

# Formation à l'utilisation d'un séquenceur Minlon

Session informatique



# L'équipe de la plateforme génomique de l'IBENS



Catherine  
Senamaud-  
Beaufort



Corinne  
Blugeon



Morgane  
Thomas-  
Chollier



Stéphane  
Le Crom



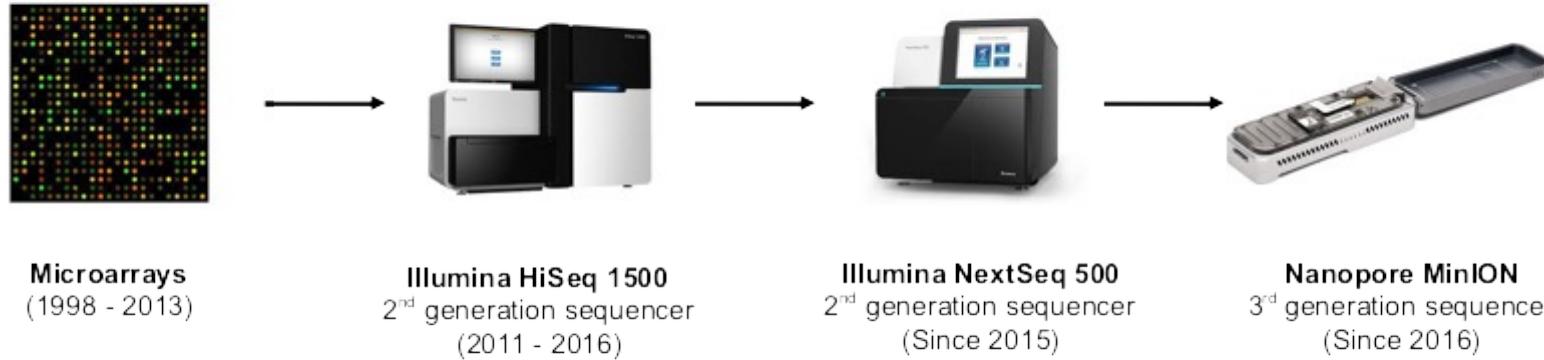
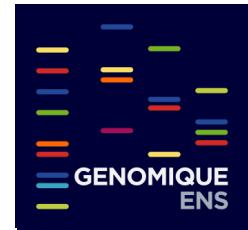
Laurent  
Jourdren



Sophie  
Lemoine

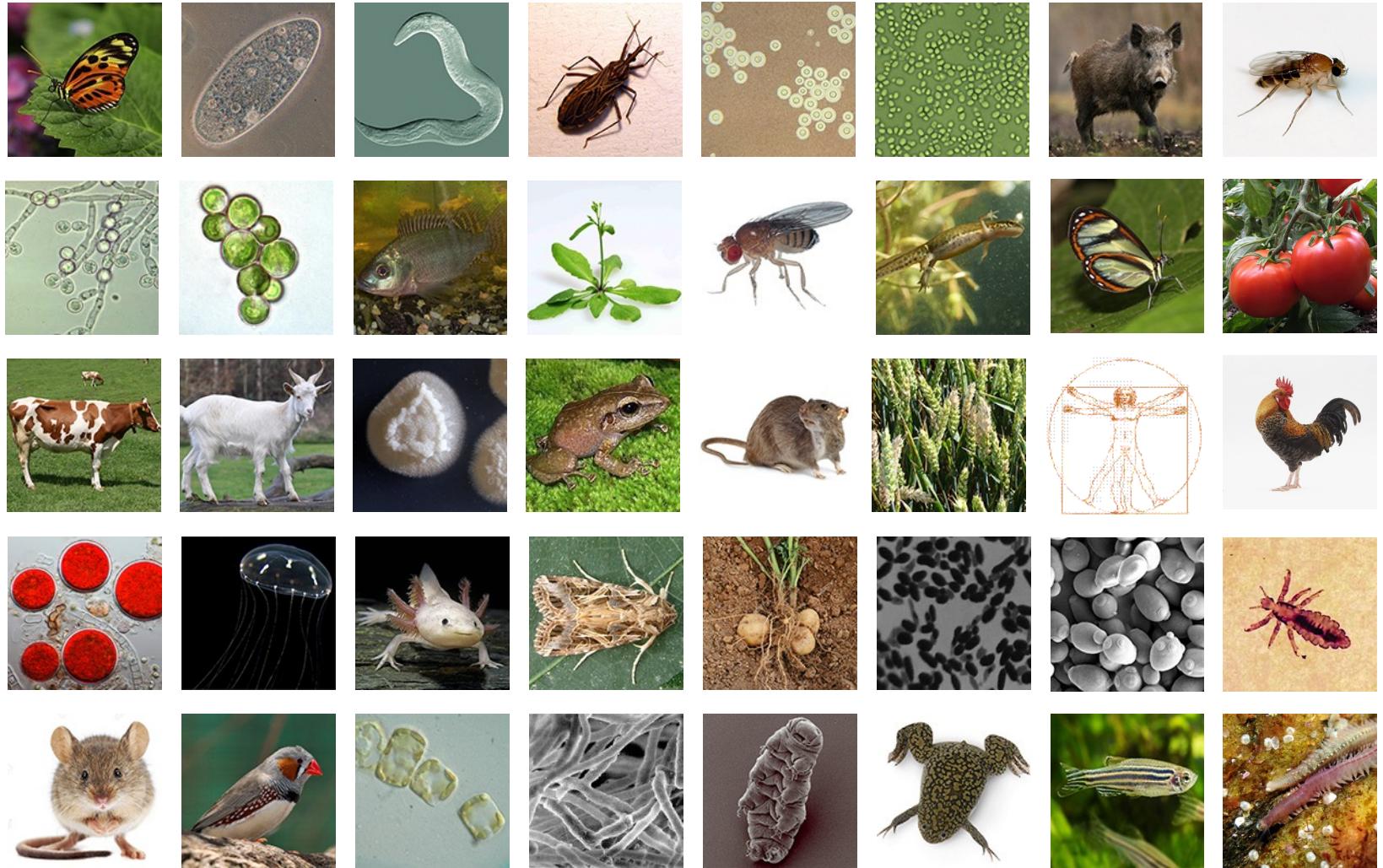
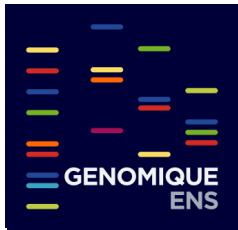
La plateforme comprend deux expérimentateurs et deux bioinformaticiens  
Nous prenons en charge les projets du contrôle des échantillons à l'analyse différentielle

# Evolution technologique

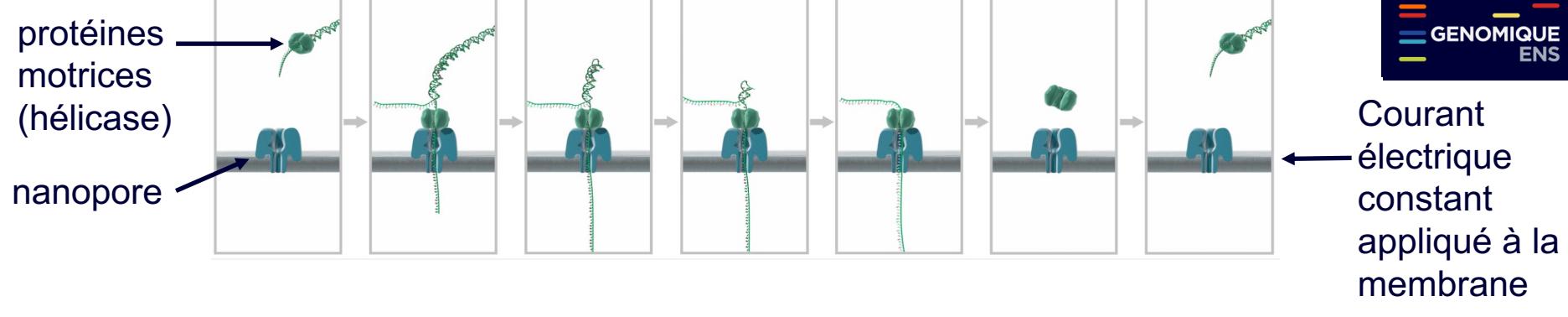
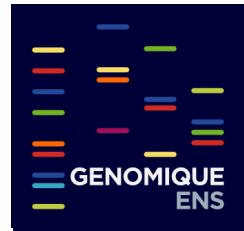


La plateforme existe depuis 1999 et a suivi le changement des technologies notamment en génomique fonctionnelle

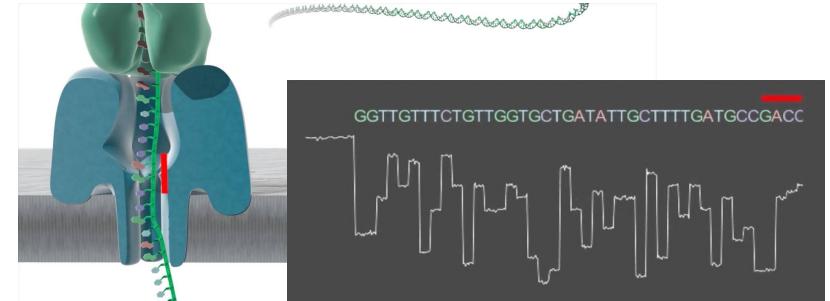
# Spécialisée sur les organismes eucaryotes



