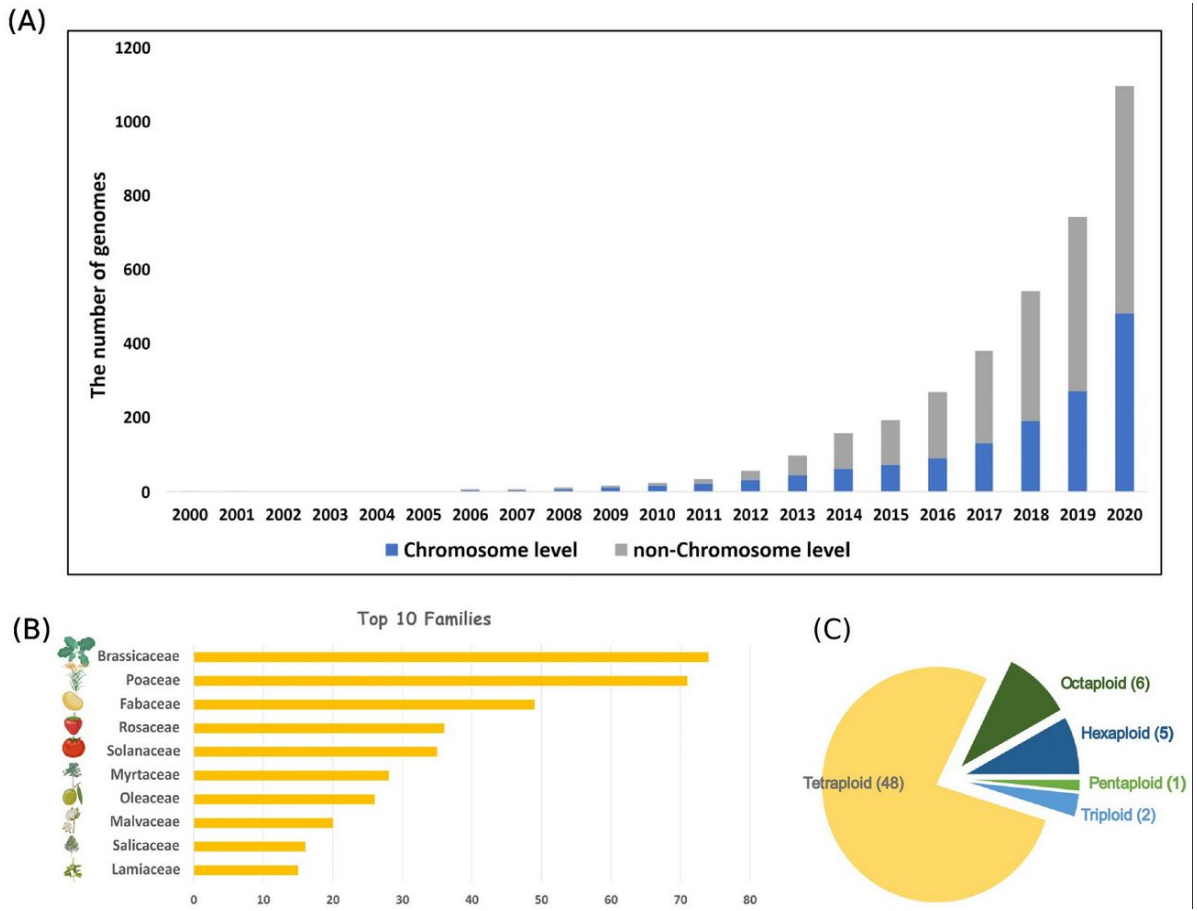
From: Representation and participation across 20 years of plant genome sequencing



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.

Marks et al, 2021

Published plant genomes from 2000



Sun et al, 2022

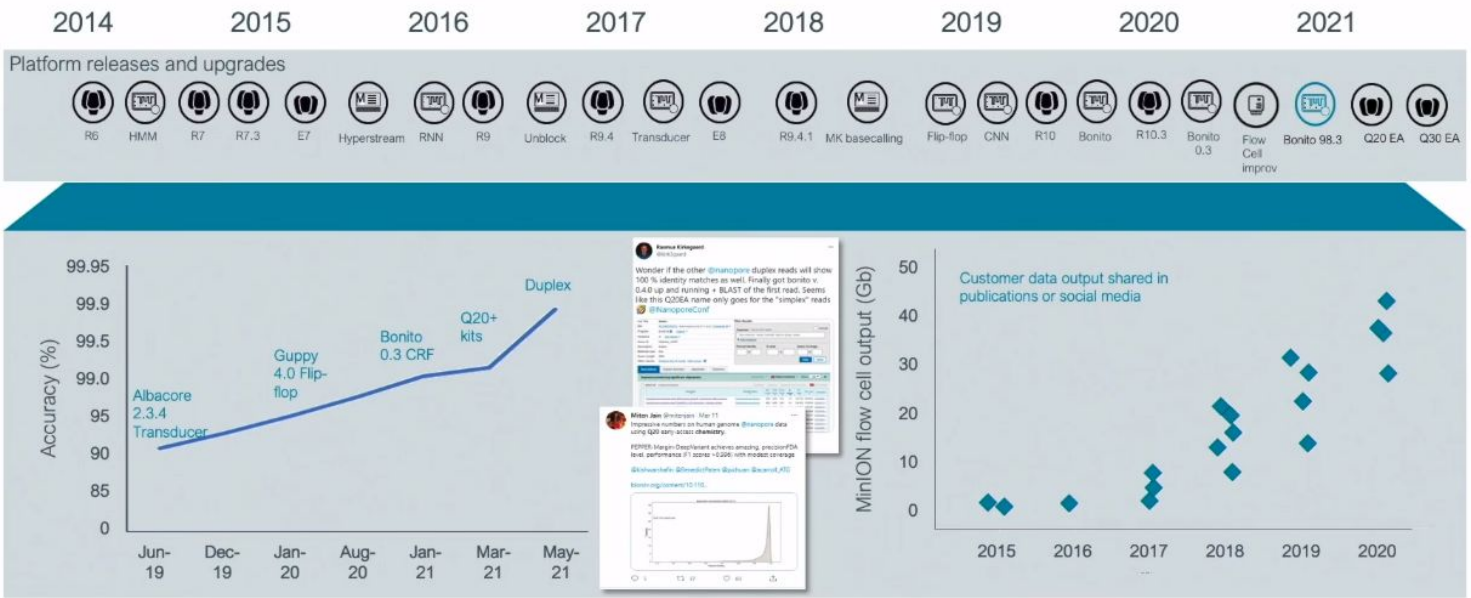
(Plant) genome project workflow from DNA extraction over ONT sequencing to data submission

		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		

Boas Pucker et al. 2022

Upgrades drive performance enhancements

…and core ones ship in consumables and software



Oxford NANOPORE Technologies

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

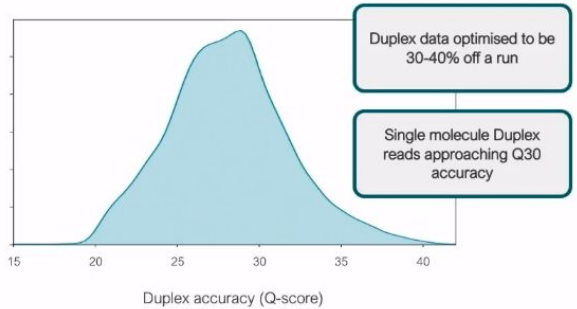
Last upgrades !

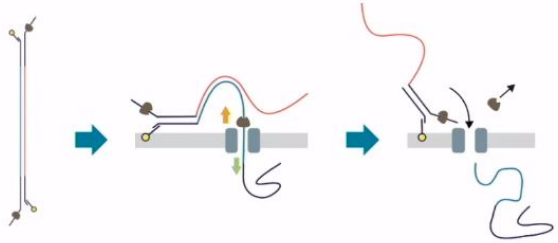
Nanopore accuracy

When we last spoke…

Duplex reads

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





Generating duplex data

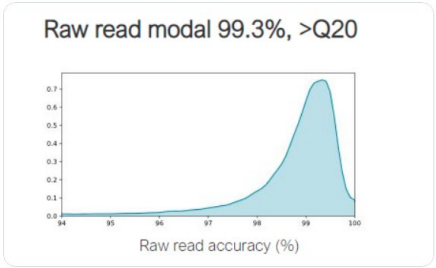
- Chances of seeing the complement follow template increased with Q20+ chemistry
- Early protocols available in EA community
- Longest Duplex Q30 read to date: 156 kbase

