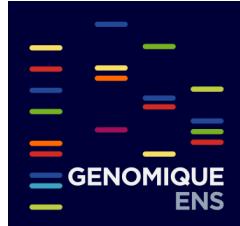


Formation à l'utilisation d'un séquenceur Oxford Nanopore

Session informatique



L'équipe de la plateforme génomique de l'IBENS



Catherine
Senamaud-
Beaufort



Corinne
Blugeon



Morgane
Thomas-
Chollier



Laurent
Jourdren



Sophie
Lemoine



Tiphaine
Marvillet



Oumy
Seidy



Stéphane
Le Crom



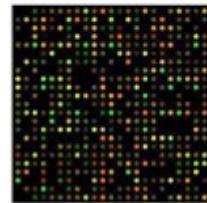
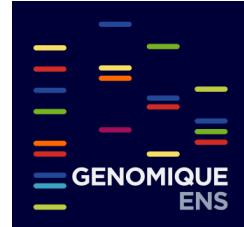
Ali
Hamraoui



Salomé
Brunon

La plateforme comprend des expérimentateurs et des bioinformaticiens. Nous prenons en charge les projets du contrôle des échantillons à l'analyse différentielle

Evolution technologique



Microarrays
(1998 - 2013)



Illumina HiSeq 1500
2nd generation sequencer
(2011 - 2016)



Illumina NextSeq 500
2nd generation sequencer
(2015 - 2023)



Illumina NextSeq 2000
2nd generation sequencer
(Since 2021)



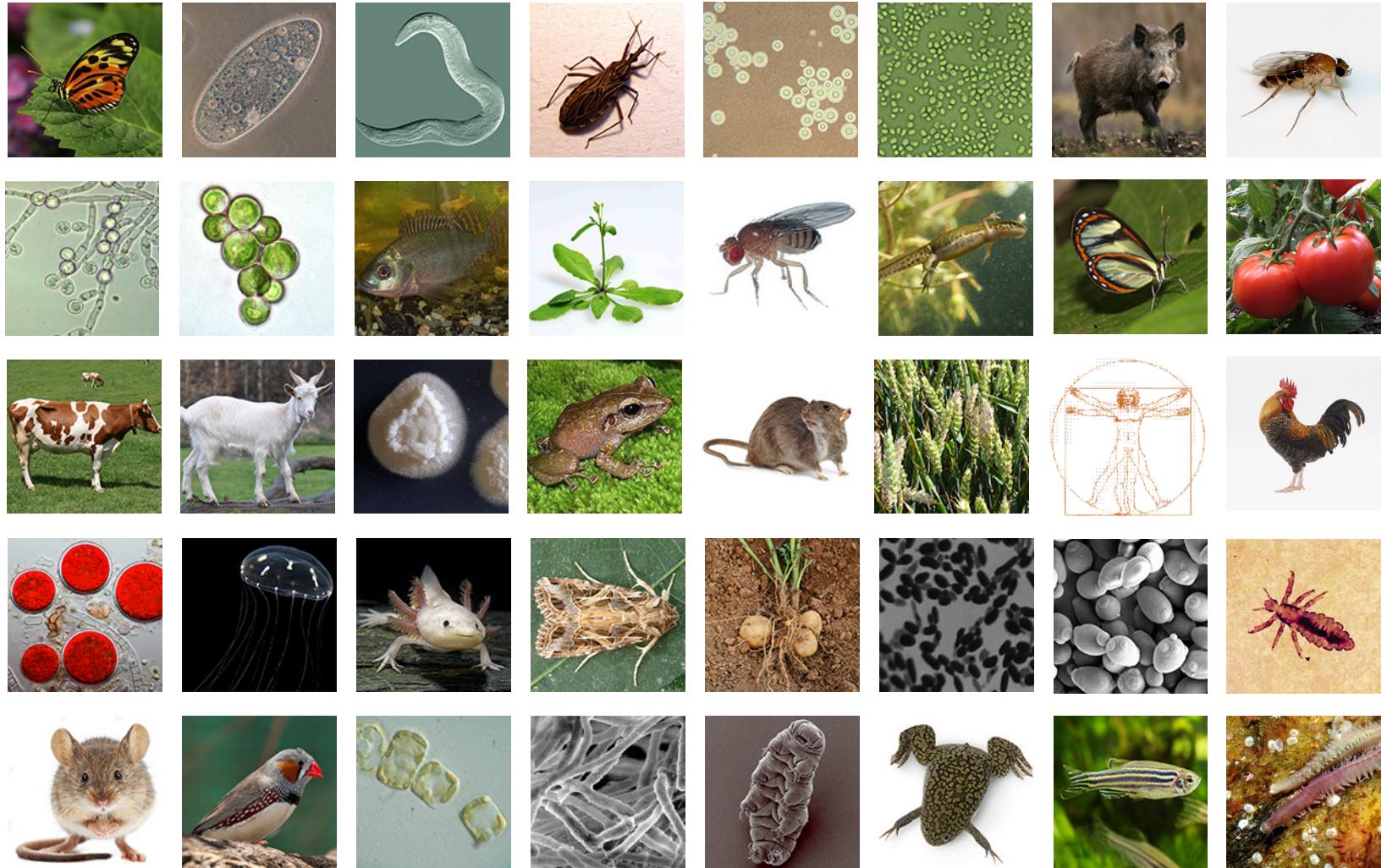
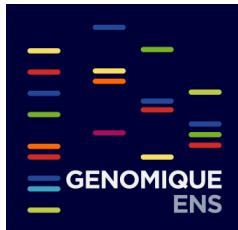
Nanopore MinION
3rd generation sequencer
(Since 2016)



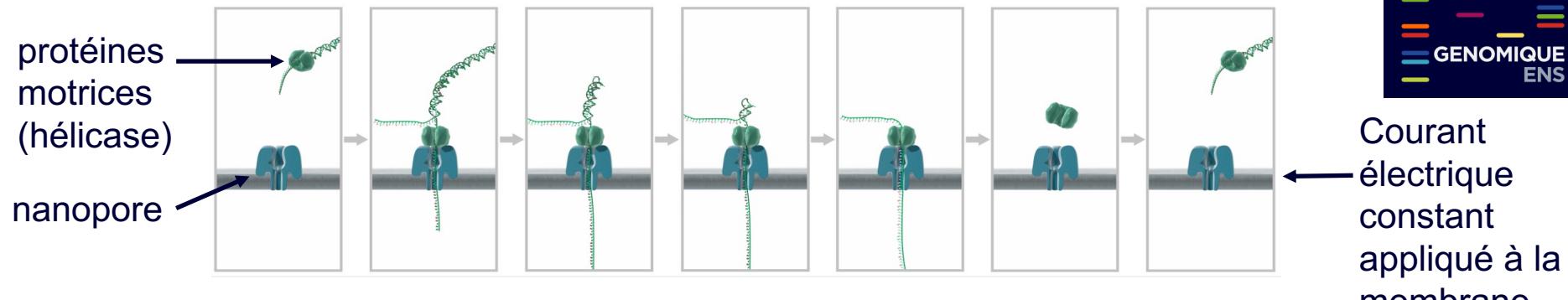
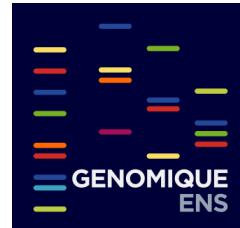
Nanopore PromethION
3rd generation sequencer
(Since 2022)

La plateforme existe depuis 1999 et a suivi le changement des technologies notamment en génomique fonctionnelle

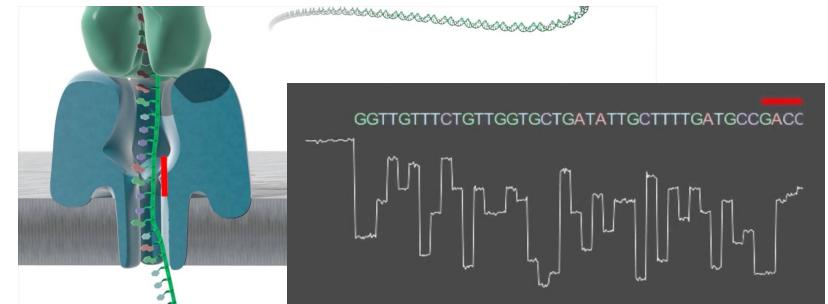
Spécialisée sur les organismes eucaryotes



Le séquençage sur nanopores



- Les variations du courant membranaire suite au passage la séquence au travers du pore sont évaluées et transcrites en bases



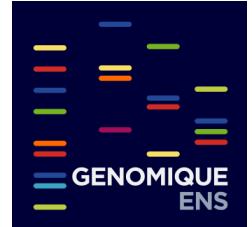
La séquence n'est alors pas déduite suite à une synthèse enzymatique
Pas de limitation de taille due à la technique

- Assemblage de génome simplifié
- ARN séquencés en pleine longueur

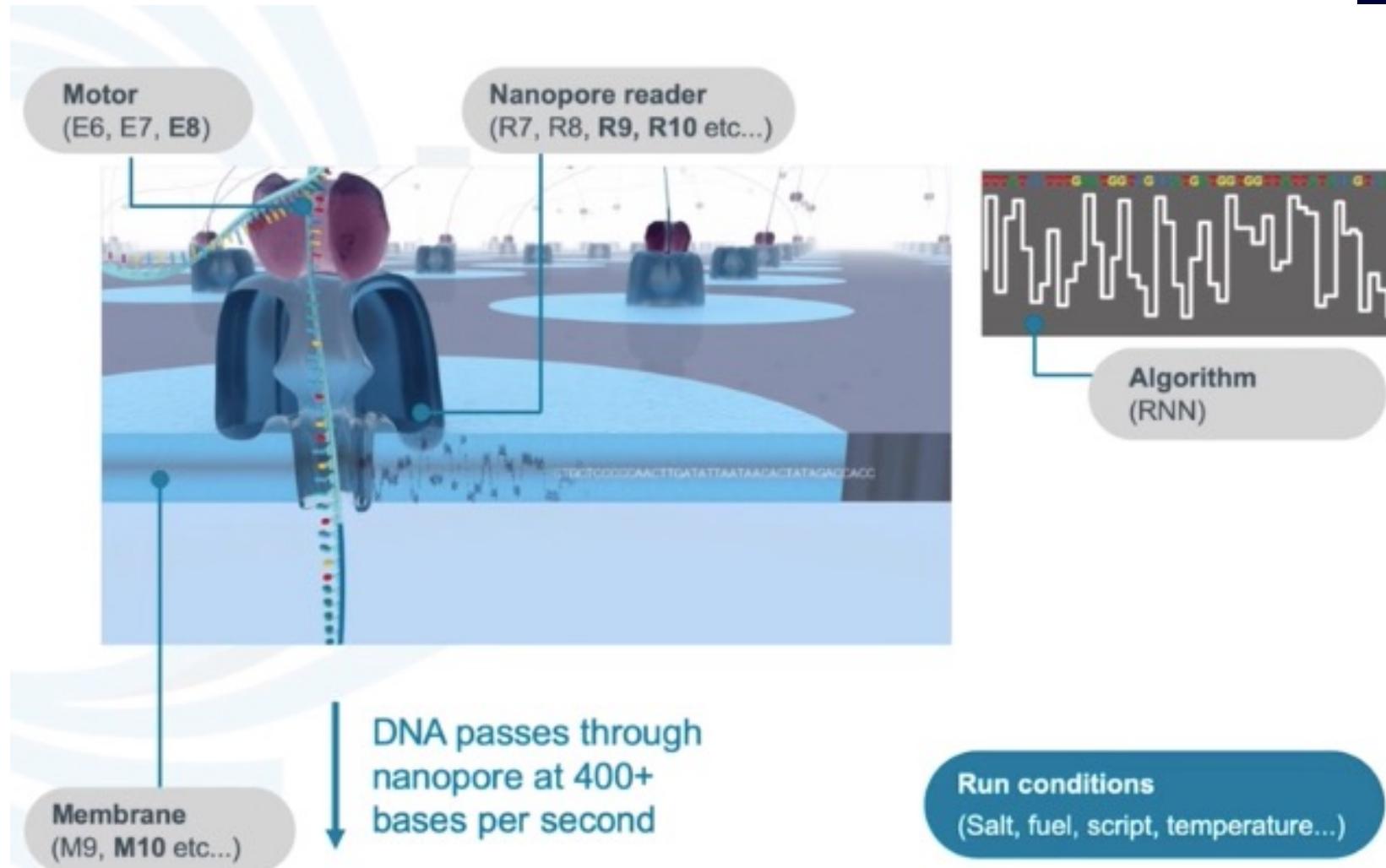
Possibilité de séquençage sans amplification de l'ADN et de l'ARN

- Détections de bases non canoniques possibles
- Algorithme d'appel de base à inventer...

<https://nanoporetech.com/>



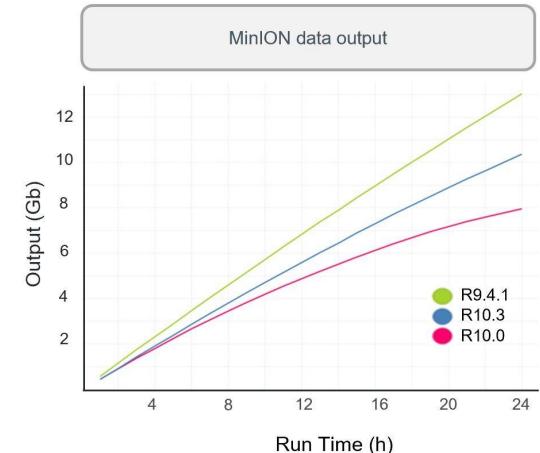
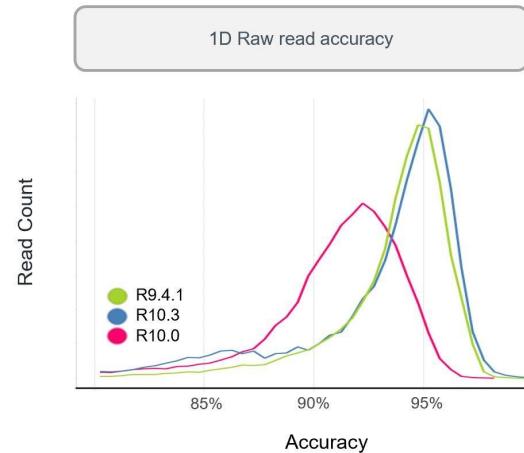
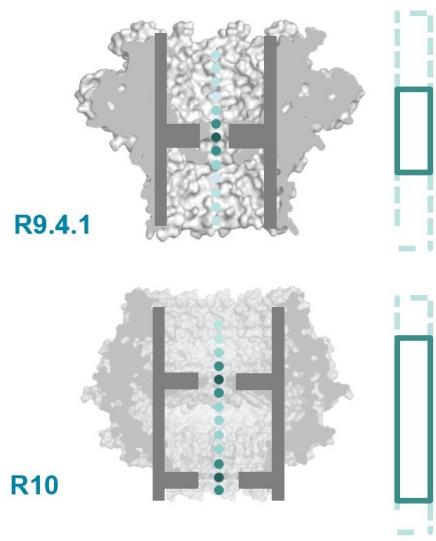
Les différents éléments d'un run



Pores types

There are **two types** of flowcell pores :

- **R9.4.1** with one reader head (01/2017, 13 Gb/24h, 95 % accuracy, 5-50 fmol)
- **R10.4.1** with two reader zones (03/2022, ?? Gb/24h, 99 % accuracy, ?? fmol)



Input Requirements:	R9.4.1	R10.3	R10.0
	5 – 50 fmol	25 – 75 fmol	50 – 100 fmol

Available Flowcells



	Flongle	MinION	PromethION
Max yield	2 GB	44 Gb	290 Gb
Channel count	126	512	2,675
Type	R10.4.1	R9.4.1 / R10.4 / RNA	R9.4.1 / R10.4.1 / RNA

ONT Sequencers

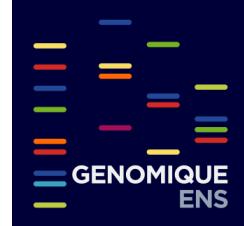


	MinION Mk1B	MinION Mk1d	GridION Mk1	PromethION	PromethION P2i	PromethION P2 Solo
Flowcell slots	1	1	5	24 or 48	2	2
Flowcell	Flongle/ MinION	Flongle/ MinION	Flongle/ MinION	PromethION	PromethION	PromethION
CPU	N/A	?	Intel i7	Intel CPU	1x Intel Core i7	N/A
GPU	N/A	?	1 x Nvidia GV100	4x NVIDIA Ampere-series GPU	1x NVIDIA Ampere-series GPU	N/A
Memory	N/A	?	64 GB	512 GB	64 GB DDR4	N/A
Storage	N/A	?	4 TB	60 TB	15 TB SSD	N/A
Support	Soon discontinued?	?	Full support	Full support	Full support	Full support

MinION/GridION/PromethION OS

- Oxford Nanopore device are powered by a **standard Ubuntu Linux** system with additional packages :
 - **MinKNOW** suite
 - **Nvidia tools** for GPU computing
 - **Docker** for running containerized applications
- ONT uses **Ubuntu 20.04** for GridION and PromethION.
- ONT gives users full administrative rights, so you can modify the operating system as you want. However its at your own risks!
- All low level administration tasks are performed in **command line** through a **SSH** connection.

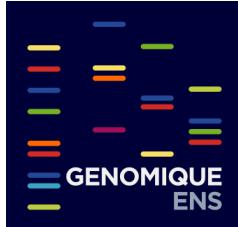




Utilisation de MinNOW

The screenshot shows the MinNOW UI interface. The top bar displays the title "MinNOW" and the date "24 mai 14:10". The left sidebar, titled "My device", contains links for "Start", "Sequencing overview", "Experiments", "System messages", and "Host settings". The main area features four large buttons: "Start sequencing" (green), "Analysis" (blue), "Flow cell check" (blue), and "Hardware check" (blue). A "More" button is located at the bottom of this section. The bottom left of the sidebar has a "Connection manager" link.

