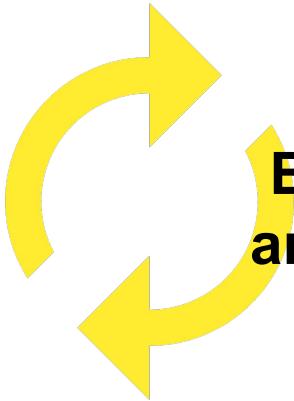




Session de formation 2021



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

comparative genomics
phylogenomics
GWAS
population genetics
polyploidy

genome assembly
transcriptome assembly
metagenomics

SNP detection
structural variation
differential expression



Rice



Banana



Palm



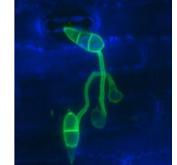
Sorghum



Coffee



Cassava



Magnaporthe

South Green

bioinformatics platform



Larmande Pierre
Sabot François
Tando Ndomassi
Tranchant Christine
Orjuela Julie



Ravel Sébastien
Mahé Frédéric
Dereeper Alexis



Bocs Stephanie
De Lamotte Fredéric
Droc Gaetan
Dufayard Jean-François
Hamelin Chantal
Martin Guillaume
Pitollat Bertrand
Ruiz Manuel
Sarah Gautier
Summo Marilyne



Rouard Mathieu
Guignon Valentin
Catherine Breton



Sempere Guilhem



South Green

bioinformatics platform

Workflow manager



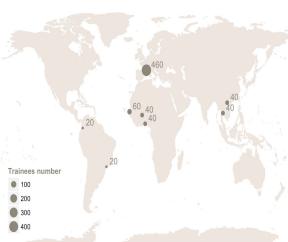
Toolbox for generic NGS analyses



HPC and trainings....



37 courses organized last 7 years



Genome Hubs & Information System

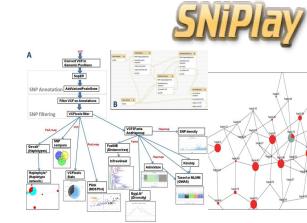


SNPs and Indels



Family ID	Family Name	Number of sequences	Status
GPF00001	Cytochrome P450 superfamily	5442	Green
GPF00001	AT5E26BP transcription factor family	5142	Green
GPF00001	NAC transcription factor family	4574	Green
GPF00001	MADS transcription factor superfamily		
GPF00001	General zufalpha-like transcription factor family		
GPF00001	Sulfatase like Serine Proteases family		
GPF00001	NPF/NPFYTF1 FAMILY		

Gene families



<https://github.com/SouthGreenPlatform>



@green_bioinfo

The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics, Current Plant Biology, 2016

Modules de formation 2021

- Toutes nos formations :
<https://southgreenplatform.github.io/trainings/>
- Topo & TP : **Lien de la FORMATION**
- Environnement de travail : [Logiciels à installer](#)

Initiation à l'analyse de données Oxford Nanopore



Alliance



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 (47)

Appliances

Outils

Topics

Appliance éditable

Ajouter



CoursAnalysesNanoporeSG

- bandage, Jupyter
- Data architecture, analysis and design, Mathematics, Statistics

CentOS 7

- Ansible, bioconda, Docker
- Bioinformatics, Informatics

Askomics

- AskOmics
- Data integration and warehousing, Data visualisation

Cytoscape

- Bureau virtuel, Cytoscape, X2Go, XFCE
- Bioinformatics, Data visualisation, Molecular interaction

Bacterial Genomics

- HMMER, Insyght, SGE - GridEngine, Ubuntu, Web interface
- Protein folds and structural domains, Sequence comparison, Sequence conservation

Bioimage

- Bureau virtuel, Icy, ImageJ-Fiji, X2Go, XFCE
- Informatics, Data visualisation, Imaging

Debian 10

- Ansible, bioconda, Docker
- Bioinformatics, Informatics

Debian 9

- Ansible, bioconda, Docker
- Bioinformatics, Informatics

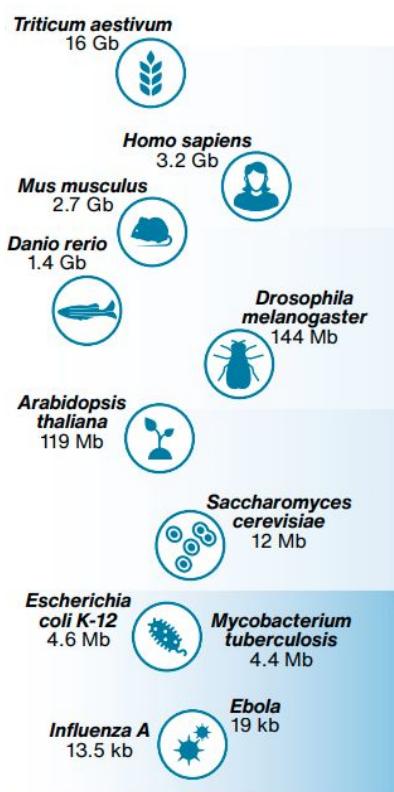


First of all!



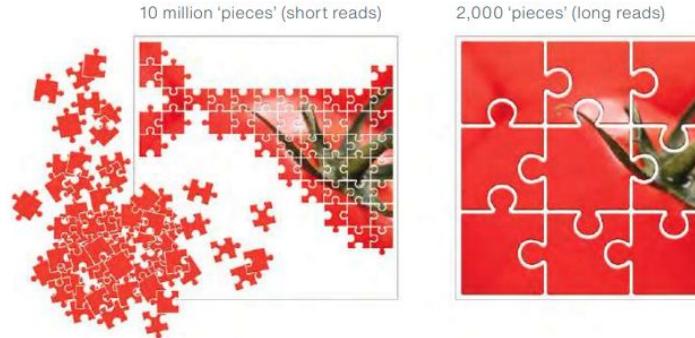
- Launch IFB virtual machines using 8 threads and 32G RAM
- https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2021/0.running_an_appliance_biosphere.ipynb
- Download data !!

Why use Long reads ?



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

- they 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.
- they provide methylation information as well