Oxford NANOPORE Technologies

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/3CEiJI9



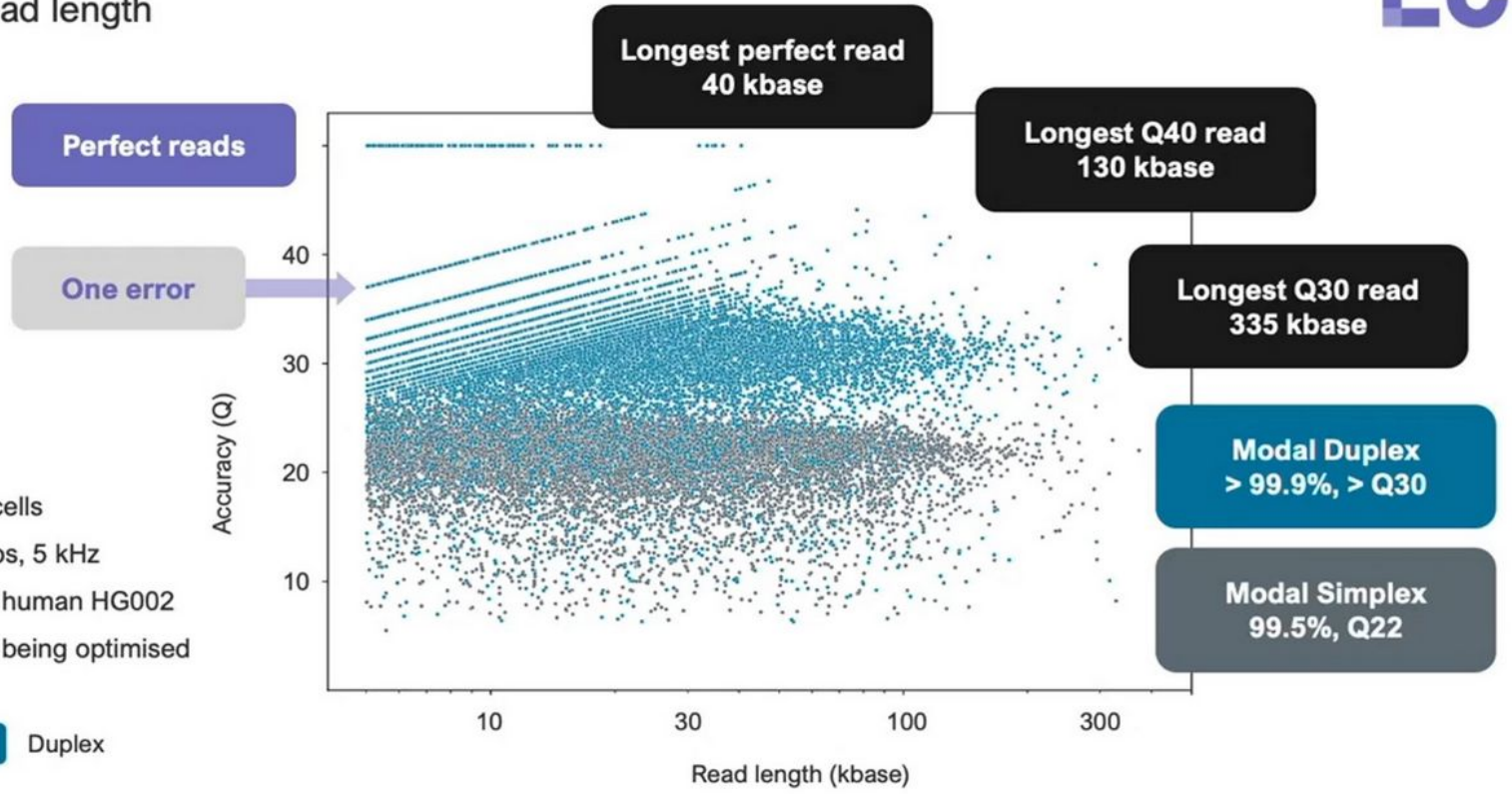
8:30 AM · Sep 23, 2021 · HubSpot

33 Retweets **1** Quote Tweet **62** Likes

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

Duplex

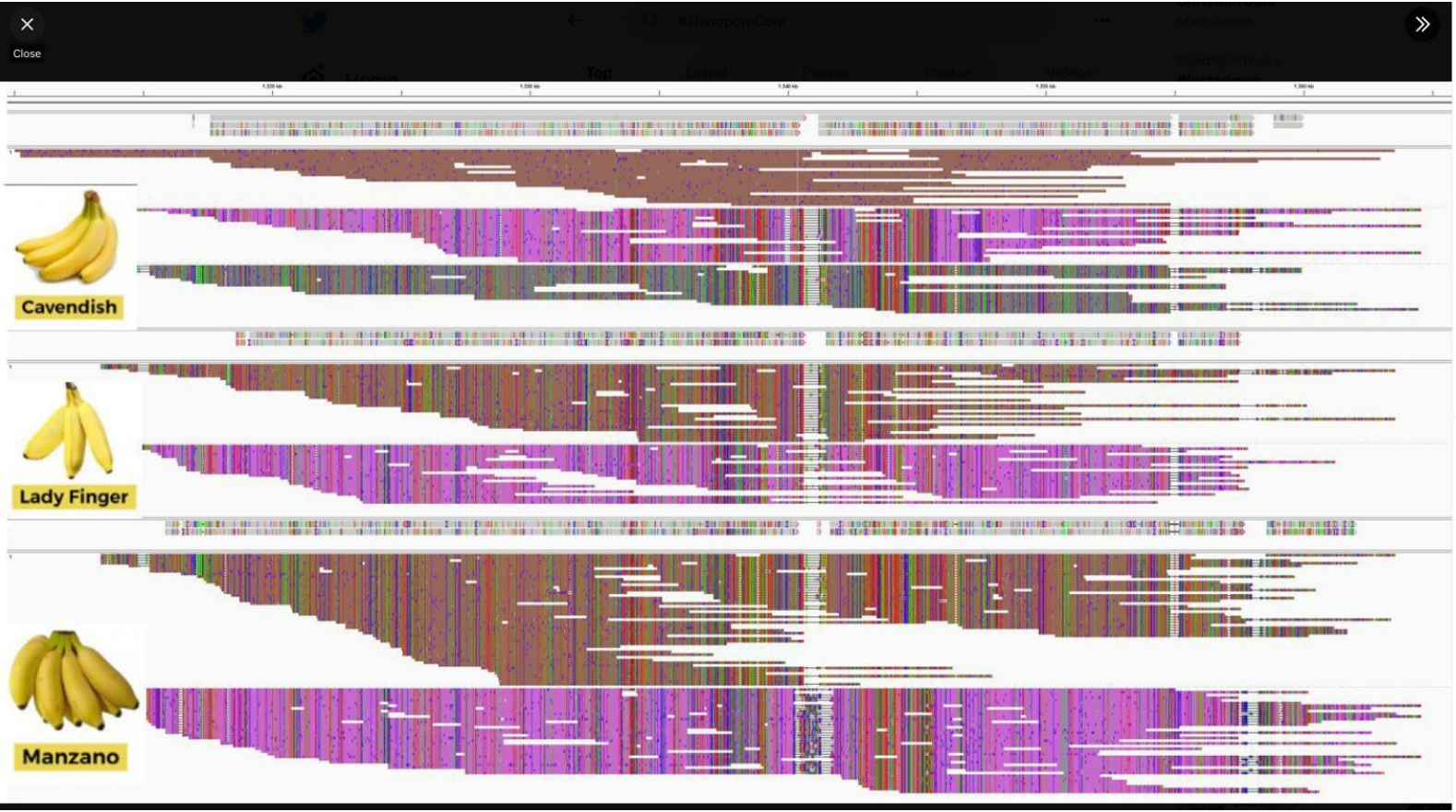
Accuracy and read length



Ultra-long Duplex:

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

#NanoporeConf



Alexander Wittenberg
@AW_NGS

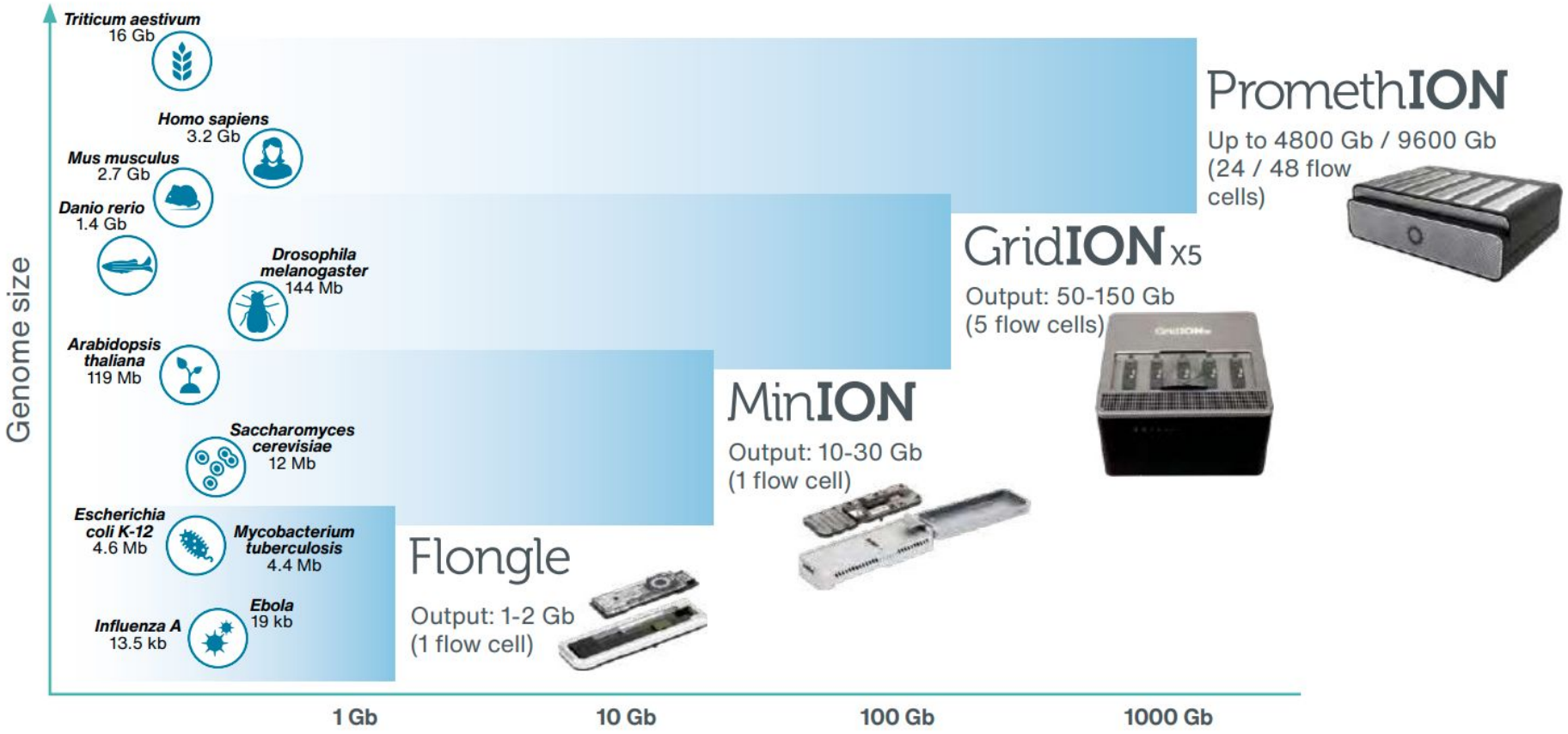
@keygene Unlocking the Power of Haplotype Based Molecular Breeding using Long Read @nanopore adaptive sequencing screening resistance gene regions in banana 🍌 #nanoporeconf Poster#53 and online #114

10:17 AM · May 18, 2023 · **271** Views

2 Retweets **6** Likes

Tweet your reply! Reply

A lot of data !