Section Sequencing overview



Start

Sequencing overview

Experiments

System messages

Host settings >

Connection manager

Sequencing overview

MinION Mk1B · MN17734 (Removable device) | PromethION · P2S-01171

MinION Mk1B · MN17734 (Removable device)

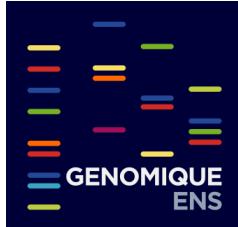
MN17734 FAK02081

Flow cell not checked
[Run error](#)

PromethION · P2S-01171

P2S-01171-A		Flow cell not checked	Run complete
P2S-01171-B		CTC85179	Flow cell not checked
			Run complete

Section Experiment



Screenshot of the NextFlow software interface showing the 'Experiments' section.

The sidebar on the left includes:

- Start
- Sequencing overview
- Experiments (selected)
- System messages
- Host settings

The main area displays:

Experiments (7)

Experiments active in the last 7 days.

Experiment details for Bidon_A2024 (Inactive):

- Reads: 0
- Estimated / basecalled bases: 0 b / 0 b
- Active runs: 0
- Total runs: 2

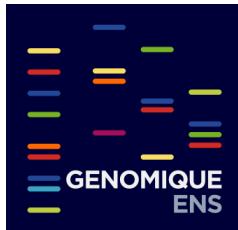
Run controls buttons: Start pore scan, Export run report, Refresh.

Experiment summary table:

Position	Flow cell ID	Sample ID	Health	Available pores	Run time	Run state	Reads	Estimated bases	Basecalled bases
MN17734	FAK02081	no_sample	—	—	3m 49s	Run error	0	0 b	—
P2S-01171-A	CTC84999	no_sample	—	—	0m 16s	Complete (no data)	0	0 b	—

- **Run statistics** : Nombre total de lectures, les bases estimées et appelées au cours d'un run, ainsi que le nombre de runs actifs et totaux.
- **Run controls** : important pour la mise en pause du run
- **Run report**
- **Experiment summary** : Accès aux graphes de suivi du run
- **Tableau de liste des runs**

Section Software : Host settings

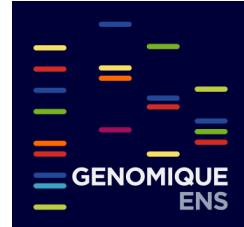


The screenshot shows the SKADI software interface under 'Host settings'. The left sidebar includes 'Back to main menu', 'Software' (which is selected), 'API access tokens', and 'Beta features'. The main area has a 'Software' heading. Under 'MinKNOW', it says 'Installed version: 24.02.10'. Below that is a 'Firmware' section listing: MinION FPGA 1.2.1, USB 1.2.7, P2 FPGA 2.1.0, and P2 USB-FW 2.5.1. To the right, a callout box titled 'Sub packages' lists: MinKNOW core: 5.9.7, Dorado: 7.3.9, Bream: 7.9.4, and Script Configuration: 5.9.12.

Section qui permet d'avoir une vue sur les versions des logiciels :

- La gestion du séquençage : MinKNOW
 - Dorado : Basecalling
 - Bream : FC check
- La communication séquenceur-FC-PC

Section Software : Beta features



SKADI Host settings Local user

Back to main menu Software API access tokens Beta features

Beta features

Beta features are early releases of new functionality. We are always looking to improve functionality, please provide any feedback on the Nanopore Community.

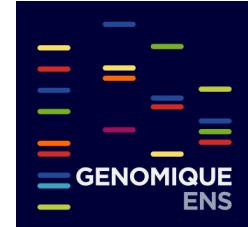
Beta features enabled

Barcode balancing adaptive sampling
Barcode balancing uses adaptive sampling to preferentially sequence underrepresented barcodes when they appear in your sample. This can increase the number of reads sequenced for these barcodes. However, in some cases, the overall data output for all reads may be reduced.

Remote transfer
Remote transfer allows transfer of data to online cloud storage locations. After a one-time setup, these locations become accessible as data output locations in run setup and from file manager.
We are actively seeking feedback on the performance of this feature in the wide range of real-world configurations, environments and use cases.

- Autorise MinKNOW à proposer ses nouveautés en cours de validation

Section Système message



Start

Sequencing overview

Experiments

System messages

Host settings >

System messages

i MinION Mk1B · MN17734 24 May, 14:26:28
Flow cell detected.

i PromethION · P2S-01171 · P2S-01171-A 24 May, 14:26:20
Sequencing device with flow cell ready.

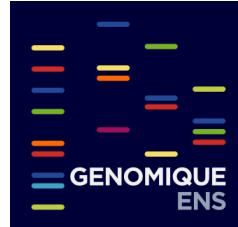
i PromethION · P2S-01171 · P2S-01171-B 24 May, 14:26:20
Sequencing device with flow cell ready.

! MinION Mk1B · MN17734 24 May, 14:20:47
Script failed. Please check for error messages below or restart the experiment. If the problem persists please contact support.

i MinION Mk1B · MN17734 24 May, 14:20:46
Flow cell disconnected.

i MinION Mk1B · MN17734 24 May, 14:17:17
Waiting up to 300 seconds for temperature to stabilise at 34.0°C

Section Connection manager



The screenshot shows the SKADI software interface. On the left, a sidebar menu includes 'Start', 'Sequencing overview', 'Experiments', 'System messages', and 'Host settings'. A large central area contains four main buttons: 'Start sequencing' (green), 'Analysis' (blue), 'Flow cell check' (blue), and 'Hardware check' (blue). Below these buttons is a 'More' button. At the bottom left of this central area is a small icon labeled 'Connection manager'. A thick black arrow points from this icon to a detailed view of the 'Connection manager' section on the right.

Connection manager

- + Add host
- ⟳ Refresh
- Filters
- ⋮

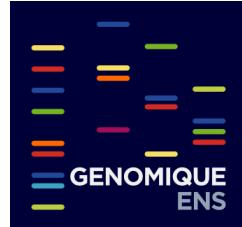
Saved hosts

My device

2 FREE | 0 RUNNING | 1 WITH ERROR

Une liste des séquenceurs disponibles sur le PC et leur état

Section Start



MinKNOW ▾ 24 mai 14:10 MinKNOW UI

MinKNOW UI View SKADI Local user

My device

- Start
- Sequencing overview
- Experiments
- System messages
- Host settings

- Connection manager

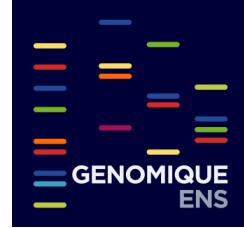
Start sequencing

Analysis

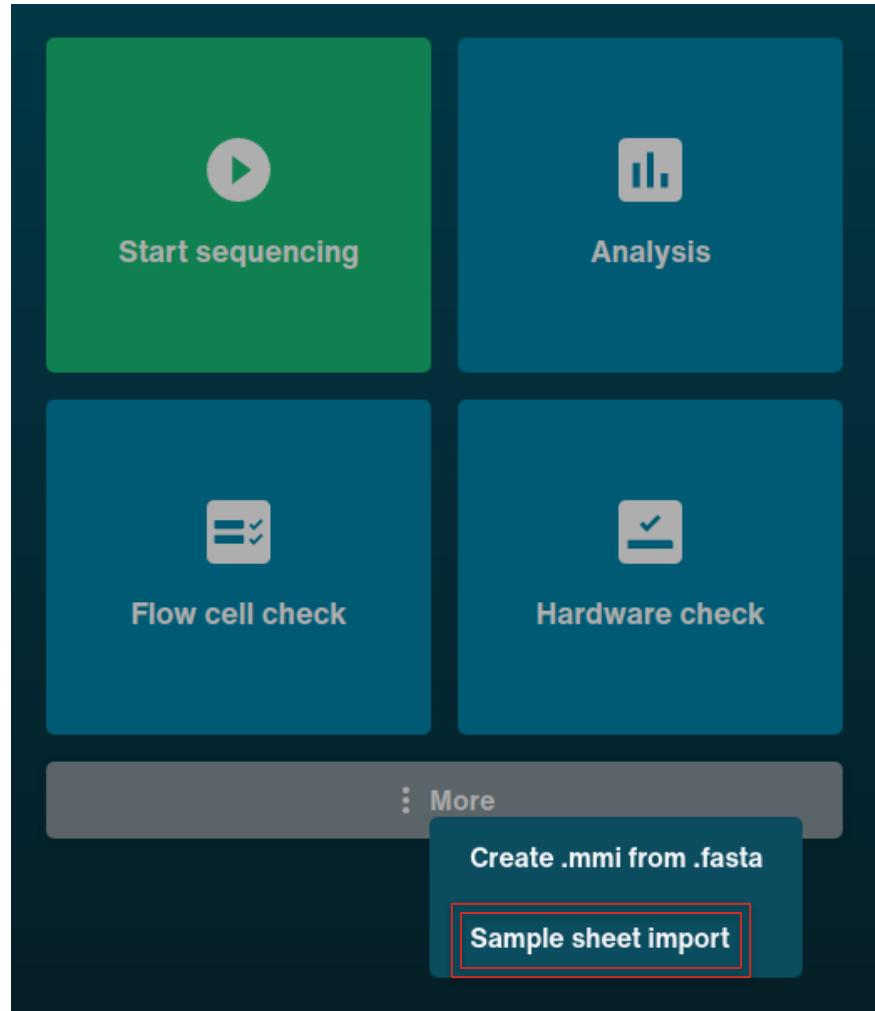
Flow cell check

Hardware check

⋮ More

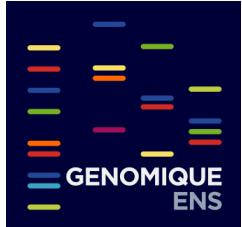


Demarrage du run avec une samplesheet et un fichier settings (format json)



Définition des feuilles de route : Samplesheet

- Permet de faire un lien entre des ID d'échantillons, les codes barres utilisés et une description des échantillons



Import a sample sheet CSV to start an experiment

Select a CSV file to import (this will overwrite the current sample sheet)

Make sure all flow cells you plan to start are inserted before importing the sample sheet.

No data loaded.

Import run settings Export run settings as template

Import a sample sheet CSV to start an experiment

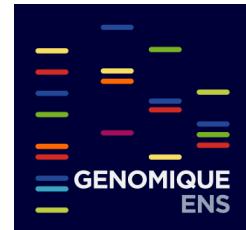
Select a CSV file to import (this will overwrite the current sample sheet)

/home/nanopore/Bureau/sample-sheet-MinION.csv

flow_cell_id	experiment_id	flow_cell_product_code	kit
FAK02081	Bidon_A2024	FLO-MIN106	SQK-PCB111-24

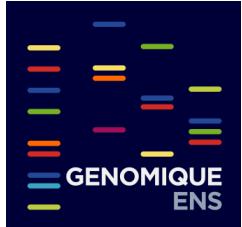
Import run settings Export run settings as template View applied run settings

Définition des feuilles de route : Samplesheet

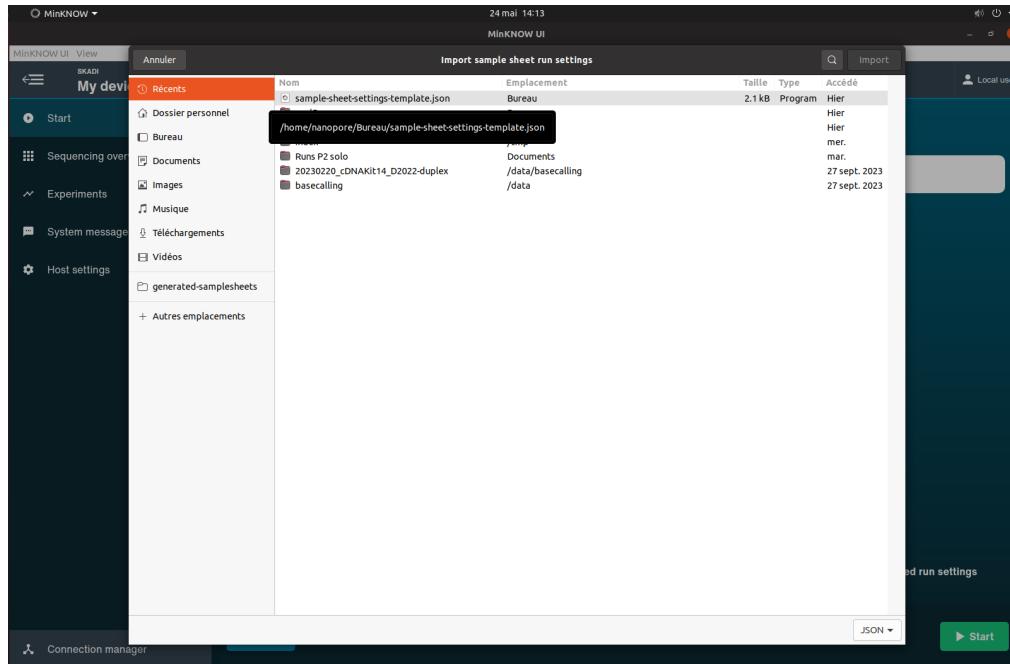


	A	B	C	D	E	F	G	H
1	position_id	flow_cell_product_code	kit	sample_id	experiment_id	barcode	alias	
2	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode01	Test1	
3	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode02	Test2	
4	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode03	Test3	
5	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode04	Test4	
6	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode05	Test5	
7	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode06	Test6	
8	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode07	Test7	
9	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode08	Test8	
10	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode09	Test9	
11	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode10	Test10	
12	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode11	Test11	
13	X1	FLO-MIN106	SQK-LSK109 EXP-NBD196	Test	Sample_sheet	barcode12	Test12	

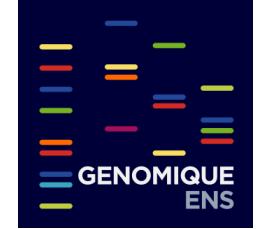
Définition des feuilles de route : Run settings



- Permet de refaire à l'identique un run déjà défini sans avoir à redéfinir manuellement toutes les étapes de MinKNOW
 - Fichier json : texte avec une architecture particulière

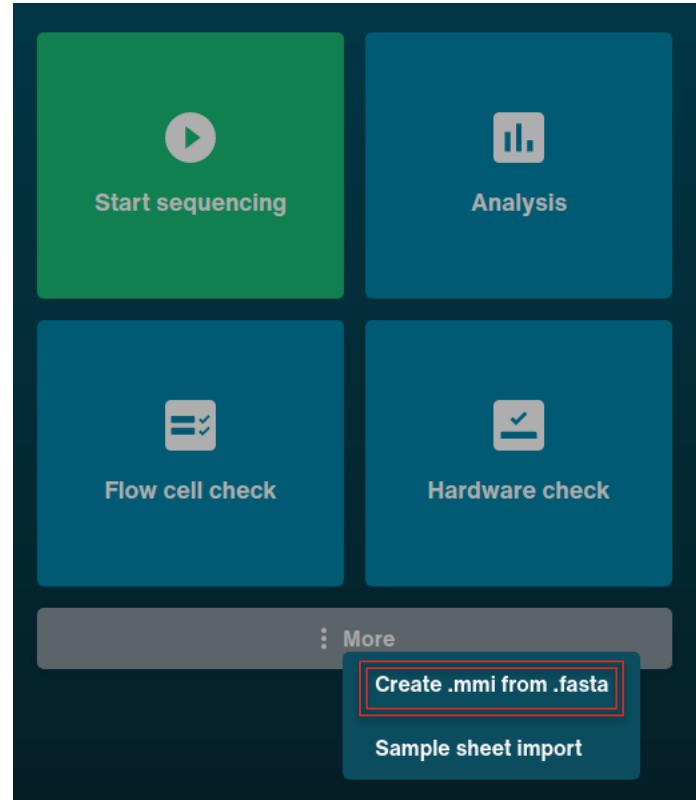
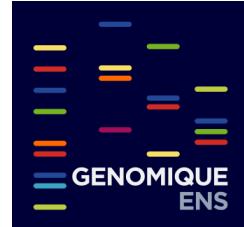


=> Voir le fichier samplesheet et le fichier json dans les documents de la formation

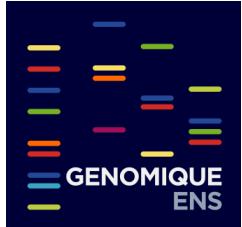


Création de l'index du génome

- Si l'on souhaite lancer l'alignement des séquences via MinKNOW en simultané ou après le séquençage



Création de l'index du génome



Create .mmi from .fasta

Select an input .fasta file

/home/nanopore/Bureau/hsapens105-chr1.fasta



Select a folder location for the output .mmi file

/home/nanopore/Bureau



Ready to process the input .fasta file to the selected folder as output file hsapens105-chr1.mmi. This will take an estimated 8 s and the reference includes an estimated 253.11 Mb.