# Le séquençage sur nanopores



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



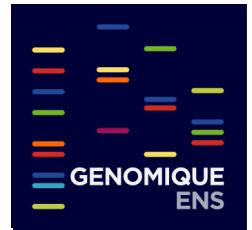
La séquence n'est alors pas déduite suite à une synthèse enzymatique  
Pas de limitation de taille due à la technique

- Assemblage de génome simplifié
- ARN séquencés en pleine longueur

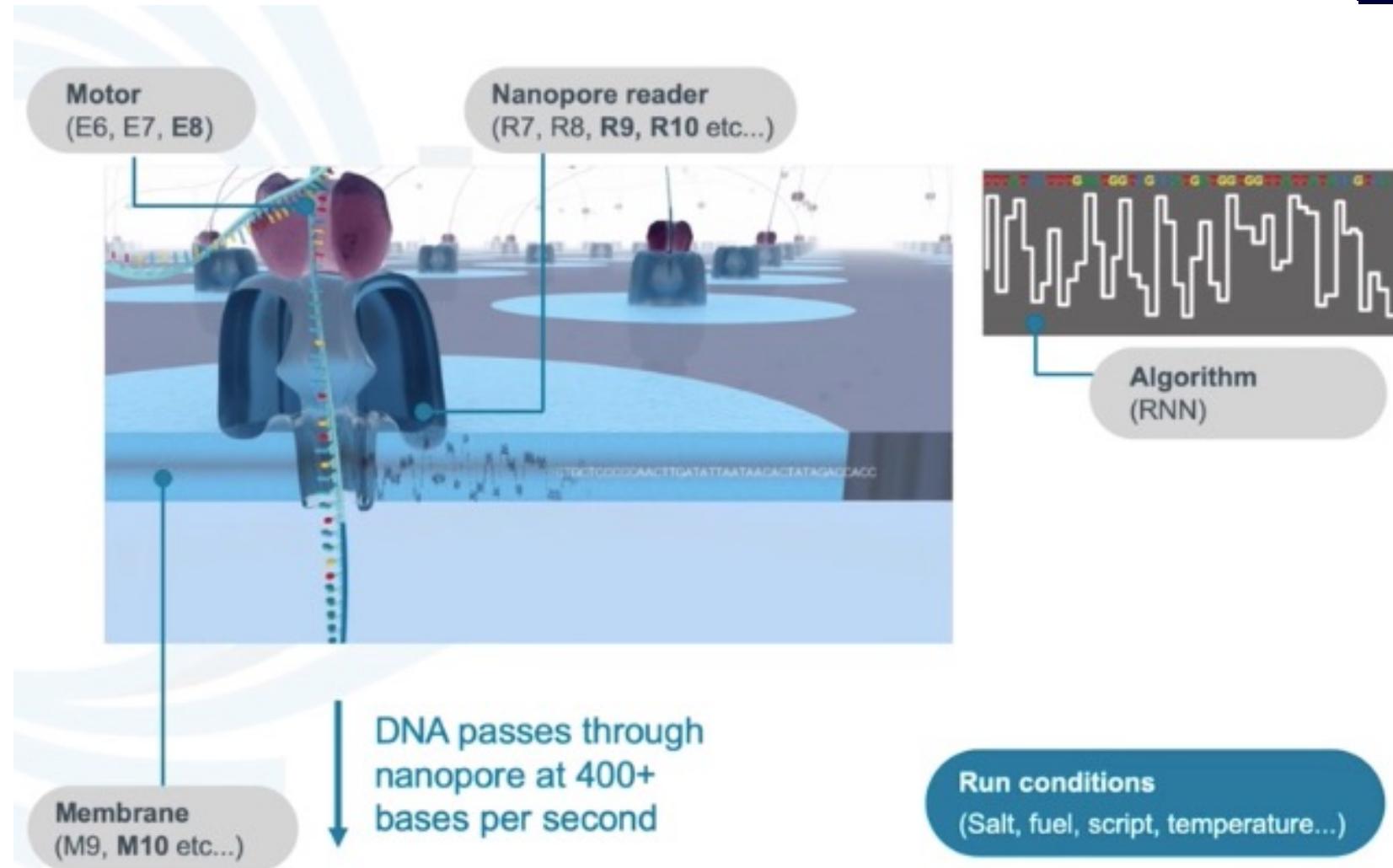
Possibilité de séquençage sans amplification de l'ADN et de l'ARN

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

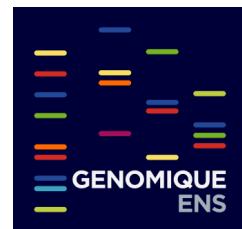
<https://nanoporetech.com/>



# Les différents éléments d'un run

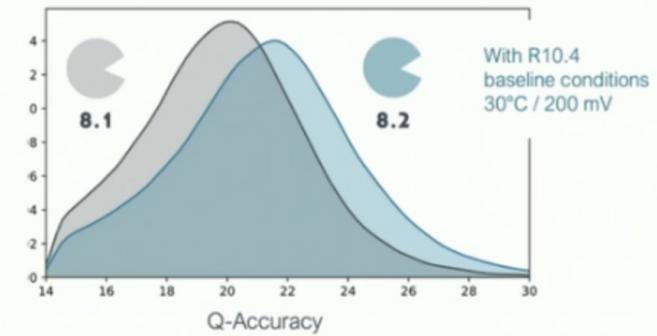
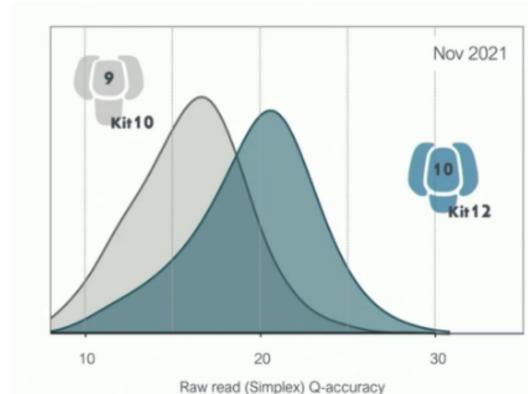
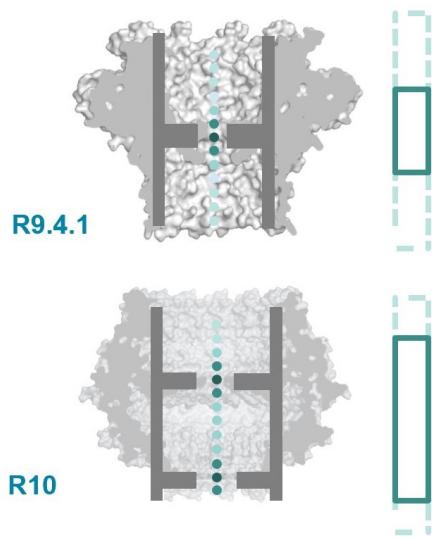


# Pores types

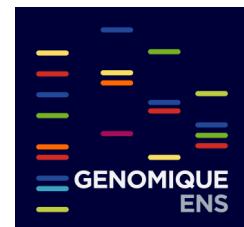


There are **three types** of flowcell pores :

- **R9.4.1** with one reader head, kit 10 (01/2017, Q13 / 95 % accuracy)  
→ discontinuation December 2022
- **R10.4** with dual reader head, kit 12 (12/2021, Q20 / 99 % accuracy)  
→ discontinuation June 2022
- **R10.4.1** with dual reader zones, kit 14 (09/2022, Q22 / 99,3 % accuracy), **early access**

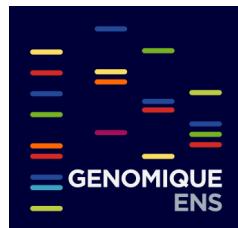


# Available Flowcells



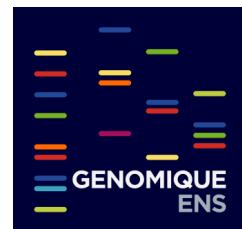
	<b>Flongle</b>	<b>MinION</b>	<b>PromethION</b>
<b>Max yield</b>	2 GB	44 Gb	242 Gb
<b>Channel count</b>	126	512	3,000
<b>Type</b>	R 9.4.1	R9.4.1 / R10.4	R9.4.1 / R10.4

# ONT Sequencers



	<b>MinION Mk1B</b>	<b>MinION Mk1C</b>	<b>GridION Mk1</b>	<b>PromethION solo</b>	<b>PromethION P2</b>	<b>PromethION</b>
<b>Flowcell slots</b>	1	1	5	2	2	24 or 48
<b>Flowcell</b>	Flongle/MinION	Flongle/MinION	Flongle/MinION	PromethION	PromethION	
<b>CPU</b>	N/A	Nvidia Jetson TX2 (Arm)	Intel i7	N/A	16 CPU cores	2 x Intel CPU 56 cores 2 x Intel CPU 80 cores
<b>GPU</b>	N/A	Nvidia Jetson TX2	2 x Nvidia 1080Ti 1 x Nvidia GV100	N/A	Nvidia A100	2/4 x Nvidia GV100 4 x Nvidia A100
<b>Memory</b>	N/A	8 GB	64 BG	N/A	64 GB	384 GB 512 GB
<b>Storage</b>	N/A	1 To SSD		N/A	16 TB	32/64 TB 60 TB
<b>Support</b>	Soon discontinued?	Full support	Full support	Soon available	Soon available	Full support

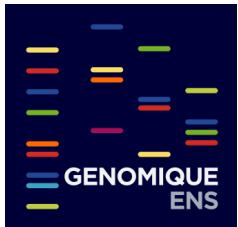
# MinION/GridION/PromethION OS



- 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 use actually **Ubuntu 16.04** to power its devices, however as Ubuntu 16.04 reaches end of life in April 2021, it will **soon be upgraded to Ubuntu 18.04**.
- ONT give 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.



