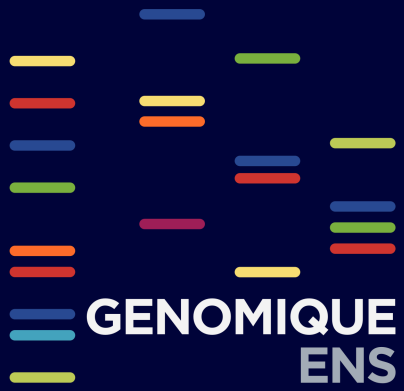


Formation à l'utilisation d'un séquenceur Minlon

Session informatique

Paris

Avril 2023

L’équipe de la plateforme génomique de l’IBENS

Catherine Senamaud-Beaufort

Corinne Blugeon

Morgane Thomas-Chollier

Laurent Jourdren

Sophie Lemoine

Mehdi Ait Djoudi Oufella

Oumy Seidy

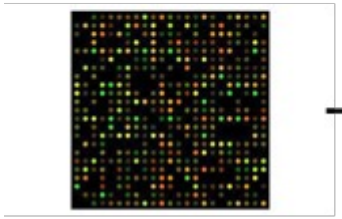
Stéphane Le Crom

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)

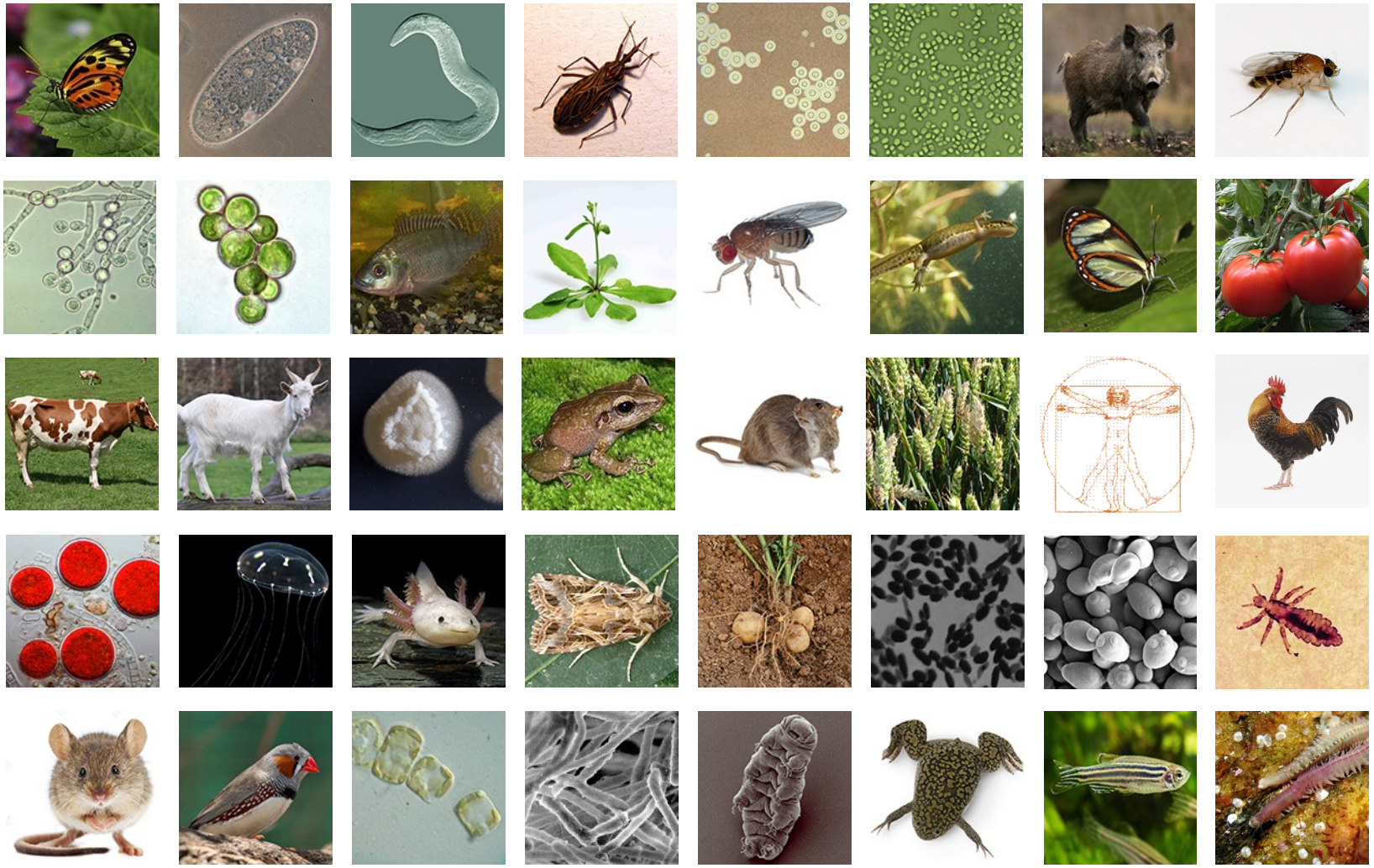
Illumina NextSeq 2000
2nd generation sequencer
(Since 2021)

Nanopore MinION
3rd generation sequencer
(Since 2016)

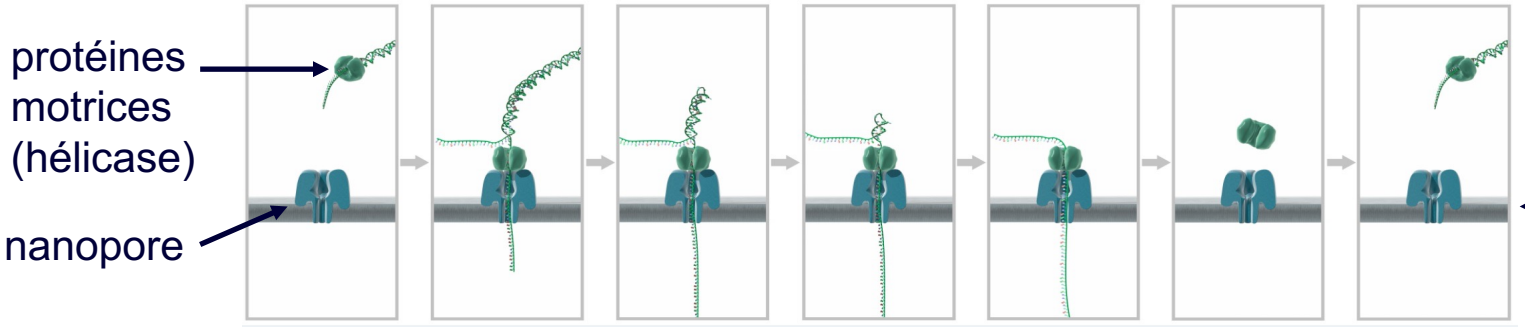
Nanopore PromethION
3rd generation sequencer
(Since 2022)

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

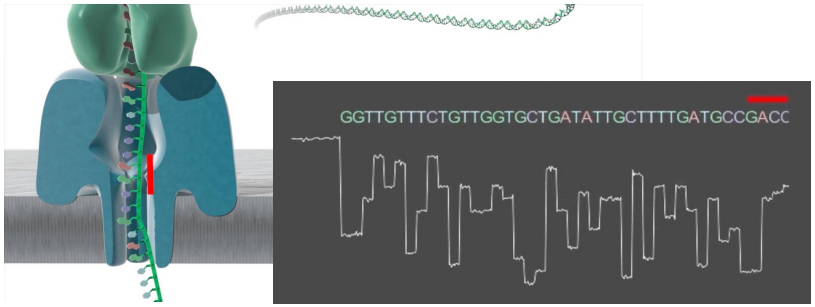
Spécialisée sur les organismes eucaryotes



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

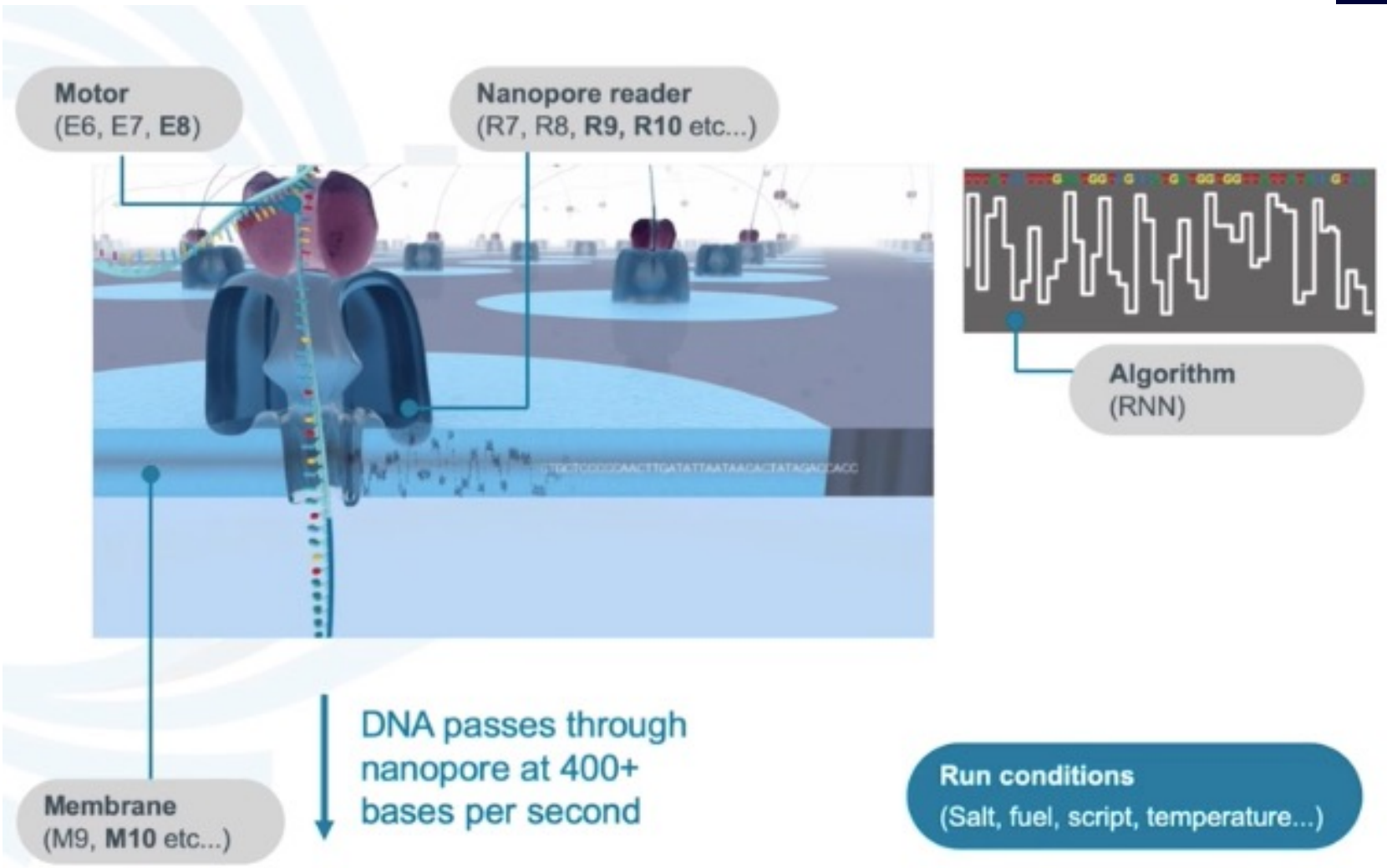
- Assemblage de génome simplifié
- ARN séquencés en pleine longueur