Two technologies

Oxford Nanopore



MinION

GridION

PromethION

Pacific BioScience



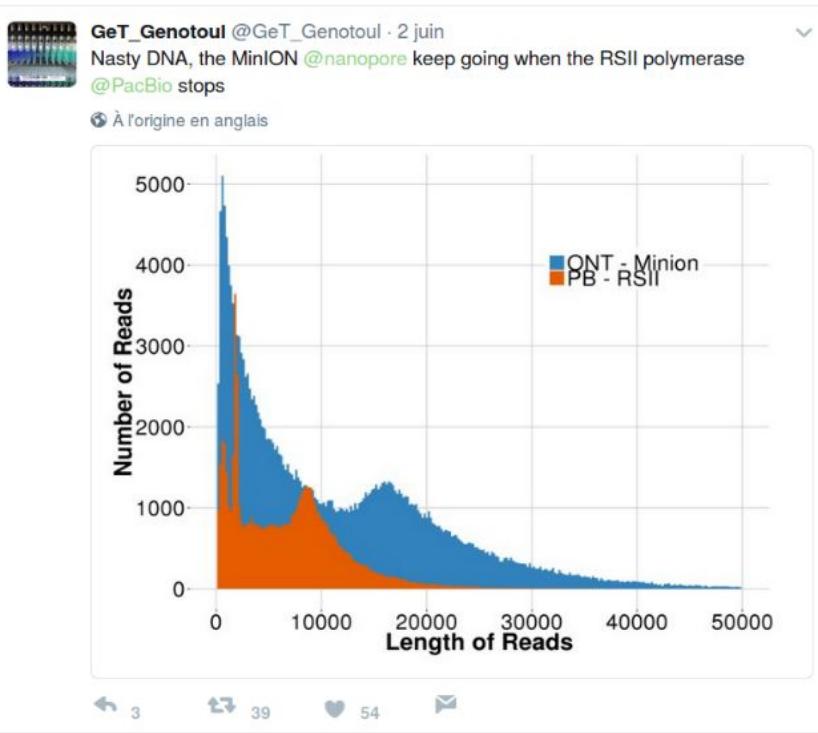
RSII



Sequel

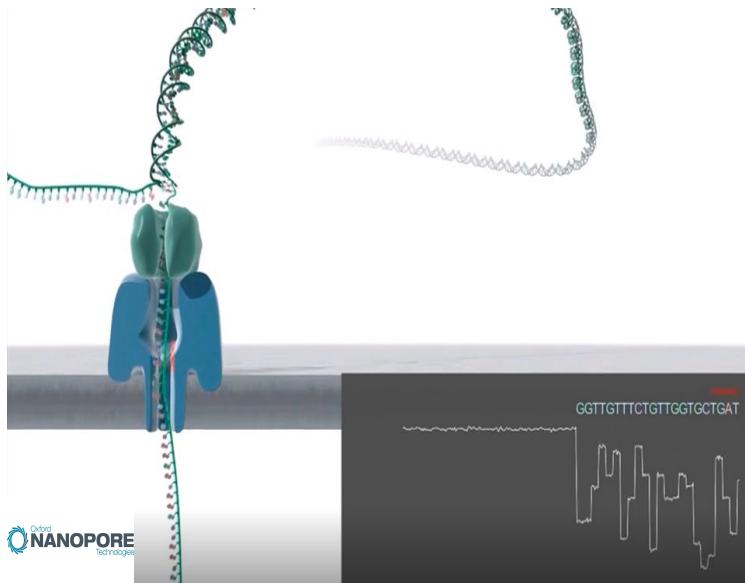
from Elixir GAAS 2018

Same sample / RSII vs MinION



- SMRT limited by the longevity of the polymerase. A faster polymerase for the Sequel sequencer (chemistry v3, 2018) increased the read lengths to an average 30-kb polymerase read length.
- ONT from 500 bp to the current record of 2.3Mb, with 10–30kb genomic libraries being common. Mostly limited by the ability to deliver very high-molecular weight DNA to the pore.

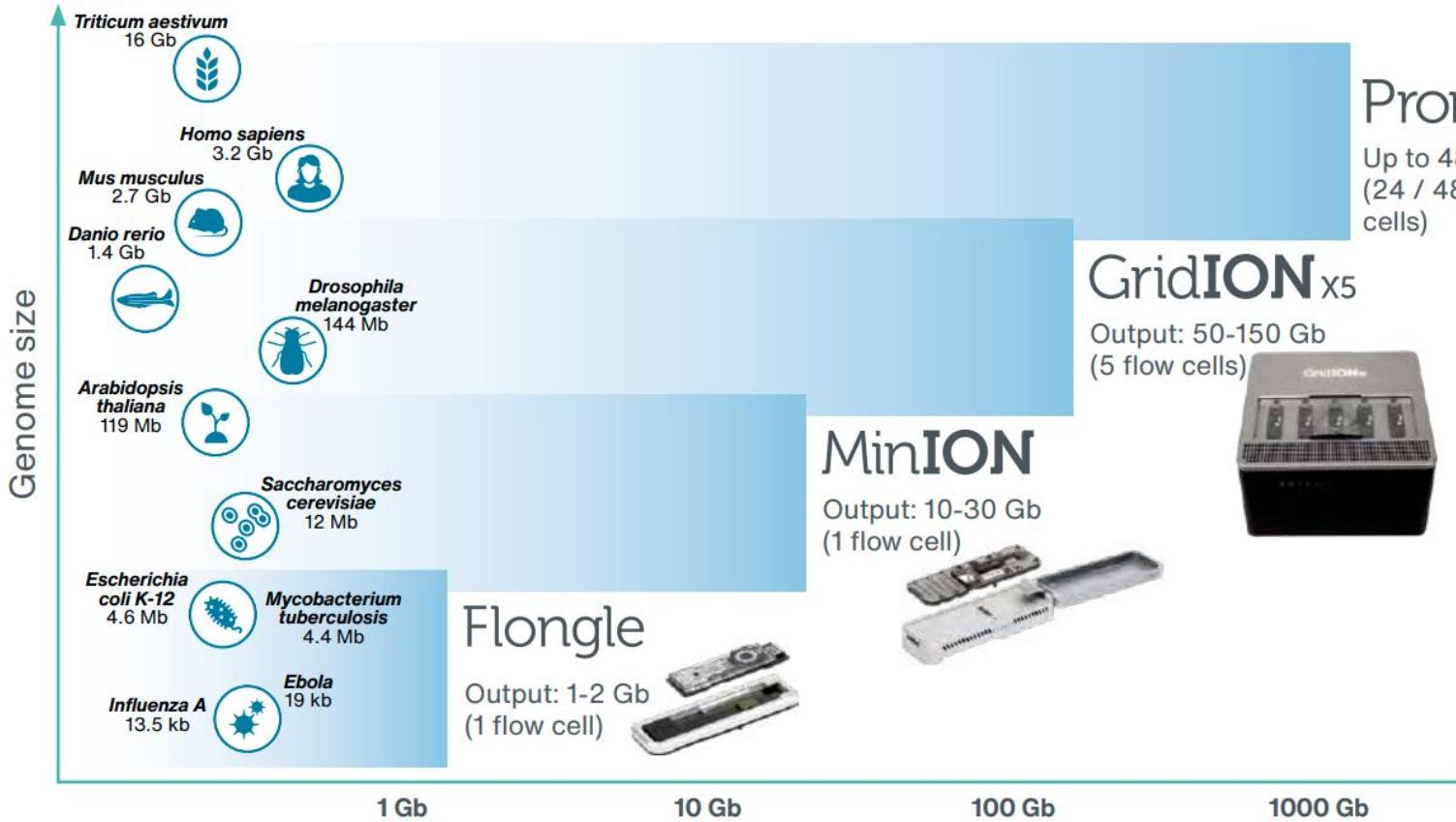
Oxford Nanopore Technology



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

- + Long reads 2-300 kb++ (record 4Mb!!)
- + Portability and sequencing speed
- Error rate (1-5% as compared to 0.5% for Illumina)
- Homopolymers in reads : Follow caller version updates !
- Some DNAs are harder to sequence because they do not go easily through the pores : Lab!

A lot of data !



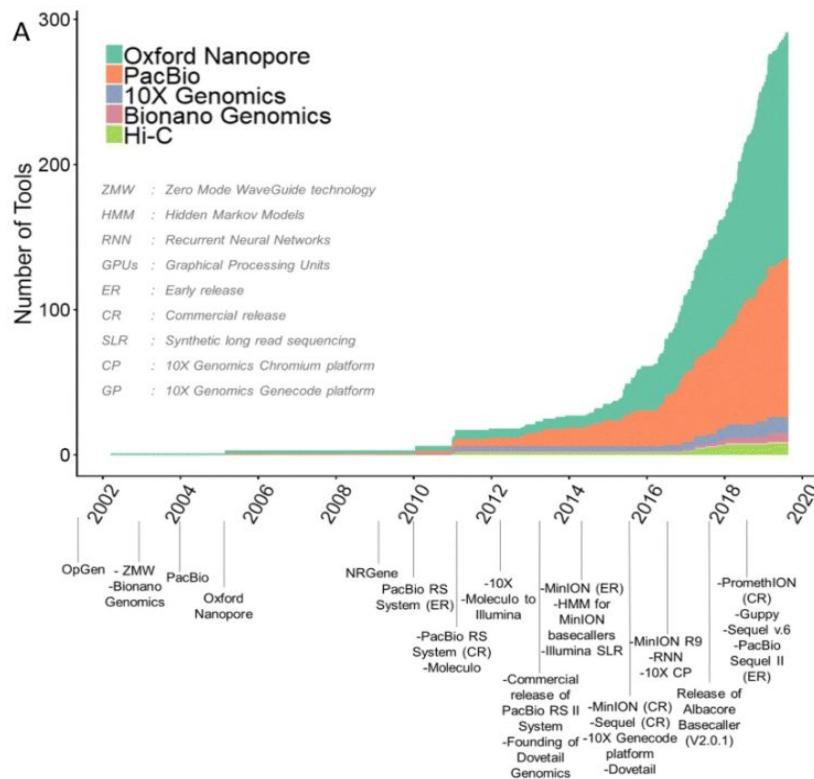
The data that these platforms produce differ qualitatively from second-generation sequencing, thus necessitating tailored analysis tools