Genome size
Triticum aestivum
16 Gb
Homo sapiens
3.2 Gb
Mus musculus
2.7 Gb
Danio rerio
1.4 Gb
Drosophila melanogaster
144 Mb
Arabidopsis thaliana
119 Mb
Saccharomyces cerevisiae
12 Mb
Escherichia coli K-12
4.6 Mb
Mycobacterium tuberculosis
4.4 Mb
Influenza A
13.5 kb
Ebola
19 kb
Flongle
Output: 1-2 Gb
(1 flow cell)
MinION
Output: 10-30 Gb
(1 flow cell)
GridION X5
Output: 50-150 Gb
(5 flow cells)
PromethION
Up to 4800 Gb / 9600 Gb
(24 / 48 flow cells)
1 Gb
10 Gb
100 Gb
1000 Gb

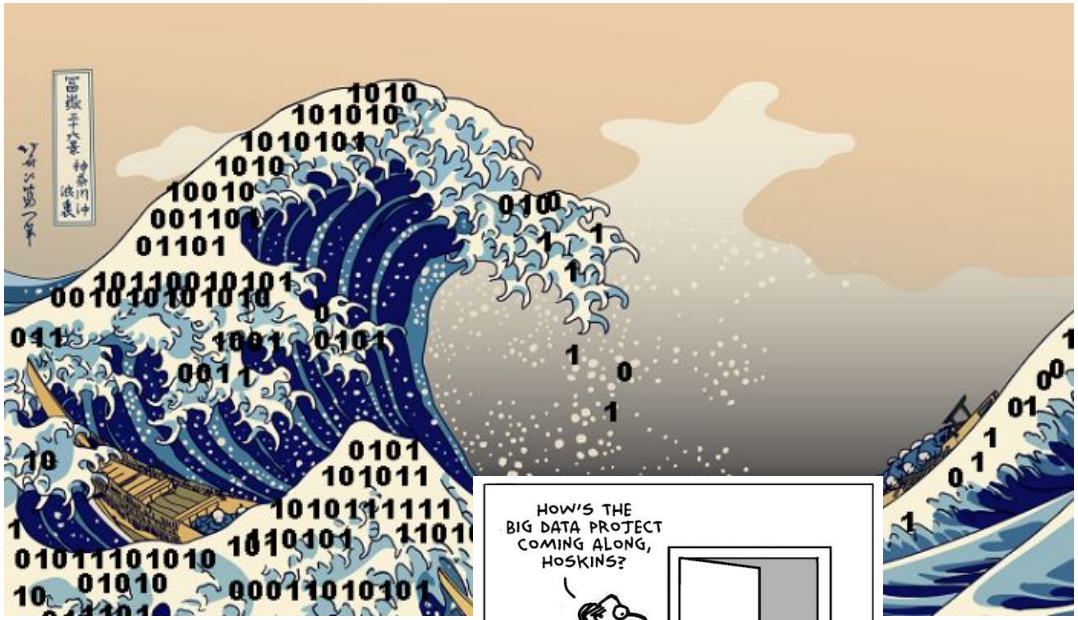
A lot of data !

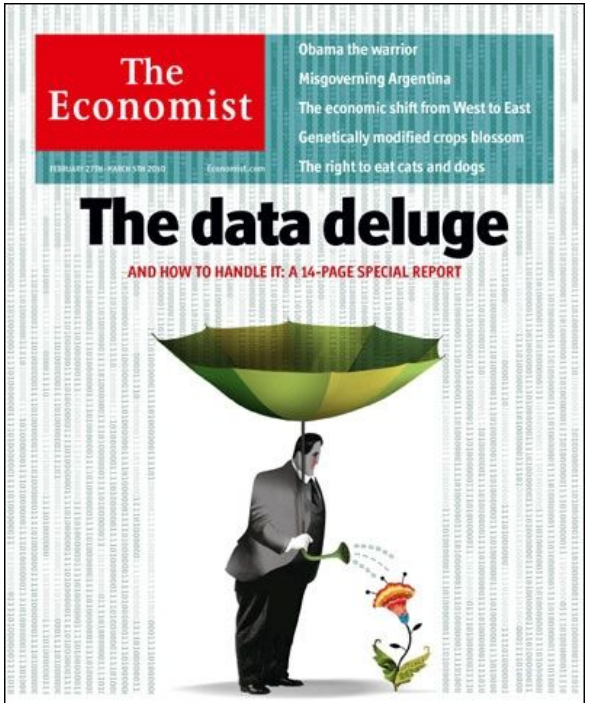
Configuration	Platform		Techniques	Tech specifications			
Number of flow cells per device	1	1	5	2	2	24	48
Maximum number of channels per flow cell	512	512	512	2,675	2,675	2,675	2,675
Run time	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours
Device TMO†	50 Gb	50 Gb	250 Gb	580 Gb	580 Gb	~7 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

Nanopore web site 2021

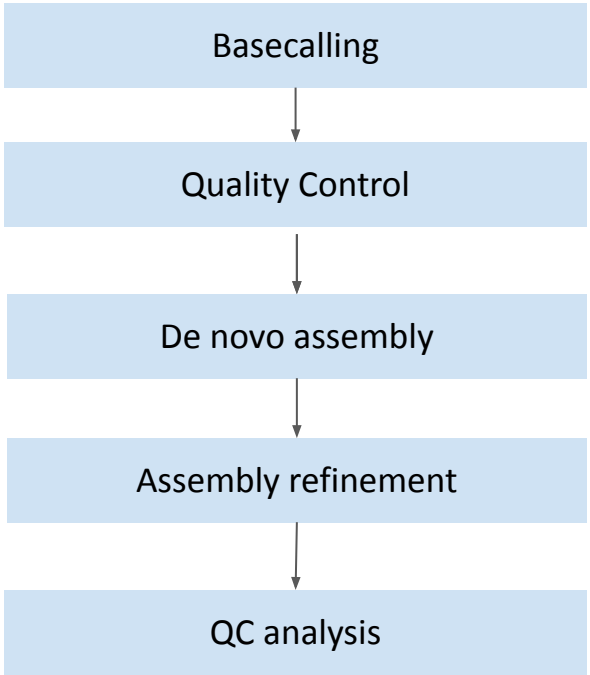
The data that these platforms produce differ qualitatively from second-generation sequencing, thus necessitating tailored analysis tools

From data rarity to data deluge



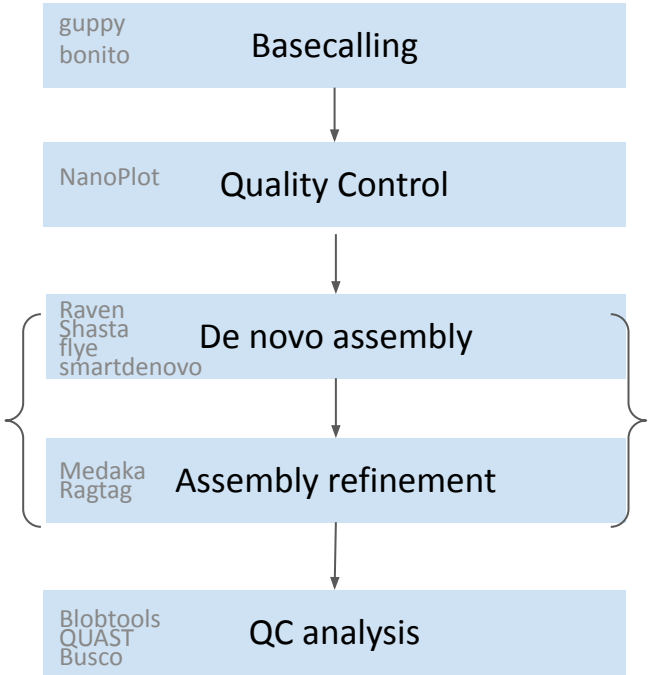


Typical long-read analysis pipelines for ONT data



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

Typical long-read analysis pipelines for ONT data



[Optional]

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

The Data !

The Pan-Genome of the cosmopolitan picophytoplankton *Bathycoccus prasinos*

A first step towards understanding adaptation to latitude and seasons

Louis Dennu, LOMIC

supervised by

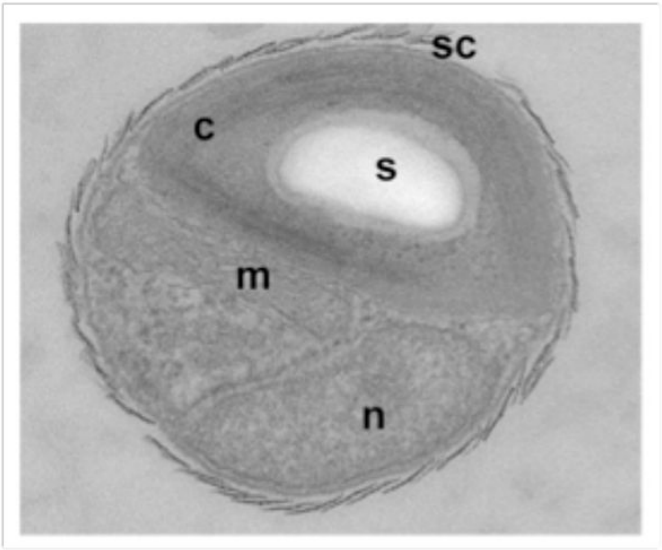
François-Yves Bouget, LOMIC - Martine Devic, LOMIC - François Sabot, IRD Montpellier

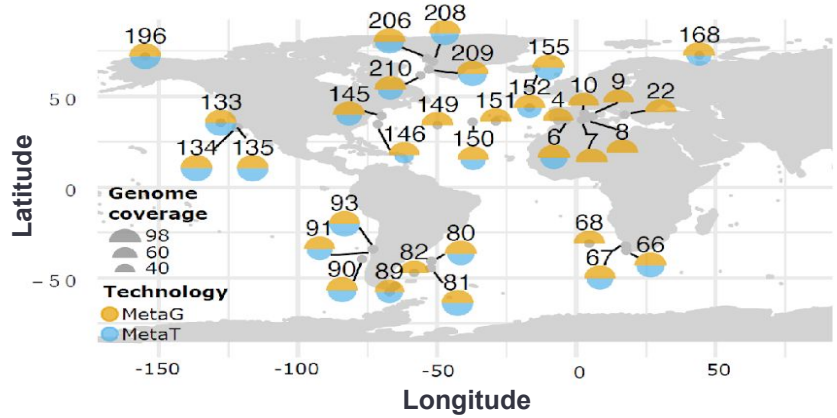
Observatoire Océanologique de Banyuls-sur-mer