Create .mmi from .fasta

Select an input .fasta file

/home/nanopore/Bureau/hsapens105-chr1.fasta



Select a folder location for the output .mmi file

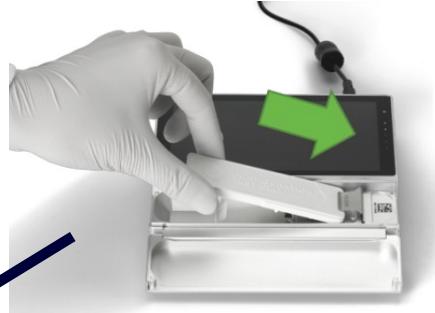
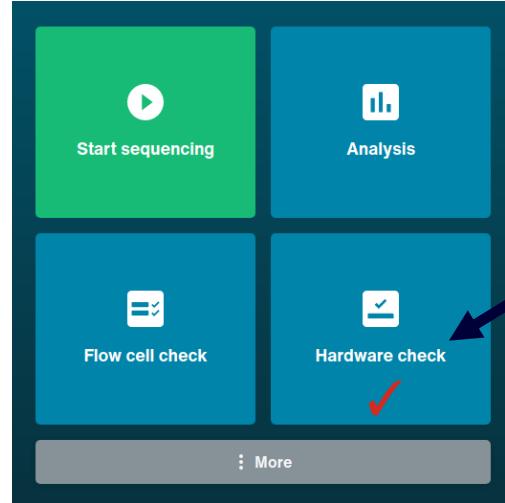
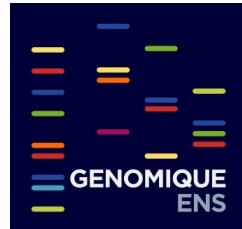
/home/nanopore/Bureau



.fasta file has been processed.

Vérification initiale du séquenceur

A la réception du séquenceur, test via la flowcell de configuration (CTC)



↓

Hardware check

MinION positions: MinION Mk1B · MN17734

PromethION positions: PromethION · P2S-01171

PromethION · P2S-01171

Select all available positions

Ready to start

Please ensure that the configuration test cell has been inserted correctly.

Sequencing overview

MinION Mk1B · MN17734 (Removable device) | PromethION · P2S-01171

MinION Mk1B · MN17734 (Removable device)

MN17734 FAK02081

Flow cell not checked

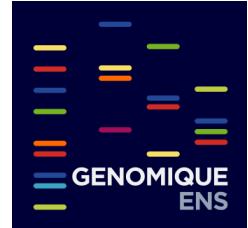
Run error

PromethION · P2S-01171

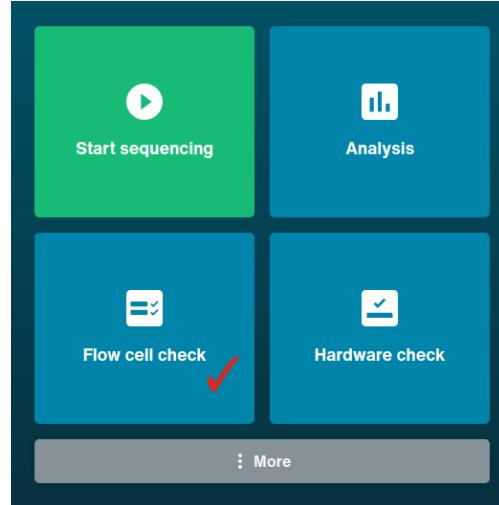
P2S-01171-A CTC84999 Flow cell not checked Run complete

P2S-01171-B CTC85179 Checking hardware

Vérification initiale de la flowcell avant de lancer le run



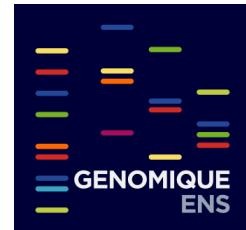
Il est nécessaire de vérifier le nombre de pores disponibles sur la flowcell avant de charger les échantillons



Le nombre de pores disponibles doit être supérieur à :

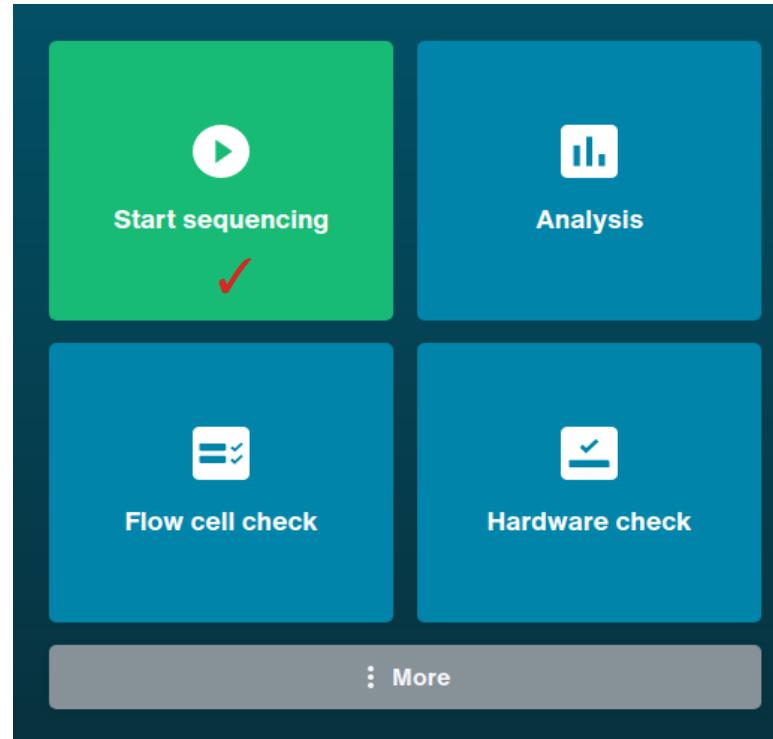
- 50 dans le cas d'une flowcell Flongle
 - 800 dans le cas d'une flowcell MinION
 - 5000 dans le cas d'une flowcell PromethION
- ⇒ Elles sont remplacées si ce n'est pas le cas !

Paramétrage et lancement d'un run

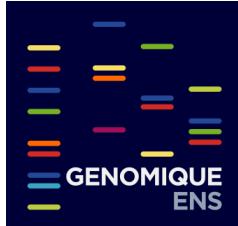


Il faut définir

- Votre expérience
- Le kit utilisé
- Le basecalling (à la volée ou après de run dans la section Analysis)
- Les format de sortie de vos données
- L'alignement (à la volée ou après) et par conséquent, les séquences références



Définition de l'expérience



1. Positions 2. Kit 3. Run options 4. Analysis 5. Output 6. Review

Positions

Experiment details

Ensure the experiment name does not contain any personally-identifiable information.

formationONT

+ Join existing Load settings from template ⋮

Device positions

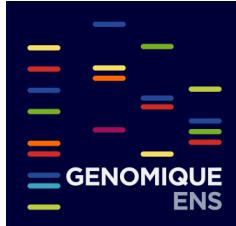
MinION positions: MinION Mk1B · MN17734

No positions available to select
All positions are either currently calibrating, running or do not have a flow cell attached.

PromethION positions: PromethION · P2S-01171

* Missing required values Continue > Skip to final review »

Définition de l'expérience



1. Positions 2. Kit 3. Run options 4. Analysis 5. Output 6. Review

Position

Experiment

Device

1. Positions 2. Kit 3. Run options 4. Analysis 5. Output 6. Review

Positions

Experiment details

Experiment name *

Device positions

MinION positions: MinION Mk1B · MN17734

PromethION positions: PromethION · P2S-01171

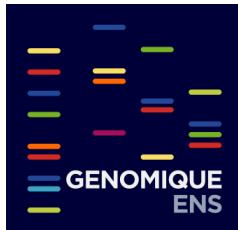
PromethION · P2S-01171

Select all available positions

No positions selected
You must select at least 1 position to start an experiment.

◀ Back * Missing required values Continue ➤ Skip to final review ➞

Définition de l'expérience



1. Positions 2. Kit 3. Run options 4. Analysis 5. Output 6. Review

Positions

Experiment details

Experiment name *
FormationONT + Join existing Load settings from template ⋮

Device positions

MinION positions: MinION Mk1B · MN17734

PromethION positions: PromethION · P2S-01171

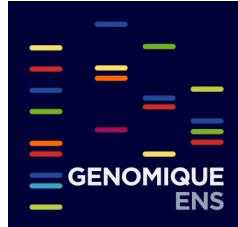
PromethION · P2S-01171
[Select all available positions](#)

Position	Flow cell ID	Flow cell type	Sample ID
P2S-01171-A	PAO42194	FLO-PRO002	FormationONT_Sample

◀ Back ◀ Back ◀ Back Continue > Skip to final review ➞

Choix du kit de séquençage

Tous les kits sont disponibles, il est possible de les filtrer selon ce que l'on séquence, selon les banques faites...



1. Positions 2. Kit 3. Run options 4. Analysis 5. Output 6. Review

Kit selection

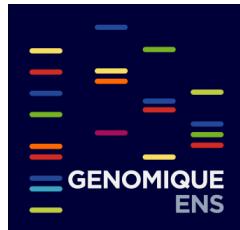
Sample type: DNA RNA PCR-free: PCR PCR-free Multiplexing: Yes No Control: Filtre  Reset filters

Ligation Sequencing Kit SQK-LSK109	<input type="radio"/>	Ligation Sequencing Kit SQK-LSK110	<input type="radio"/>	Ligation Sequencing Kit XL SQK-LSK110-XL	<input type="radio"/>	Rapid Barcoding Kit SQK-RBK004	<input type="radio"/>
Rapid Barcoding Kit 96 SQK-RBK110-96	<input type="radio"/>	Direct RNA Sequencing Kit SQK-RNA002	<input type="radio"/>	16S Barcoding Kit (BC1-24) SQK-16S024	<input type="radio"/>	CAS109 Sequencing Kit SQK-CS9109	<input type="radio"/>
Direct cDNA Sequencing Kit SQK-DCS109	<input type="radio"/>	Field Sequencing Kit SQK-LRK001	<input type="radio"/>	Ligation Sequencing Kit (48 reactions) SQK-LSK109-XL	<input type="radio"/>	Ligation Sequencing Kit SQK-LSK112	<input type="radio"/>
Ligation Sequencing Kit XL SQK-LSK112-XL	<input type="radio"/>	Multiplex Ligation Sequencing Kit XL SQK-MLK111-96-XL	<input type="radio"/>	Native Barcoding Sequencing Kit 24 SQK-NBD112-24	<input type="radio"/>	Native Barcoding Sequencing Kit 96 SQK-NBD112-96	<input type="radio"/>
PCR Barcoding Kit SQK-PBK004	<input type="radio"/>	PCR cDNA Barcoding Kits SQK-PCB109	<input type="radio"/>	PCR cDNA Barcoding Kit - 24 SQK-PCB111-24	<input type="radio"/>	PCR-cDNA Sequencing Kit SQK-PCS109	<input type="radio"/>
PCR cDNA Sequencing Kit SQK-PCS111	<input type="radio"/>	PCR Sequencing Kit SQK-PSK004	<input type="radio"/>	16S Barcoding Kit SQK-RAB204	<input type="radio"/>	Rapid Sequencing Kit SQK-RAD004	<input type="radio"/>
Rapid Sequencing Kit	<input type="radio"/>	Rapid PCR Barcoding Kit	<input type="radio"/>	Ultra-Long DNA Sequencing Kit	<input type="radio"/>	VolTRAX PCR Tiling 1-12 COVID-	<input type="radio"/>

< Back Continue > Skip to final review >

Choix du kit de séquençage

Tous les kits sont disponibles, il est possible de les filtrer selon ce que l'on séquence, selon les banques faites...



1. Positions 2. Kit 3. Run options 4. Analysis 5. Output 6. Review

Kit selection

Sample type: DNA RNA

PCR-free: PCR PCR-free

Multiplexing: Yes No

Control

[Reset filters](#)

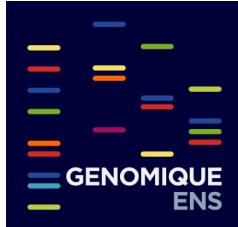
Ligation Sequencing Kit SQK-LSK109	<input type="radio"/>	Ligation Sequencing Kit SQK-LSK110 <input checked="" type="radio"/>	Ligation Sequencing Kit XL SQK-LSK110-XL <input type="radio"/>	Rapid Barcoding Kit SQK-RBK004 <input type="radio"/>
Rapid Barcoding Kit 96 SQK-RBK110-96	<input type="radio"/>	Direct RNA Sequencing Kit SQK-RNA002 <input type="radio"/>	16S Barcoding Kit (BC1-24) SQK-16S024 <input type="radio"/>	CAS109 Sequencing Kit SQK-CS9109 <input type="radio"/>
Direct cDNA Sequencing Kit SQK-DCS109	<input type="radio"/>	Field Sequencing Kit SQK-LRK001 <input type="radio"/>	Ligation Sequencing Kit (48 reactions) SQK-LSK109-XL <input type="radio"/>	Ligation Sequencing Kit SQK-LSK112 <input type="radio"/>
Ligation Sequencing Kit SQK-LSK112-XL	<input type="radio"/>	Multiple Ligation Sequencing Kit YI <input type="radio"/>	Native Barcoding Sequencing Kit 24 <input type="radio"/>	Native Barcoding Sequencing Kit 96 <input type="radio"/>

Select expansion pack

Native Barcoding Expansion 1-12 (PCR-free) EXP-NBD104 <input checked="" type="radio"/>	Native Barcoding Expansion 13-24 (PCR-free) EXP-NBD114 <input type="radio"/>	Native Barcoding Expansion 96 EXP-NBD196 <input type="radio"/>	PCR Barcoding Expansion 1-12 EXP-PBC001 <input type="radio"/>
PCR Barcoding Expansion 1-96 EXP-PBC096 <input type="radio"/>			

[< Back](#) [< Back](#) [Continue >](#) [Skip to final review >>](#)

Configuration des options du run



1. Positions 2. Kit 3. Run options **4. Analysis** 5. Output 6. Review

Run options

Run until ⓘ

Run limit: 72 hours duration
Flow cell data target: None set

Options

Minimum read length ⓘ

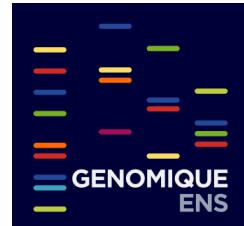
20 bp 200 bp 1000 bp

Adaptive sampling ⓘ

Enrich sequences
 Deplete sequences
 Barcode balancing Beta

Show advanced user options >

Back Continue > Skip to final review »



Les conditions d'arrêt normal du run :

- La durée
- Le volume de données produit
- Le nombre de pores à utiliser

Run options

Run until ⓘ

Run limit: 72 hours duration

Flow cell data target: None set

Options

Run until defines the rules that will cause a run to stop.

Rules can be based on time, target data, or flow cell pores.

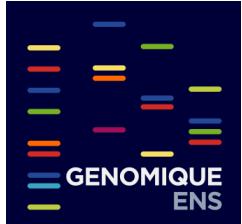
All runs must have a run limit.

< Back

Continue >

Skip to final review »

Les conditions d'arrêt normal du run : La durée



Run until

Run limit

Action

Stop run when

Flow cell condition

Action

Stop run when

OU

Run limit

Action

Stop run when

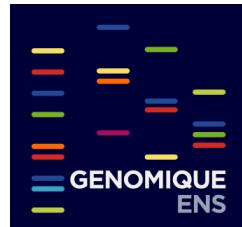
Condition	Value
Time equals	72 Hrs

Must be between 0.05–1,000

Condition	Value
Flow cell is	End of life

- En RNASeq, les runs durent 72h ou jusqu'à épuisement de la FC

Les conditions d'arrêt normal du run : Le volume de données produit



Run until

Run limit ⓘ

Action	Condition	Value
Stop run when	Flow cell is	End of life

Flow cell data target (optional) ⓘ

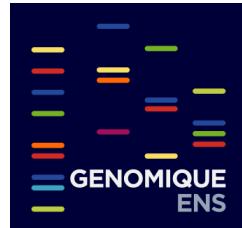
Action	Condition	Value
Stop run when	Select	<input type="text" value="Input"/> Gb <input type="button" value="Gb"/>
	Estimated bases equal	
	Basecalled bases equal	

Flow cell data target (optional) ⓘ

Action	Condition	Value
Stop run when	Estimated bases equal	<input type="text" value="100"/> Gb <input type="button" value="Gb"/>

Must be greater than 0 and no greater than 5000Gb.

Les conditions d'arrêt normal du run : Le volume de données produit



Run until

Run limit ⓘ

Action	Condition	Value
Stop run when	Flow cell is	End of life

Flow cell data target (optional) ⓘ

Action	Condition	Value
Stop run when	Select	Input Gb
	Estimated bases equal	
	Basecalled bases equal	

Apply rules

A target base rule will stop the run once the defined amount of estimated or basecalled bases are detected.

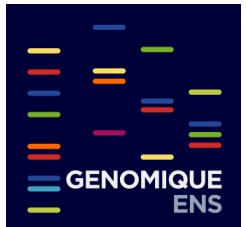
Targets are set per flow cell.

Flow cell data target (optional) ⓘ

Action	Condition	Value
Stop run when	Basecalled bases equal	100 Gb

ⓘ Performance warning
In pass folder. Must be greater than 0 and no greater than 5000Gb.

Les conditions d'arrêt normal du run : Le volume de données produit



Run until

Run limit ?

Action	Condition	Value
Stop run when	Flow cell is	End of life

Flow cell data target (optional) ?

Action	Condition	Value
Stop run when	Select	Input Gb
	Estimated bases equal	
	Basecalled bases equal	

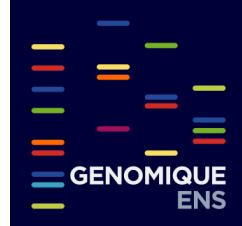
Flow cell data target (optional) ?

Action	Condition	Value
Stop run when	Basecalled bases equal	100 Gb

ⓘ Performance warning
In pass folder. Must be greater than 0 and no greater than 5000Gb.

Basecalling must keep up with data production for a Target bases rule to be effective.

Consider using "Estimated bases" if your device or basecall model is unlikely to allow keep up.