# Setup your MinION Mk1C

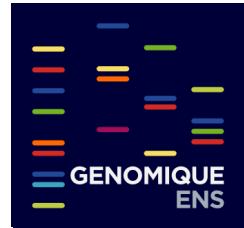


To properly setup your MinION Mk1C, you need :

- Power outlet to avoid bad electrical contact
- An Ethernet cable
- An external computer with WiFi and ssh command
- Some basics command line skills



# Utilisation du MinION Mk1C via son interface graphique



## 1- Sélection et connexion au séquenceur à utiliser

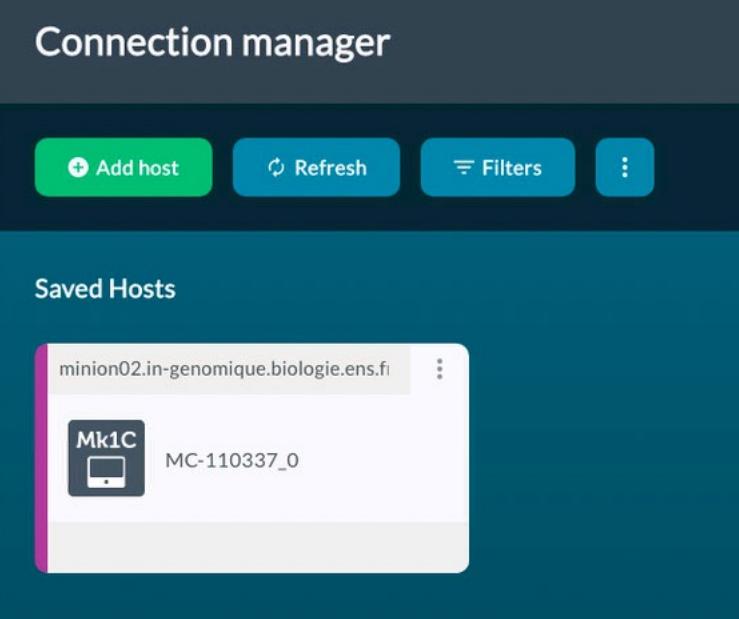
Connection manager

Add host Refresh Filters :

Saved Hosts

minion02.in-genomique.biologie.ens.fr MC-110337\_0

Mk1C



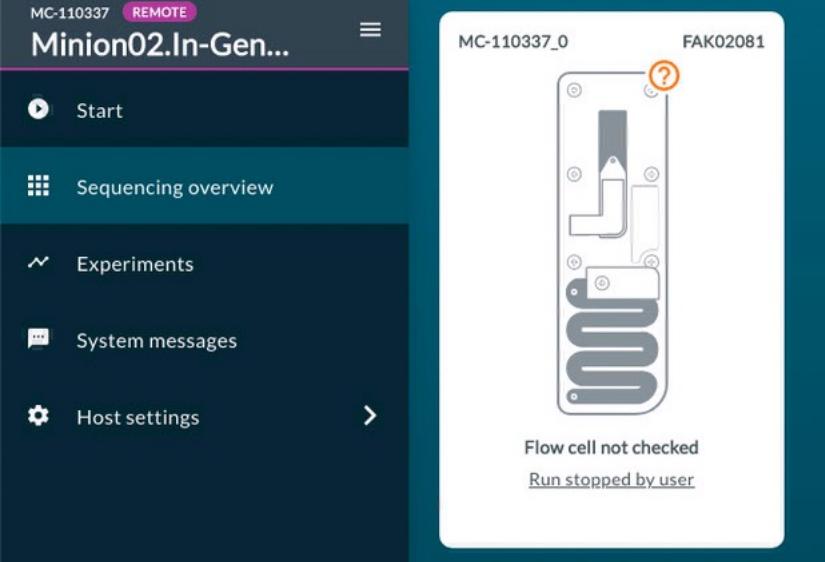
## 2- Sélection de la Flowcell en place

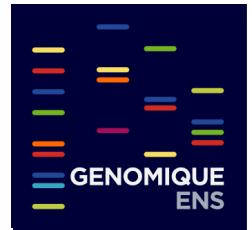
MC-110337 REMOTE Minion02.In-Gen...

Start Sequencing overview Experiments System messages Host settings

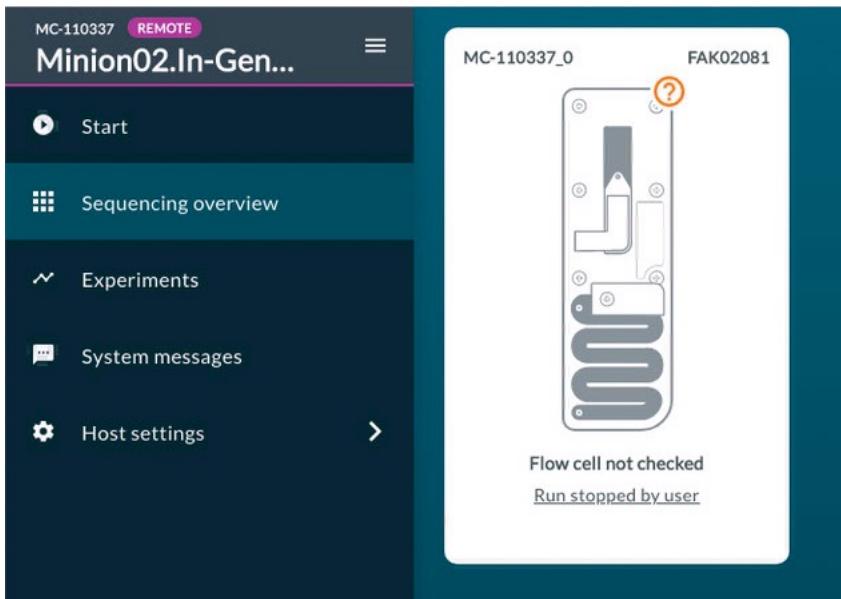
MC-110337\_0 FAK02081

Flow cell not checked Run stopped by user

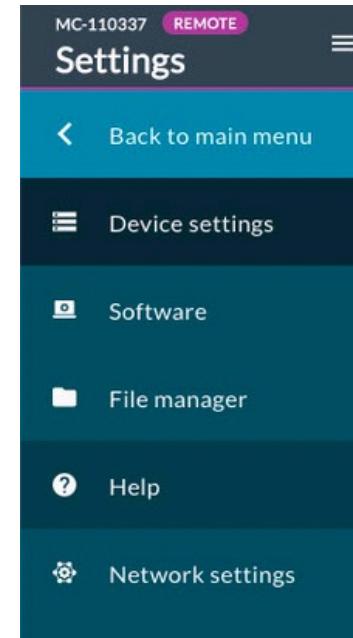




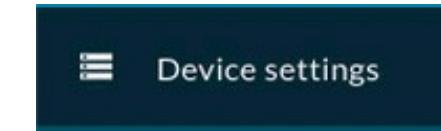
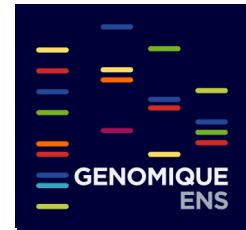
# Paramètres accessibles, file manager...



Le menu permet d'accéder au séquençage mais également aux paramètres, fichiers enregistrés...



# La section Device settings



Device settings

**System**  
MC-110337

[Shutdown](#) [Reboot](#)

**Date and time**  
Thu 2021-03-18 14:25:49 Europe/Paris

[Change date and time](#) [Change time zone](#)

**Disk management**

[Add network drive](#)

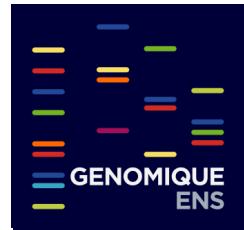
**Local disk**

Used	Free
10.2 GB used	3.6 GB free
557.6 GB used	381.1 GB free

## Permet

- De juger de l'espace disponible sur le disque local
- De redémarrer ou d'arrêter le système d'exploitation

# La section Software



Software

MinKNOW      Installed version: 21.02.2

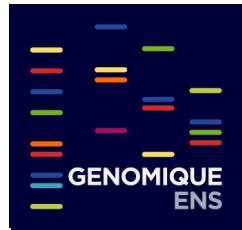
System      Total packages: 7      Security packages: 7  
If you see some system packages remain after a system update, please update MinKNOW and the number should return to zero.