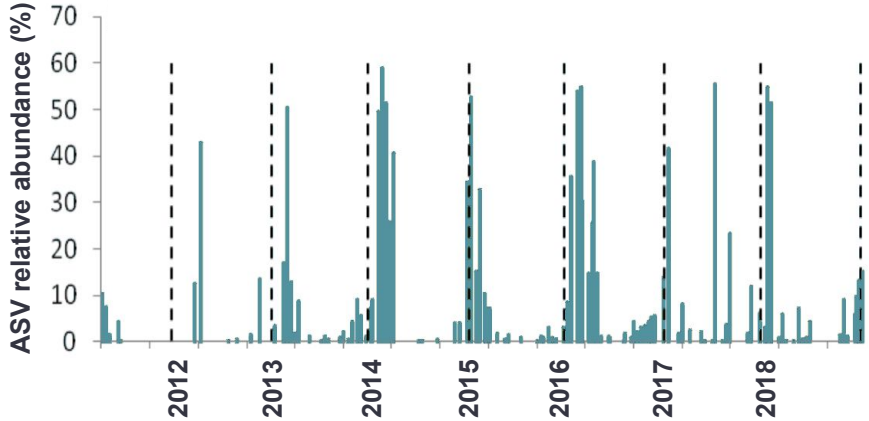
Laboratoire d'Océanographie Microbienne

Bathycoccus prasinos
A cosmopolitan model to study adaptation to latitude and seasons

Bathycoccus prasinos





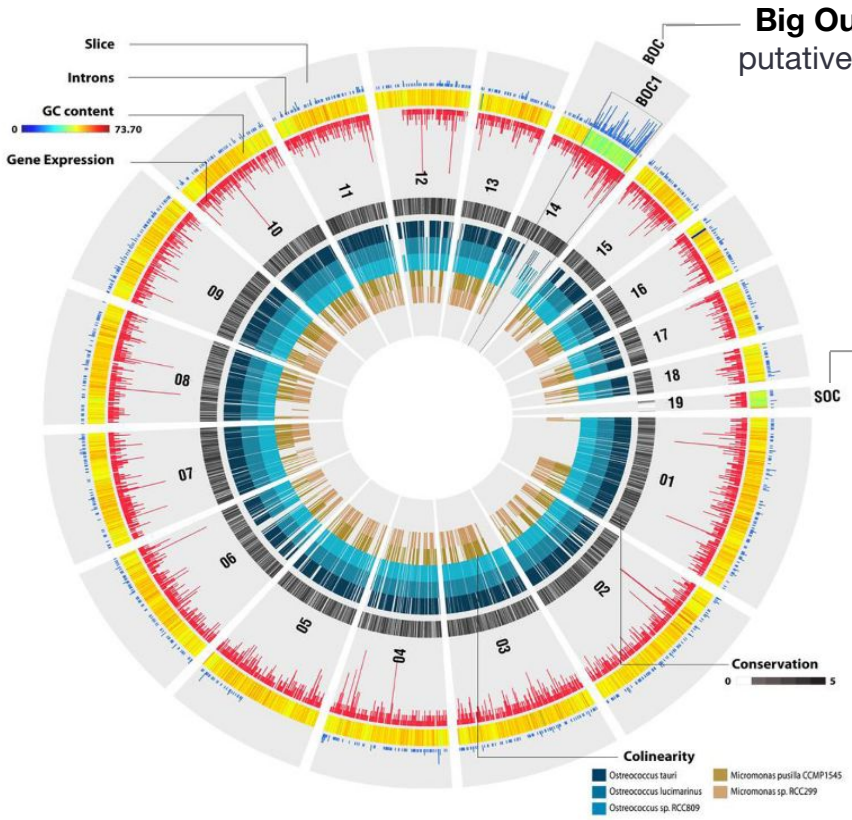


Bathycoccus prasinos single reference genome

RCC1105
Banyuls-sur-Mer, France

Size
15 Mbp

19
Chromosomes



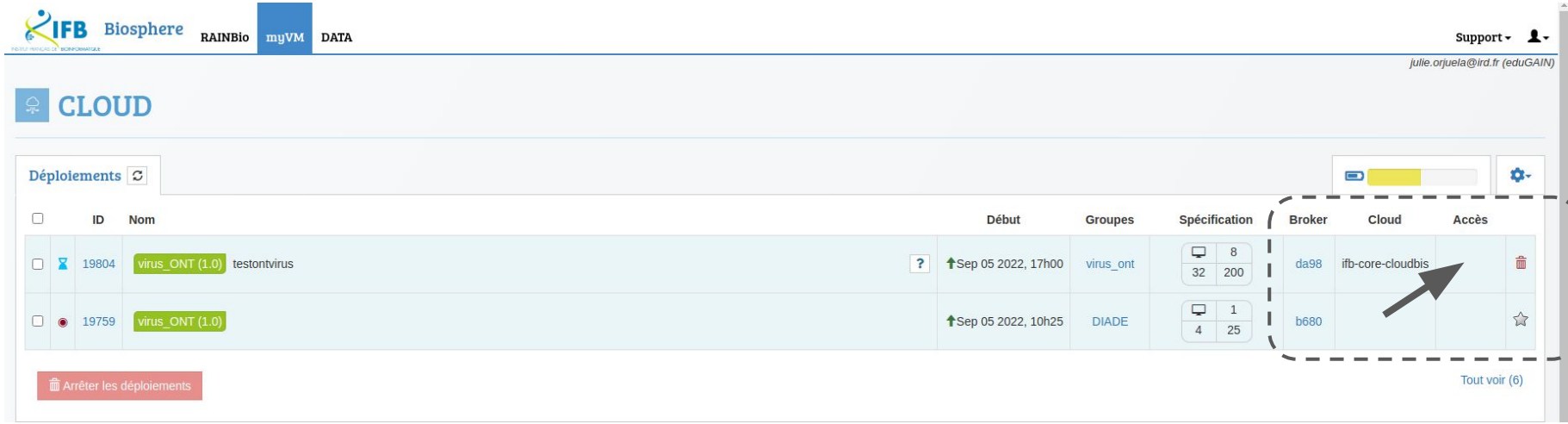
Big Outlier Chromosome
putative sexual chromosome

Small Outlier Chromosome
hypervariable chromosome
virus resistance

Back to Machines!

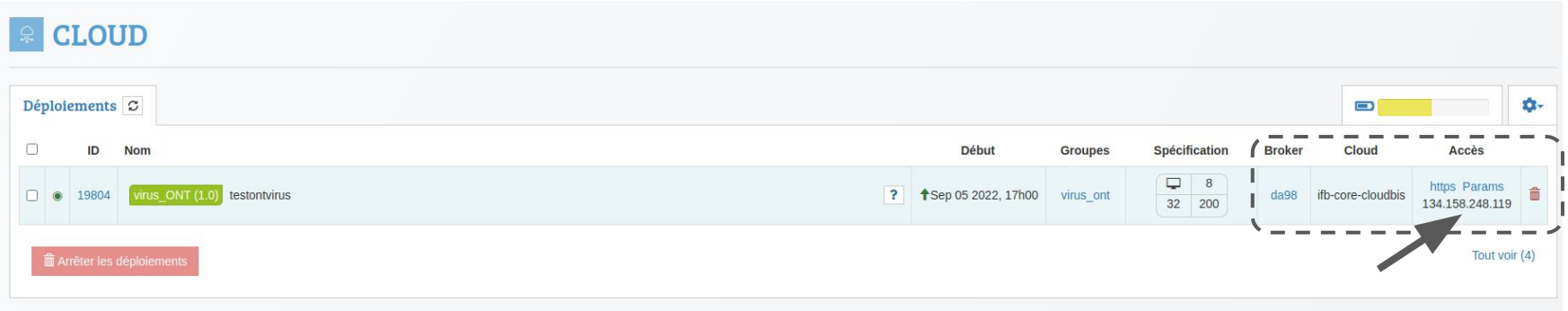
Let’s run your vm through the cloud

Loading…



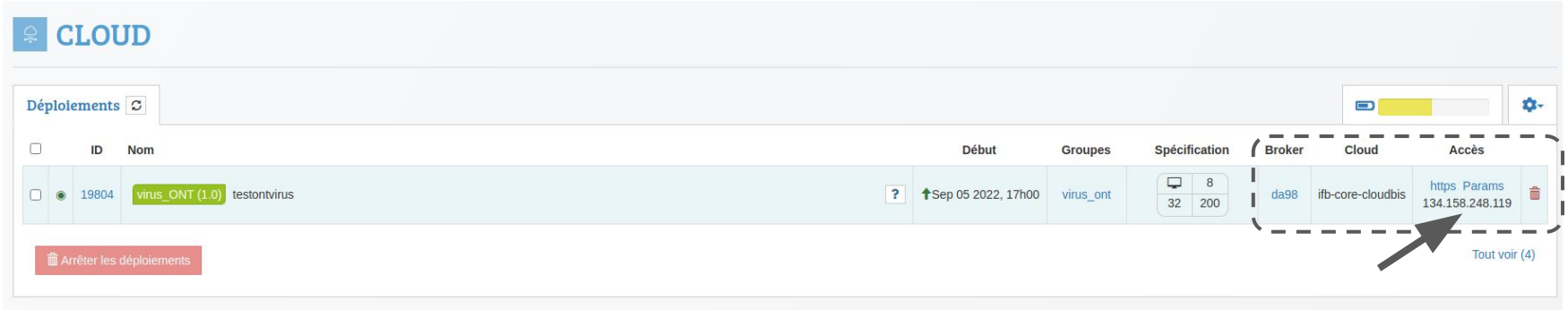
Let’s run your vm through the cloud

ready !