Filtre sur les données produites en préambule

- La taille des séquences (attentes variables selon les thématiques)

1. Positions 2. Kit 3. Run options 4. Analysis 5. Output 6. Review

Run options

Run until [?](#)

Run limit: 72 hours duration

Flow cell data target: None set

Options

Minimum read length [?](#)

20 bp 200 bp 1000 bp

Adaptive sampling [?](#)

Enrich sequences
 Deplete sequences
 Barcode balancing

Show advanced user options

The minimum length of a read that will be written out. This can be decreased if, for example, sequencing very short fragments. Decreasing the minimum read length can increase the amount of data produced, which may increase overall data processing times. The minimum read length does not take into account any barcodes or adapters used in sample preparation, therefore the reads may appear longer than the minimum set length.

Beta

< Back Continue > Skip to final review »

Filtre sur les données produites en préambule

- Il est possible d'enrichir les nos séquences selon un catalogue

The screenshot shows the 'Run options' step of a sequencing pipeline. The top navigation bar includes tabs for 1. Positions, 2. Kit, 3. Run options, 4. Analysis, 5. Output, and 6. Review. The 'Run options' tab is active.

Run until (72 hours duration, None set)

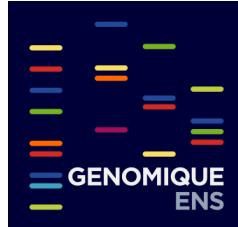
Minimum read length (200 bp selected, 20 bp and 1000 bp are also shown)

Adaptive sampling (Beta)

- Enrich sequences
- Deplete sequences
- Barcode balancing

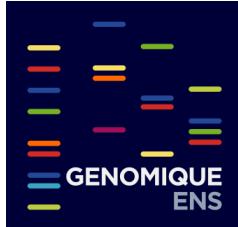
Show advanced user options

Buttons at the bottom: < Back, Continue >, Skip to final review >



Enrichissement

- Catalogue des séquences que l'on souhaite garder à fournir



Enrich sequences

Alignment reference *

No reference

Specify sequence coordinates of interest

No reference

Accepted file format: .bed

Advanced options

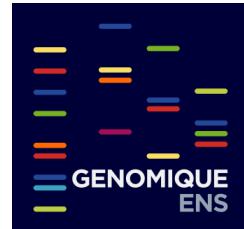
Channels to use for adaptive sampling

Start channel

End channel



Il est possible de mobiliser seulement une partie de la FC dans cette tâche



Appauvrissement

- Catalogue des séquences que l'on ne souhaite pas garder à fournir

Deplete sequences

Alignment reference *

No reference

Specify sequence coordinates of interest

No reference

Accepted file format: .bed

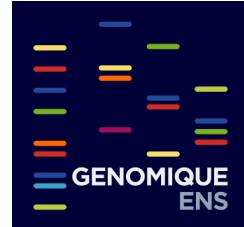
Advanced options

Channels to use for adaptive sampling

Start channel - 1 +

End channel - 3000 +

Equilibrage des codes barres utilisés en cours de séquençage



Barcode balancing Beta

Balance all barcodes in specified kit(s)

Custom selection

Enter barcode numbers or ranges separated by commas or spaces (e.g. 1-2, 5)

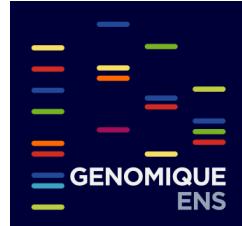
Advanced options

Channels to use for barcode balancing

Start channel: 1

End channel: 3000

Paramètres avancés



▼ Show advanced user options

Select best performing channel

Time between pore scans

- 1,5 hours +

Reserve pores ?

Simulated playback

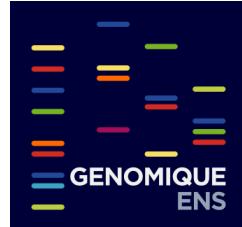
?

Delay using some of the pores until the voltage has dropped later in the run, such that other pores can catch up. This can be used when e.g. wanting to save pores until their accuracy and output is optimal for the voltage across the array. Switch this feature off to fully front-load sequence data acquisition.

Delay using some of the pores until the voltage has dropped later in the run, such that other pores can catch up. This can be used when e.g. wanting to save pores until their accuracy and output is optimal for the voltage across the array. Switch this feature off to fully front-load sequence data acquisition.

Définition de l'appel de base

- Passage du signal électrique au fichier séquence



Analysis

Basecalling

Basecalling ON [?](#) →

Modified bases OFF [?](#)

Models [Edit](#)
High-accuracy basecalling

Barcoding [?](#)

Barcoding ON

Options [Edit](#)
Enabled

Alignment [?](#)

Reference sequence - OFF

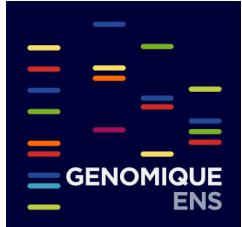
Options [Edit](#)
Disabled

Basecall the reads produced in this experiment during the run. If there are queued reads left at the end of the run, the software will continue basecalling until it has caught up with all the reads.

Trois possibilités de basecalling :

- Fast : Pratique pour le diagnostic parce rapide
- High-accuracy : Plus long mais moins d'erreur
- Super-accurate : Encore plus long mais moins d'erreur

Définition de l'appel de base : cas des bases modifiées



- Passage du signal électrique au fichier séquence
- Dictionnaires de bases possibles incluent certaines bases modifiées
 - Attention car le catalogue dépend de la FC et du protocole de fabrication de banque

Basecalling

Basecalling ON

Modified bases ON

Basecalling options

Basecalling model

High-accuracy basecalling

Modified bases

5mC CG contexts

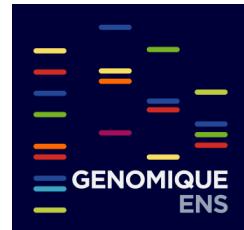
5mC CG contexts

5mC & 5hmC CG contexts

Include modified base calls in your data.

Modified base calls are disabled when basecalling is turned off.

Gestion des codes barres



Analysis

Basecalling

Basecalling ON [?](#)

Modified bases OFF [?](#)

Models [Edit](#)

High-accuracy basecalling

Barcoding [?](#) Barcoding ON

Options [Edit](#)

Enabled

Split basecalled data into folders by barcode. The reads in the output files will also have a designated barcode ID.

Alignment [?](#)

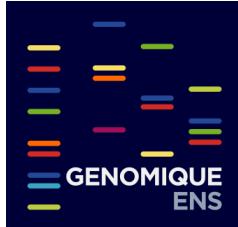
Reference sequence · OFF



Options [Edit](#)

Disabled

Gestion des codes barres : le triming



Analysis

Basecalling

Basecalling ON [?](#)

Modified bases OFF [?](#)

Models [Edit](#)

High-accuracy basecalling

Barcoding [?](#)

Barcoding ON

Options [Edit](#)

Enabled

Alignment [?](#)

Reference sequence · OFF

Options [Edit](#)

Disabled

Barcode options

- Trim barcodes [?](#)
- Mid-read barcode filtering [?](#)
- Barcode both ends [?](#)
- Override minimum barcoding score [?](#)

60

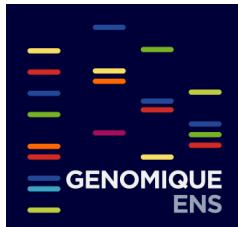
+

Cancel

Apply

Minimum score to be considered a valid barcode. The value should be between 40 and 100, inclusive. A perfect barcode match will score 100, but lower scores will be assigned to imperfect matches based on the number of errors. A higher score threshold increases specificity at the expense of more unclassified reads.

Alignement des séquences en simultané



- Peut être fait à postérieur
- Peut avoir un intérêt pour du séquençage de procaryotes, des séquençages courts...
- Réalisé avec Minimap2
 - Index à faire en début de paramétrage (fichier .mmi) ou à charger en lieu d'un fasta

Analysis

Basecalling

Basecalling ON [?](#)

Modified bases OFF [?](#)

Models [Edit](#)
High-accuracy basecalling

Barcoding [?](#)

Barcoding ON

Options [Edit](#)
Enabled

Alignment [?](#)

Reference sequence · OFF

Options [Edit](#)
Disabled

Alignment

Reference sequence [?](#) Align the reads to a specific reference.

Options [Edit](#)
Disabled

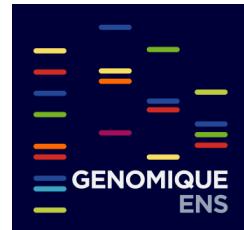
Alignment options

Use .bed file [?](#)

Options [Edit](#)
Disabled

Cancel Apply

Alignement des séquences en simultané



Analysis

Basecalling

Basecalling ON [?](#)

Modified bases OFF [?](#)

Models [Edit](#)

High-accuracy basecalling

Barcode [?](#)

Barcoding ON

Options [Edit](#)

Enabled

Alignment [?](#)

Reference sequence · ON

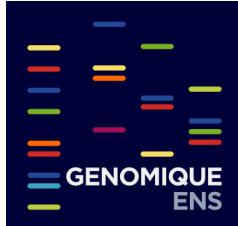
/tmp/index/hsapens105.fasta X 🔍

Options [Edit](#)

None applied

Processing this file will take an estimated 1 m 27 s and the reference includes an estimated 3.14 Gb.

Les fichiers de sortie



1. Positions

2. Kit

3. Run options

4. Analysis

5. Output

6. Review

Output

Data saved as

FormationONT/PA042194/FormationONT_Sample/

Experiment/FC/Samples
(Défini au début)

Saved output file location

/var/lib/minknow/data/.



Output formats



Basecalled reads: .BAM, .FASTQ | Every 10 minutes | Split files by barcode | Compression



Raw reads: .FAST5



Filtering

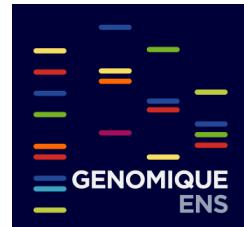
Qscore: 9 | Readlength: Unfiltered | Fail unaligned reads: Disabled



› Show advanced user options

< Back

Continue to final review >



Les fichiers de sortie : Les formats des données après le basecalling

Basecall output options

Format [?](#)

.BAM

.FASTQ

File output frequency [?](#)

Every 10 minutes

Barcode

Split files by barcode [?](#)

FASTQ options

Compression [?](#)

Downstream informatics applications may not work with FASTQ Gzip compression

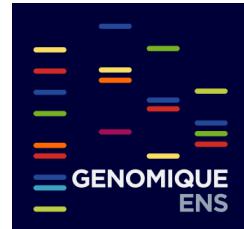
Cancel [Apply](#)

Format binaire

Format non binaire sur 4 lignes

Time between files being generated and made available for downstream workflows.

- Every 10 minutes: smaller files allowing for faster real time results from downstream workflows
- Hourly: balance between file size and frequency
- End of run: a single large file at the end of the run



Les fichiers de sortie : Les formats des données après le basecalling

Basecall output options

Format [?](#)

.BAM

.FASTQ

File output frequency [?](#)

Every 10 minutes

Barcoding

Split files by barcode [?](#)

FASTQ options

Compression [?](#)

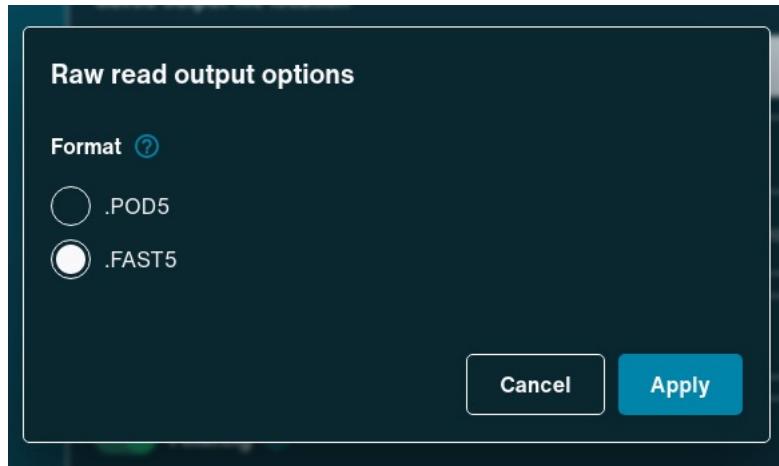
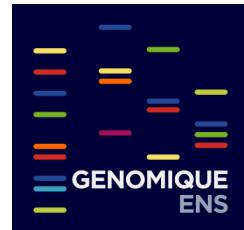
Downstream informatics applications may not work with
FASTQ Gzip compression

Reads will be split into separate
directories by barcode.

Gzip is used to compress files to ~55%
of their original size.

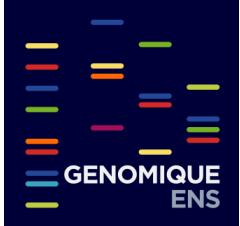
Cancel **Apply**

Les fichiers de sortie : Les formats des données brutes



Les données brutes sont importantes si l'on veut relancer le basecalling

- Historiquement, sortie en FAST5 (ouverture possible avec HDFView)
 - Beaucoup d'outils d'analyse de la qualité du run, de recherche de méthylation utilisent encore le FAST5
 - Si la FC est une R9 : le format de sortie par défaut est le FAST5
- Le POD5 est plus optimal concernant le volume des données brutes
 - Si la FC est une R10 : le format de sortie par défaut est le POD5
 - Beaucoup d'outil ne le prennent pas en charge
 - Il est possible de générer un FAST5 à partir du POD5



Les fichiers de sortie : Les filtres appliqués

Output

Data saved as
FormationONT/PA042194/FormationONT_Sample/

Saved output file location
`/var/lib/minknow/data/.`

Output formats

Basecalled reads: .BAM, .FASTQ | Every 10 minutes | Split files by barcode | Compression

Raw reads: .FAST5

Filtering

Qscore: 9 | Readlength: Unfiltered | Fail unaligned reads: Disabled

▶ Show advanced user options

Par défaut avec le modèle high-accuracy :

- Qscore 9 = séquences PASS
- Pas de filtre sur la taille de la séquence
- Pas de conservation des séquences non alignées



Les fichiers de sortie : Les filtres appliqués sur les séquences brutes

Filtering options

Pass / fail filtering

Setting a minimum and maximum read length will determine the cut-off of reads for the pass and fail folders.

Qscore ⓘ Min readlength (kb) ⓘ Max readlength (kb) ⓘ

9 Min readlength Max readlength

Minimum Q score for a read to pass.

Minimum read length for a read to pass.

Maximum read length for a read to pass.



Les fichiers de sortie : Les filtres appliqués sur les alignements

Filtering options

Pass / fail filtering

Setting a minimum and maximum read length will determine the cut-off of reads for the pass and fail folders.

Qscore ⓘ

Min readlength (kb) ⓘ

Max readlength (kb) ⓘ

9

Min readlength

Max readlength

Alignment

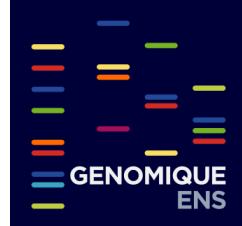
Fail unaligned reads ⓘ

Cancel

Apply



Selecting this option will send all reads
that do not align with your specified
reference to the fail folder.



Les fichiers de sortie : Les options avancés

Output

Data saved as
FormationONT/PA042194/FormationONT_Sample/

Saved output file location
/var/lib/minknow/data/.

Output formats

- Basecalled reads: .BAM, .FASTQ | Every 10 minutes | Split files by barcode | Compression
- Raw reads: .FAST5

Filtering ?

Qscore: 9 | Readlength: Unfiltered | Fail unaligned

Show advanced user options

▼ Show advanced user options

Output additional information about a run for debugging. This will produce a very large file.