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…

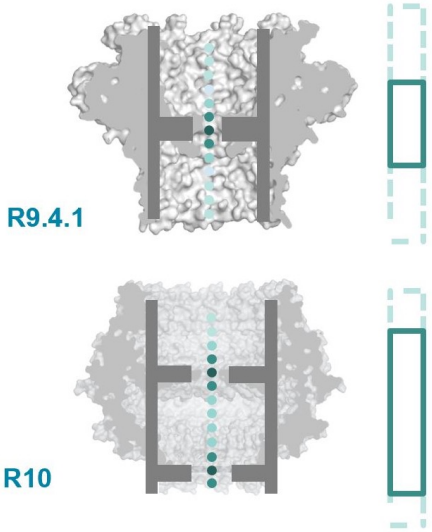
https://nanoporetech.com/

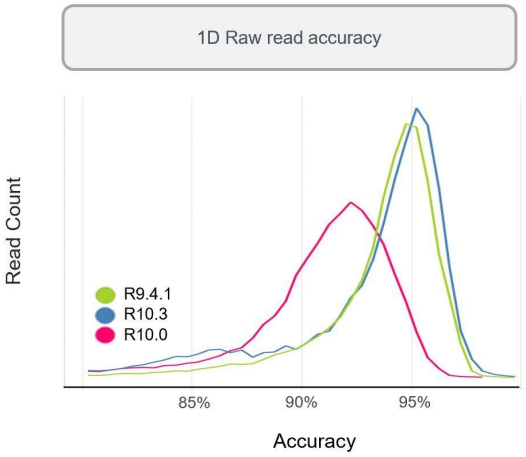
Les différents éléments d'un run

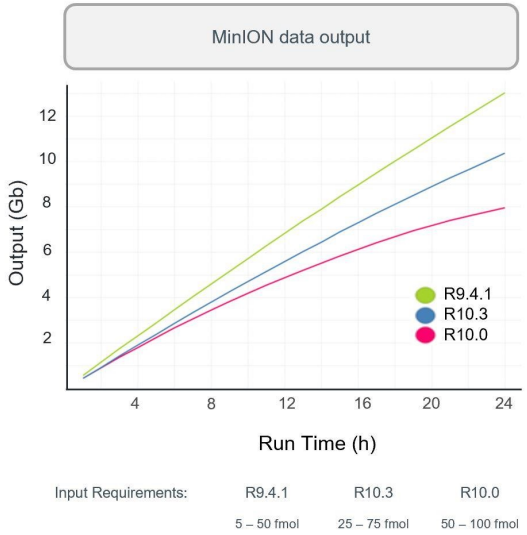


Pores types

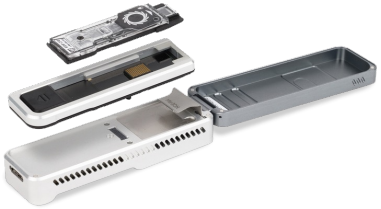
- There are **two types** of flowcell **pores** :
  - **R9.4.1** with one reader head (01/2017, 13 Gb/24h, 95 % accuracy, 5-50 fmol)
  - **R10.4.1** with two reader zones (03/2022, ?? Gb/24h, 99 % accuracy, ?? fmol)







Available Flowcells



	Flongle	MinION	PromethION
Max yield	2 GB	44 Gb	242 Gb
Channel count	126	512	2,675
Type	R 9.4.1 / R10.4.1	R9.4.1 / R10.4	R9.4.1 / R10.4 / R10.4.1

ONT Sequencers






	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION	PromethION P2	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	Nvidia Jetson TX2 (Arm)	Intel i7	Intel CPU	?	N/A
GPU	N/A	Nvidia Jetson TX2	1 x Nvidia GV100	4 x Nvidia A100	1 x Nvidia A100	N/A
Memory	N/A	8 GB	64 GB	512 GB	128 GB	N/A
Storage	N/A	1 To SSD	4 TB	60 TB	16TB SSD	N/A
Support	Soon discontinued?	Full support	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 use actually **Ubuntu 18.04** for MinION Mk1C and **Ubuntu 20.04** for GridION and PromethION.

- 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.

PromethION P2 solo

- The PromethION P2 solo works like a MinION Mk1B.

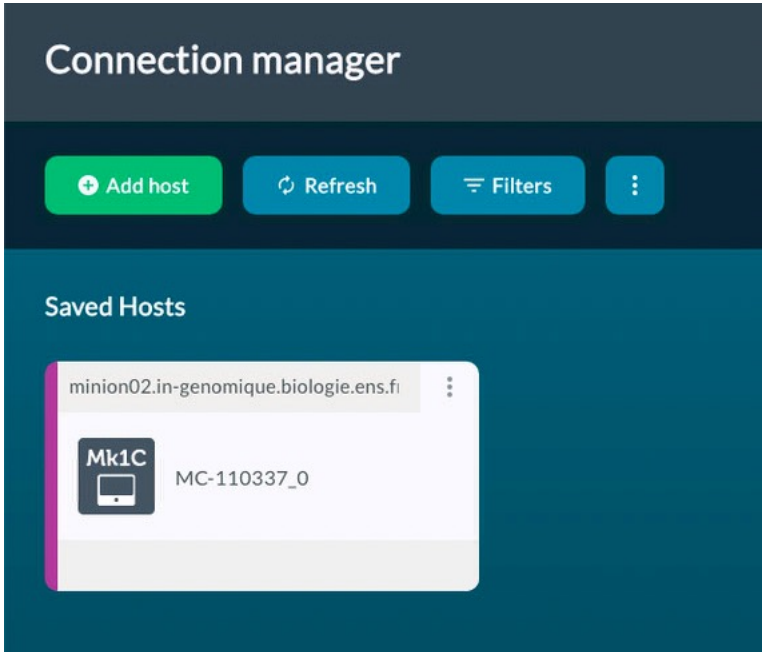
  This device is currently in **early-access** for both **hardware** and **software**. The MinION MinKNOW version is not compatible with the P2 solo.

- Requirements:
  - Ubuntu 20.04, macOS 10.14 or Windows 10
  - USB-C 5 Gb/s
  - 2 To internal SSD + 6-8 To SDD for data storage
  - NVIDIA RTX 3080 avec 12 Go / NVIDIA RTX A6000
  - 64 Go RAM, 8-12 cores (Intel i7 7 generation or AMD Ryzen)

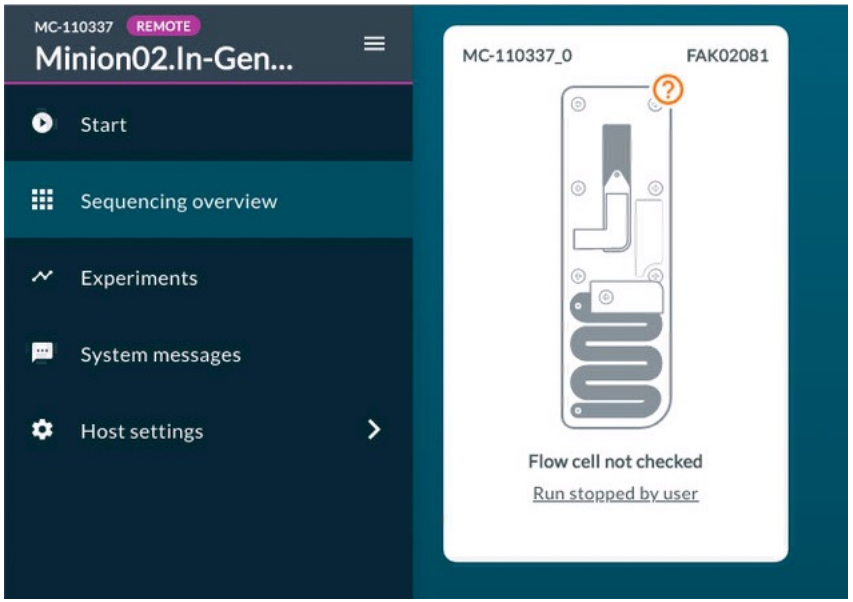
- Each run will produce **2-3 TB** of raw data, and need 2-3 days for basecalling with 2 NVIDIA RTX A6000 (“sup” quality).