Install packages

Deux types de mises à jour possible :

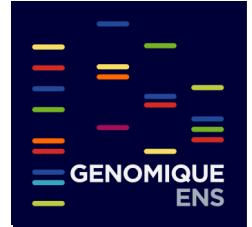
- Les mises à jour de MinKNOW
- Les mises à jour du système d'exploitation (il y en a tout le temps)

# La section File Manager



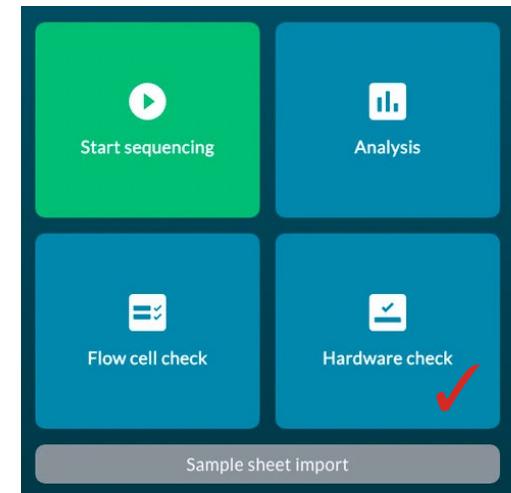
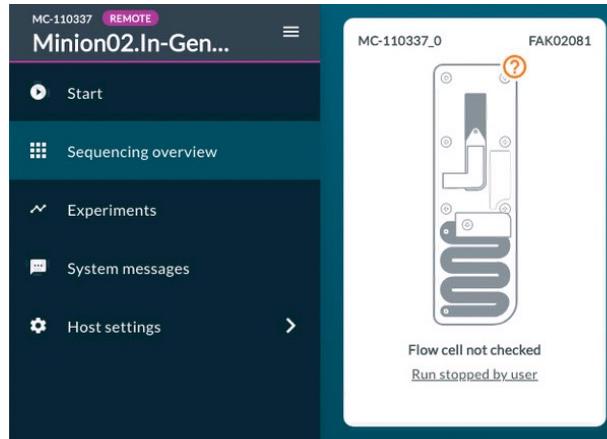
Internal		Removable		Network	
Name		Directory Count	File Count	Size	Date Modified
20210120_ScNaUMI_A2020		1	0		20 Jan 2021 15:26
20210128_RETARD_C2020		1	0		28 Jan 2021 15:13
20210201_test		2	0		01 Feb 2021 21:39
20210308_PlattyLong_A2020		1	0		08 Mar 2021 14:02
20210310_PlattyLong_A2020		1	0		10 Mar 2021 11:39

Permet la navigation dans les runs enregistrés en local mais aussi sur disque externe ou en réseau selon la configuration du labo

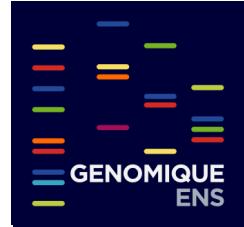


# Vérification initiale du séquenceur

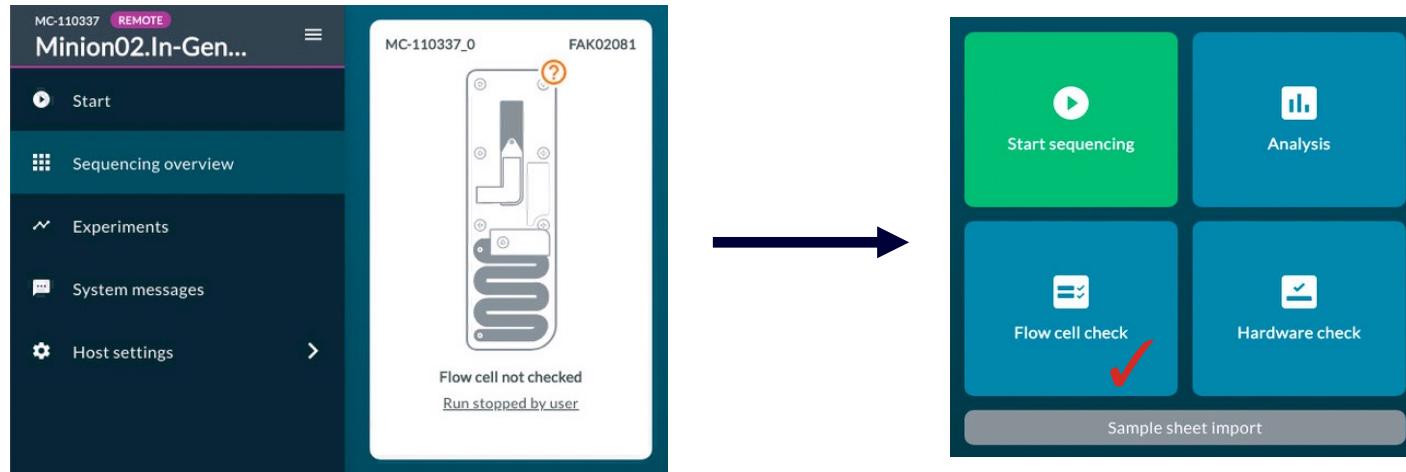
A la réception du séquenceur, test via la flowcell de configuration (CTC)



# Vérification initiale de la flowcell avant de lancer la manip



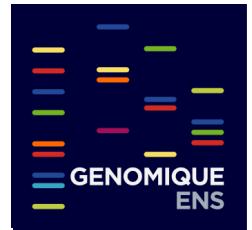
Il est nécessaire de vérifier le nombre de pores disponibles sur la flowcell avant de charger les échantillons



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

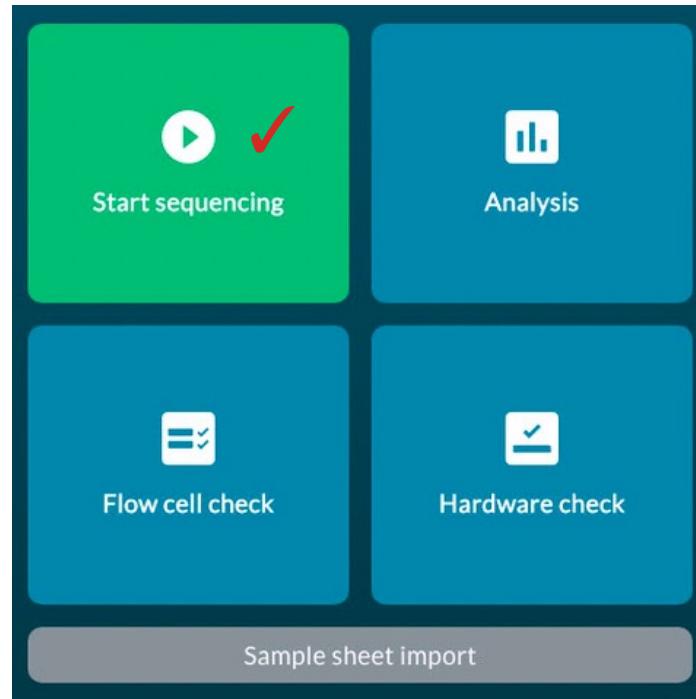
⇒ Elles sont remplacées si ce n'est pas le cas !



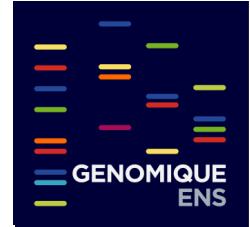
# Paramétrage et lancement d'un run

Il faut définir

- Votre expérience
- Le kit utilisé
- Le type de basecalling (choix du modèle de réseau de neurone) s'il est fait à la volée
- Les format de sortie de vos données
- L'alignement à la volée ou non et par conséquent, les séquences références



# Définition de l'expérience



Select positions

Bidon

Join existing Load saved settings :

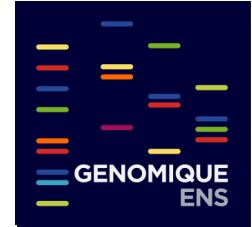
Position	Flow cell ID	Flow cell type	Sample ID
MC-110337_0	FAK02081	FLO-MIN106	Sample ID

Dans le cas d'un Mk1C ou d'un Flongle,  
il n'y a qu'une position possible.