Bulk output items Bulk output selected channels

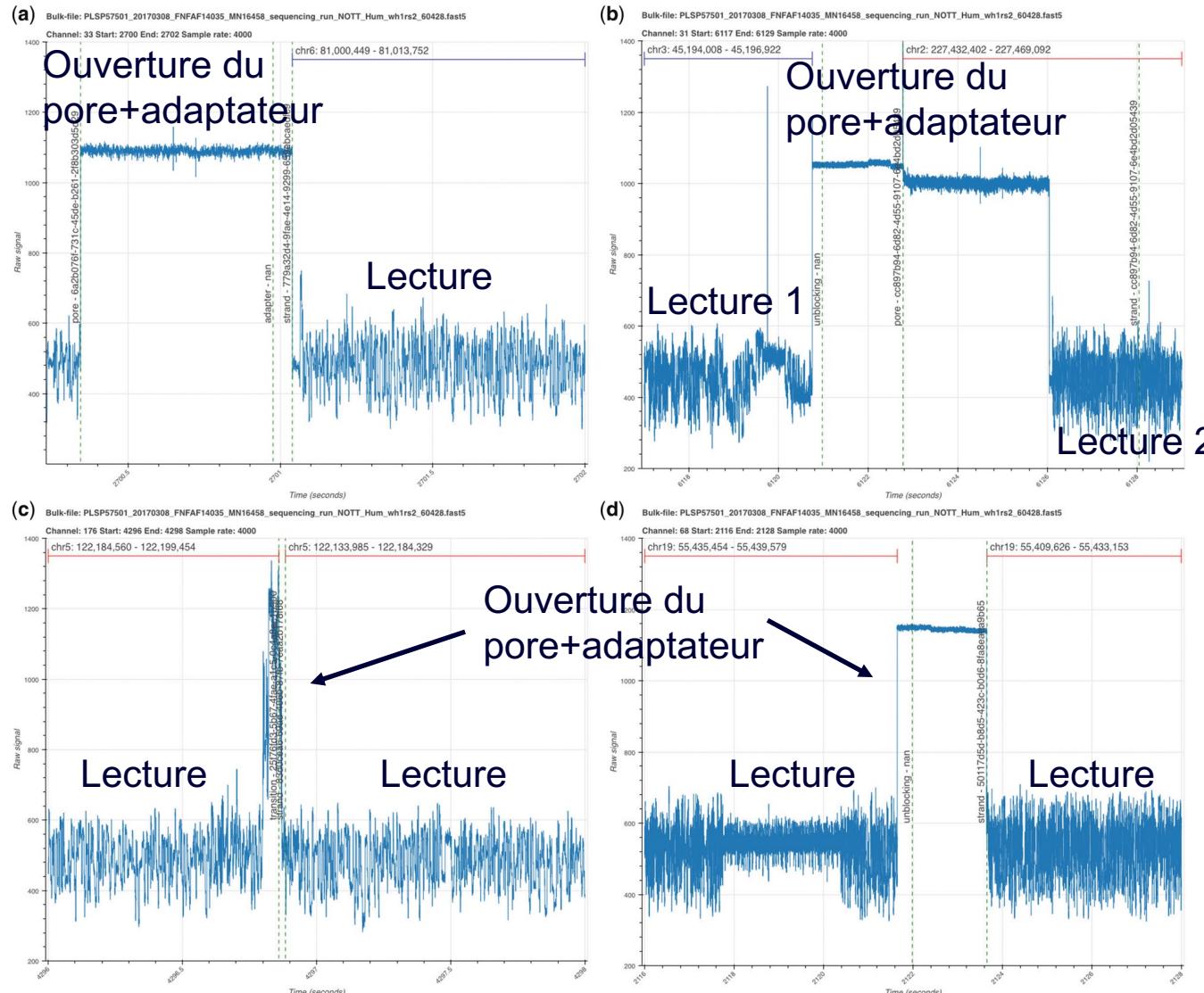
Events channel range
 Events 1-3000

Read table channel range
 Read table 1-3000

Raw channel range
 Raw 1-3000

Fichier Bulk :
Pas de coupure entre chaque lecture d'un pore
=>Il est possible de visualiser le signal et de voir les coupures déterminant les lectures (dans BulkVis par exemple)

Données bulk visualisées dans BulkVis



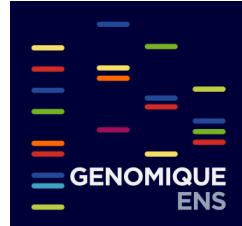
Les lectures adjacentes s'alignent à des positions non adjacentes => Le découpage de MinKNOW est justifié

Les lectures adjacentes s'alignent à des positions adjacentes => Le découpage de MinKNOW est erroné

BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files,
Alexander Payne et al, Bioinformatics, Volume 35, Issue 13, 1 July 2019, Pages 2193–2198

Récapitulatif avant lancement du run

- Ils peuvent être édités
- Ils peuvent être sauvegardés pour un prochain séquençage



1. Positions 2. Kit 3. Run options 4. Analysis 5. Output 6. Review

FormationONT

Selected positions

Kit

Selected kit:	SQK-LSK110	
Expansion packs:	EXP-NBD104	

Run options

Run limit:	72 hours	
Minimum read length:	200 bp	
Adaptive sampling:	Off	

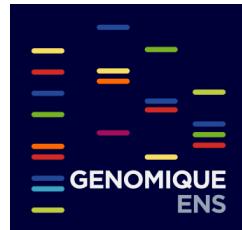
[Advanced run options](#)

Analysis

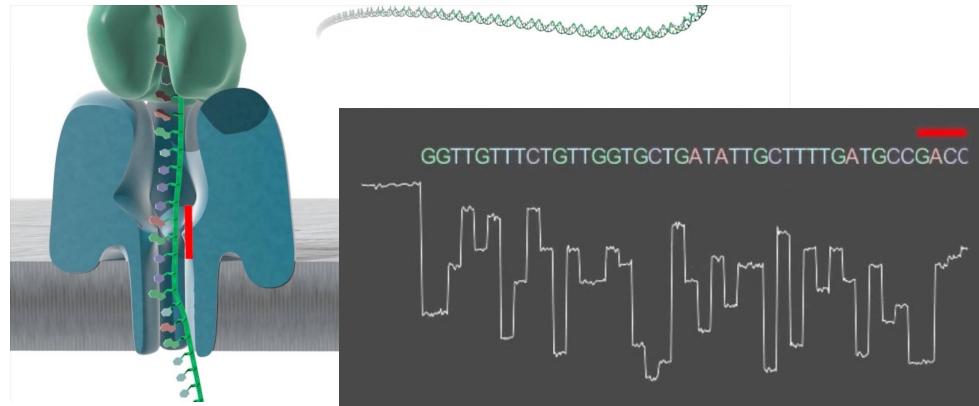
Basecalling:	On (High-accuracy basecalling, 450bps)	
Modified basecalling:	Off	
Barcode:	On	
Alignment:	hsapens105.fasta	

Back Save settings as template Start

L'appel de base permet la transcription du signal électrique en séquence nucléotidique

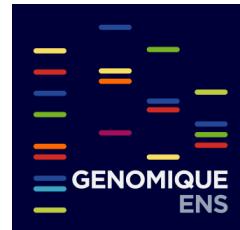


Le séquençage nanopore ne se limite pas aux bases A,T,G,C
Il permet la détection des U et de certaines bases modifiées



- Capture de la perturbation du courant membranaire induite par le passage 5 bases au niveau du reader du pore
- Progression dans le pore nucléotide par nucléotide
- Déduction à la volée ou a posteriori de la séquence via des réseaux de neurones (Dorado)

Chaque base est associée à un score de qualité



Chaque base est associée à un score de qualité

C'est le score **Phred** = probabilité d'identifier une base par erreur

$$Q = -10 \log_{10} P$$

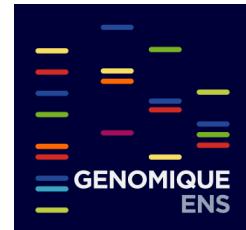
Score Qualité Phred Probabilité d'identifier une base par erreur

Les scores de qualité phred sont reliés de façon logarithmique à la probabilité d'erreur d'identification d'une base

Score de qualité phred	Probabilité d'une identification incorrecte	Précision de l'identification d'un base
10	1 pour 10	90 %
20	1 pour 100	99 %
30	1 pour 1000	99.9 %
40	1 pour 10000	99.99 %
50	1 pour 100000	99.999 %

- Avec ONT : Q15 en moyenne en R9.4.1, Q20 promis en R10.4
- Avec Illumina >Q33

Comment coder la qualité d'un nucléotide appelé ?



GSM1150340: NT_Input_control; Mus musculus; ChIP-Seq (SRR868906)

Metadata Reads Download

Filter: Find Filtered Download [What does it do?](#)

[What can the filter be applied to?](#)

1. SRR868906.1 SRS429786

name: HWI-1KL138:5:1101:6219:2000,
member: CGATGTCGATGT
x: 6219, y: 2000

2. SRR868906.2 SRS429786

name: HWI-1KL138:5:1101:8606:1998,
member: CGATGTCGATGT
x: 8606, y: 1998

3. SRR868906.3 SRS429786

name: HWI-1KL138:5:1101:9544:2000,
member: CGATGTCGATGT
x: 9544, y: 2000

4. SRR868906.4 SRS429786

name: HWI-1KL138:5:1101:10437:1997,
member: CGATGTCGATGT
x: 10437, y: 1997

5. SRR868906.5 SRS429786

name: HWI-1KL138:5:1101:10564:1996,
member: CGATGTCGATGT
x: 10564, y: 1996

Read

View: biological reads technical reads quality scores [advanced options](#)

>gnl|SRA|SRR868906.8 HWI-1KL138:5:1101:13384:1996
NCTCTAGTTCCAAGATTAAAGNGATTGGTTGAGAATACTGATGTATAAT

One channel quality score

33 46 57 62 55 61 61 59 50 55 62 60 62 60 58 61 61 61 61 61
61 33 48 57 62 62 61 58 62 62 61 59 62 62 61 61 61 61 61 61
61 61 61 61 59 61 61 61 61 56

→ Séquence de la lecture

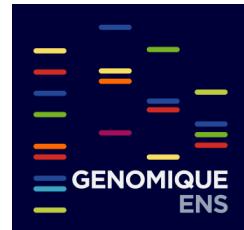
- **Une base = une lettre**

→ Score Q

- **Une base = deux chiffres**

- Ecrire la qualité sur un nombre à 2 chiffres est lourd et ne peut pas correspondre directement à la séquence en fasta
- Astuce pour coder la qualité d'une base sur 1 caractère

➤ **Utilisation de la table ASCII**



La table ASCII pour coder la qualité des séquences

American Standard Code for Information Interchange

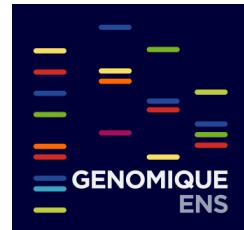
= Code américain normalisé pour l'échange d'information

Début de codage Phred
= Phred+33

De 0 à 32, caractères « invisibles »

Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]	48	30	1100000	60	0	96	60	1100000	140	'
1	1	1	1	[START OF HEADING]	49	31	1100001	61	1	97	61	1100001	141	a
2	2	10	2	[START OF TEXT]	50	32	1100010	62	2	98	62	1100010	142	b
3	3	11	3	[END OF TEXT]	51	33	1100011	63	3	99	63	1100011	143	c
4	4	100	4	[END OF TRANSMISSION]	52	34	1101000	64	4	100	64	1100100	144	d
5	5	101	5	[ENQUIRY]	53	35	1101001	65	5	101	65	1100101	145	e
6	6	110	6	[ACKNOWLEDGE]	54	36	1101100	66	6	102	66	1100110	146	f
7	7	111	7	[BELL]	55	37	1101110	67	7	103	67	1100111	147	g
8	8	1000	10	[BACKSPACE]	56	38	1110000	70	8	104	68	1101000	150	h
9	9	1001	11	[HORIZONTAL TAB]	57	39	1110001	71	9	105	69	1100001	151	i
10	A	1010	12	[LINE FEED]	58	3A	1110010	72	:	106	6A	1101010	152	j
11	B	1011	13	[VERTICAL TAB]	59	3B	1110011	73	;	107	6B	1101011	153	k
12	C	1100	14	[FORM FEED]	60	3C	1111000	74	<	108	6C	1101100	154	l
13	D	1001	15	[CARRIAGE RETURN]	61	3D	1111001	75	=	109	6D	1101101	155	m
14	E	1110	16	[SHIFT OUT]	62	3E	1111100	76	>	110	6E	1101110	156	n
15	F	1111	17	[SHIFT IN]	63	3F	1111110	77	?	111	6F	1101111	157	o
16	10	10000	20	[DATA LINK ESCAPE]	64	40	100000000	100	@	112	70	1110000	160	p
17	11	100001	21	[DEVICE CONTROL 1]	65	41	100000001	101	A	113	71	1110001	161	q
18	12	100100	22	[DEVICE CONTROL 2]	66	42	100000010	102	B	114	72	1110010	162	r
19	13	10011	23	[DEVICE CONTROL 3]	67	43	10000011	103	C	115	73	1110011	163	s
20	14	10100	24	[DEVICE CONTROL 4]	68	44	10000100	104	D	116	74	1110100	164	t
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]	69	45	10000101	105	E	117	75	1110101	165	u
22	16	10110	26	[SYNCHRONOUS IDLE]	70	46	10000110	106	F	118	76	1110110	166	v
23	17	10111	27	[ENG OF TRANS. BLOCK]	71	47	10000111	107	G	119	77	1110111	167	w
24	18	11000	30	[CANCEL]	72	48	10001000	110	H	120	78	1111000	170	x
25	19	11001	31	[END OF MEDIUM]	73	49	10001001	111	I	121	79	1111001	171	y
26	1A	11010	32	[SUBSTITUTE]	74	4A	10001010	112	J	122	7A	1111010	172	z
27	1B	101011	33	[ESCAPE]	75	4B	10001011	113	K	123	7B	1111011	173	{
28	1C	11100	34	[FILE SEPARATOR]	76	4C	1001100	114	L	124	7C	1111100	174	
29	1D	11101	35	[GROUP SEPARATOR]	77	4D	1001101	115	M	125	7D	1111110	175	}
30	1E	11110	36	[RECORD SEPARATOR]	78	4E	1001110	116	N	126	7E	1111110	176	~
31	1F	11111	37	[UNIT SEPARATOR]	79	4F	1001111	117	O	127	7F	1111111	177	[DEL]
32	20	100000	40	[SPACE]	80	50	1010000	120	P					
33	21	1000001	41	!	81	51	1010001	121	Q					
34	22	1000010	42	"	82	52	1010010	122	R					
35	23	1000011	43	#	83	53	1010011	123	S					
36	24	1001000	44	\$	84	54	1010100	124	T					
37	25	1001001	45	%	85	55	1010101	125	U					
38	26	1001100	46	&	86	56	1010110	126	V					
39	27	1001110	47	'	87	57	1010111	127	W					
40	28	1010000	50	(88	58	1011000	130	X					
41	29	1010011	51)	89	59	1011001	131	Y					
42	2A	1010100	52	*	90	5A	1011010	132	Z					
43	2B	1010111	53	+	91	5B	1011011	133	I					
44	2C	1011000	54	,	92	5C	1011100	134	\					
45	2D	1011001	55	-	93	5D	1011101	135	J					
46	2E	1011010	56	.	94	5E	1011110	136	^					
47	2F	1011111	57	/	95	5F	1011111	137	-					

La correspondance entre les colonnes permet le passage de 2 à 1 caractère



Le format FASTQ (exemple d'ADN)

La séquence est codée sur 4 lignes

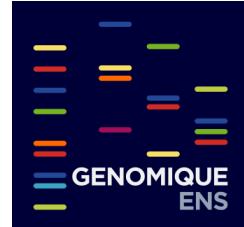
@Ligne d'identifiant

Séquence de la lecture

+Ligne d'identifiant (répétition de la ligne1)

Qualité de la lecture codée en ASCII

```
@6b109bc4-8bbd-408e-9089-7e4b530ce5e6 runid=91ceec1ff67a0b21bf1ab0fce80e77cbd2a6da8
sampleid=EMTome_A2020 read=27472 ch=469 start_time=2020-11-24T08:41:45Z barcode=barcode04
GGTATGCTTCGTTGGTTCAAGGTGGGTGTTCTTGATCCATCATCGTACTTCCAGTTCTATCGTGTTCCTATTCTGTTGGTGTGAT
ATTGCGGGTCTGCTGGGTGTTAACCTAACGCGATGGTATCAACATGAGTACGGGGCCTAGCCTCCGCCTCCAGCCAAGCTCTCCGCCGTCG
GCTCCCGGGCGCCGCCAACCGACGTGGAAGACGGAGAAGGAAACCTGCGCCCTGGCCTCTCCTCAAGGGCTCAGGCTCCAAGTCGGAGGCGA
CAAGATATT...
+
,,,,,%1348::+9$.06/.('##&%$$'15+,-,#$%&,&'(%'$(*'%'%&/+'&&(%&(3=8=@F<9A<DE?;)><:28<967-
;9;9@?:>DBB8>@D8:<;?9B<9?DC>9@A>?;6397;9976++/.259<;0*)%&&-//1)*2-*,&0-1006576*-111-
)+'*+$%7&&+,96578A<--/9<<:*-*-3--.-
++%$&&0169;2>A@@@:49698;0668,786866:8:;<:9>6/4983;()6',*&$ $$%(&37999128=:%'.1.13122)+54('9>C=;8
88%2/(--/,/=%(0++(+0062...
```



Le format FASTQ (exemple d'ARN direct)

La séquence est codée sur 4 lignes

@Ligne d'identifiant

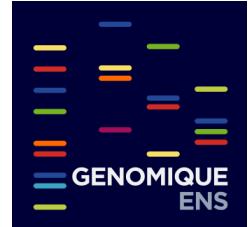
Séquence de la lecture

+Ligne d'identifiant (répétition de la ligne1)

Qualité de la lecture codée en ASCII

```
@1f4360c4-59d5-4387-9532-909dba74acae runid=3351522d21d04b833ec675d19fbf23f85bfceb28
sampleid=directRNA_validation_WT2 read=61 ch=413 start_time=2018-09-04T13:39:29Z
CGCUGAAAAGAGCUCCGUUGCUUCCUUAGCUACCUCGGUGGCCUUUGCAAAGCCCUCUGCAACAAACCCAUUCAGCCCGCAGGCGCUGG
UUAGGAGUGGCUGGACAGGCCUGAGUGGCCAUCCGUUCAUACAGUUUCAUUCAGGGCCGGGCUCCUCCCCUUUGGCAUUUA
UACUUGGUGGGGGCAUUGAAGGCCUGCGGUGGAAGGCUUGGGACUGGUCCUGUAACUGUGUGUCCCCUGAAGGCUGAGUUGGCCAUUUGGGUUAG
UCCCAGGUGACAGAAGGGAGAAACCAAAAGAACAGAAGAAAUGCAGAACAGAUUUCGAUCAGUGAACUUGACAGGAAGAACUU
UAAAAAAUGCACAGUCACCUGGUGGCCUUGUAGGGGCCUACCCUGGUGUGGGUCAGGCCACUGCUCAGCCACUUGGGUCGUGGUUGU
AGUACCAGUUCUGGGAGGAAGACAUGUUGGAGACUGCCAGAAAUCCCUGUACAUCGUUUUACGUGCUUGUGUCCAAGUGAAUUGUAUUGGUUC
GGUGUUUGCUAAACAAAGUGACUUAACAGCUAAAAAAAAAUUGUGC
+
'%%&() /3, (*((( ('&+()&&&((-0%&&'((**'.21.,,1.,531-,+,0-))..))**)))(*54*-0.*(')))15,))).-*.*-
*), ')& '&(&(3))+.)*)*,- -**(( '...1**) + -3-+ +2/2021+,*()) () - +''' (+,&(11/- .540+))++)*-33,+*-,0--
*.(%)('+. /-0+,0+*-.-+) ' &*)) + ' %() ' &*(+,) + 2083-.-.,1,))) ) - ) * **+, , +**252-00-+.* **+, , /- * -- -
*(*'* )+ ) ( ' ) 0+ ' * + (*.,, + - * ( ++ ), 1811, --01-, 1/, *) * ' + * 1750.+), - , 0.2, - / .3/23- , , -4 --
, , + * .13146300, 4./-* ) 0 (+/-0763,), - * & ( ( * ) & + - * + ., /.) ) 0.*,-
., ) ( , 2. ( 1, (&)+, +*, ) ( ( ' ) + ) ' ) , . + 0 + *** ) ( &, - 63213/*20***--0201( * ), , - 3 /-* . - ' * + ) - % ) ) ( *-
0'(&'(/*.,, + * 0-15//, 54+ + 27// /- ( )) ) ' ) + .. *** , , . . . 0040, // - 2- / + + - . ) * *. - + ( ' . . . /41-, &) -- + - 5, +, .-
)*(( + * + , * ) + 5663360,) * ( '%$
```