Utilisation du MinION Mk1C via son interface graphique

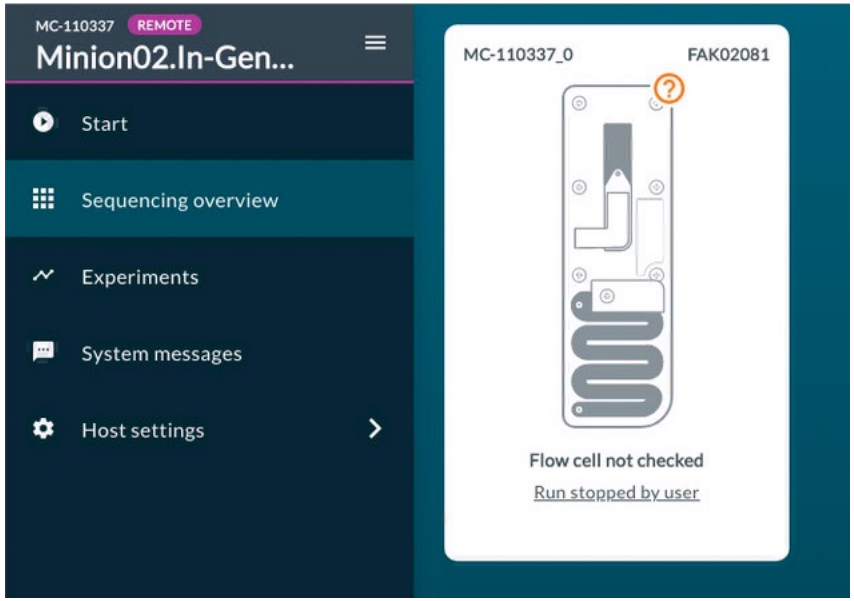
1- Sélection et connexion au séquenceur à utiliser



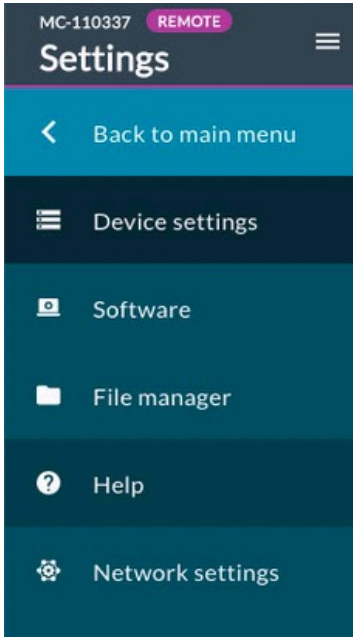
2- Sélection de la Flowcell en place



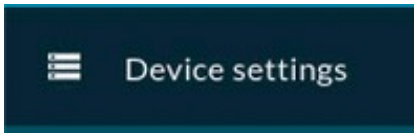
Paramètres accessibles, file manager…

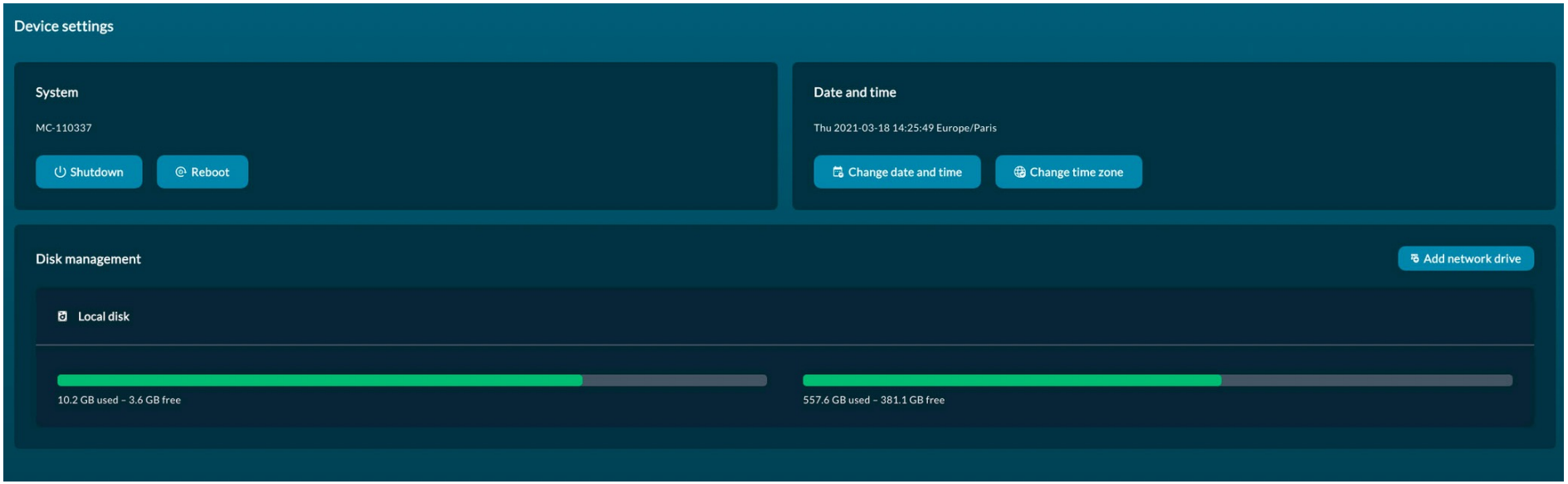


Le menu permet d'accéder au séquençage mais également aux paramètres, fichiers enregistrés…



La section Device settings

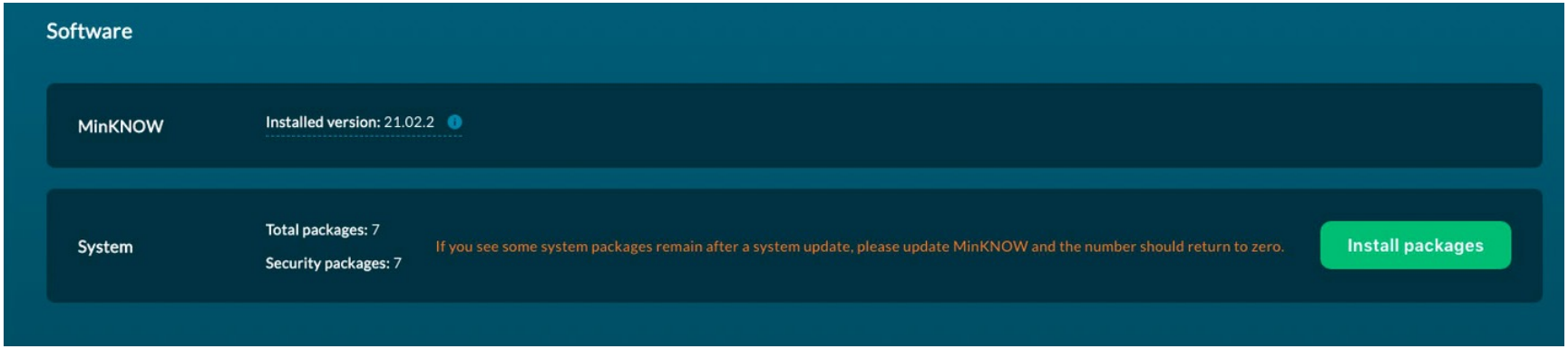




Permet

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

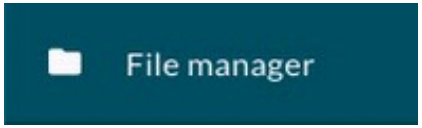
La section Sofware



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)

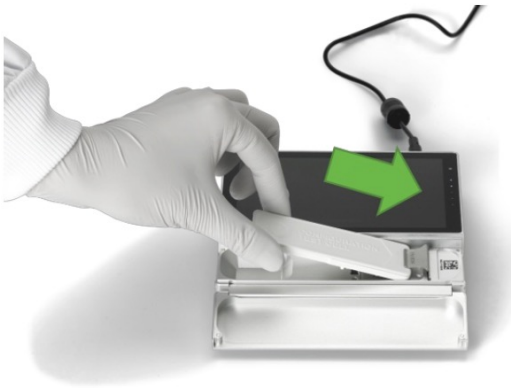
La section File Manager

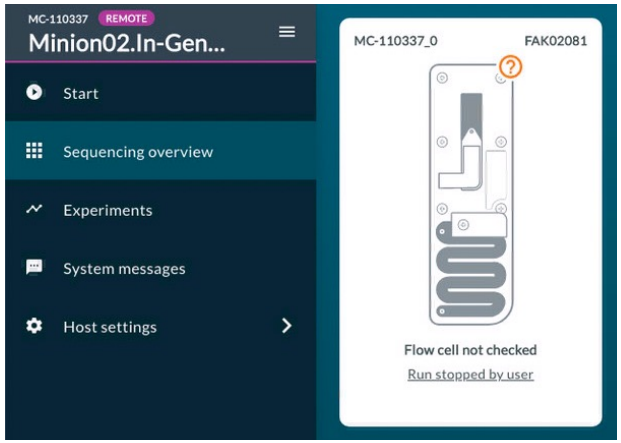


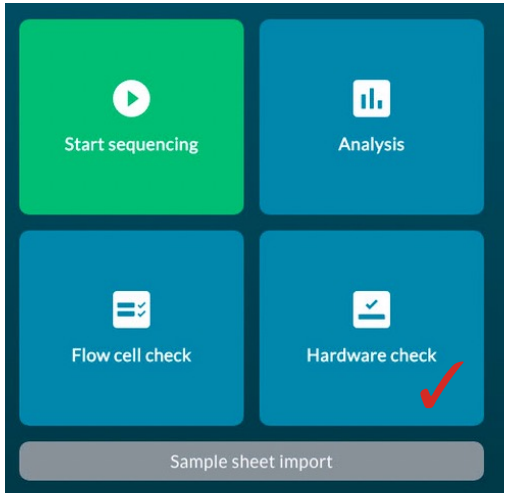
Permet la navigation dans les runs enregistrés en local mais aussi sur disque externe ou en réseau selon la configuration du labo