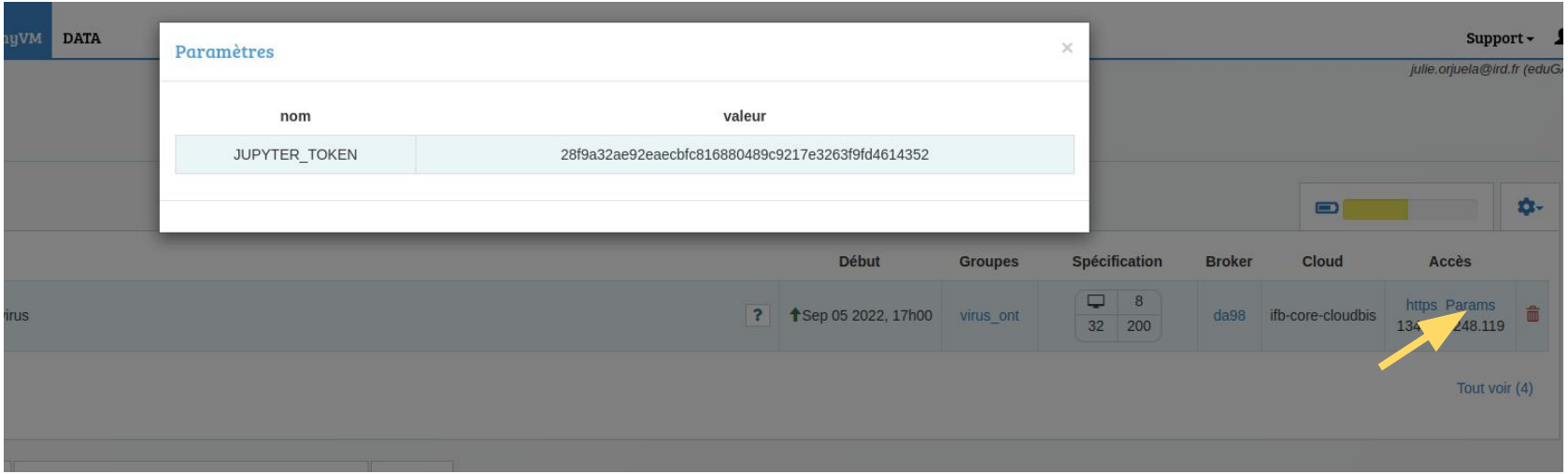
Let’s run your vm through the cloud

get the url… link “https”



Let’s run our vm through the cloud

Get the token identifiant… link “Params”

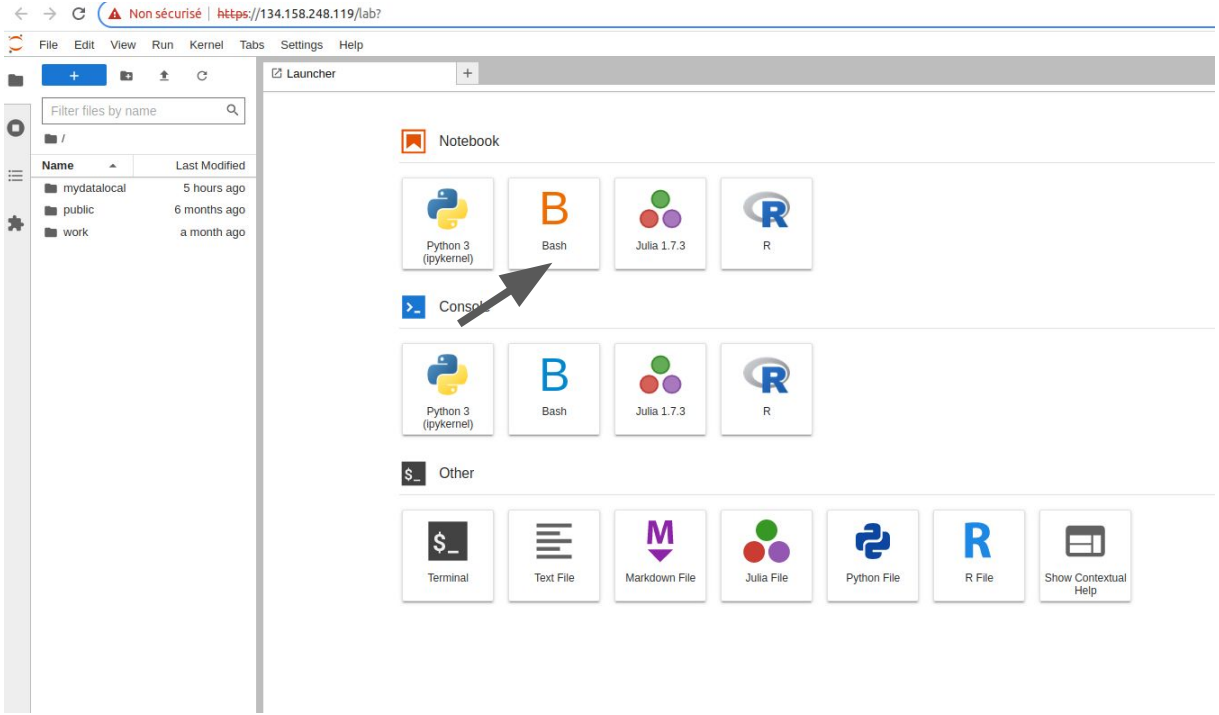


Let's run our vm through the cloud

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

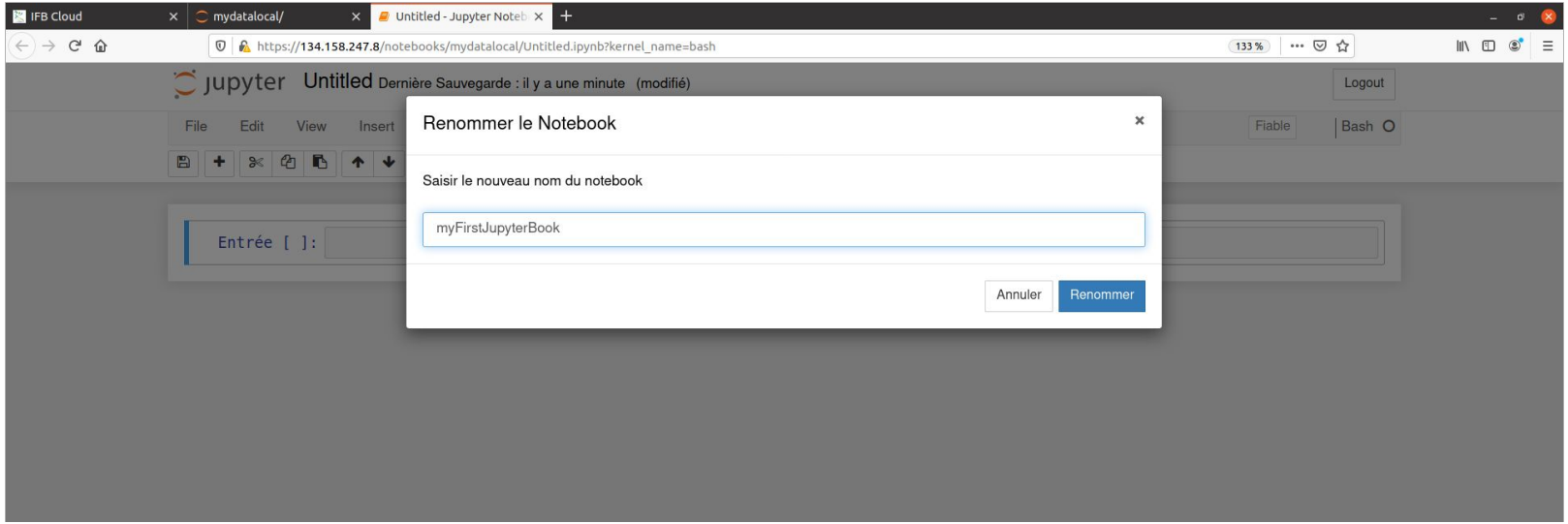
Create your first jupyter book

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



Rename your first jupyter book

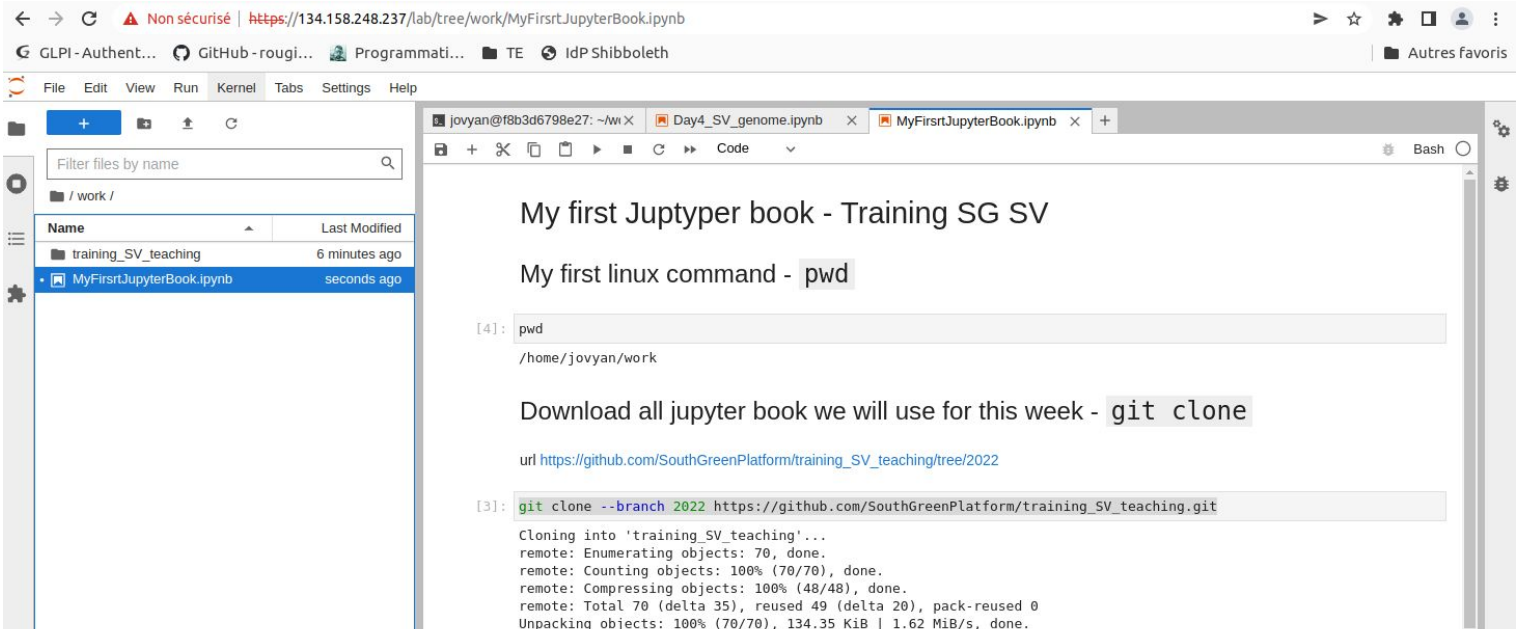
myFirstJupyterBook



Run your first bash command - *git clone*

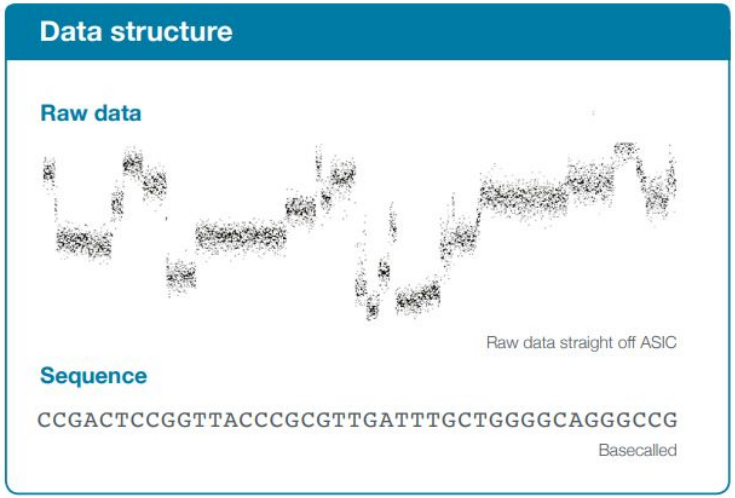
All jupyterbook used for practice are here :
https://github.com/SouthGreenPlatform/training_ONT_teaching