Les outils pour manipuler les fichiers FASTQ



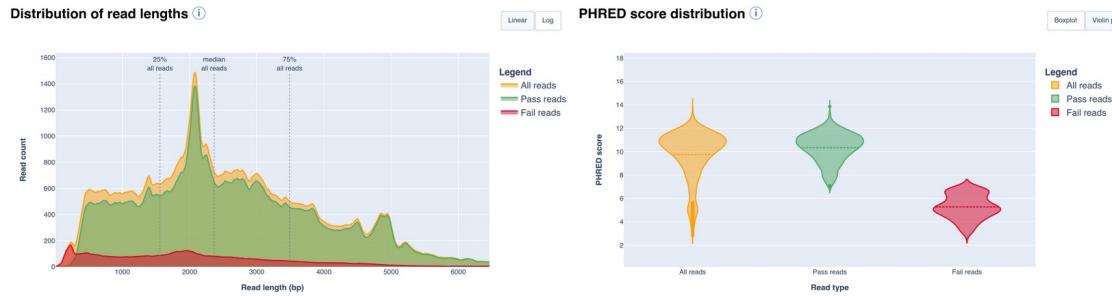
- **Les outils de QC**

- **Nanopack** : <https://github.com/wdecoster/nanopack>
- **ToulligQC** : <https://github.com/GenomiqueENS/toulligQC>



ToulligQC is dedicated to the QC analyses of Oxford Nanopore runs. This software is written in Python and developed by the [GenomiqueENS core facility](#) of the [Institute of Biology of the Ecole Normale Supérieure \(IBENS\)](#).

Click on [following image](#) to see an report example. An [online help](#) is available to better understand graphics generated with ToulligQC when clicking on the ⓘ icon.

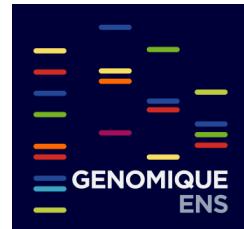


- **Les assembleurs**

- **RNA-Bloom** : <https://github.com/bcgsc/RNA-Bloom>

- **Les aligneurs**

Quels outils pour l'alignement ?

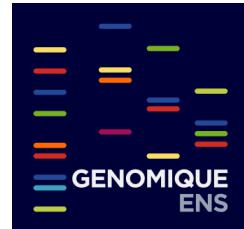


Il existe très peu d'outils fonctionnels pour l'alignement de longues lectures

- **BWA** (<https://github.com/lh3/bwa>)
 - n'évolue plus beaucoup (14/04/2024)
 - ne prend pas en charge les jonctions (ARN)
- **GMAP/GSNAP** (<http://research-pub.gene.com/gmap/>)
 - maintenu (dernière version le 20/05/2024)
 - dédié aux ARNs.
 - Versions anciennes non fonctionnelles sur ARN longues lectures, versions récentes non testées
- **Minimap2** (<https://github.com/lh3/minimap2>)
 - le standard depuis sa sortie
 - Maintenu (dernière version le 27/03/2024)
 - même développeur que BWA
 - prend en charge les données de tout type (ADN, ARN)

La spécificité d'un alignement de RNASeq :

La jonction exon-exon

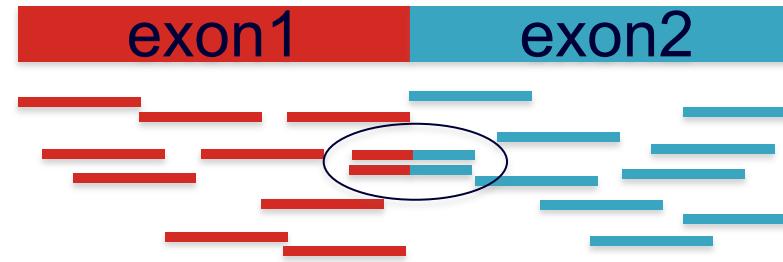


L'alignement est idéalement fait sur le génome

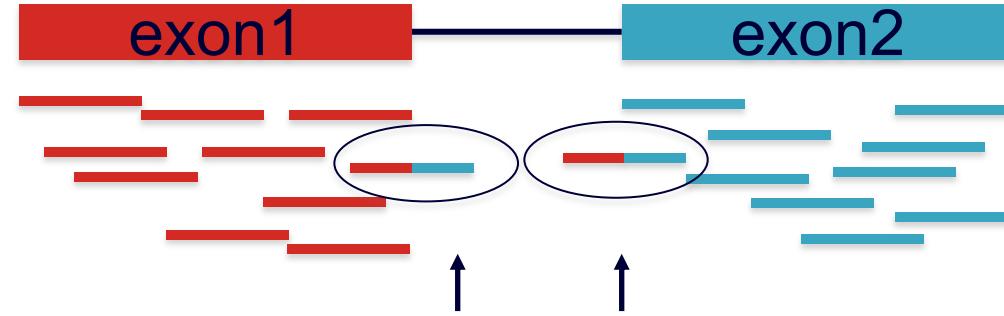
- Présence des introns sur la référence

Les lectures issues des transcrits s'alignent sur les exons

Si on aligne sur le transcriptome :

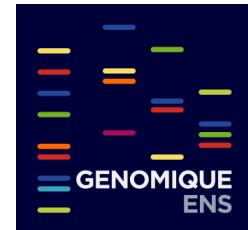


Si on aligne sur le génome :



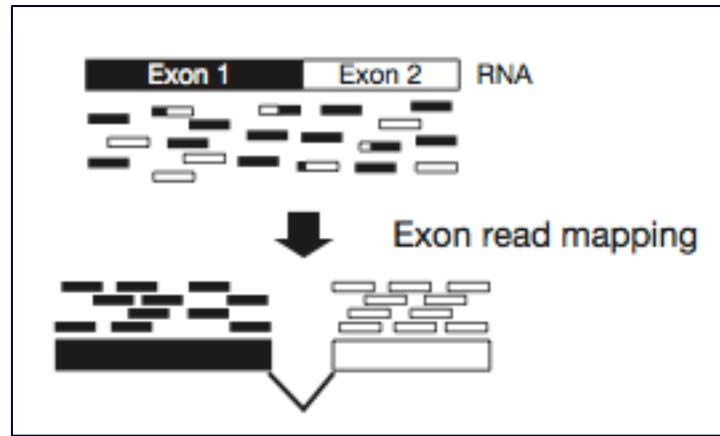
Alignement au niveau de la jonction exon-exon
problématique car pas de gaps possibles

BWA



Aligneur développé pour l'alignement de séquences génomiques

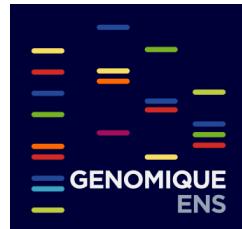
- Pas d'alignement sur les jonctions prévu



S'ils sont utilisés en RNASeq

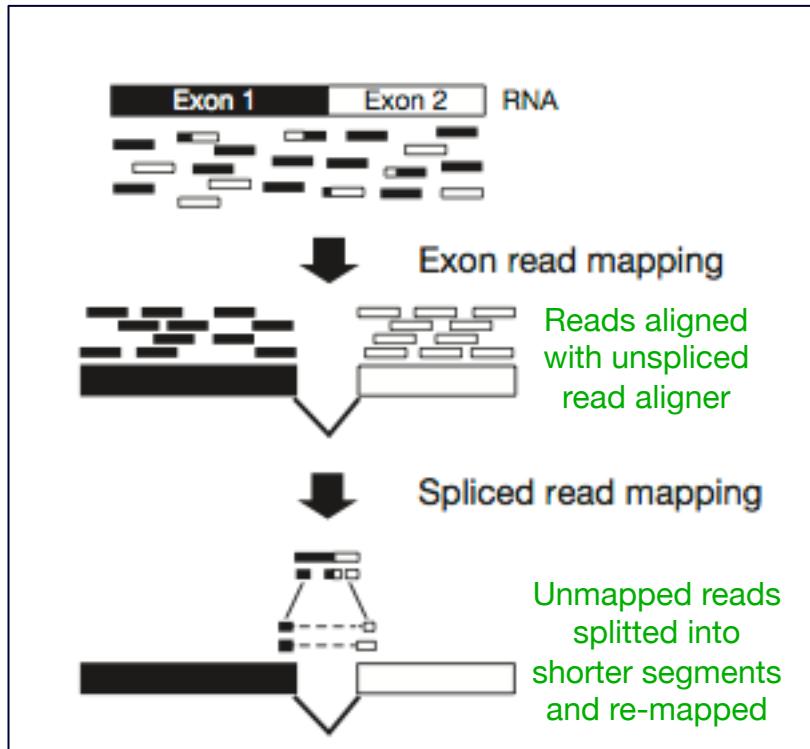
- Les séquences qui s'alignent à cheval sur 2 exons donc au niveau des jonctions sont « jetées »

TopHat, GSNA...



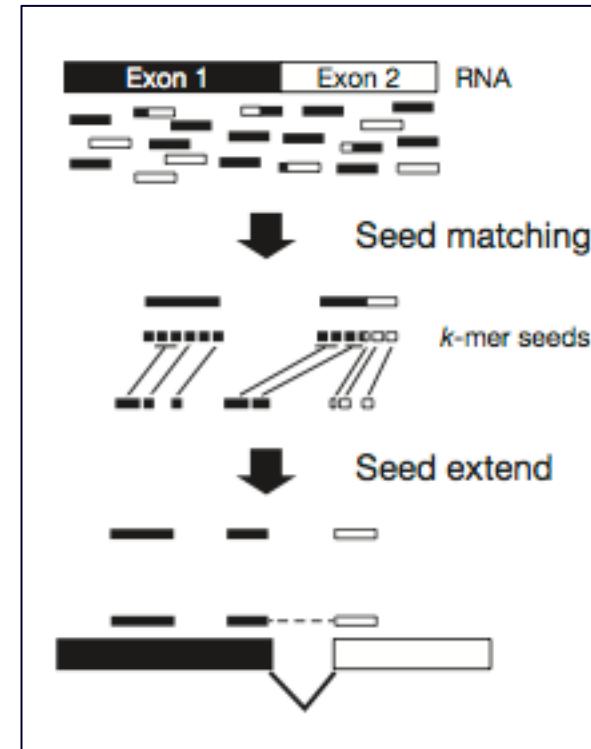
Aligneurs développés pour l'alignement d'ARNm

- Les jonctions sont prises en compte



Stratégie TopHat : exon first

Récupération des lectures correspondant aux jonctions, non alignées dans un 1^{er} temps, découpage pour réalignement



Stratégie GSNA : junction first

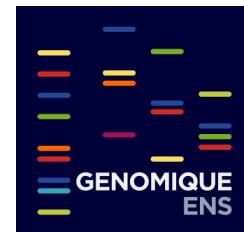
Découpage pour alignement/extension en une seule fois

Nat Methods 2011 Garber

Minimap2

<https://github.com/lh3/minimap2>

Dernière version 27/03/24



- ❑ Outil versatile :
 - ARN, ADNc, ADN génomiques
 - Lectures PacBio, lectures ONT
- ❑ Outil peu gourmand en mémoire (sauf indexation du génome) et très rapide
- ❑ Fichier de sortie au format SAM
- ❑ Outil exon-first : il peut favoriser les processed-pseudogenes
 - Possibilité de joindre à l'alignement un fichier au format bed12 (jonctions exon-exon) pour aider l'alignement des données épissées

Alignment des cDNA :

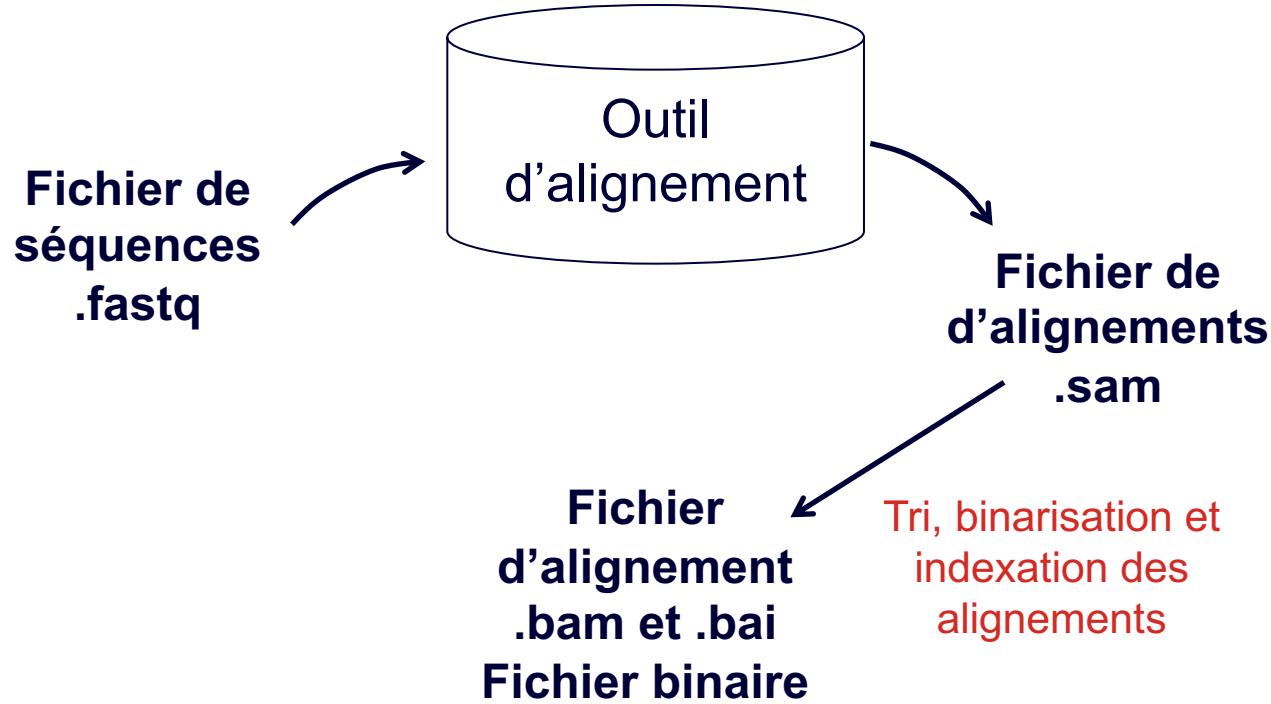
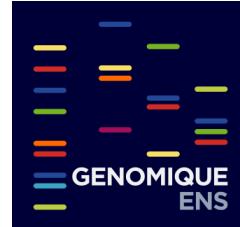
```
minimap2 -ax splice ref.fa nanopore-cdna.fa > aln.sam
```

Alignment des ARN :

```
minimap2 -ax splice -uf -k14 ref.fa direct-rna.fq > aln.sam
```

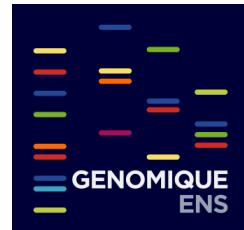
Les formats de fichiers en sorties d'alignment

Les formats de sorties standards sont les fichiers **.sam** et **.bam**



Les fichiers binaires sont lisibles très efficacement par la machine mais plus par l'œil humain

- Inutile de tenter de les ouvrir = utilisation d'un visualiseur de génome



Les sections du format SAM

Sequence Alignment/Map format

Header

Dictionnaire des séquences références utilisées dans l'alignement

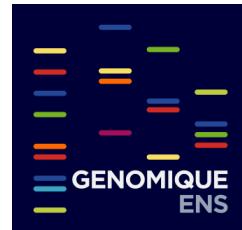
```
@HD    VN:1.4  SO:unsorted
@SQ    SN:1    LN:248956422
@SQ    SN:2    LN:242193529
@SQ    SN:3    LN:198295559
@SQ    SN:4    LN:190214555
@SQ    SN:5    LN:181538259
@SQ    SN:6    LN:170805979
@SQ    SN:14   LN:107043718
...
@SQ    SN:X   LN:156040895
@SQ    SN:Y   LN:57227415
@SQ    SN:MT  LN:16569
```

Ligne de commande ayant généré le fichier

```
@PG    ID:minimap2    VN:2.17-r941    CL:minimap2 -a -t 16 -x splice --junc-bed
/import/rhodos01/shares-net/ressources/sequencages/bed12/only_chr_Homo_sapiens_ens96.bed
/import/pontos02/analyses/EMTome_A2020/minimap2indexgenerator_output/minimap2indexgenerator_output_minimap2index_genomefile/genome.idx /import/pontos02/analyses/EMTome_A2020/eoulsan-20201217-190452/working/filterreads_output_reads_2020385_file0.fq      PN:minimap2
```