Continue to kit selection >

# 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...



Kit selection

Sample Type:  DNA  RNA  PCR  PCR-Free Multiplexing:  Yes  No  Control [Reset Filters](#)

Sample ID	Kit ID	Sample ID	Kit ID
SQK-LSK109	SQK-RBK004	SQK-RAD004	SQK-RNA002
SQK-PBK004	SQK-165024	SQK-CAS109	SQK-CS9109
SQK-DCS109	SQK-LRK001	SQK-LSK109-XL	SQK-LSK110
SQK-LSK110-XL	SQK-NBD110-24	SQK-NBD110-96	SQK-PCB109
SQK-PCB110	SQK-PCS109	SQK-PCS110	SQK-PRC109
SQK-PSK004	SQK-RBK110-96	SQK-RNA003	SQK-RPB004
SQK-ULK001	VSK-VMK002	VSK-VMK003	VSK-VSK002
VSK-VSK003	SQK-DCS108	SQK-LSK108	SQK-PCS108
SQK-RAB204			

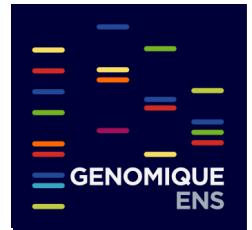
Filtre  
⟳

Kit selection

Sample Type:  DNA  RNA  PCR  PCR-Free Multiplexing:  Yes  No  Control [Reset Filters](#)

Sample ID	Kit ID	Sample ID	Kit ID
SQK-LSK109	SQK-PBK004	SQK-165024	SQK-LSK109-XL
SQK-LSK110	SQK-LSK110-XL	SQK-PSK004	SQK-RPB004
SQK-LSK108	SQK-RAB204		

[Continue to run options ➤](#)



# Configuration des options du run

Selon les projets, le temps de run peut varier. En RNASeq, le run peut durer jusqu'à 72h

Contrôle actif des canaux

Sauvegarde d'un nombre de pore pour les faire intervenir dans la durée du run

Run options

Run length ⑦ Bias voltage ⑦

- 72 hours + - -180 mV +

Hide Advanced User Options

Active channel selection ⑦ Time between MUX scans ⑦

ON - 1,5 hours +

Reserved pores ⑦

0%

Continue to basecalling >

Le voltage initial de la flowcell peut être modifié mais il vaut mieux être expert pour cela

Temps entre chaque changement des canaux

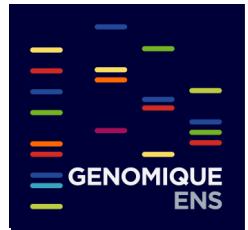
# Configuration de l'appel de base



## Basecalling

## Trois possibilités de basecalling

- Fast : Pratique pour le diagnostic parce rapide
- High-accuracy : Plus long mais moins d'erreur
- Modified : Dictionnaires de bases possibles incluent certaines bases modifiées



# Configuration du traitement des barcodes

Basecalling

Basecalling ?

Config: Fast basecalling Options

**Barcoding ?** Enabled Options

Alignment

Disabled Options

Barcoding options X

Trim barcodes ? OFF

Barcode both ends ? OFF

Mid-read barcodes ? OFF

Minimum barcoding score ? 60 - +

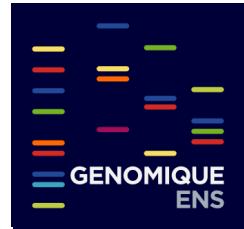
Cancel Save

Dans le cas d'utilisation de code barres :

- Suppression des code barres aux extrémités des données basecallées
- Recherche des code barres à chaque extrémité de la lecture pour classifier la lecture
  - => Si un seul des code barres est trouvé, la lecture est perdue
- Recherche de code barre au milieu de la lecture
  - => Elimination de la lecture si un code barre est trouvé

# Configuration de l'alignement à la volée

## Utilisation de minimap2



### Basecalling

#### Basecalling

Config: Fast basecalling

⚙ Options

#### Barcode

Enabled

⚙ Options

#### Alignment

Disabled

⚙ Options

### Basecalling

#### Basecalling

Config: High-accuracy basecalling

⚙ Options

#### Barcode

Enabled

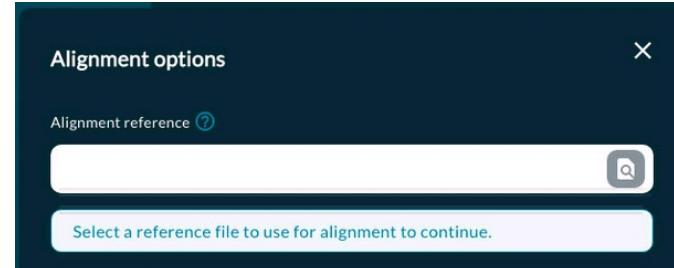
⚙ Options

#### Alignment

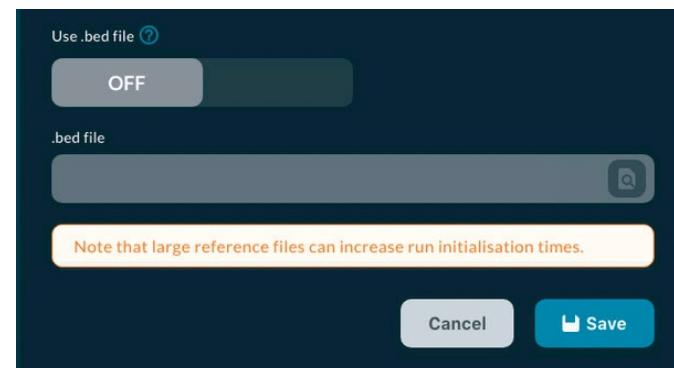
No reference selected

⚙ Options

Fichier fasta à fournir pour tout type de données à aligner

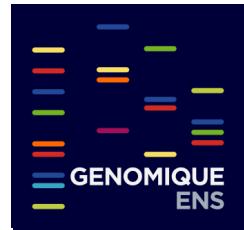


Fichier bed12 à fournir dans si les lectures à aligner sont des ADNc ou des ARN



Continue to output >

# Choix du type de fichiers de sortie et du seuil pass/fail



Données basecallées,  
demultiplexées (si besoin)  
et classées en pass/fail

Données brutes :  
Important si l'on  
veut relancer le  
basecalling

Output

Output location ? /data/.

Output format ?

FAST5  FASTQ  BAM

Filtering ?

Qscore: 7 | Readlength: Unfiltered  Options

Données alignées si  
l'alignement  
à la volée a  
été demandé

Filtering options

Qscore ? 7

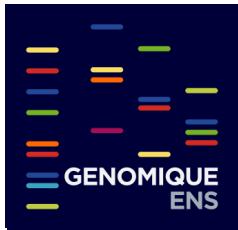
Min readlength (Kb) ? Min readlength

Max readlength (Kb) ? Max readlength

Cancel Save

Paramètres de filtres des lectures  
pass/fail  
Les lectures peuvent être filtrées sur  
leur qscore minimal et/ou leur taille

# Fichier de sortie en bulk



Output

Output location ?  
/data/.

Output format ?  
 FAST5  FASTQ  BAM

Filtering ?  
 Qscore: 7 | Readlength: Unfiltered  Options

▼ Hide Advanced User Options

Bulk file ?

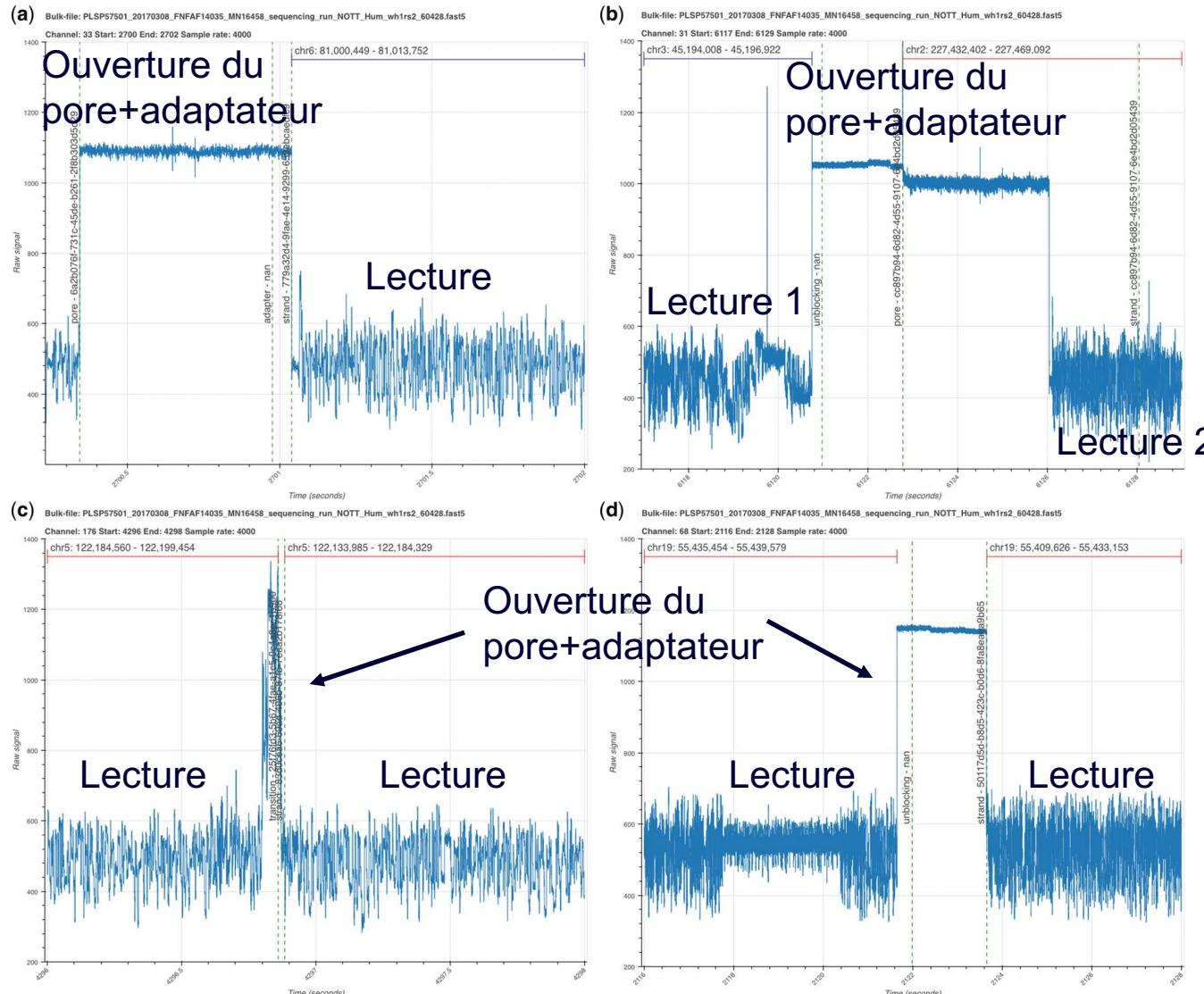
Fichier FAST5 Bulk :