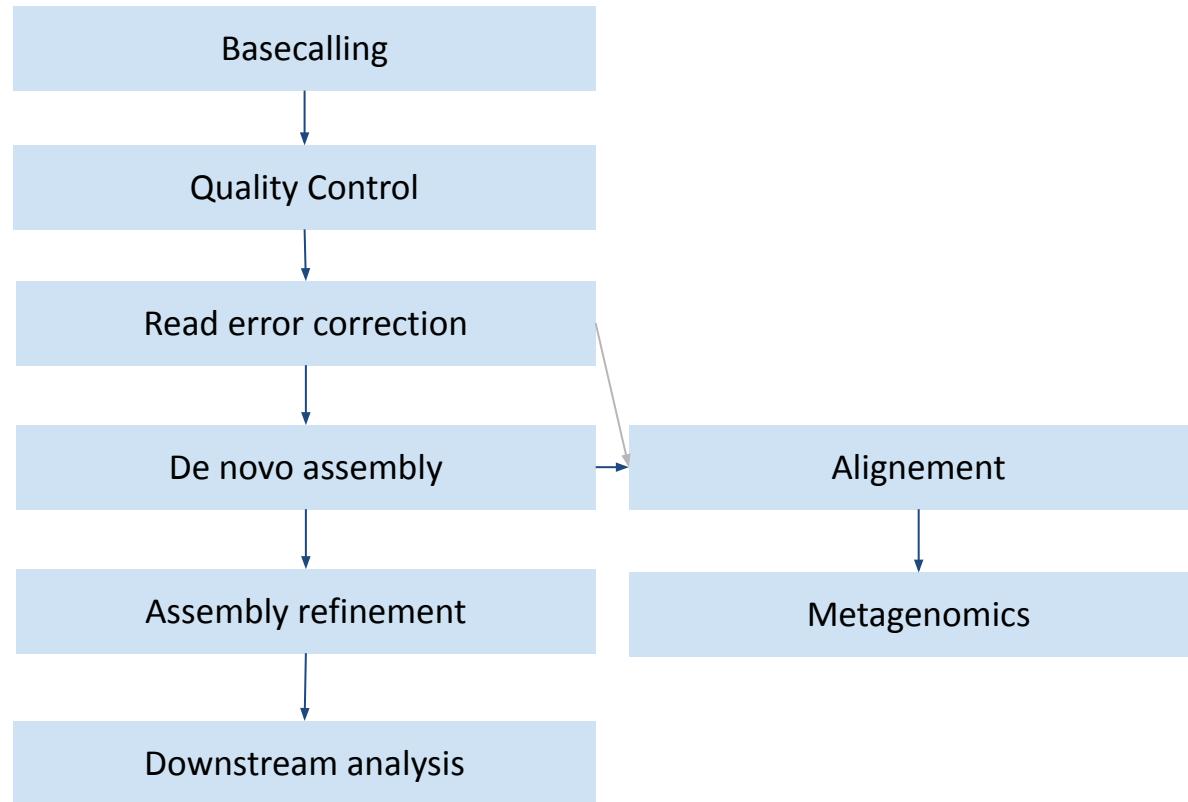
<https://long-read-tools.com>



A lot of tools are being developed and upgrade frequently !