Vérification initiale du séquenceur

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

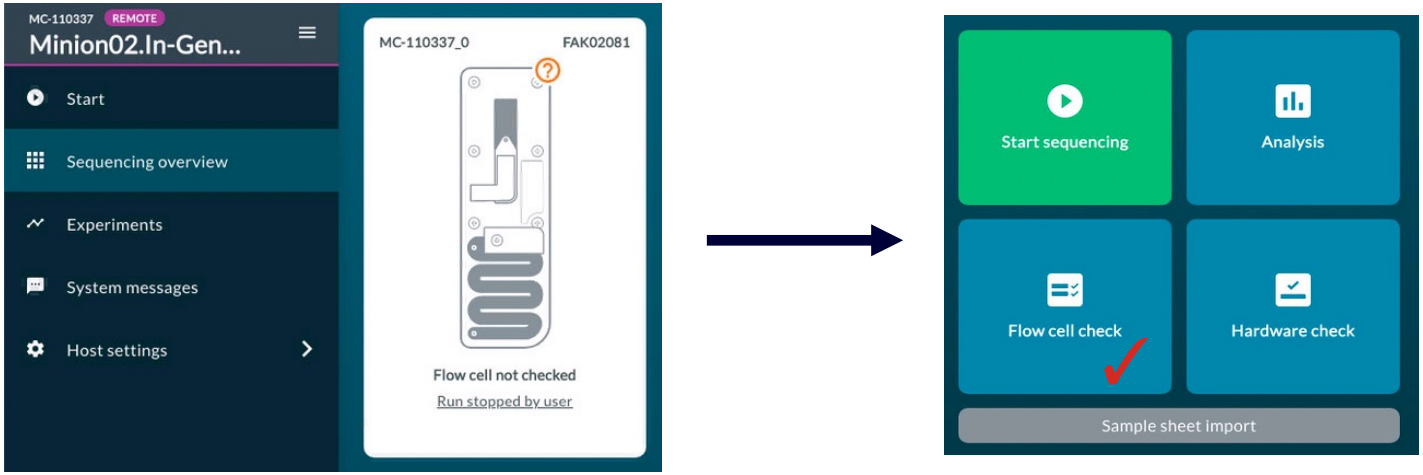






Vérification initiale de la flowcell avant de lancer la manip

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 :

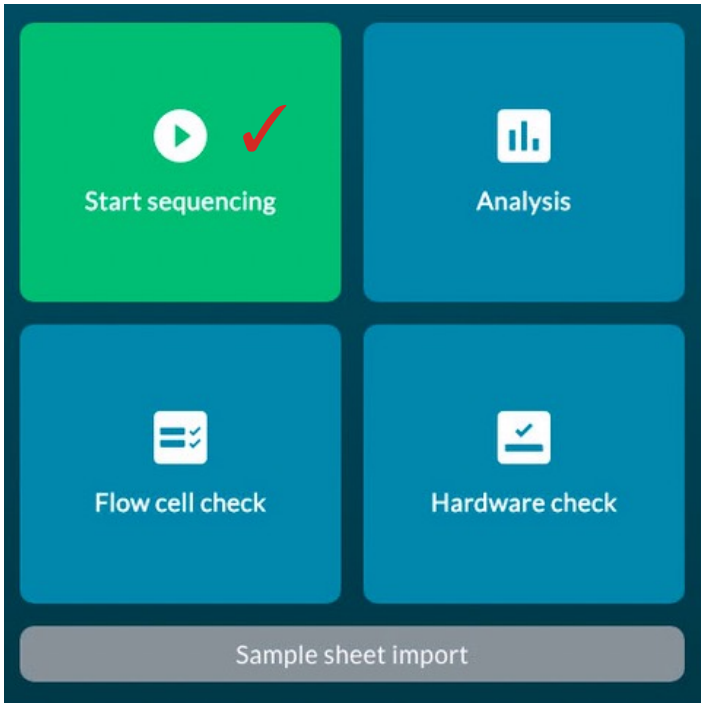
- 50 dans le cas d'une flowcell Flongle
- 800 dans le cas d'une flowcell MinION

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

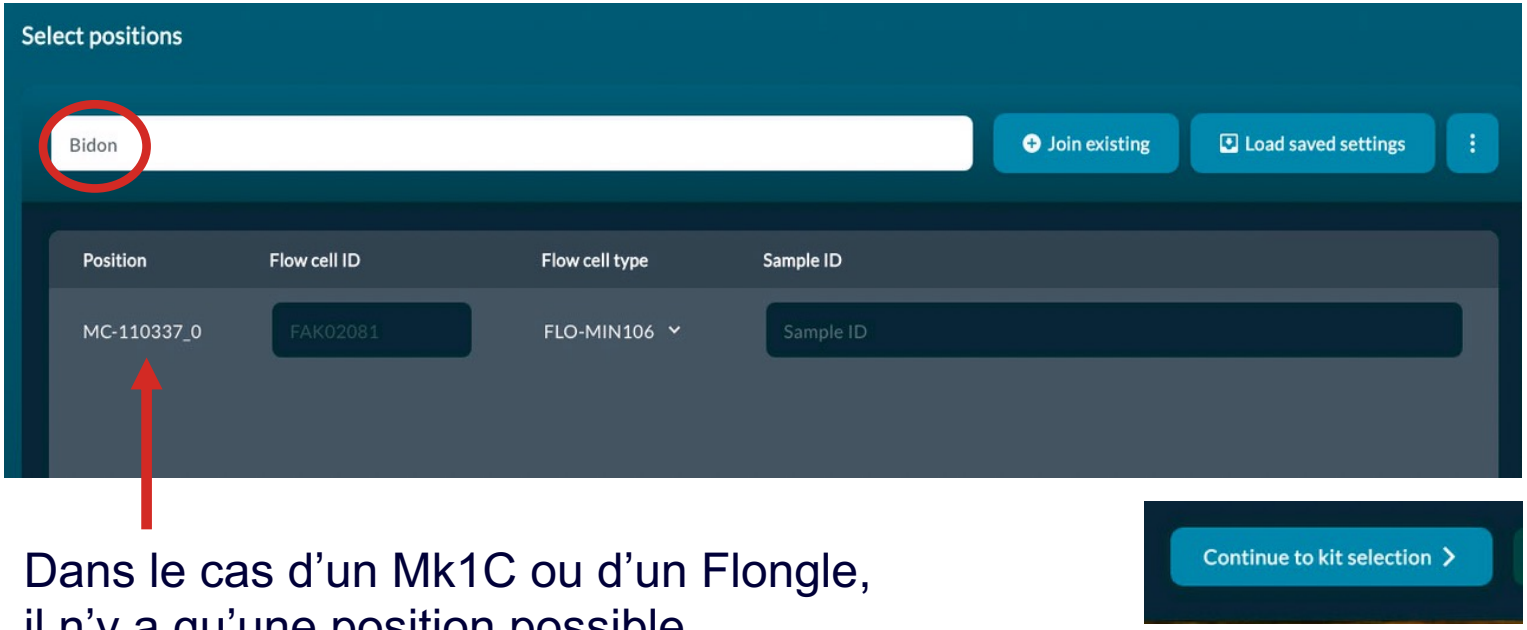
Paramétrage et lancement d'un run

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

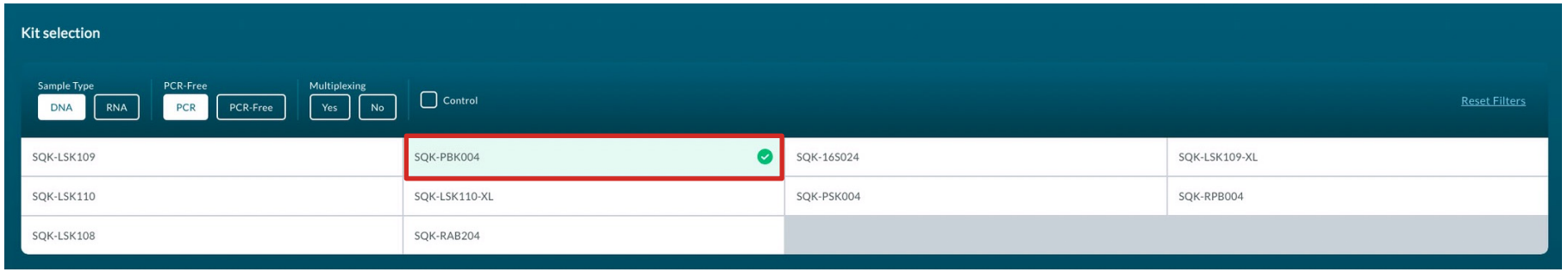


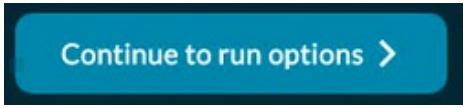
Dans le cas d'un Mk1C ou d'un Flongle,
il n'y a qu'une position possible.