Download all the jupyter books with the command *git clone*
git clone https://github.com/SouthGreenPlatform/training_ONT_teaching.git
git checkout 2023_MTP

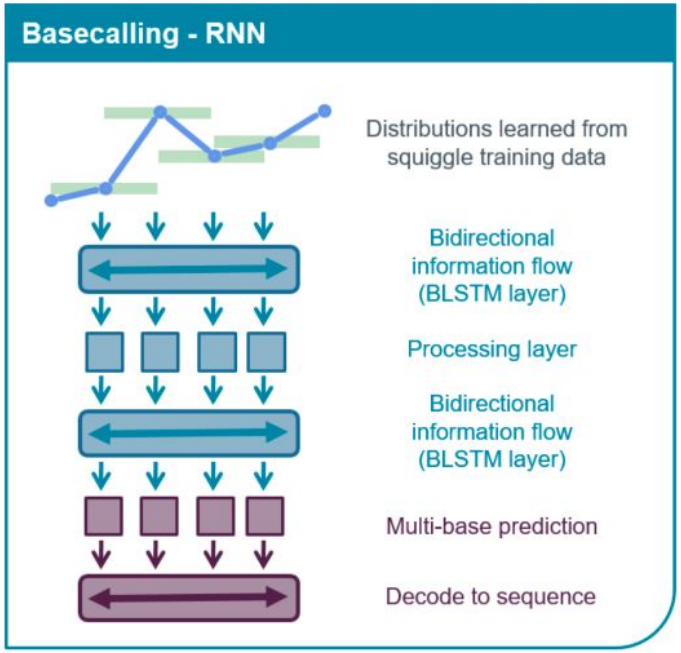


Chapitre 1
Reads Quality Control

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

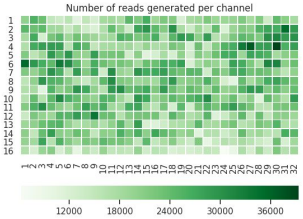
FASTQ FORMAT

```
@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
+
@CFFFFFGGHHHHIJJJJIEE<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

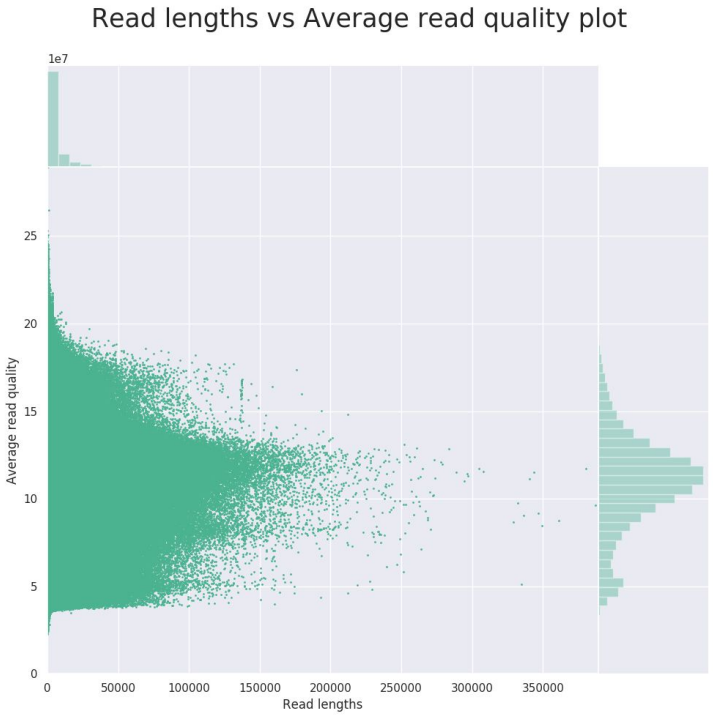
- @identifiant de la séquence
- Séquence
- + (id séquence).
- Qualité de la séquence = un caractère ASCII pour chaque base

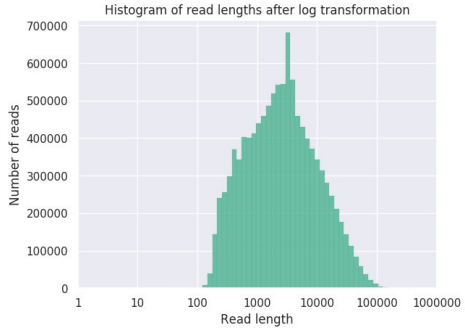
Reads Quality control : *NanoPlot*

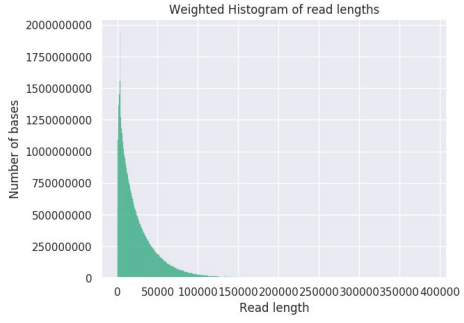


Summary statistics

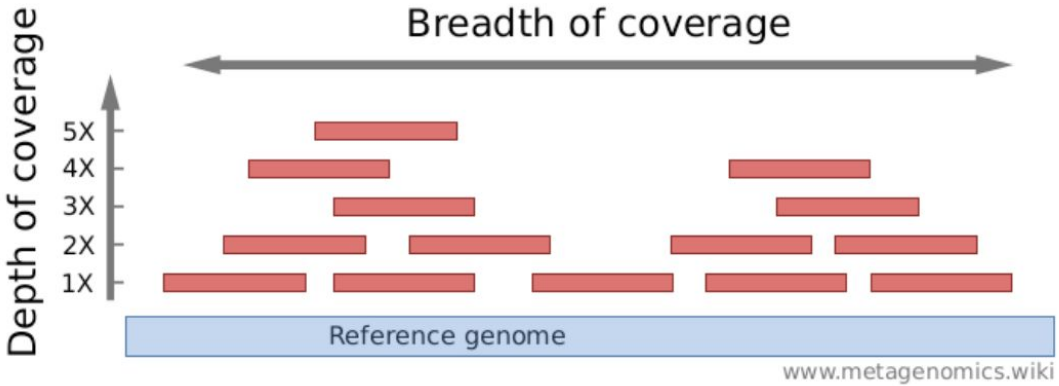
General summary	
Active channels	512.0
Mean read length	6,315.6
Mean read quality	10.9
Median read length	2,517.0
Median read quality	11.1
Number of reads	10,847,854.0
Read length N50	16,816.0
Total bases	68,510,227,164.0







Calculate depth of coverage

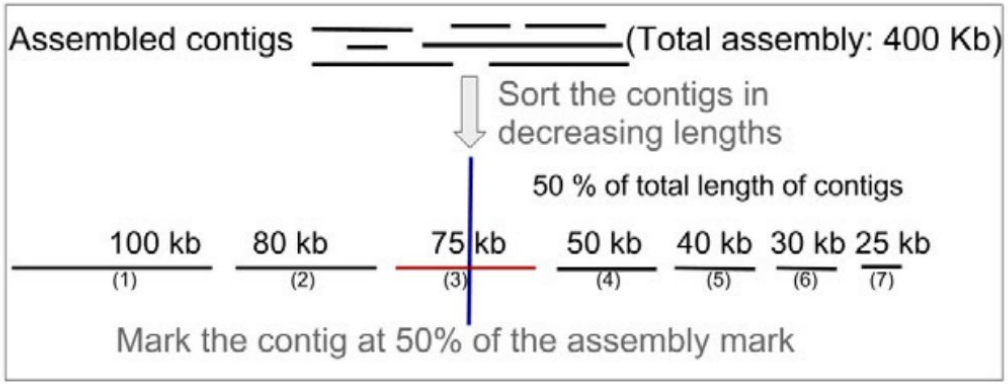


depth of coverage estimation :

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

depth_of_coverage = total_pb/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

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

Reads Quality control

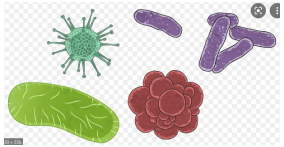
NanoPlot : https://github.com/wdecoster/NanoPlot
NanoComp : https://github.com/wdecoster/nanocomp
(mini_qc : https://github.com/roblanf/minion_qc)

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

Chapitre 2
Assemblies

Which assembler to use over my favorite organism?

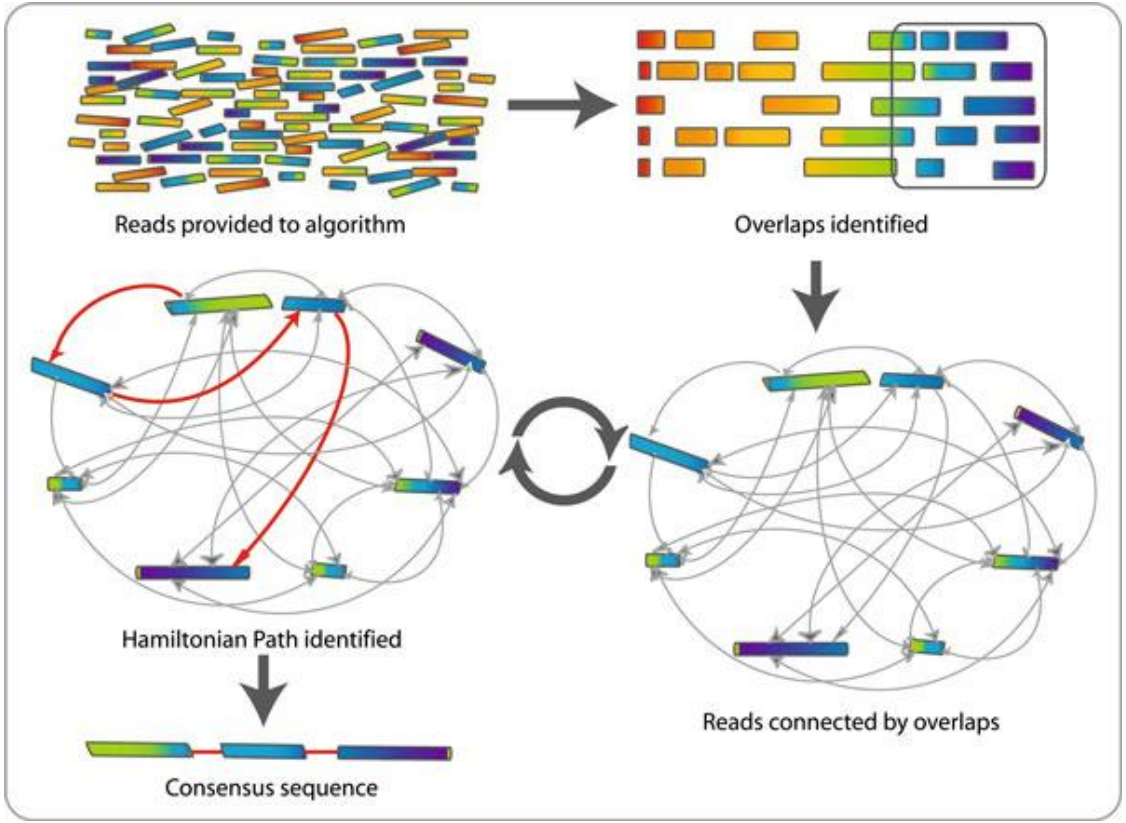
Long reads simplify genome assembly, with the ability *to* span repeat-rich sequences (characteristic of antimicrobial resistance genes) and structural variants. Nanopore sequencing also shows a lack of bias in GC-rich regions, in contrast to other sequencing platforms. To perform microbial genome assembly, 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