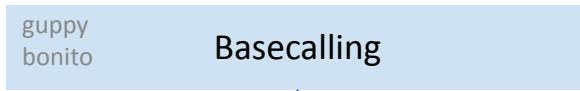


Typical long-read analysis pipelines for ONT data



Typical long-read analysis pipelines for ONT data

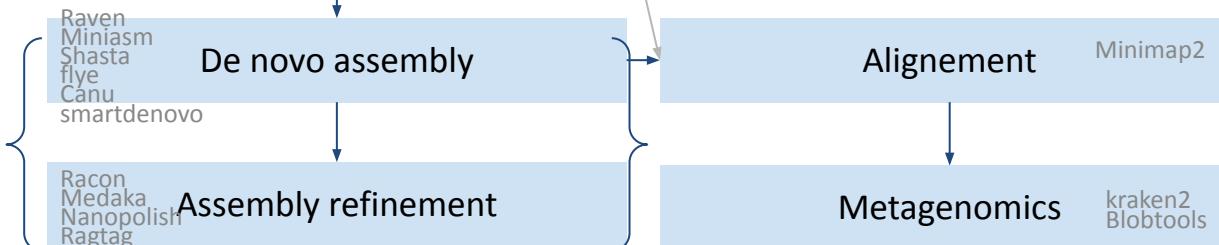
Demo



Practical 1

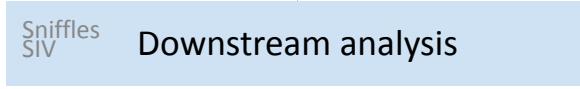


Practical 2



Practical 3

Practical 4

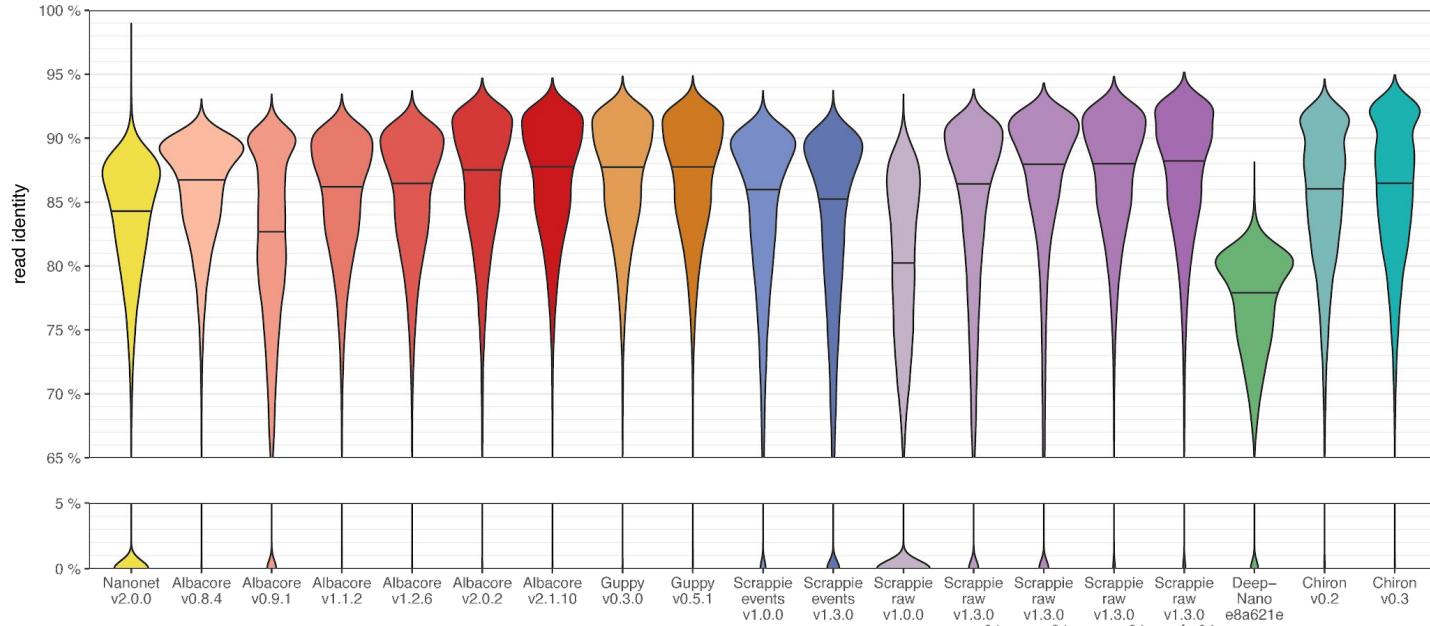


Reads Quality Control

ONT Read calling, cleaning and filtering

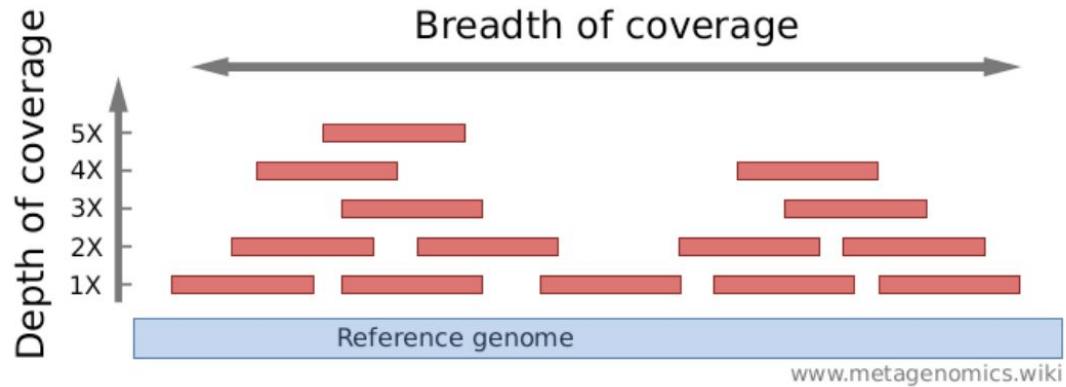
Sequencer ONT : raw fast5 files

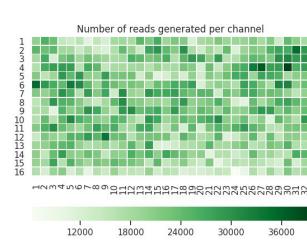
- Transform fast5 signal in fastq standard format *Guppy, Bonito*
- Optional Demultiplexing and removing adapters *Guppy options*
- Optional Quality filtering using the *sequencing_summary.txt* information : *Guppy options, filtlong, nanofilt*



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

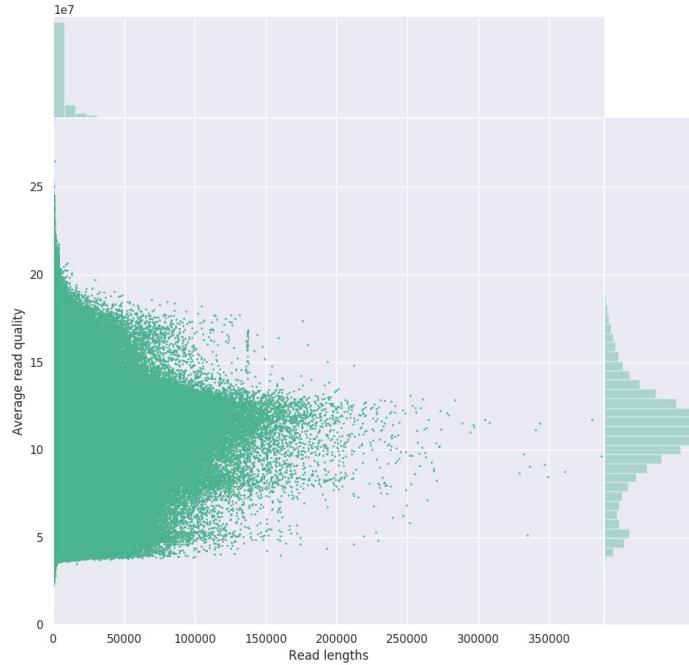
Compare estimated and obtained depth of coverage





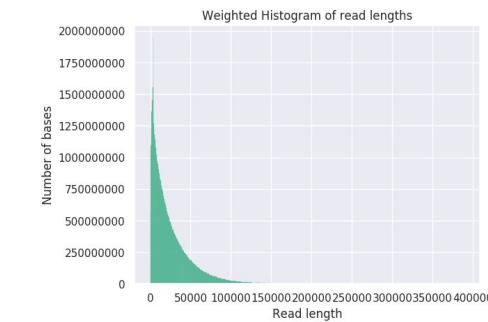
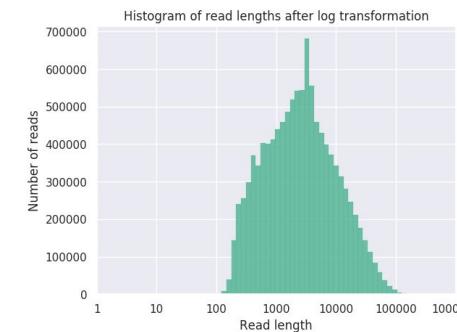
Reads Quality control : NanoPlot

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



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 !

Reads Correction or not?

Reads Correction process

Correction strategies

- External reads : Illumina
- Internal reads :
 - Long reads corrected by short ones
 - All versus all

Correction pipeline

- Read alignment
- Consensus calling

Assembly without reads correction

- Miniasm, Smartdenovo, Flye are members of this “new” family
- Improves speed (no correction).
- Can work with less read depth.
- Can also assemble corrected reads

	PacBio	ONT/PacBio	PacBio
	HGAP	CANU	Falcon
Read correction aligner	Read splitting for correction. blasr	MHAP	1 read per film selection daligner
Selection		40X of longest reads	Size selection criteria
Assembly of corrected reads	First steps of WGS-assembler + Specific module	WGS-assembler	First steps of WGS-assembler + Specific module

from Elixir GAAS 2018

TP1. Reads Quality Control

- TP1

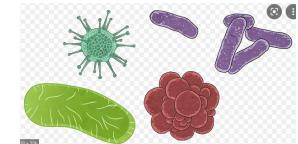
[https://github.com/SouthGreenPlatform/training_ONT_teaching/
blob/2021/1.raw_quality_control.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2021/1.raw_quality_control.ipynb)

Assemblies

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.

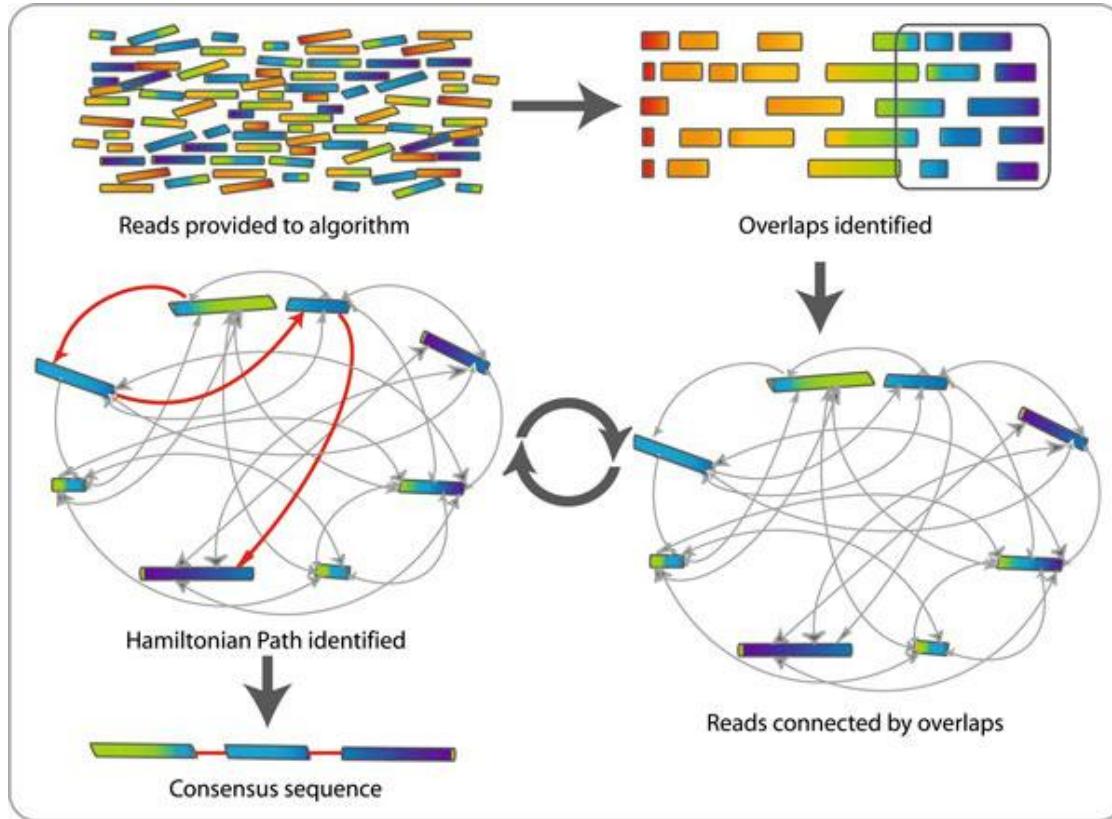
<https://nanoporetech.com/sites/default/files/s3/literature/microbial-genome-assembly-workflow.pdf>