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…

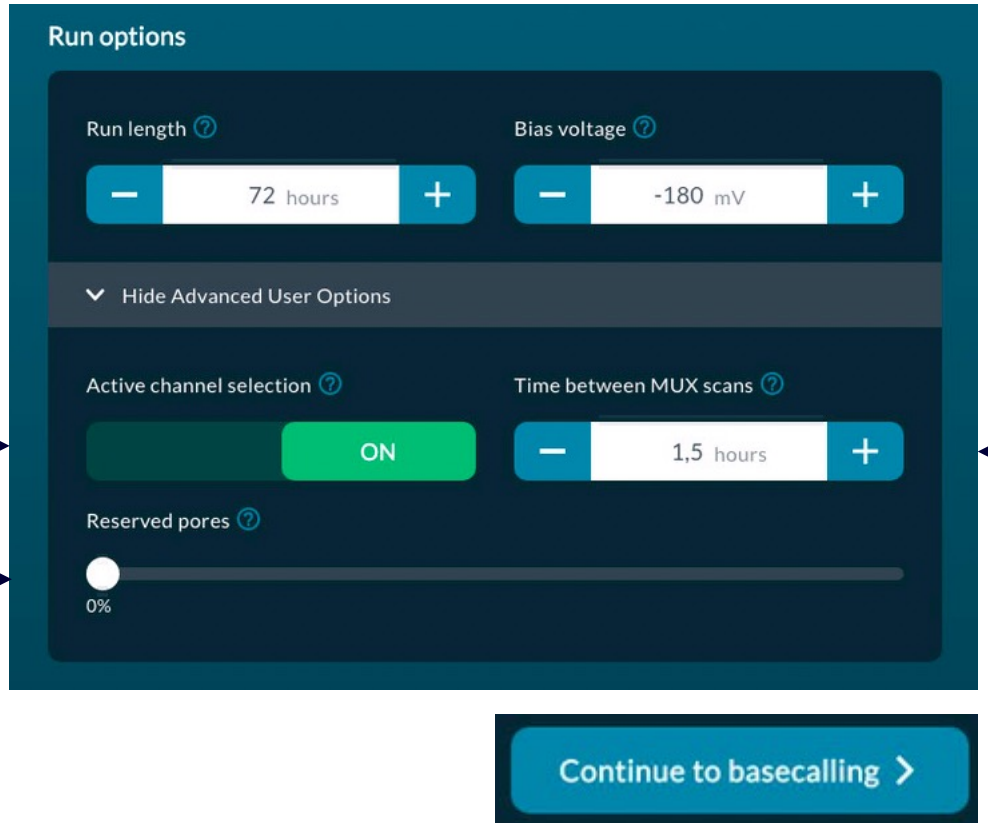
Filtre





Configuration des options du run

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



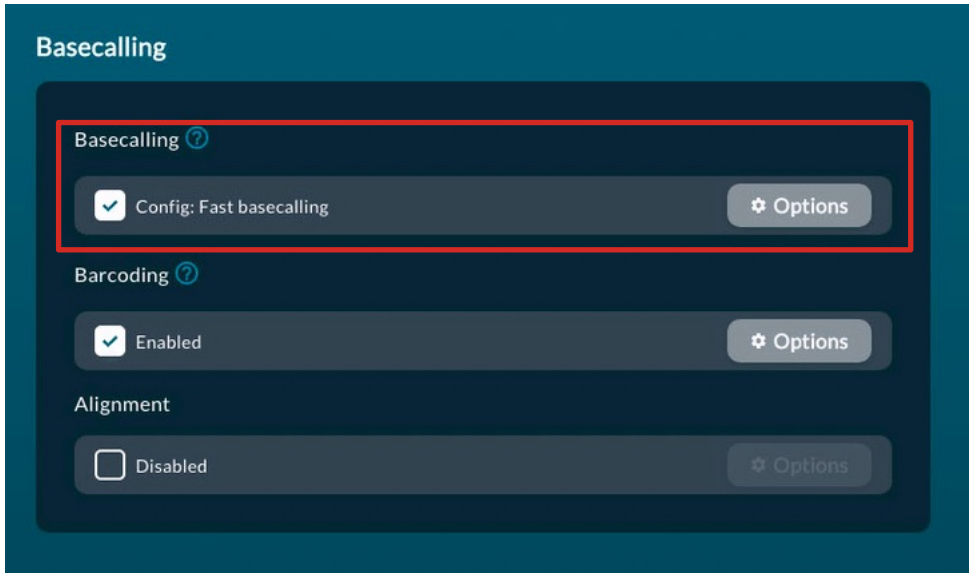
Le voltage initial de la flowcell peut être modifié mais il vaut mieux être expert pour cela

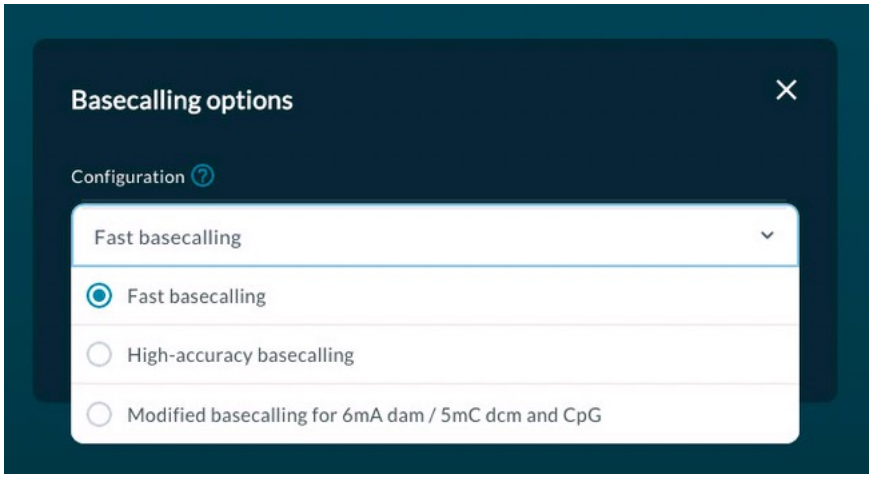
Contrôle actif des canaux

Temps entre chaque changement des canaux

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

Configuration de l'appel de base

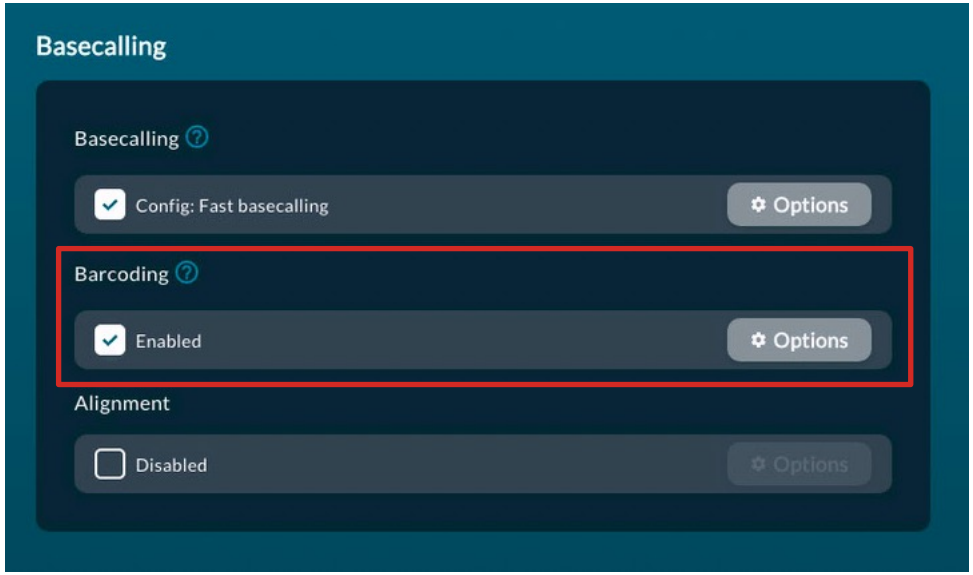


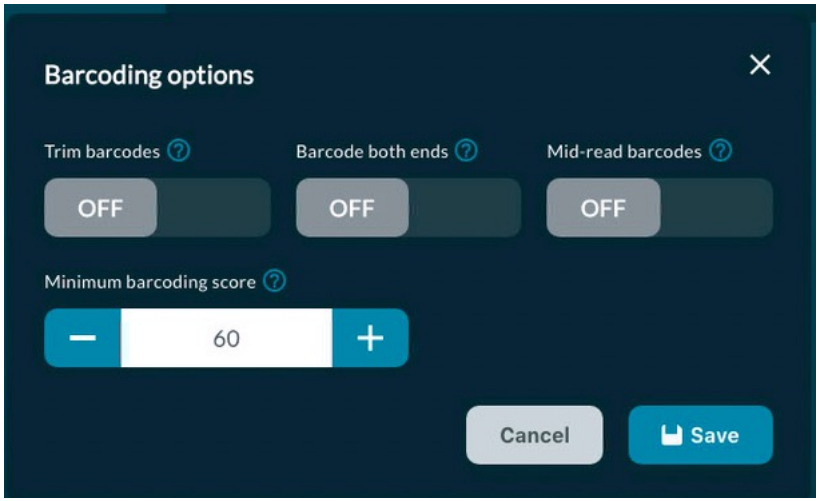


Trois possibilités de basecalling

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

Configuration du traitement des barcodes

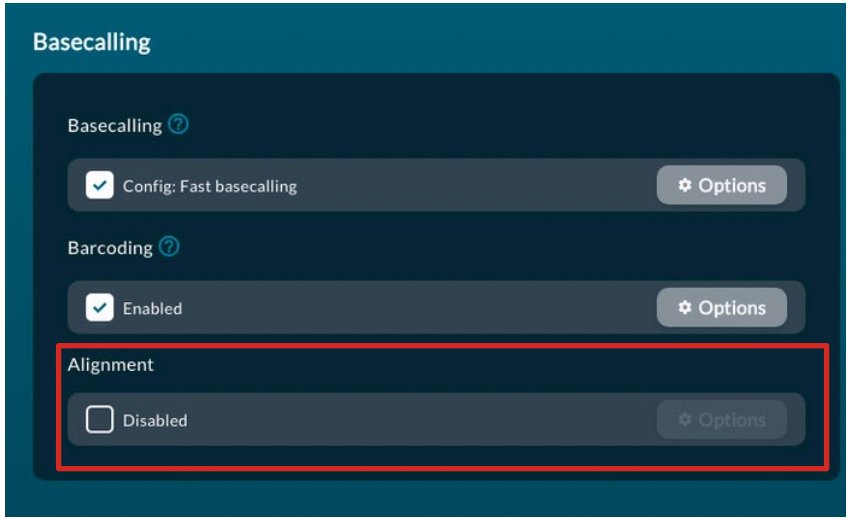




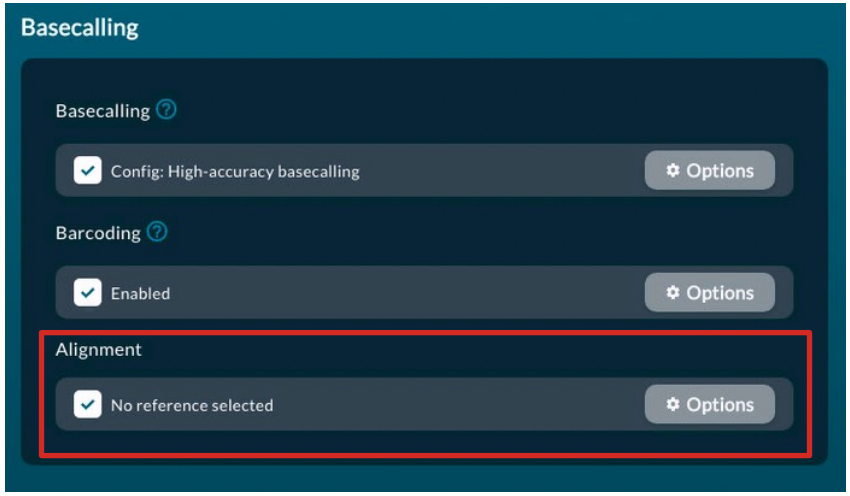
Dans le cas d'utilisation de code barres :