Résultats de l'alignement

```
6b109bc4-8bbd-408e-9089-
7e4b530ce5e6    0        3        141738285        60        152S16M1D33M3I8M2D19M1I24M1D7M1D45M2I1M1I66M1D7M4992
N48M1602N37M1D79M2D8M3I25M2I35M4I36M1I4M1D49M1I5M2I4M1D67M3I1M1D2M1I5M1I13M2D17M1I28M1D9M1D64M2D5M1D26M1I12M
117S    *        0        0        GGTATGCTTCGTTCGGTT...GAAGTCA    ,,,,%1348::+9$.06..4552322442:7%    s1:i:601
          s2:i:299    NM:i:65 AS:i:614    de:f:0.0587    rl:i:25
          cm:i:135    nn:i:0  tp:A:P  ms:i:642    ts:A:+
```



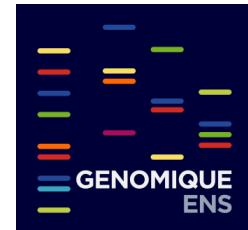
La section alignment du format sam

Chaque alignement est décrit par une ligne tabulée

Certains champs sont obligatoires :

QNAME	Id Lecture	6b109bc4-8bbd-408e-9089-7e4b530ce5e6
FLAG	016 sens / 16 reverse	0
RNAME	Ref Sequence	3
POS	Start alignment	141 738 285
MAPQ	QScore d'alignement (0-255)	60
CIGAR	M = Match/Mismatch N = Gap S = SoftClipped I = Insertion	152S16M1D33M3I8M2D19M1I24M1D7M1D45M2I1M1I66M1D7M49 92N48M1602N37M1D79M2D8M3I25M2I35M4I36M1I4M1D49M1I5M 2I4M1D67M3I1M1D2M1I5M1I13M2D17M1I28M1D9M1D64M2D5M1 D26M1I12M117S
RNEXT	Id Lecture suivante	* Si pas d'info
PNEXT	Start de la lecture suivante	0
TLEN	Taille de la lecture	0 Si pas d'info
SEQ		GGTATGCTTCGTTGGT...GAAGTCA
QUAL		,,,,%1348::+9\$.06...4552322442:7%

Le format BAM pour stocker les données non alignées (format Binary Alignment Map)



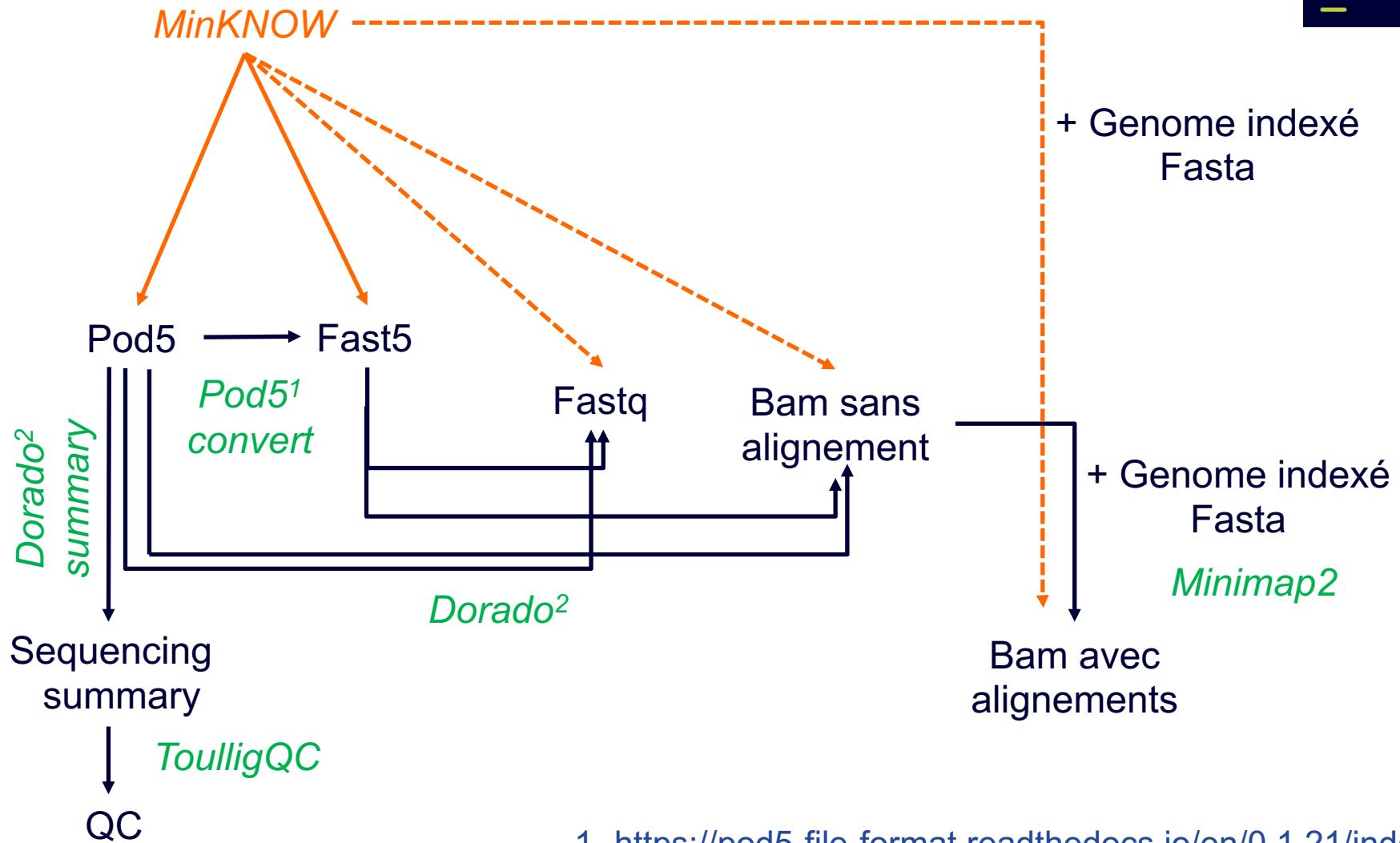
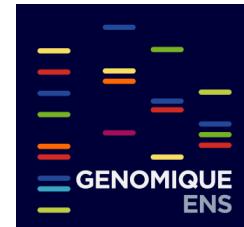
Les fichiers BAM « non alignés » peuvent être utilisés pour stocker les fichiers FASTQ.

Les fichiers BAM présentent de **nombreux avantages** par rapport à FASTQ.

Ils sont :

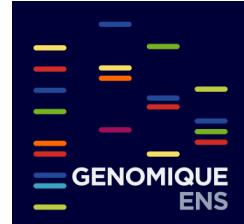
- Compressés,
- Toutes les informations sur la séquence sont sur une seule ligne (4 lignes pour le fastq),
- Peuvent stocker des informations sur les échantillons via des tag
- Il existe de nombreux outils pouvant opérer sur les fichiers BAM (extraire par tag, filtrer par tag, etc)

Avec et sans MinKNOW



1- <https://pod5-file-format.readthedocs.io/en/0.1.21/index.html>
2- <https://github.com/nanoporetech/dorado>

Les outils pour manipuler les fichiers SAM/BAM



- Samtools : <https://www.htslib.org/>

Samtools

Home Download Workflows Documentation Support

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

- Samtools** Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
- BCFtools** Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
- HTSlib** A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htslib so they can be built independently.

- Picard : <https://broadinstitute.github.io/picard/>

Picard
build passing

A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

Latest Jar Release Source Code ZIP File Source Code TAR Ball View On GitHub

Picard is a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. These file formats are defined in the [Hts-specs](#) repository. See especially the [SAM specification](#) and the [VCF specification](#).

Note that the information on this page is targeted at end-users. For developers, the source code, building instructions and implementation/development resources are available on [GitHub](#).

The Picard toolkit is open-source under the [MIT license](#) and free for all uses.

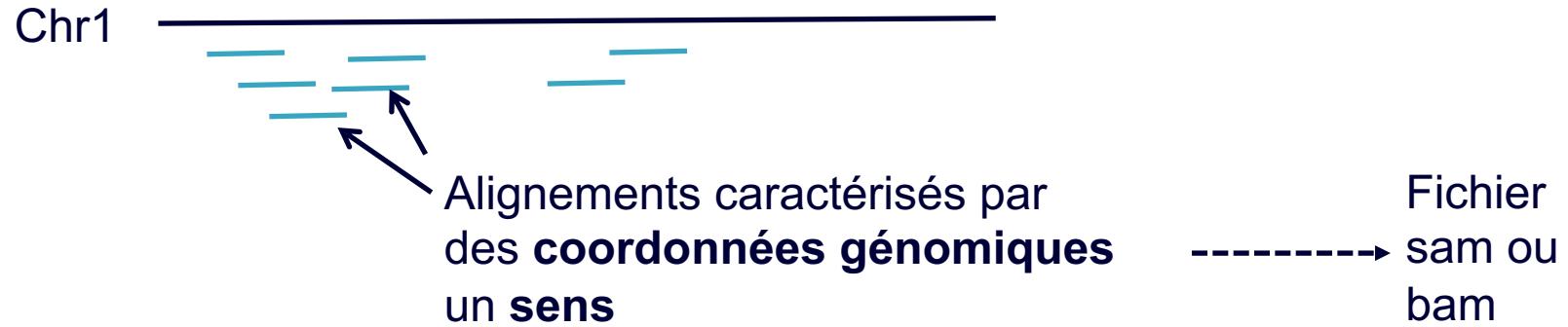
Enjoy!

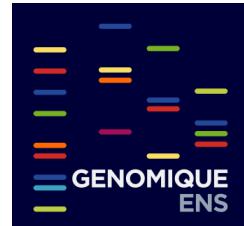


De l'alignement au nom de gène

Le fichier sam ne contient pas de référence au nom de gène

- Il contient simplement des coordonnées

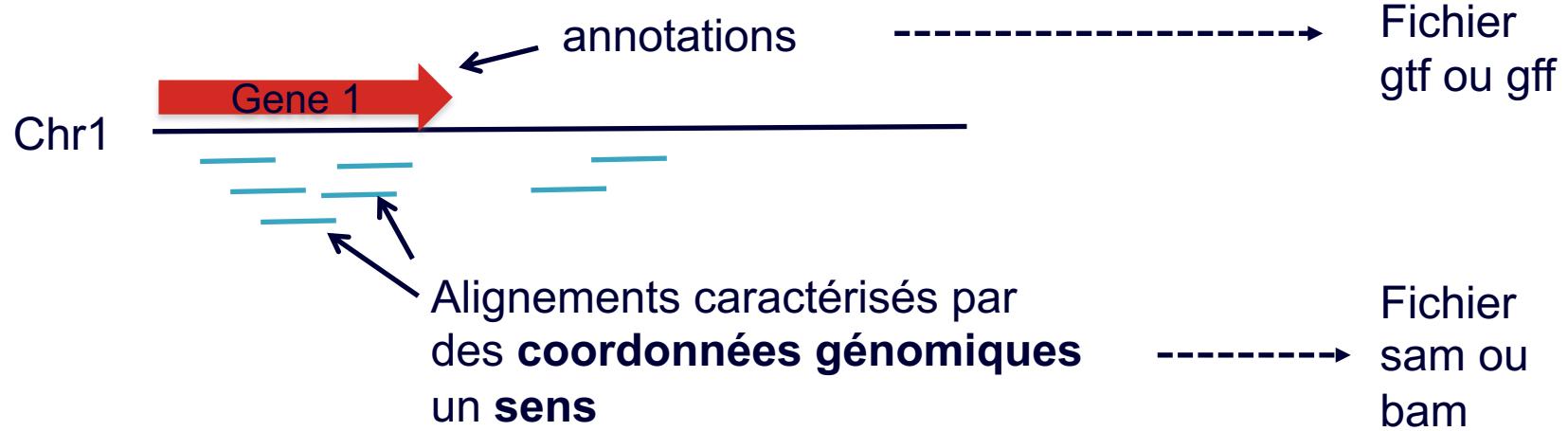




De l'alignement au nom de gène

Le fichier sam ne contient pas de référence au nom de gène

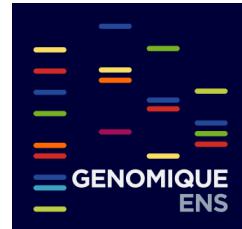
- Il contient simplement des coordonnées



Le fichier sam doit être croisé avec un fichier d'annotations (gtf ou gff)

- Il contient les coordonnées de chaque élément génomique connu et le nom de cet élément

Les fichiers d'annotation au format GFF ou GTF



- Les formats GFF/GTF sont des **fichiers qui décrivent les éléments génomiques**
➤ **Gene, exon, ARNm, ARNnc, transposons, UTR...**
- Ce sont des fichiers **texte** organisés en **9 colonnes** séparés par des **tabulations**.
- Ils sont associés au fichier génome **.fasta**
- Ils sont trouvables sur les **sites officiels tels qu'Ensembl, UCSC, TAIR10, JGI**



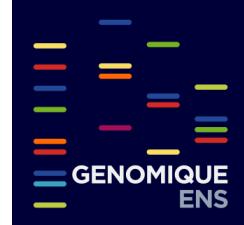
Génomes et transcriptomes



★	Species	Show/hide columns								Filter							
		DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM	
Y	Human <i>Homo sapiens</i>	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM					
Y	Mouse <i>Mus musculus</i>	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM					
Y	Zebrafish <i>Danio rerio</i>	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	-	-	BAM					
	Alpaca <i>Vicugna pacos</i>	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-					
	Anole lizard <i>Anolis carolinensis</i>	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	BAM					
	Armadillo <i>Dasyprocta novemcincta</i>	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	BAM					
	Bushbaby <i>Otolemur garnettii</i>	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-					
	C.intestinalis <i>Ciona intestinalis</i>	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-					
	C.savignyi <i>Ciona savignyi</i>	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-					
	Caenorhabditis elegans <i>Caenorhabditis elegans</i>	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-					

Annotations





Description du format gtf

Coordonnées de l'élément sur le génome

Le seqid doit correspondre au >seqid du fichier fasta correspondant

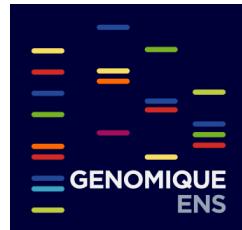
seqid	source	type	start	end	score	strand	phase
1	ensembl	gene	3102016	3102125	.	+	.
1	ensembl	transcript	3102016	3102125	.	+	.
1	ensembl	exon	3102016	3102125	.	+	.
1	ensembl_havana	gene	3205901	3671498	.	-	.
1	havana	transcript	3205901	3216344	.	-	.
1	havana	exon	3213609	3216344	.	-	.
1	havana	exon	3205901	3207317	.	-	.

attributes

gene_id	ENSMUSG00000064842;	gene_version	1;	gene_name	Gm26206;	gene_source	ensembl;	gene_biotype	snRNA;		
gene_id	ENSMUSG00000064842;	gene_version	1;	transcript_id	ENSMUST00000082908;	transcript_version	1;	gene_name	Gm26206;	gene_source	ensembl;
gene_id	ENSMUSG00000064842;	gene_version	1;	transcript_id	ENSMUST00000082908;	transcript_version	1;	exon_number	1;	gene_name	Gm26206;
gene_id	ENSMUSG00000051951;	gene_version	5;	gene_name	Xkr4;	gene_source	ensembl_havana;	gene_biotype	protein_coding;	havana_gene	"OTTMUSG0000026353";
gene_id	ENSMUSG00000051951;	gene_version	5;	transcript_id	ENSMUST00000162897;	transcript_version	1;	gene_name	Xkr4;	gene_source	ensembl_havana;
gene_id	ENSMUSG00000051951;	gene_version	5;	transcript_id	ENSMUST00000162897;	transcript_version	1;	exon_number	1;	gene_name	Xkr4;
gene_id	ENSMUSG00000051951;	gene_version	5;	transcript_id	ENSMUST00000162897;	transcript_version	1;	exon_number	2;	gene_name	Xkr4;

Annotations correspondant aux positions sur le génome

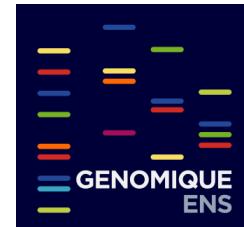
Les fichiers d'annotation au format BED (Browser extensible Data)



- Ce sont des fichiers **texte** organisés en 3 colonnes au minimum et **12 colonnes au maximum** séparés par des **tabulations**.

Column number	Title	Definition
1	chrom	Chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671) name
2	chromStart	Start coordinate on the chromosome or scaffold for the sequence considered (the first base on the chromosome is numbered 0 i.e. the number is zero-based)
3	chromEnd	End coordinate on the chromosome or scaffold for the sequence considered. This position is non-inclusive, unlike chromStart (the first base on the chromosome is numbered 1 i.e. the number is one-based).
4	name	Name of the line in the BED file
5	score	Score between 0 and 1000
6	strand	DNA strand orientation (positive ["+"] or negative ["-"] or "." if no strand)
7	thickStart	Starting coordinate from which the annotation is displayed in a thicker way on a graphical representation (e.g.: the start codon of a gene)
8	thickEnd	End coordinates from which the annotation is no longer displayed in a thicker way on a graphical representation (e.g.: the stop codon of a gene)
9	itemRgb	RGB value in the form R, G, B (e.g. 255,0,0) determining the display color of the annotation contained in the BED file
10	blockCount	Number of blocks (e.g. exons) on the line of the BED file
11	blockSizes	List of values separated by commas corresponding to the size of the blocks (the number of values must correspond to that of the "blockCount")
12	blockStarts	List of values separated by commas corresponding to the starting coordinates of the blocks, coordinates calculated relative to those present in the chromStart column (the number of values must correspond to that of the "blockCount")

Les outils pour manipuler les fichiers BED



- Bedtools : <https://bedtools.readthedocs.io/en/latest/index.html>

bedtools v2.31.0 »



bedtools is a fast, flexible toolset for genome arithmetic.