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>

Overlap–layout–consensus genome assembly algorithm (OLC)



Overlap–layout–consensus genome assembly algorithm (OLC)

- [Canu](#) performs assembly by using corrected and trimmed reads. Its assembly strategy uses a overlap-layout-consensus (OLC) approach.
- [Flye](#) generates the concatenation of multiple disjoint genomic segments called disjointigs to build a repeat graph. Reads are mapped to this repeat graph to resolve conflicts (unbridged repeats) and output contigs.
- [Miniasm](#) performs the layout step of OLC. Read overlapping is done separately with Minimap2.
- [Raven](#) also uses an OLC approach. The overlapping step is similar to Miniasm and performed by Minimap2. The initial consensus step uses also Racon. Layout step removes spurious overlaps from the graph, improving contiguity.
- [Smartdenovo](#) performs read overlapping, rescues missing overlaps, identifies low-quality regions and produces a consensus.
- [Shasta](#) is a computationally efficient assembler, relying on a reduced representation of marker k-mers used to find overlaps and to build an assembly graph.

listes diff entre certains assembleurs : flye, shasta, etc,
temps ram (exemples), algo? params interessants
(--meta pour flye, polish round, ? => tableau

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.

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

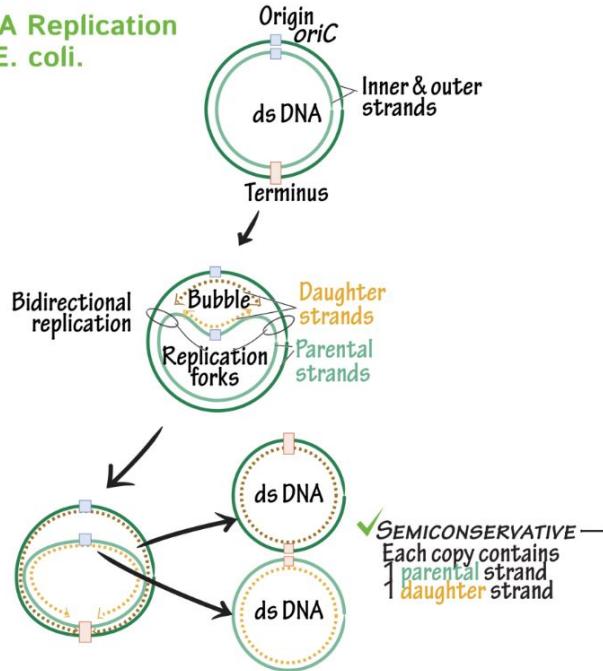
[**Pilon**](#) correct assemblies using illumina reads. (recursive use)

Autres : [NeuralPolish](#)



Circularisation ?

DNA Replication
in E. coli.



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

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

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

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

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

TP2. Assemblies

- TP2

[https://github.com/SouthGreenPlatform/training_ONT_teaching/
blob/2021/2.assemblies.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2021/2.assemblies.ipynb)

Contigs Quality

QUAST

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

26 March 2021, Friday, 07:37:40

[View in Icarus contig browser](#)

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

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

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

[Extended report](#)

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

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

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

less contigs : flye+racon puis raven+racon

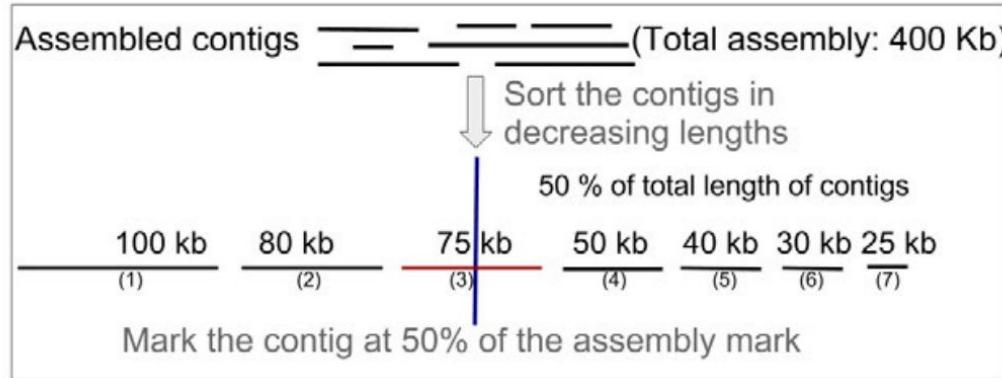
largest contig : flye+racon

largest N50 : flye

largest L50 : flye

what is N50 and L50?

What is N50 and L50?



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

QUAST

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

26 March 2021, Friday, 07:37:40

[View in Icarus contig browser](#)

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

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

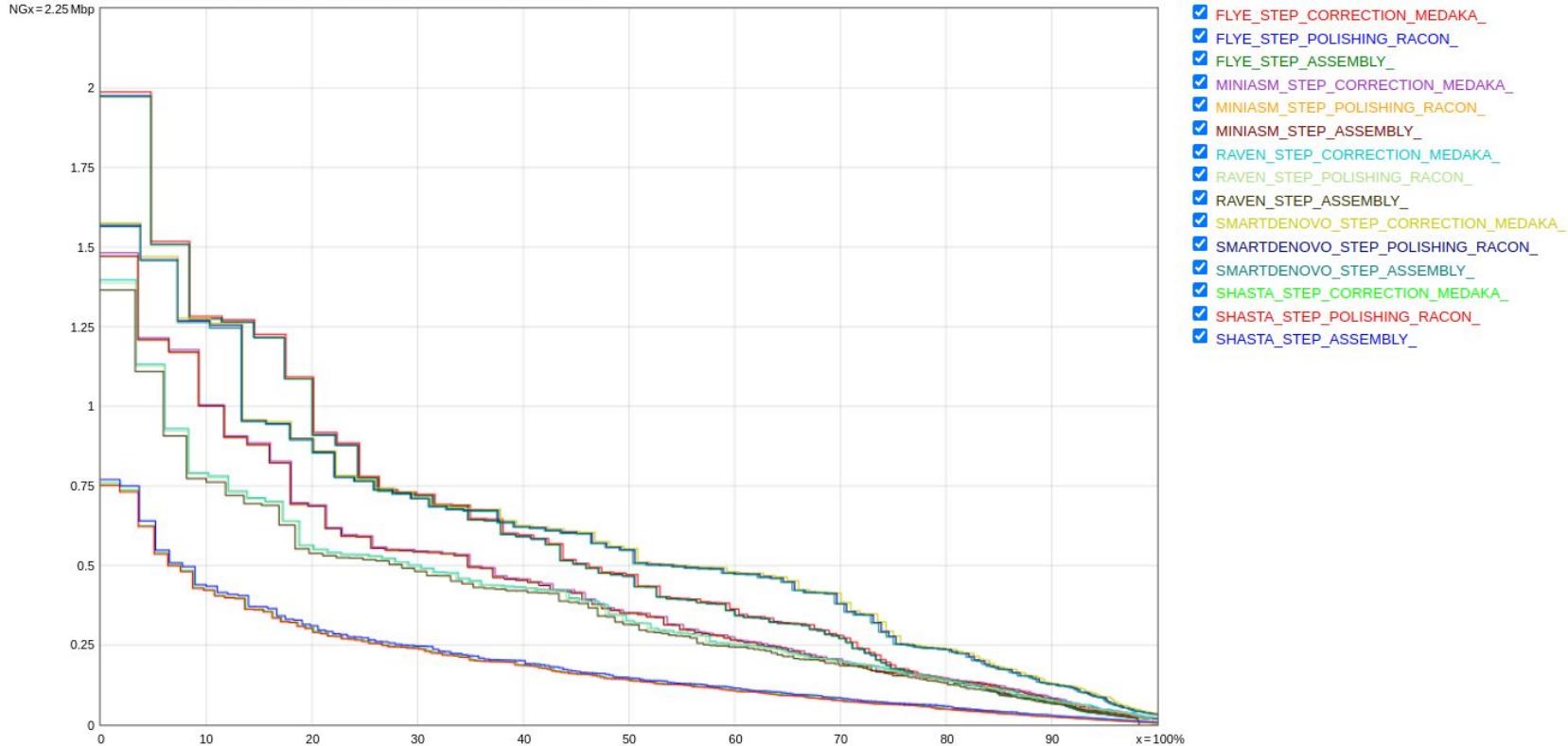
Worst	Median	Best	<input type="checkbox"/> Show heatmap
Genome statistics			
Genome fraction (%)	65.801	65.916	65.417
Duplication ratio	1.036	1.041	1.041
Largest alignment	2 503 013	2 501 477	1 739 590
Total aligned length	255 403 246	257 194 821	255 339 839
NGA50	48 559	48 062	42 714
LGA50	1338	1333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