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)

Attention, cette option génère un gros volume de données

# Données FAST5 bulk visualisés 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 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

**Bidon**

Kit

Selected kit: SQK-PBK004

Run options

Run length: 72 hours  
Bias voltage: -180 mV  
Adaptive sampling: Off

Advanced run options

Time between MUX scans: 1.5 hours  
Reserved pores: 0%

Basecalling

Basecalling: On (High-accuracy basecalling)  
Barcode: On  
Minimum barcode score: 60  
Alignment: Off

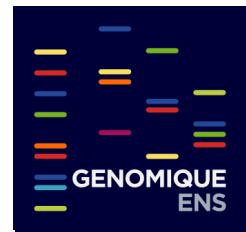
Output

Location: /data/  
FAST5: On (Raw, FASTQ Record, zlib, 4000 reads per f...  
FASTQ: On (4000 reads per file)  
Read filtering: Qscore: 7 | Readlength: Unfiltered

Advanced run options

Bulk file: Off

# Quels outils pour l'alignement ?



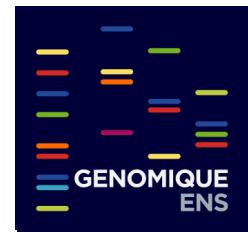
Il existe très peu d'outils fonctionnels pour l'alignement de longues lectures

- BWA
  - n'évolue plus
  - ne prend pas en charge les jonctions
- GMAP
  - maintenu (dernière version le 17/12/21)
  - dédié aux ARNs.
  - Versions anciennes non fonctionnelles sur ARN longues lectures, versions récentes non testées
- Minimap2
  - le standard depuis sa sortie
  - même développeur que BWA
  - prend en charge les données de tout type (ADN, ARN)

# Minimap2

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

Dernière version 26/12/21



-Outil versatile :

- ARN, ADNc, ADN génomiques
- Lectures PacBio, lectures ONT

-Outil peu gourmand en mémoire et très rapide

-Fichier de sortie au format SAM

-Possibilité de joindre à l'alignement un fichier au format bed12 (jonctions exon-exon) pour aider l'alignement des données épissées

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

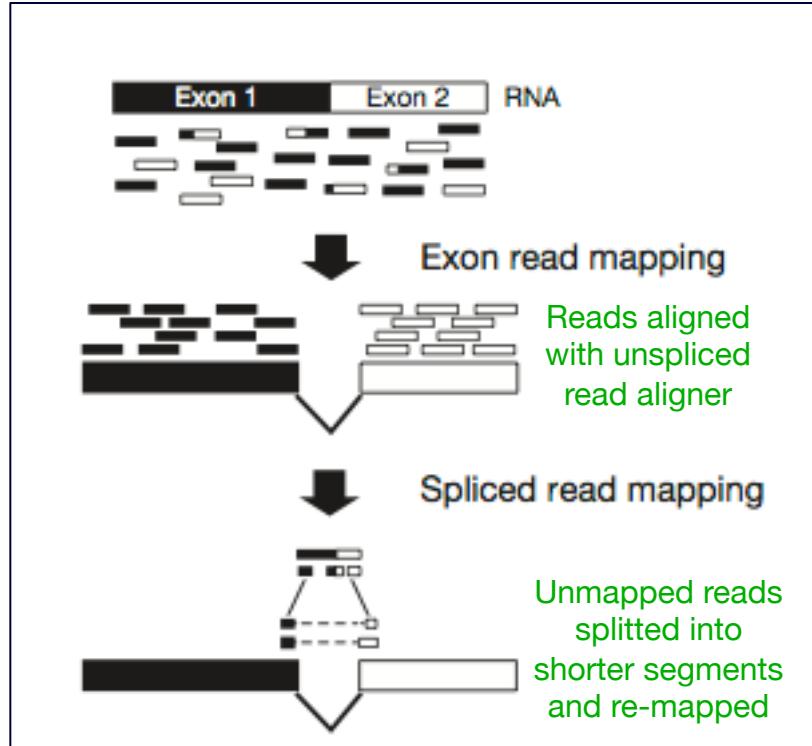
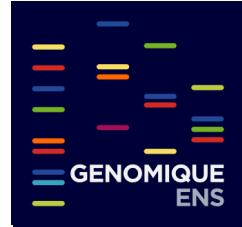
```
minimap2 -ax splice -uf -k14 ref.fa direct-rna.fq > aln.sam
```

Attention => Outil exon-first : il peut favoriser les processed-pseudogenes

# Les stratégies d'alignement...

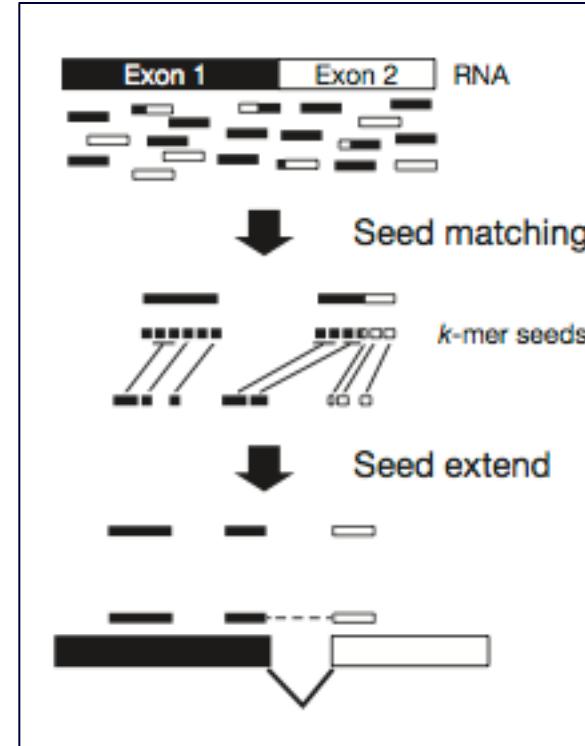
Alineurs développés pour l'alignement d'ARNm

- Les jonctions sont prises en compte



Stratégie exon first

Récupération des lectures correspondant aux jonctions, non alignées dans un 1<sup>er</sup> temps, découpage pour réalignement

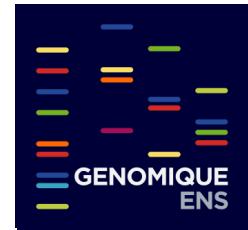


Stratégie junction first

Découpage pour alignement/extension en une seule fois

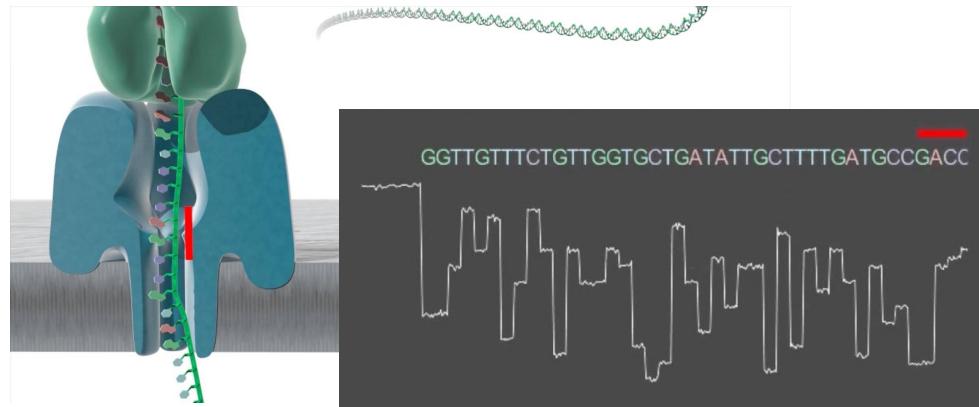
Nat Methods 2011 Garber

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



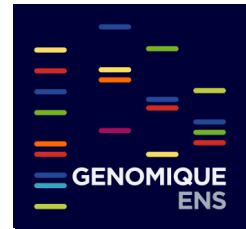
Le séquençage nanopore ne se limite pas aux bases A,T,G,C