- 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

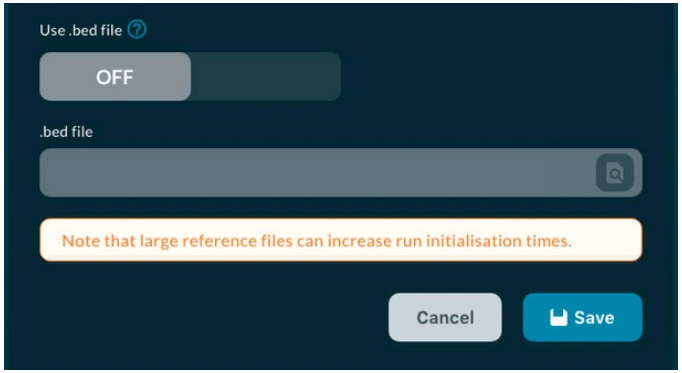
Configuration de l'alignement à la volée
Utilisation de minimap2

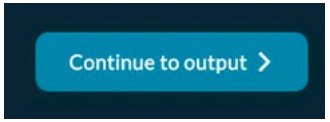


Fichier fasta à fournir pour tout type de données à aligner



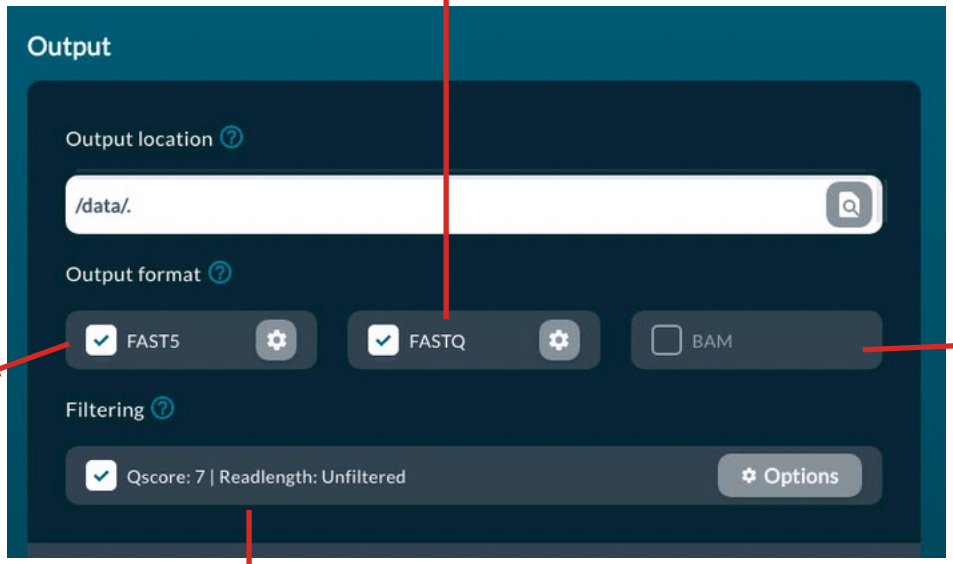
Fichier bed12 à fournir dans si les lectures à aligner sont des ADNc ou des ARN