[Extended report](#)

**Check misassemblies and N percentage.
BE CAREFUL! A misassembly for QUAST is a structural variation!**

Nx graph

Plots: Cumulative length Nx NAx NGx NGAx Misassemblies GC content



Assembly is better, greater is the area under curve. 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/
```



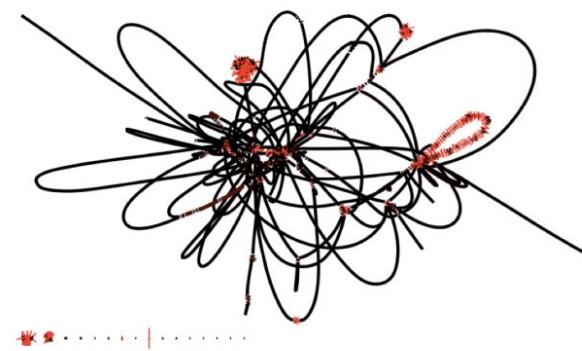
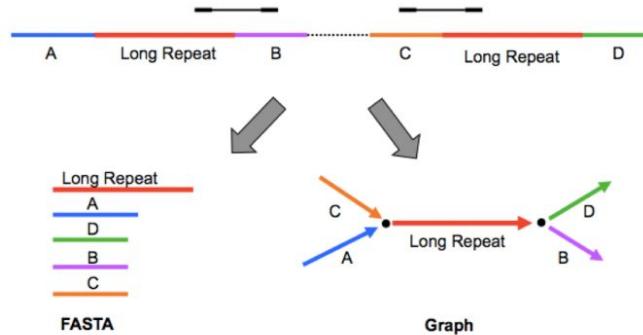
Bandage

a Bioinformatics Application for Navigating *De novo* Assembly Graphs Easily

Bandage is a tool for visualizing assembly graphs with connections.

You can zoom in to specific areas of the graph and interact with it by moving nodes, adding labels, changing colors and extracting sequences.

Several assemblers such Spades, Miniasm and Raven outputs the assembly graph in GFA format.



Read alignment statistics

Read congruency is an important measure in determining assembly accuracy.

Clusters of read pairs that align incorrectly are strong indicators of mis-assembly.

Comparaison 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/>

CANU

FLYE

MINIASM

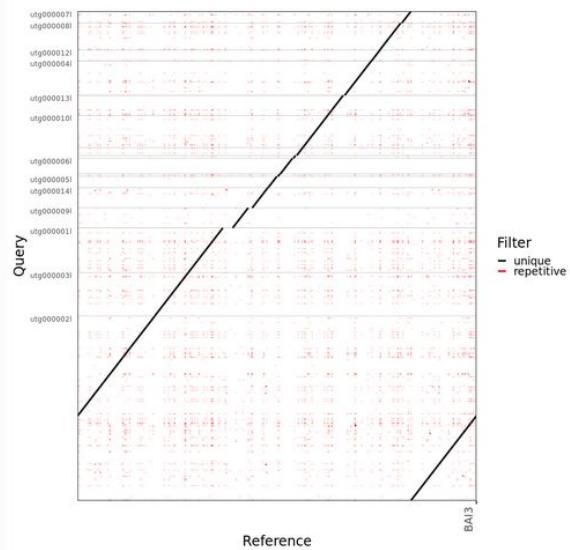
RAVEN

SMARTDENOVO

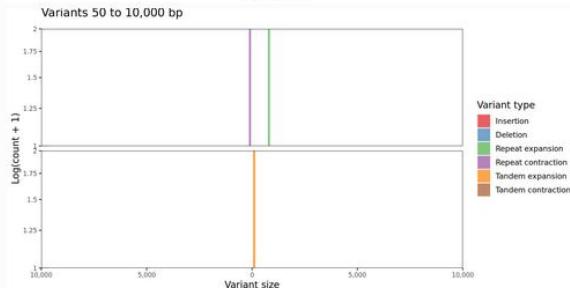
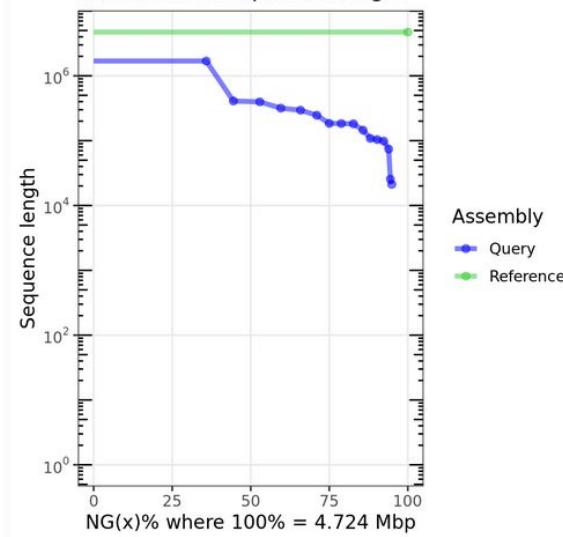
SHASTA

STEP_CORRECTION_NANOPOLISH_STARTFIXED

Dot plot of Assemblytics filtered alignments



Cumulative sequence length



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.

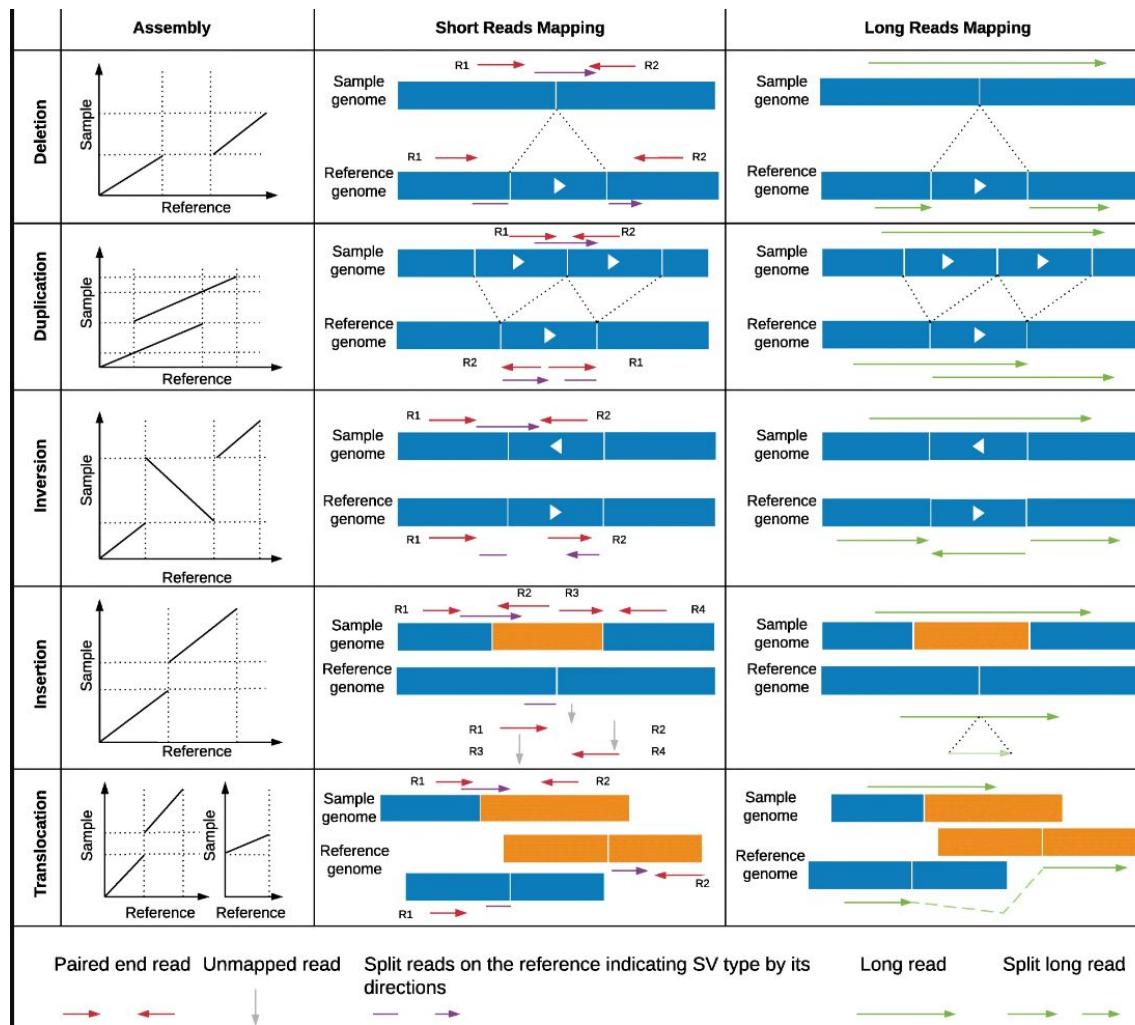
TP3. Contigs Quality

- TP3

[https://github.com/SouthGreenPlatform/training_ONT_teaching/
blob/2021/3.contigs_quality.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2021/3.contigs_quality.ipynb)

Variants Detection

Structural variant Detection



Paired end read 1:

mapped read Split

leads on the reference indicating SV type by its
relations

Long read

Split long reads



Variants Detection

- TP4

[https://github.com/SouthGreenPlatform/training_ONT_teaching/
blob/2021/4.variants_detection.ipynb](https://github.com/SouthGreenPlatform/training_ONT_teaching/blob/2021/4.variants_detection.ipynb)

Conclusions

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

Marketing moment for our tools

CulebrONT: a streamlined long reads
multi-assembler pipeline for prokaryotic and
eukaryotic genomes



Snakemake

Open-Source

Modulable

Scalable

Traceable

The 7 steps of the CulebrONT pipeline