Il permet la détection des U et de bases modifiées (seulement m6A pour le moment en séquençage d'ARN)



- 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 (guppy)

## 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

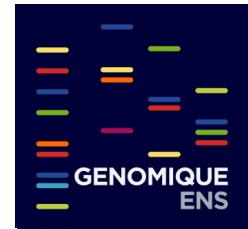
## Probabilité d'identifier une base par erreur

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 %

# Ordre de grandeur en séquençage ONT

# 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 1 >

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

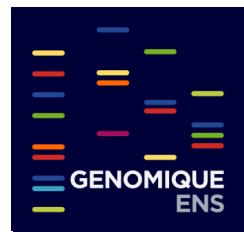
- **Une base = une lettre**

→ 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

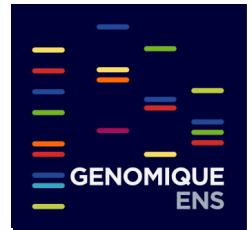
= Code américain normalisé pour l'échange d'information

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	1101001	151	i
10	A	1010	12	[LINE FEED]	58	3A	1110100	72	:	106	6A	1101010	152	j
11	B	1011	13	[VERTICAL TAB]	59	3B	1110110	73	;	107	6B	1101011	153	k
12	C	1100	14	[FORM FEED]	60	3C	1111000	74	<	108	6C	1101100	154	l
13	D	1101	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	10001	21	[DEVICE CONTROL 1]	65	41	100000001	101	A	113	71	1110001	161	q
18	12	10010	22	[DEVICE CONTROL 2]	66	42	100000010	102	B	114	72	1110010	162	r
19	13	10011	23	[DEVICE CONTROL 3]	67	43	100000011	103	C	115	73	1110011	163	s
20	14	10100	24	[DEVICE CONTROL 4]	68	44	1000000100	104	D	116	74	1110100	164	t
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]	69	45	1000000101	105	E	117	75	1110101	165	u
22	16	10110	26	[SYNCHRONOUS IDLE]	70	46	1000000110	106	F	118	76	1110110	166	v
23	17	10111	27	[ENG OF TRANS. BLOCK]	71	47	1000000111	107	G	119	77	1110111	167	w
24	18	11000	30	[CANCEL]	72	48	10000001000	110	H	120	78	1111000	170	x
25	19	11001	31	[END OF MEDIUM]	73	49	10000001001	111	I	121	79	1111001	171	y
26	1A	11010	32	[SUBSTITUTE]	74	4A	10001000000	112	J	122	7A	1111010	172	z
27	1B	11011	33	[ESCAPE]	75	4B	10001000001	113	K	123	7B	1111011	173	{
28	1C	11100	34	[FILE SEPARATOR]	76	4C	10001100000	114	L	124	7C	1111100	174	
29	1D	11101	35	[GROUP SEPARATOR]	77	4D	10001100001	115	M	125	7D	1111110	175	}
30	1E	11110	36	[RECORD SEPARATOR]	78	4E	100011000000	116	N	126	7E	1111110	176	~
31	1F	11111	37	[UNIT SEPARATOR]	79	4F	100011000001	117	O	127	7F	1111111	177	[DEL]
32	20	1000000	40	[SPACE]	80	50	1010000000000	120	P					
33	21	1000001	41	!	81	51	1010000000001	121	Q					
34	22	1000010	42	"	82	52	1010000000010	122	R					
35	23	1000011	43	#	83	53	1010000000011	123	S					
36	24	1001000	44	\$	84	54	1010000000100	124	T					
37	25	1001001	45	%	85	55	1010000000101	125	U					
38	26	1001100	46	&	86	56	1010100000000	126	V					
39	27	1001101	47	'	87	57	1010100000001	127	W					
40	28	1010000	50	(	88	58	1011000000000	130	X					
41	29	1010001	51	)	89	59	1011000000001	131	Y					
42	2A	1010100	52	*	90	5A	1011010000000	132	Z					
43	2B	1010110	53	+	91	5B	1011011000001	133	I					
44	2C	1011000	54	,	92	5C	1011100000000	134	\					
45	2D	1011001	55	-	93	5D	1011101000001	135	J					
46	2E	1011100	56	.	94	5E	1011110000000	136	^					
47	2F	1011110	57	/	95	5F	1011111000001	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

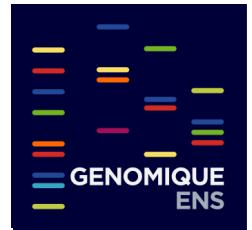
@Ligne d'identifiant

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
GGTATGCTCGTTGGTTCAAGGTGGGTGTTCTTGATCCATCATCGTACTTCCAGTTCTATCGTGTTCCTATTCTGTTGGTGTGAT
ATTGCGGGTCTGCTGGGTGTTAACCTAACGCGATGGTATCACACATGAGTACGGGGCCTAGCCTCCGCCTCCAAAGCCAACGTCTCCGCCGTCG
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

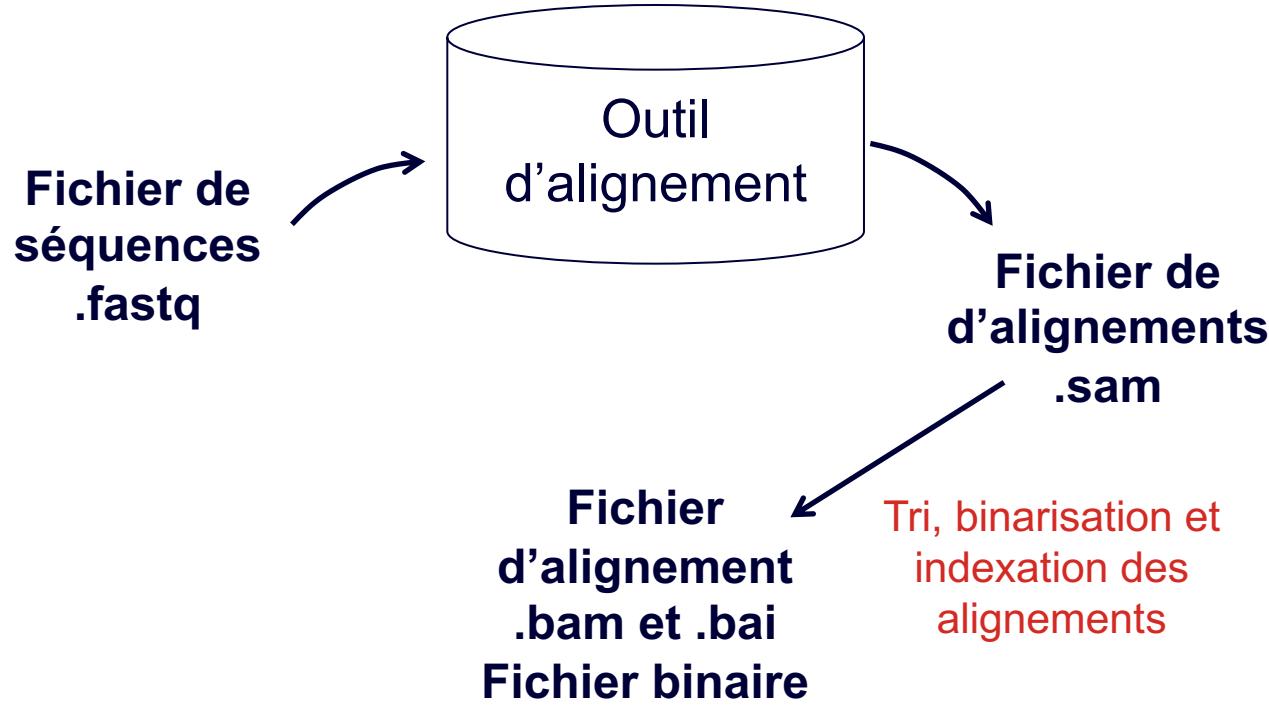
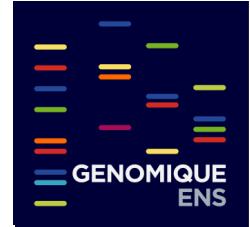
+Ligne d'identifiant (répétition de la ligne1)