Oxford NANOPORE Technologies

Overlap–layout–consensus genome assembly algorithm (OLC)



Canu, Flye, Miniasm, Raven, Smartdenovo, Shasta

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

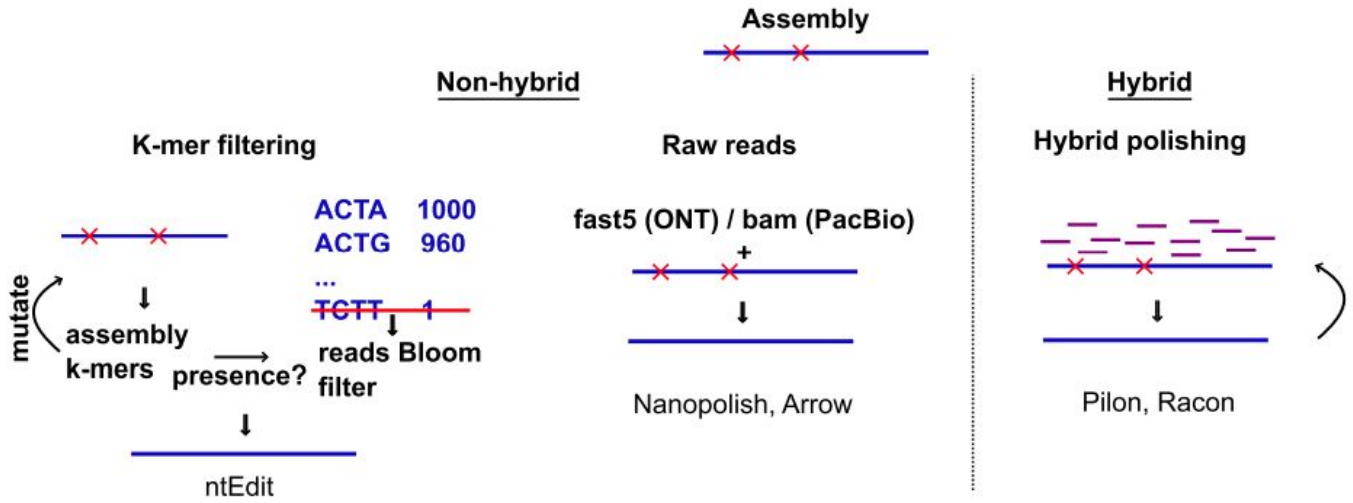
Polishing / Correction

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

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

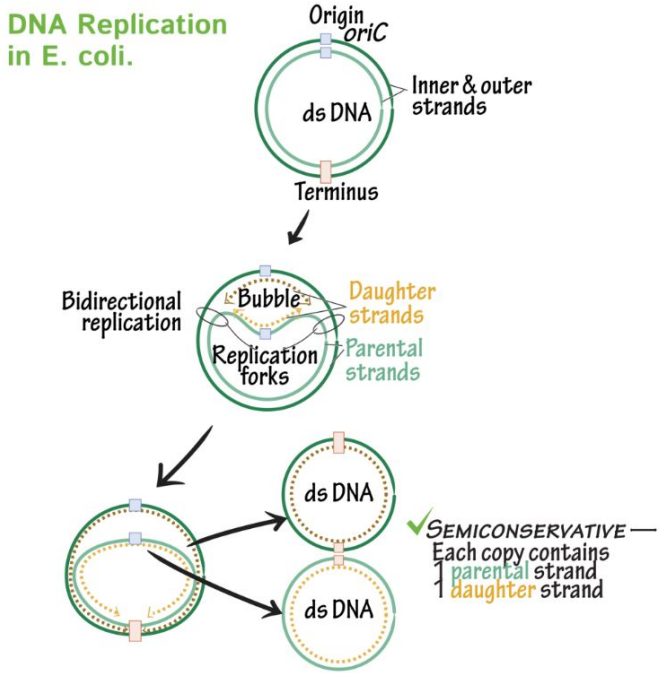
+ Medaka uses neural networks where Nanopolish uses HMMs.
+ Medaka uses basecalled reads, not the raw signal.
+ Medaka propose the ability to train one's own basecalling model

Pilon correct assemblies using illumina reads. (recursive use)
Autres : NeuralPolish , ntEdit



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

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

Worst Median Best ☐ Show heatmap

Genome statistics	FLYE_STEP_POLISHING_RACON_	FLYE_STEP_ASSEMBLY_	RAVEN_STEP_POLISHING_RACON_	RAVEN_STEP_ASSEMBLY_	SHASTA_STEP_POLISHING_RACON_	SHASTA_STEP_ASSEME
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

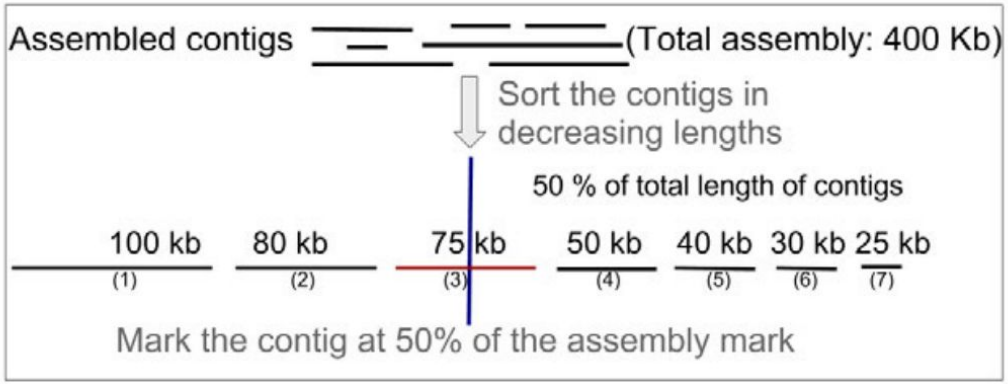
https://github.com/ablab/quast

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

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

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

26 March 2021, Friday, 07:37:40

View in Icarus contig browser

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

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

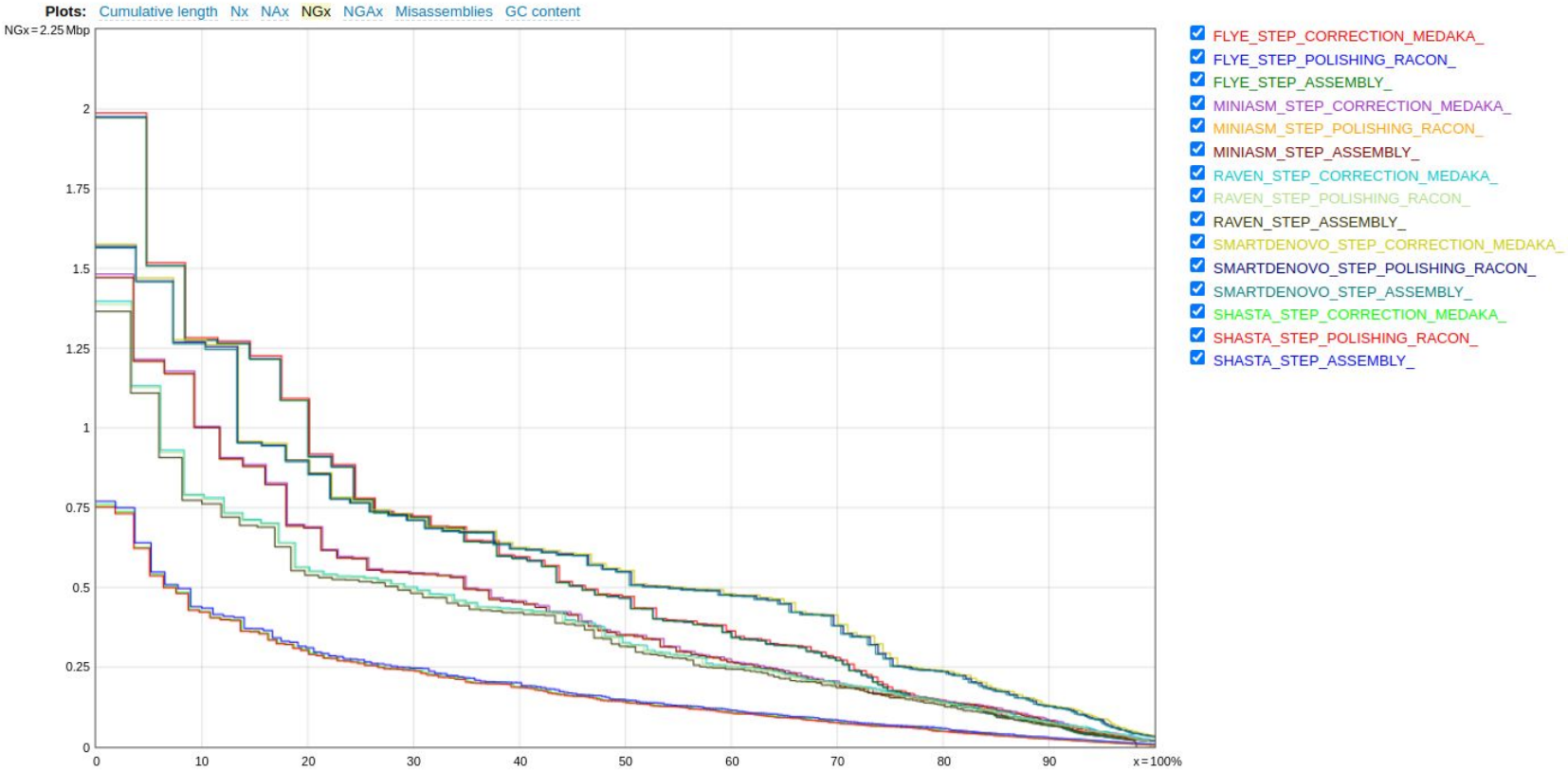
Worst Median Best ☐ Show heatmap

Genome statistics	FLYE_STEP_POLISHING_RACON_	FLYE_STEP_ASSEMBLY_	RAVEN_STEP_POLISHING_RACON_	RAVEN_STEP_ASSEMBLY_	SHASTA_STEP_POLISHING_RACON_	SHASTA_STEP_ASSEMB
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



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.

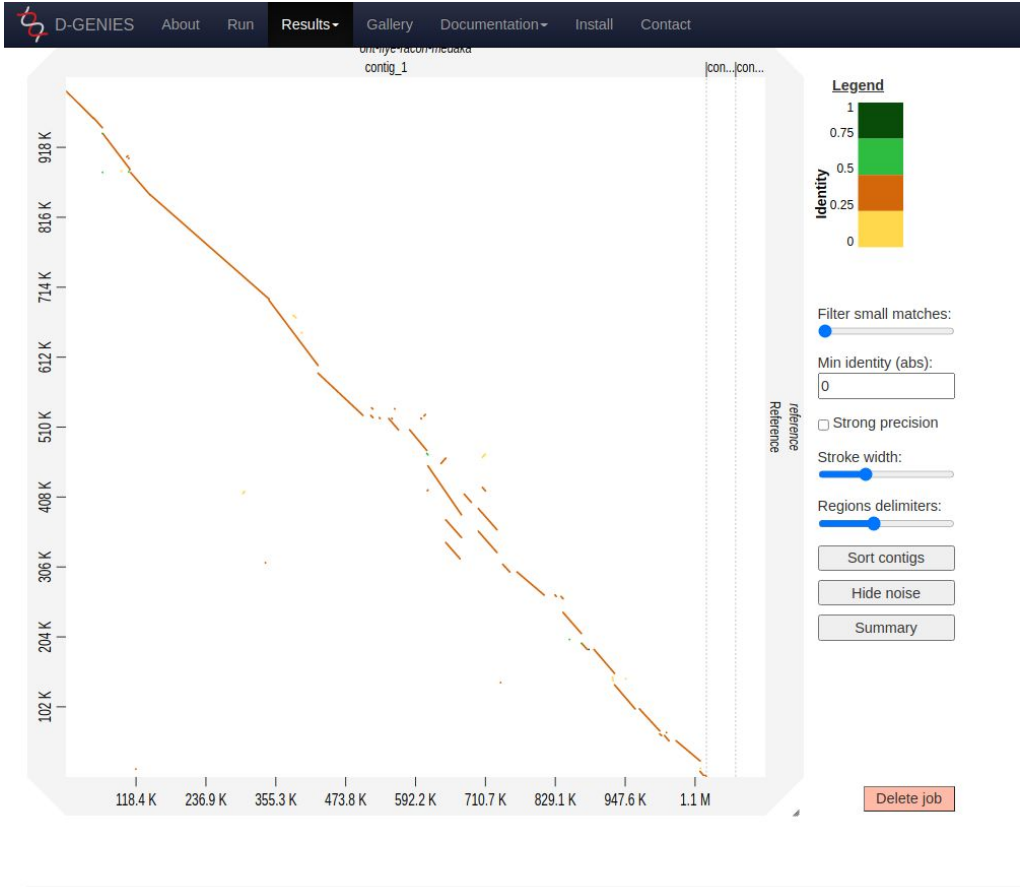
Based on evolutionarily-informed expectations of gene content of near-universal single-copy orthologs, BUSCO metric is complementary to technical metrics like N50.

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/
```