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

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

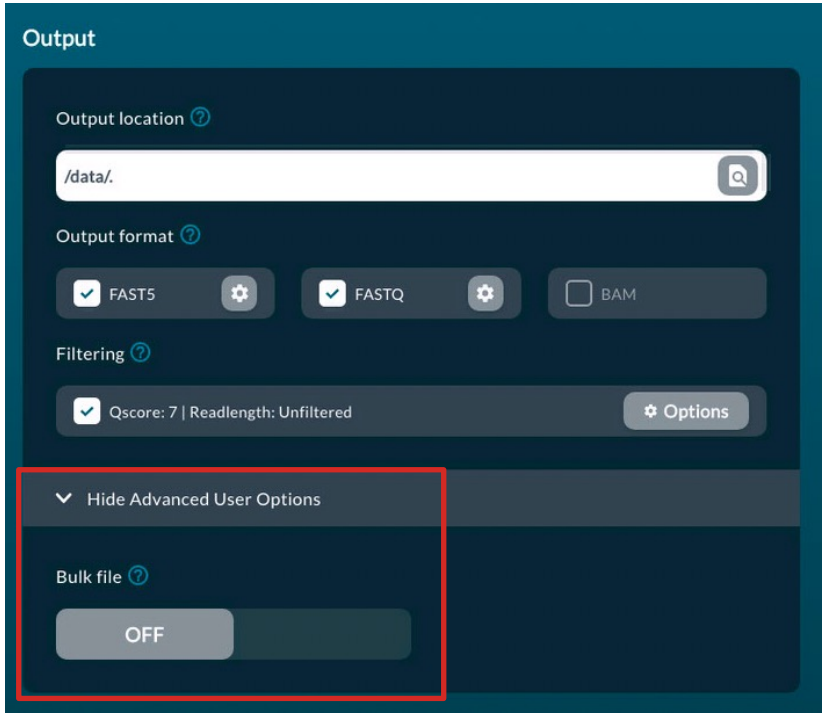


Données brutes : Important si l'on veut relancer le basecalling

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

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

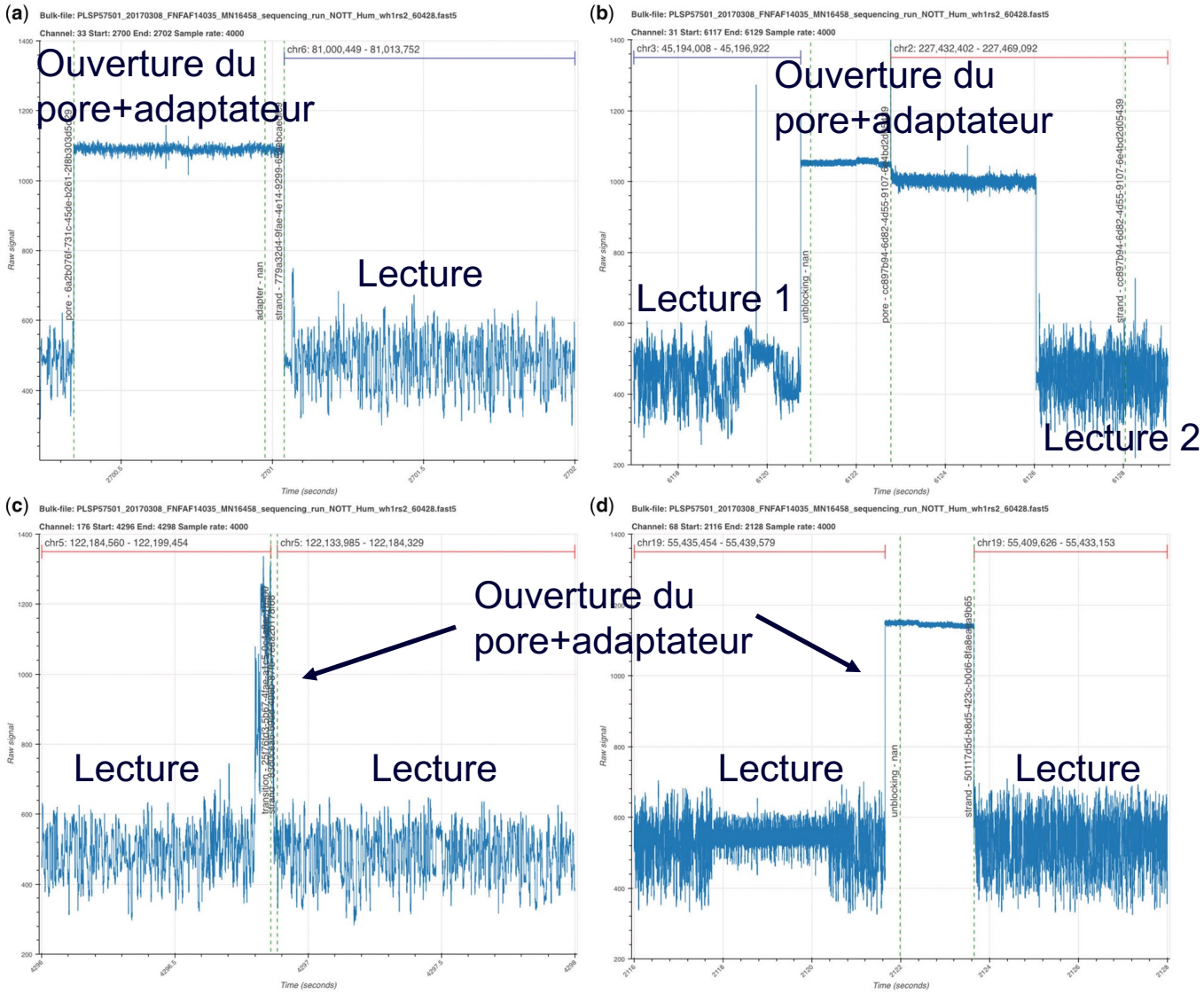
Fichier de sortie en bulk



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

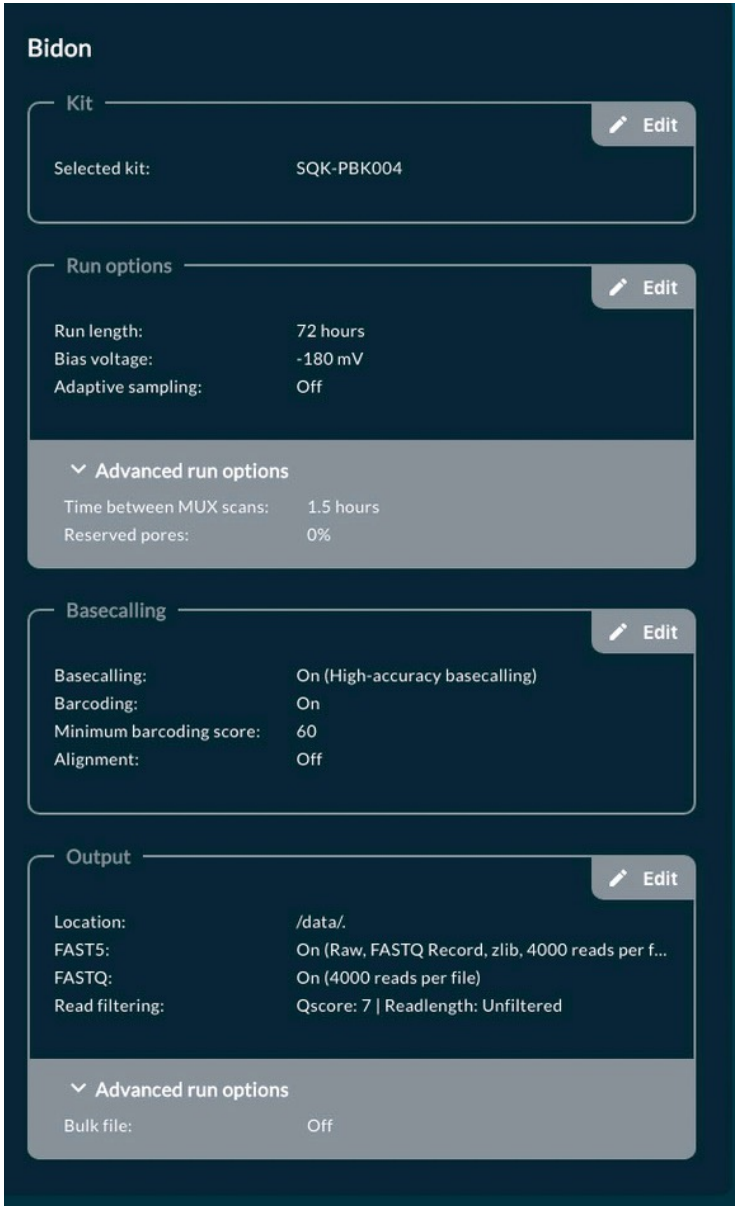
Raw data formats

- There are currently **2 raw data formats** as MinKNOW output:
  - Fast5 (default)
  - POD5 (new format)

- Fast5 is based on the **HDF5 format** (hierarchical data structure) that is very convenient to store any type of data and data structures can be easily modified. Usually there is 4,000 reads in each Fast5 files.

- POD5 is based on the **Apache Arrow format** (column data structure) that allow:
  - **Higher I/O** when reading or writing data.
  - POD5 files are also **≈10% smaller** than Fast5 files (all metadata redundancy of the Fast5 format has been removed). One file per run ?
  - POD5 will also allow **faster duplex basecalling**.

Récapitulatif avant lancement du run



Duplex basecalling

- There are **3 steps** for duplex basecalling:
  - Simplex basecalling (can be performed in fast mode)
  - Read pairing (against read length, channel and timestamp)
  - Duplex basecalling (can be performed in sup mode)

- There are **4 ways** to perform duplex basecalling:
  - Manually with Guppy by launching guppy_basecaller and then guppy_basecaller_duplex
  - Guppy-duplex-tools https://github.com/nanoporetech/duplex-tools
  - Ont-guppy-duplex-pipeline-gpu https://pypi.org/project/ont-guppy-duplex-pipeline
  - Dorado https://github.com/nanoporetech/dorado

- Great results (Q30), 2/3 duplex reads for DNA-seq but not for RNA-seq as there too much reads with almost the same size (≈ 5%).

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 ne 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

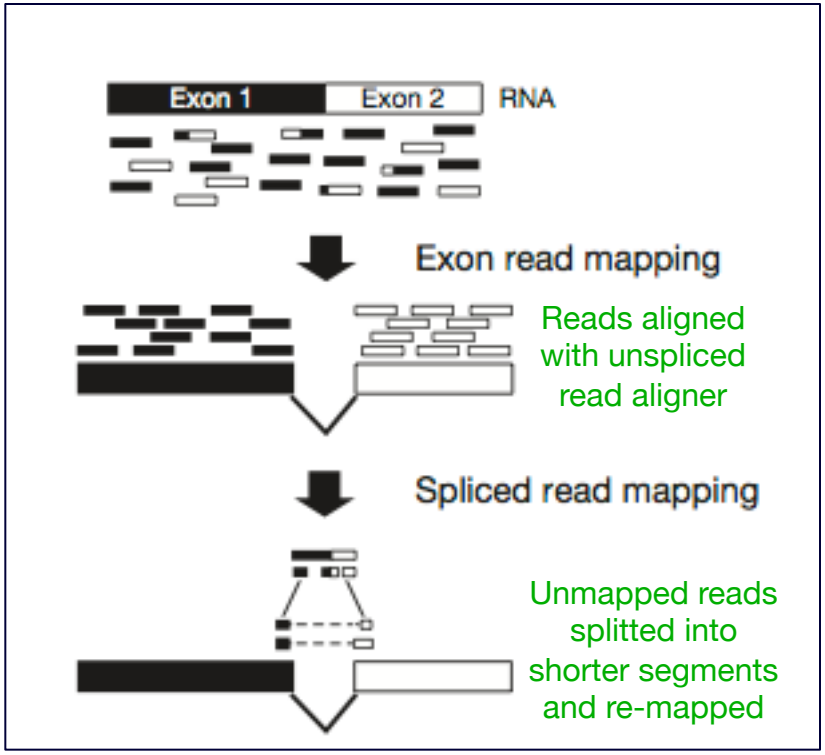
```
minimap2 -ax splice ref.fa nanopore-cdna.fa > aln.sam

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…

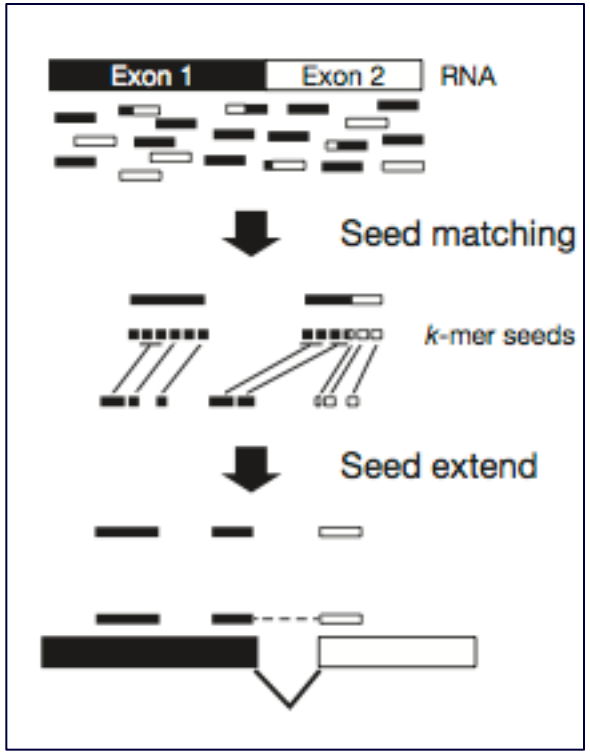
Aligneurs développés pour l'alignement d'ARNm

➢ **Les jonctions sont prises en comptes**



Stratégie exon first

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



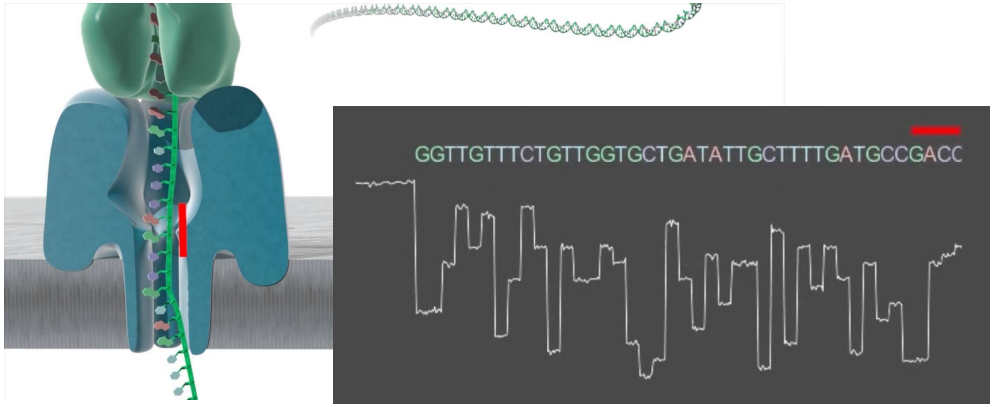
Stratégie junction first

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

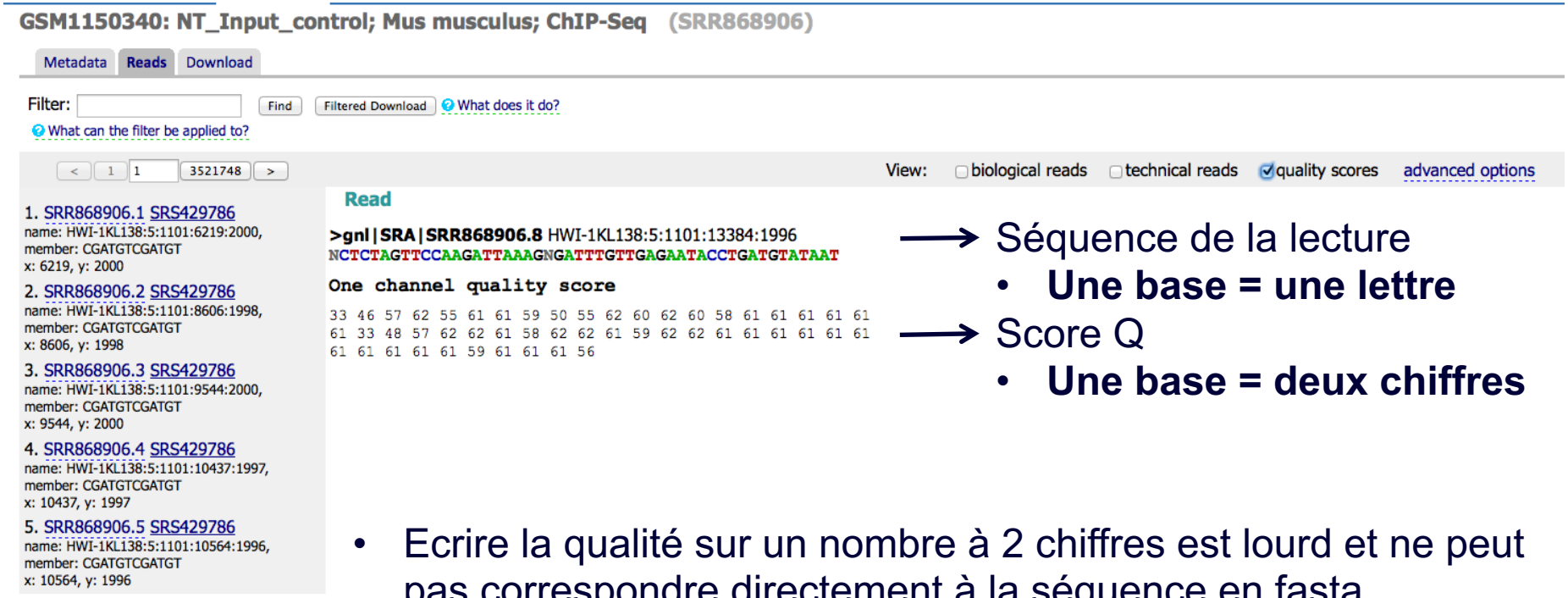
Probabilité d'identifier une base par erreur

Les score de qualité phred sont relié 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 ?



→ 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

De 0 à 32, caractères « invisibles »

Début de codage Phred = Phred+33

Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]
1	1	1	1	[START OF HEADING]
2	2	10	2	[START OF TEXT]
3	3	11	3	[END OF TEXT]
4	4	100	4	[END OF TRANSMISSION]
5	5	101	5	[ENQUIRY]
6	6	110	6	[ACKNOWLEDGE]
7	7	111	7	[BELL]
8	8	1000	10	[BACKSPACE]
9	9	1001	11	[HORIZONTAL TAB]
10	A	1010	12	[LINE FEED]
11	B	1011	13	[VERTICAL TAB]
12	C	1100	14	[FORM FEED]
13	D	1101	15	[CARRIAGE RETURN]
14	E	1110	16	[SHIFT OUT]
15	F	1111	17	[SHIFT IN]
16	10	10000	20	[DATA LINK ESCAPE]
17	11	10001	21	[DEVICE CONTROL 1]
18	12	10010	22	[DEVICE CONTROL 2]
19	13	10011	23	[DEVICE CONTROL 3]
20	14	10100	24	[DEVICE CONTROL 4]
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]
22	16	10110	26	[SYNCHRONOUS IDLE]
23	17	10111	27	[ENG OF TRANS. BLOCK]
24	18	11000	30	[CANCEL]
25	19	11001	31	[END OF MEDIUM]
26	1A	11010	32	[SUBSTITUTE]
27	1B	11011	33	[ESCAPE]
28	1C	11100	34	[FILE SEPARATOR]
29	1D	11101	35	[GROUP SEPARATOR]
30	1E	11110	36	[RECORD SEPARATOR]
31	1F	11111	37	[UNIT SEPARATOR]
32	20	100000	40	[SPACE]
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	0
49	31	110001	61	1
50	32	110010	62	2
51	33	110011	63	3
52	34	110100	64	4
53	35	110101	65	5
54	36	110110	66	6
55	37	110111	67	7
56	38	111000	70	8
57	39	111001	71	9
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	A
66	42	1000010	102	B
67	43	1000011	103	C
68	44	1000100	104	D
69	45	1000101	105	E
70	46	1000110	106	F
71	47	1000111	107	G
72	48	1001000	110	H
73	49	1001001	111	I
74	4A	1001010	112	J
75	4B	1001011	113	K
76	4C	1001100	114	L
77	4D	1001101	115	M
78	4E	1001110	116	N
79	4F	1001111	117	O
80	50	1010000	120	P
81	51	1010001	121	Q
82	52	1010010	122	R
83	53	1010011	123	S
84	54	1010100	124	T
85	55	1010101	125	U
86	56	1010110	126	V
87	57	1010111	127	W
88	58	1011000	130	X
89	59	1011001	131	Y
90	5A	1011010	132	Z
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	a
98	62	1100010	142	b
99	63	1100011	143	c
100	64	1100100	144	d
101	65	1100101	145	e
102	66	1100110	146	f
103	67	1100111	147	g
104	68	1101000	150	h
105	69	1101001	151	i
106	6A	1101010	152	j
107	6B	1101011	153	k
108	6C	1101100	154	l
109	6D	1101101	155	m
110	6E	1101110	156	n
111	6F	1101111	157	o
112	70	1110000	160	p
113	71	1110001	161	q
114	72	1110010	162	r
115	73	1110011	163	s
116	74	1110100	164	t
117	75	1110101	165	u
118	76	1110110	166	v
119	77	1110111	167	w
120	78	1111000	170	x
121	79	1111001	171	y
122	7A	1111010	172	z
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	[DEL]

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=91ceec1ff67a0b21bf1ab0fcef80e77cbd2a6da8
sampleid=EMTome_A2020 read=27472 ch=469 start_time=2020-11-24T08:41:45Z barcode=barcode04
GGTATGCTTCGTTCGGTTTTTCAAGGTGGGTGTTTCCTTGATCCATCATCGTGACTTCCAGTTCTATCGTGTTTCCCTATTTCTGTTGGTGCTGAT
ATTGCGGGTCTGCTTGGGTGTTTAACCTAAGCGATGGTATCAACATGAGTACGGGGCCTAGCCTTCCGCCTTCCCAAGCCAACGTCTCCGCCGTCG
GCTCCGCGGCGCCGCCGCCAACCGACGTGGAAGACGGAGAAGGAAACCTGCGCCCTGGCCTCTCCTCAAGGGCTCAGGCTCCAAGTCGGGAGGCGA
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/(--/,/=3/%(0++(+0062…
```