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
UUAAGGAGUGGCUGGACAGGCCUGAGUGGCCAUCCGUUCAUACAGUUUCAUUCCCAGGGGCGGGGCUCCUCCCCUUUGGCAUUUA
UACUUGGUGGGGGCAUUGAAGGCCUGCGGUGGAAGGCUUGGGACUGGUCCUGUAACUGUGUGUCCCCUGAAGGCUGAGUUGGCCAUUUGGGUUAG
UCCCAGGUGACAGAAGGGAGAAACAAAAUGAAGCAGAAGAAAAUGCAAUUGCAGAUUCAGAGAACUUGACAGGAAGAACUU
UAAAAAAUGCUACAGUCACCUGGUGGCCUUGUAGGGGCCUACCCUGGUGUGGGUCAGGCCACUGCUCAGCCACUUGGGUCGUGGUUGU
AGUACCAGUUUCUGGGAGGAAGACAUGUUGGAGACUGCCAGAAAUCCCUGUACAUCGUUUUACGUGCUUGUGUCCAAGUGAAUUGUAUUGGUUC
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,)*( '%$
```

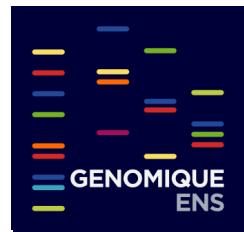
# Les formats de fichiers en sorties d'alignment

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

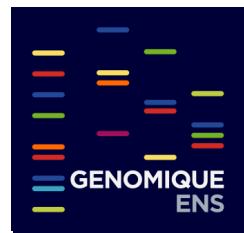
```
@HD    VN:1.4  SO:unsorted
@SQ    SN:1    LN:248956422
@SQ    SN:2    LN:242193529
@SQ    SN:3    LN:198295559
@SQ    SN:4    LN:190214555
@SQ    SN:5    LN:181538259
@SQ    SN:6    LN:170805979
@SQ    SN:14   LN:107043718
...
@SQ    SN:X    LN:156040895
@SQ    SN:Y    LN:57227415
@SQ    SN:MT  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

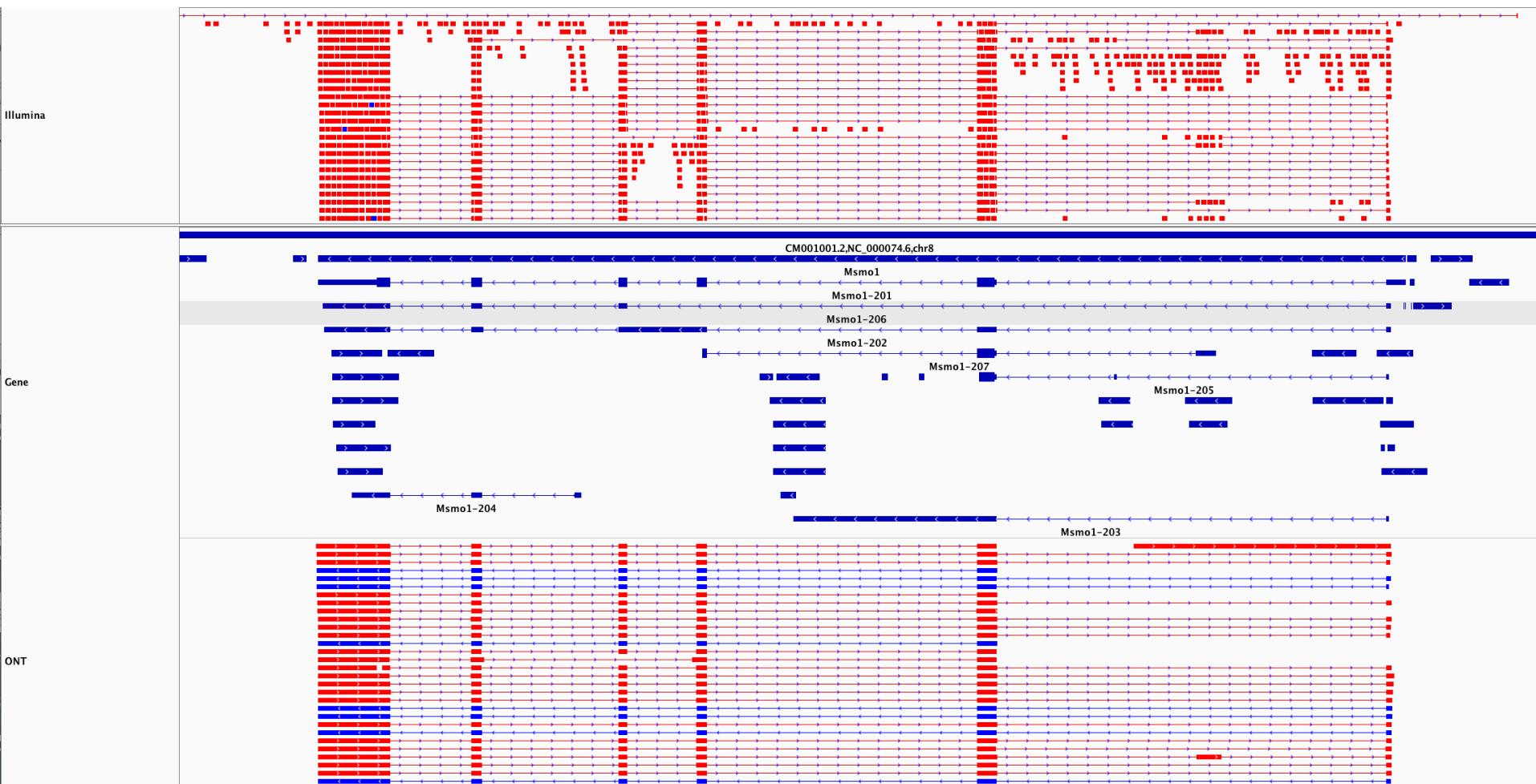
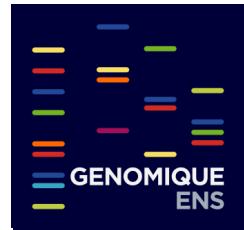
Chaque alignement est décrit par une ligne tabulée

Certains champs sont obligatoires :

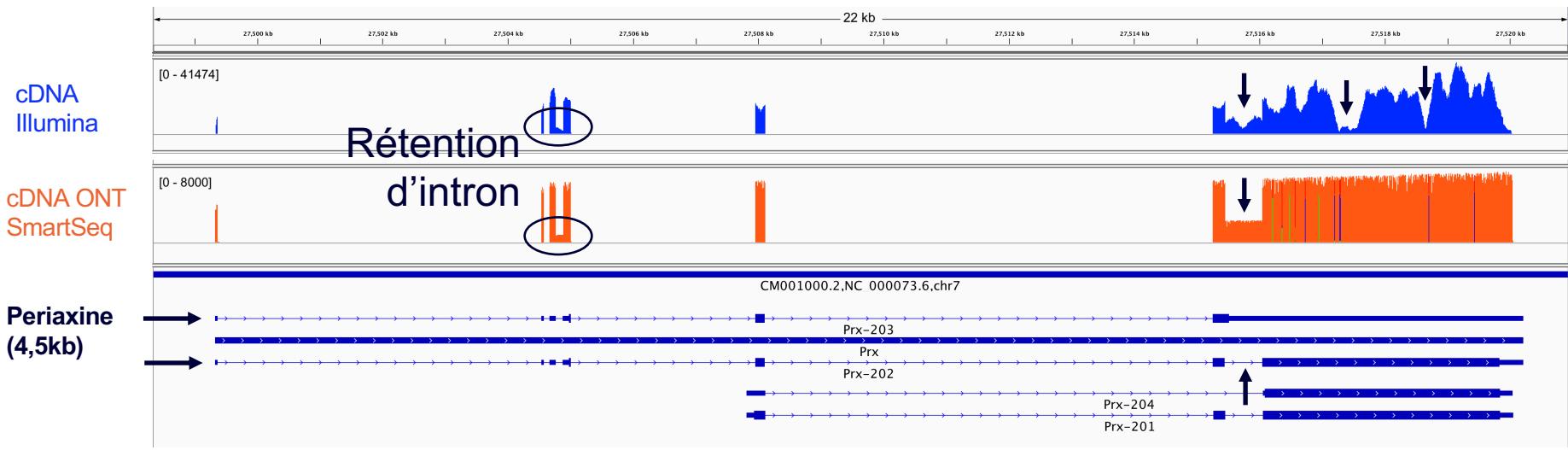
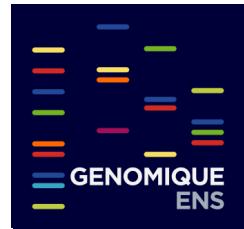
QNAME	Id Lecture	6b109bc4-8bbd-408e-9089-7e4b530ce5e6
FLAG	016 sens / 16 reverse	0
RNAME	Ref Sequence	3
POS	Start alignment	141 738 285
MAPQ	QScore d'alignement (0-255)	60
CIGAR	M = Match/Mismatch N = Gap S = SoftClipped I = Insertion	152S16M1D33M3I8M2D19M1I24M1D7M1D45M2I1M1I66M1D7M49 92N48M1602N37M1D79M2D8M3I25M2I35M4I36M1I4M1D49M1I5M 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 séquençage de transcrits sur nanopores

## Exemples de lectures alignées Illumina / ONT



# Le séquençage ONT permet une couverture homogène le long du transcript



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