Assemblies

Circularization

Quality

Polishing

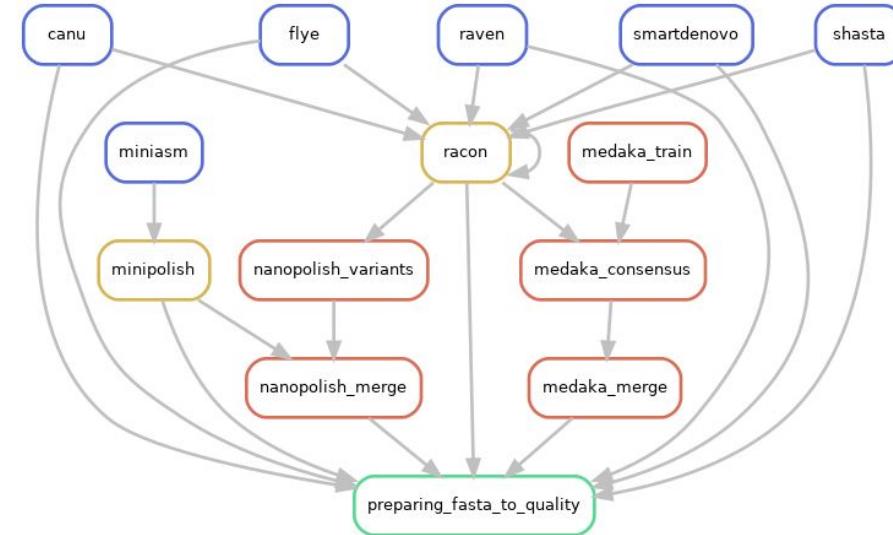
Reporting

Correction

Fixstart

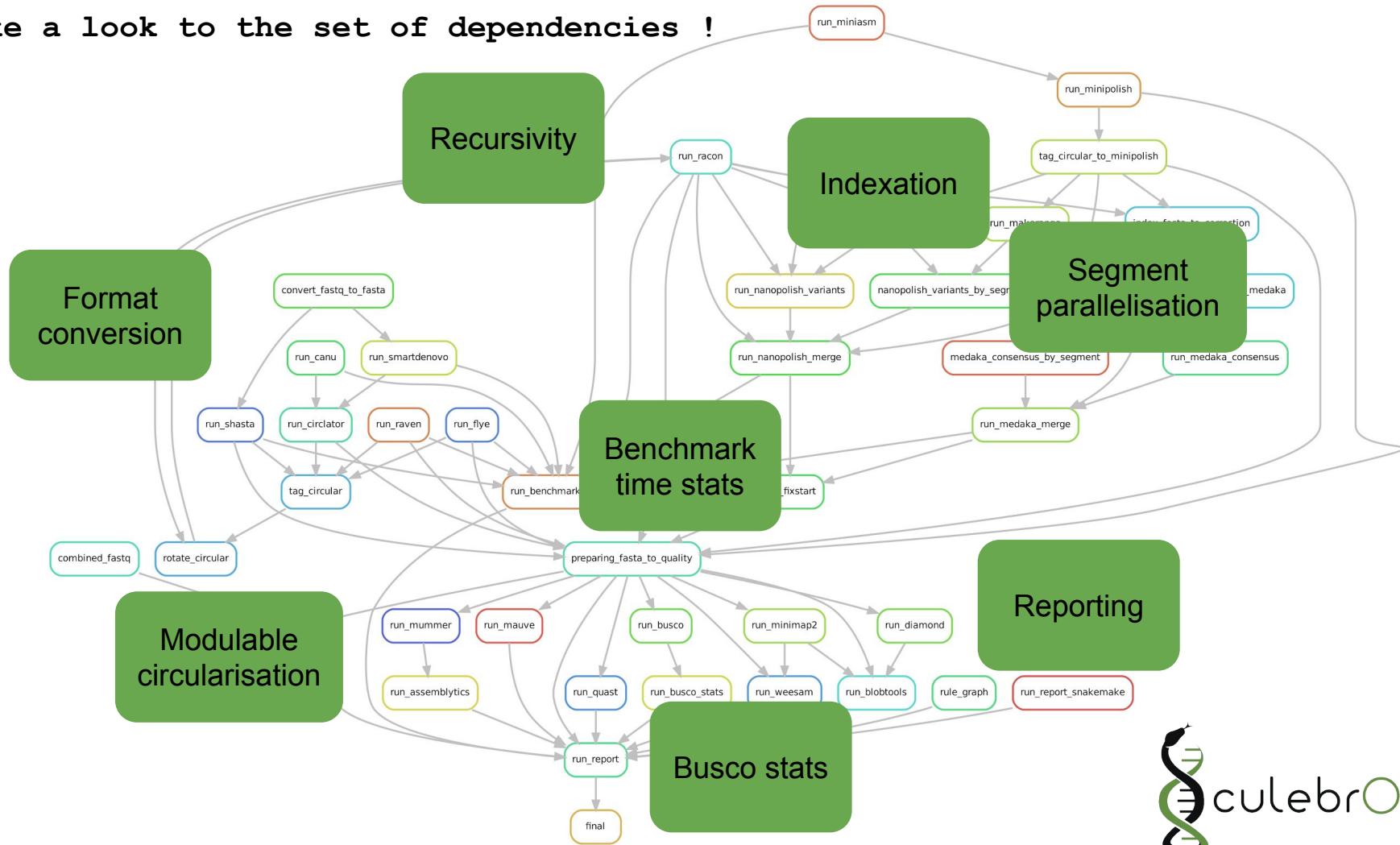


Building a workflow



RAVEN : True
SMARTDENOV : True
SHASTA : True
POLISHING :
RACON : True
CIRCULAR : True
CORRECTION :
NANOPOLISH : True

Take a look to the set of dependencies !



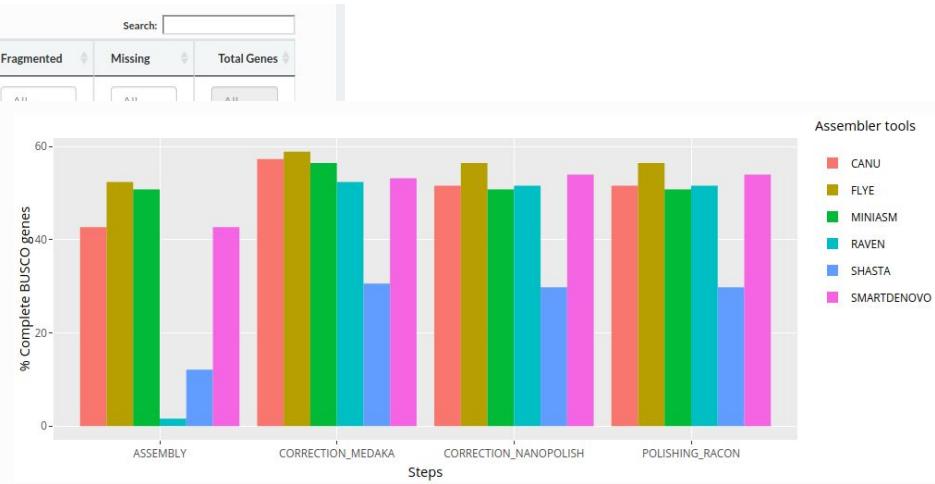
A nice html report !

Quality report

- Home
- Rulegraph
- Config file Parameters
- Benchmark
- Snakemake report
- BUSCO**
 - 6percentB1-1
 - 5percentB1-1
- QUAST

21 November, 2020

Assembler	Steps	Complete	Single	Duplicate	Fragmented	Missing	Total Genes
All	All	All	All	All	All	All	All
CANU	ASSEMBLY	42.7%	42.7%	0.0%			
CANU	CORRECTION_MEDAKA	57.3%	57.3%	0.0%			
CANU	CORRECTION_NANOPOLISH	51.6%	51.6%	0.0%			
CANU	POLISHING_RACON	51.6%	51.6%	0.0%			
FLYE	ASSEMBLY	52.4%	52.4%	0.0%			
FLYE	CORRECTION_MEDAKA	58.9%	58.9%	0.0%			
FLYE	CORRECTION_NANOPOLISH	56.5%	56.5%	0.0%			
FLYE	POLISHING_RACON	56.5%	56.5%	0.0%			
MINIASM	ASSEMBLY	50.8%	50.8%	0.0%			
MINIASM	CORRECTION_MEDAKA	56.5%	56.5%	0.0%			
MINIASM	CORRECTION_NANOPOLISH	50.8%	50.8%	0.0%			
MINIASM	POLISHING_RACON	50.8%	50.8%	0.0%	36.3%	12.9%	124
RAVEN	ASSEMBLY	1.6%	1.6%	0.0%	4.0%	94.4%	124
RAVEN	CORRECTION_MEDAKA	52.4%	52.4%	0.0%	37.1%	10.5%	124
RAVEN	CORRECTION_NANOPOLISH	51.6%	51.6%	0.0%	37.1%	11.3%	124
RAVEN	POLISHING_RACON	51.6%	51.6%	0.0%	37.1%	11.3%	124
SHASTA	ASSEMBLY	12.1%	12.1%	0.0%	13.7%	74.2%	124
SHASTA	CORRECTION_MEDAKA	30.6%	30.6%	0.0%	20.2%	49.2%	124
SHASTA	CORRECTION_NANOPOLISH	29.8%	29.8%	0.0%	17.7%	52.5%	124
SHASTA	POLISHING_RACON	29.8%	29.8%	0.0%	17.7%	52.5%	124
SMARTDENOVО	ASSEMBLY	42.7%	41.9%	0.8%	40.3%	17.0%	124
SMARTDENOVО	CORRECTION_MEDAKA	53.2%	53.2%	0.0%	37.1%	9.7%	124
SMARTDENOVО	CORRECTION_NANOPOLISH	54.0%	54.0%	0.0%	36.3%	9.7%	124



Completeness by orthology status of predicted genes : BUSCO

A nice html report !

file:///home/orjuela/Documents/2019/SNAKEMAKE/CULEBRONT/TMP/gitmerge/culebront_pipeline/output_xoo_sub_FIXSTARTCIRC-JUJU/FINAL_R ... ↻ ☆

CulebrONT report

QUAST

QUAST is a good starting point to help evaluate the quality of assemblies. It provides many helpful contiguity statistics.

6percentB1-1 5percentB1-1

[Open Quast report on new window](#)

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