https://busco.ezlab.org/busco_userguide.html

Comparison with a reference genome



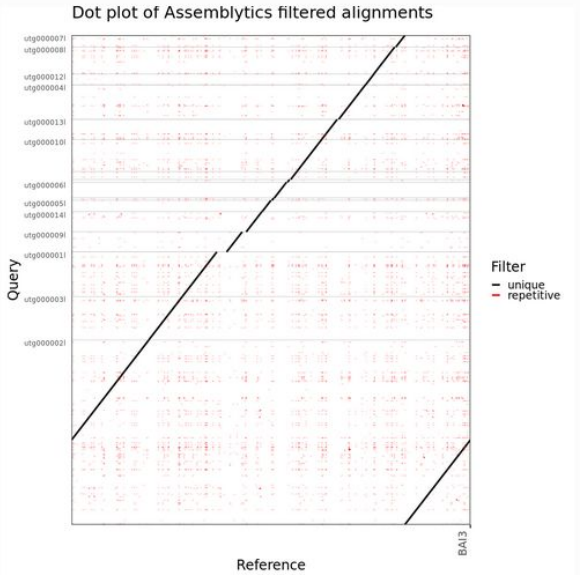
- NUCMER : Aligns a set of draft sequence contigs to a finished sequence http://mummer.sourceforge.net/
- D-Genies : Online tool to compare two genomes by dot plot method http://dgenies.toulouse.inra.fr/
- autre: *Gepard*

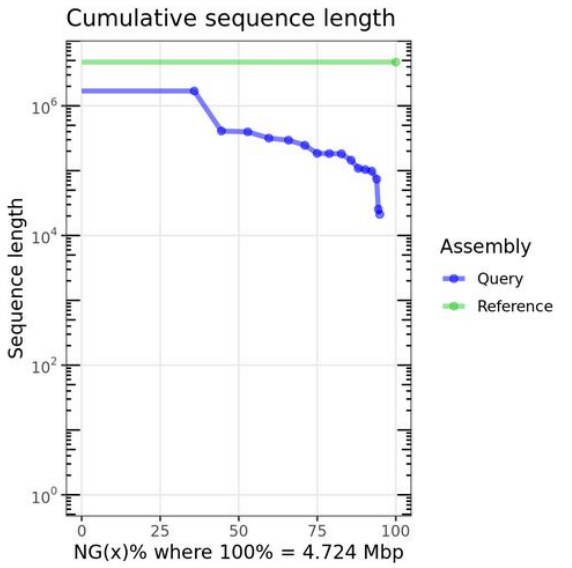
Assemblytics

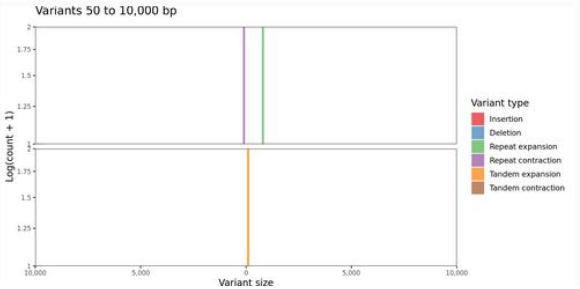
http://assemblytics.com

CANU | FLYE | MINIASM | RAVEN | SMARTDENOVO | SHASTA

STEP_CORRECTION_NANOPOLISH_STARTFIXED



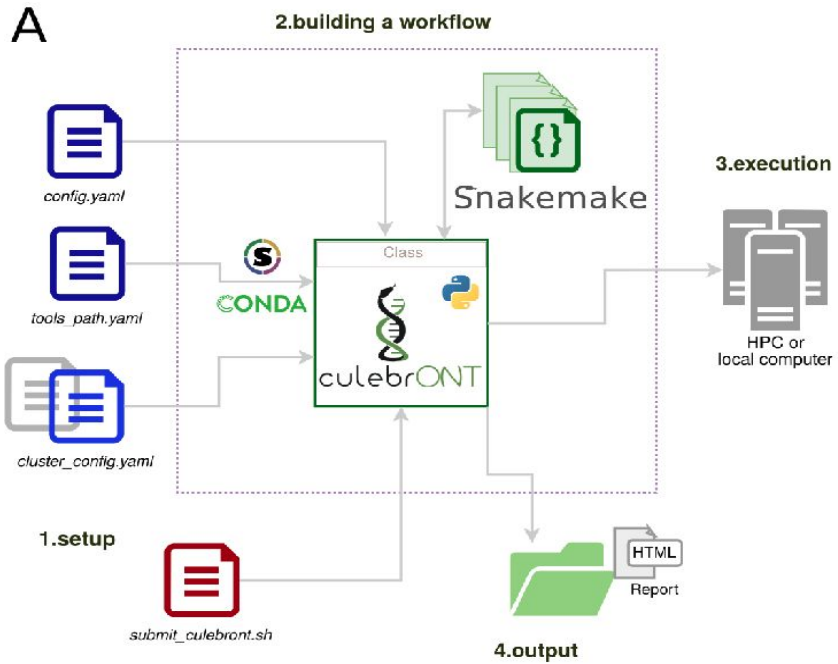




generated using culebrONT

A flexible and reproducible pipeline for LR assembly and evaluation

```
pip install culebrONT
```



- Assemblies
- Circularisation
- Polishing
- Correction
- Fixstart
- Quality
- Reporting

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

From contigs to chromosomes

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

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

Genetic map : marker assisted contig bridging

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

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 !

South Green bioinformatics platform

SUIVEZ NOUS SUR TWITTER !

South Green : @green_bioinfo

I-Trop : @ItropBioinfo