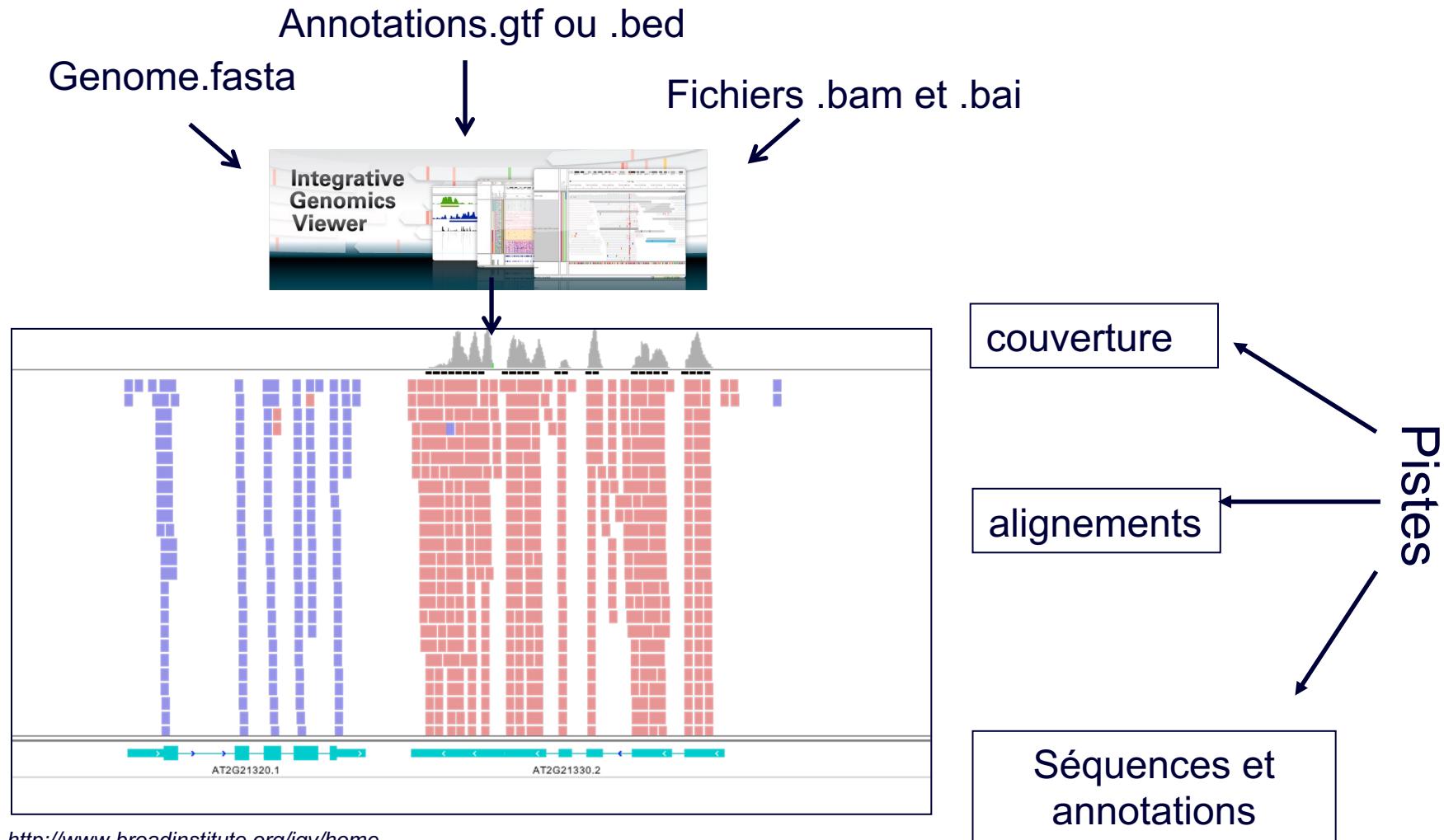
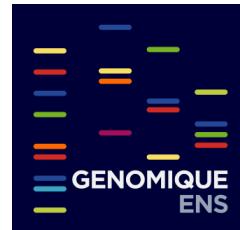
bedtools: a powerful toolset for genome arithmetic

Collectively, the **bedtools** utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable *genome arithmetic*: that is, set theory on the genome. For example, **bedtools** allows one to *intersect*, *merge*, *count*, *complement*, and *shuffle* genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF. While each individual tool is designed to do a relatively simple task (e.g., *intersect* two interval files), quite sophisticated analyses can be conducted by combining multiple bedtools operations on the UNIX command line.

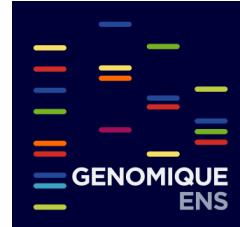
bedtools is developed in the [Quinlan laboratory](#) at the [University of Utah](#) and benefits from fantastic contributions made by scientists worldwide.

Tutorial

Visualiser ses données Integrative Genomics Viewer (IGV)



Les outils pour visualiser les données après alignement



- **Integrative Genome Viewer** : <https://igv.org/>

IGV

Integrative Genomics Viewer

 IGV

IGV desktop application

 IGV Web App

IGV in a web browser

 Juicebox Web

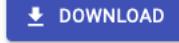
Hi-C contact map viewer

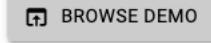
- **JBrowse2** : <https://jbrowse.org/jb2/>

JBrowse

The next-generation genome browser

JBrowse is a new kind of genome browser that runs on the web, on your desktop, or embedded in your app.

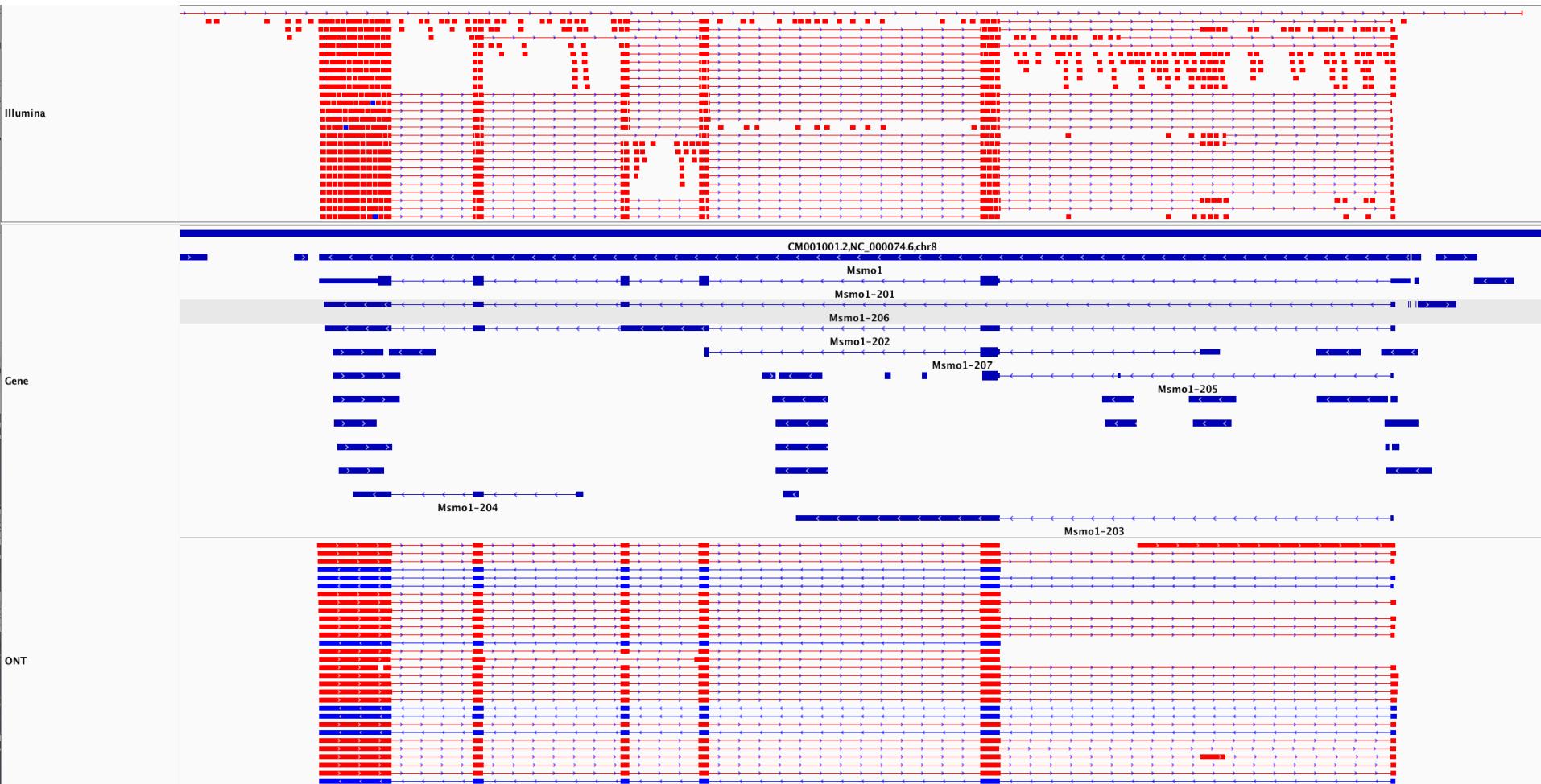
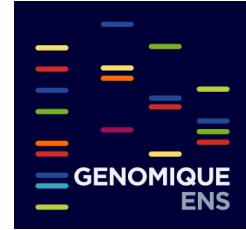
 DOWNLOAD

 BROWSE DEMO

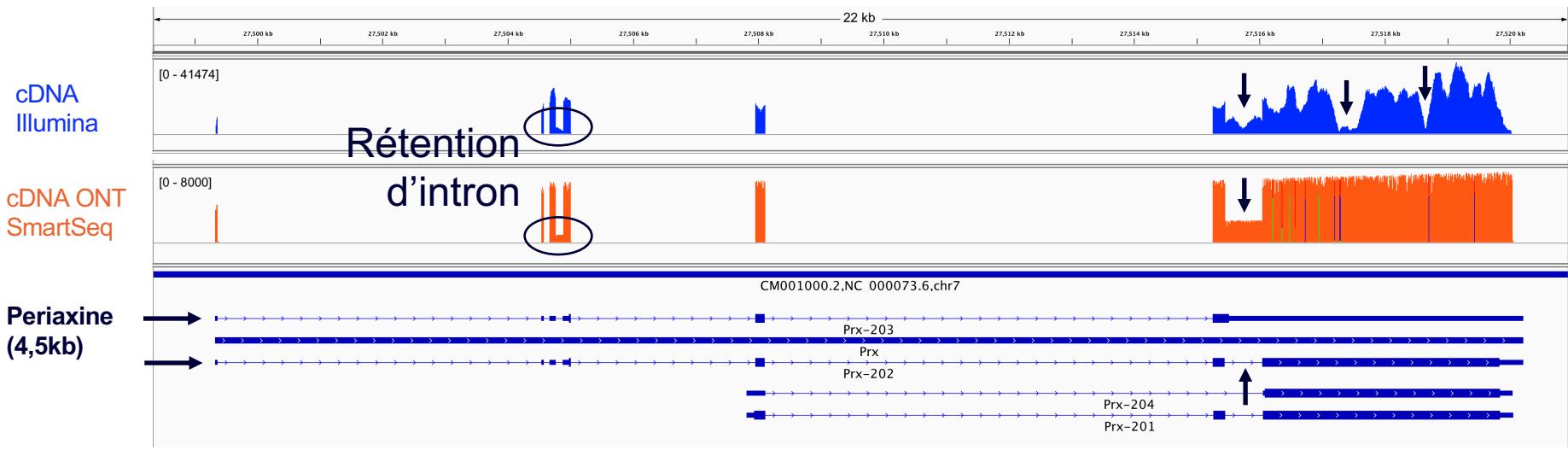
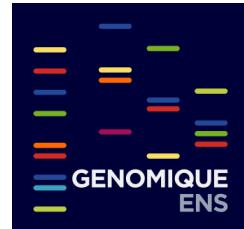
Also check out our [latest release blogpost](#), our [embedded components](#), and our [command line tools](#).

Le séquençage de transcrits sur nanopores

Exemples de lectures alignées Illumina / ONT



Le séquençage ONT permet une couverture homogène le long du transcrit

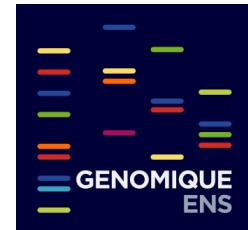


L'hétérogénéité de la couverture illumina ne permet pas de voir avec certitude les isoformes présents

- Parce que permettant une couverture homogène, le séquençage ONT est beaucoup plus clair

Le séquençage ONT et ré-annotation du génome

Application en SingleCell



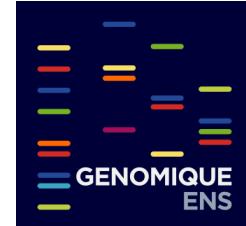
Mapping ②

Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	79.9%
Reads Mapped Confidently to Intergenic Regions	46.3%
Reads Mapped Confidently to Intronic Regions	3.6%
Reads Mapped Confidently to Exonic Regions	30.0%
Reads Mapped Confidently to Transcriptome	23.3%
Reads Mapped Antisense to Gene	0.5%

**Seulement 23% du signal
SingleCell est pris en compte...
Pourquoi ?**

Le séquençage ONT et ré-annotation du génome

Application en SingleCell



Mapping ②

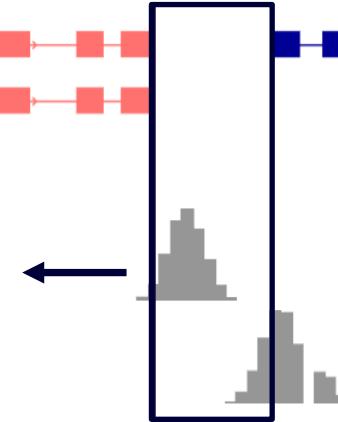
Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	79.9%
Reads Mapped Confidently to Intergenic Regions	46.3%
Reads Mapped Confidently to Intronic Regions	3.6%
Reads Mapped Confidently to Exonic Regions	30.0%
Reads Mapped Confidently to Transcriptome	23.3%
Reads Mapped Antisense to Gene	0.5%

Seulement 23% du signal SingleCell
est pris en compte...

Annotation
officielle



Signal SingleCell
non compté



Les lectures 10X ne sont pas comptées si les UTRs (ici 3') ne sont pas
présentes dans les annotations

Réannotations en utilisant Isoquant et RNA-Bloom

Annotation officielle



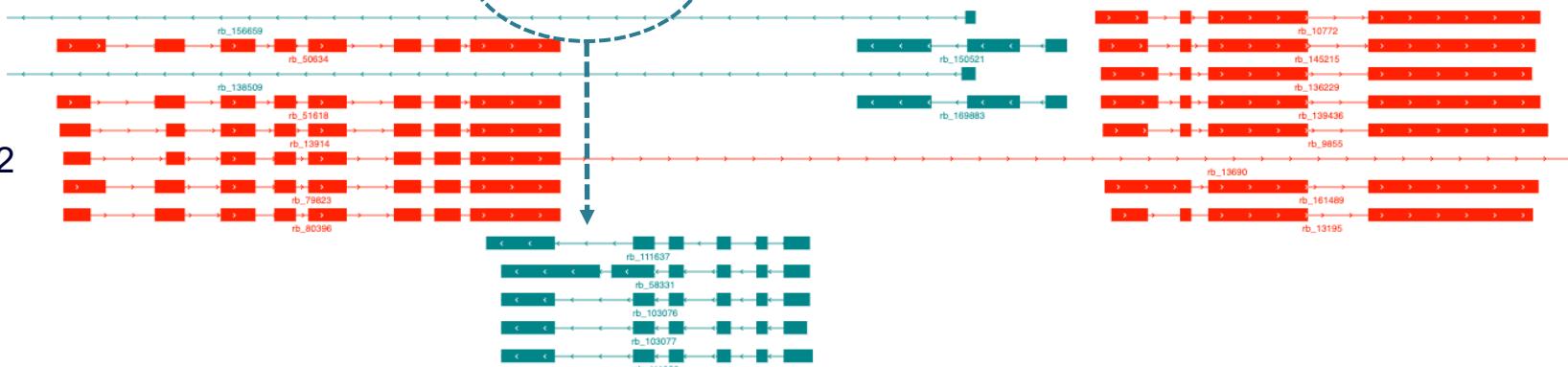
ONT Isoquant



??

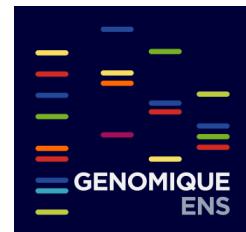
Non annoté par Isoquant

ONT RNABloom2



- Les modèles des gènes générés par l'utilisation conjointe de RNA-Bloom et d'isoquant permettent de compter les lectures alignées sur les UTRs (82% vs 23%)

Notre expertise



SingleCell en longues lectures

- Protocole Librairies 10x
- Analyse SiCeLORE

Annotation des gènes

- Protocole de librairies à ré-orienter (maison ou Smartseq)
- Ré-annotation Minimap2 + Isoquant + RNA-Bloom

Analyse différentielle

- Protocole de librairies Smartseq
- Minimap2 + Isoquant

Séquençage direct de l'ARN