21 November 2020, Saturday, 21:52:48
[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 "BAI3_Sanger" | 4 723 778 bp | 1 fragment | 63.87% G+C

Worst Median Best Show heatmap

Genome statistics	CANU_STEP_CORRECTION_NANOPOLI...	CANU_STEP_CORRECTION_MEDAKA_S...	CANU_STEP_POLISHING_I...
Genome fraction (%)	99.432	99.42	99.432
Duplication ratio	1.007	1.006	1.007
Largest alignment	1 279 326	1 280 264	1 279 326
Total aligned length	4 727 567	4 724 376	4 727 567
NGA50	1 150 846	1 151 785	1 150 846
LG50	2	2	2
Misassemblies			
# misassemblies	1	1	1
Misassembled contigs length	1 371 955	1 374 058	1 371 955
Mismatches			
# mismatches per 100 kbp	190.55	214.83	190.55
# indels per 100 kbp	197.15	183.29	197.15
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	11	11	11
Largest contig	1 371 955	1 374 058	1 371 955
Total length	4 781 737	4 789 824	4 781 737
Total length (≥ 1000 bp)	4 781 737	4 789 824	4 781 737
Total length (≥ 10000 bp)	4 781 737	4 789 824	4 781 737
Total length (≥ 50000 bp)	4 737 241	4 745 200	4 737 241

24 November, 2020

Contributors



Aurore COMTE Sébastien RAVEL Sébastien CUNNAC



Julie ORJUELA



Bao Tram VI



Florian CHARRIAT



François SABOT



culebrONT

documentation

<https://culebront-pipeline.readthedocs.io/en/latest/>

international seminary

<https://nanoporetech.com/events/nanopore-seminars-online-series>

publication

<https://www.biorxiv.org/content/10.1101/2021.07.19.452922v1.full.pdf>

Formateurs

- Julie Orjuela



- François Sabot



- Gautier Sarah



Merci pour votre attention !

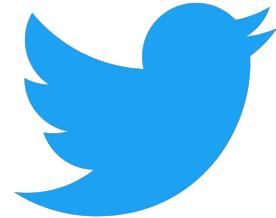


Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International:

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



SUIVEZ NOUS SUR TWITTER !



South Green : [@green_bioinfo](#)



i-Trop : [@ItropBioinfo](#)



N'oubliez pas de nous citer !

Comment citer les clusters?

"The authors acknowledge the IRD i-Trop HPC at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: <http://bioinfo.ird.fr/> "

"The authors acknowledge the CIRAD UMR-AGAP HPC (South Green Platform) at CIRAD montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: <http://www.southgreen.fr>"



Thanks!

Thanks to **i-Trop** IRD platform for support on docker VM creation and data stockage.

Special thanks to **IFB** for support and availability of VM on Biosphere !