Le format FASTQ (exemple d'ARN direct)

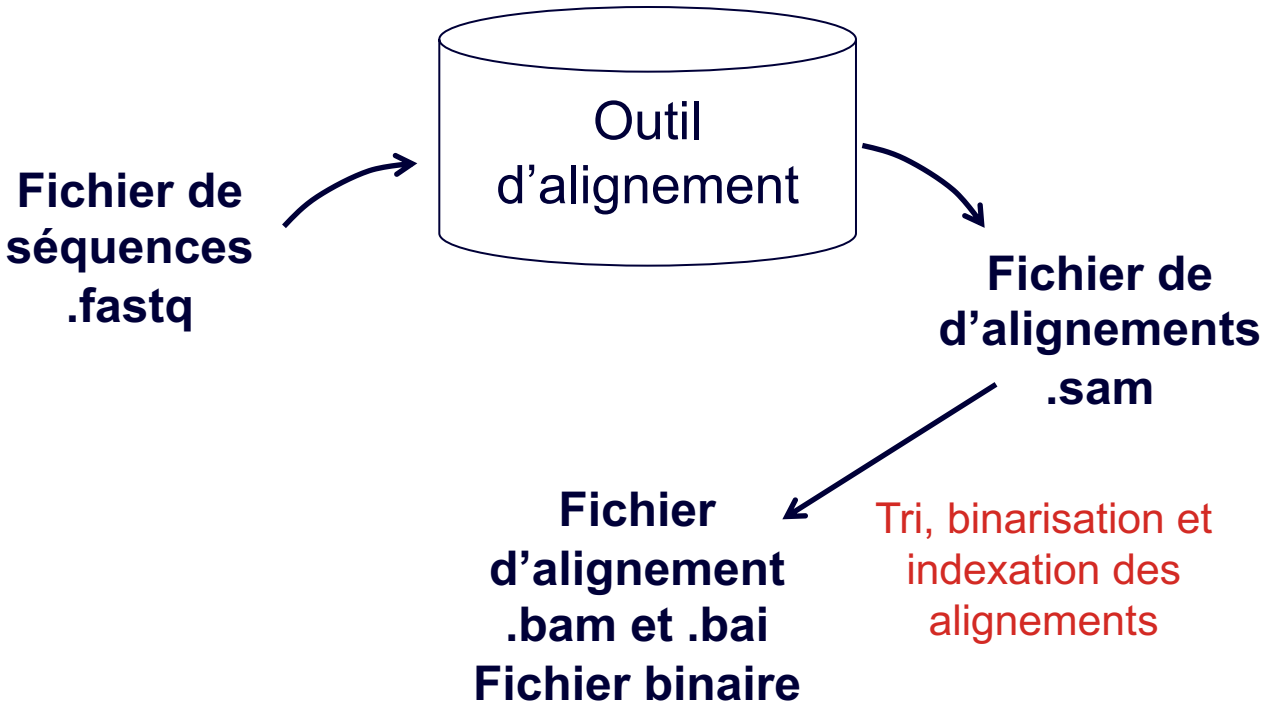
La séquence est codée sur 4 lignes

@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
CGCUGAAAAGAGCUCGGCUUGCUUUCCCUUAGCUACCUCCGGUGCCUUUUUGCAAAGCCCUCUGCAACAAAACCCAUUCAGCCCGCAGGCGCUGG
UUAAGGAGUGGCUGGACAGGCCUGAGUGGCCAUCGGCCUUCCUGUUCAUACAGUUUCAUUCCCAGGGGCCGGGGCUCCUUCCCCUUUGGCAUUUA
UACUUGGUGGGGGCAUUGAAGGCCUGCGGUGGAAGGCUUGGGACUGGUCCUGUAACUGUGUGUCCCCUGAAGGCUGAGUUGGCCAUUUGGGUUAG
UCCCAGGUGACAGAAGGGAGAAACCAAAAUGAAGCAGAAGAAAAUGCAAUUGCCAAGAUCAGAUUUCGAUCAGUGAAACUUGACAGGAAGAACUU
UAAAAAAUGCUACAGUCACCUGGUGGCUUUGAUGUAGGGGCCCUACCCUGGUGUGUGGGUCAGGCCCACUGCUCAGCCACUUGGGUCGUGGUUGU
AGUACCAGUUCUGGGAGGAAGACAUGUUGGAGACUGCCAGAAAAUCCCUGUACAUCGUUUUACGUGCUUGUGUCCAAGUGAAUUGUAAUGGUUUC
GGUGUUUGCUAAACAAAGUGACUUAACAGCUAAAAAAAAAAAUUGUGC
+
'%%&()/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'alignement

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_mini
map2index_genomefile/genome.idx /import/pontos02/analyses/EMTome_A2020/eoulsan-20201217-
190452/working/filterreads_output_reads_2020385_file0.fq      PN:minimap2
```

Résultats de l'alignement

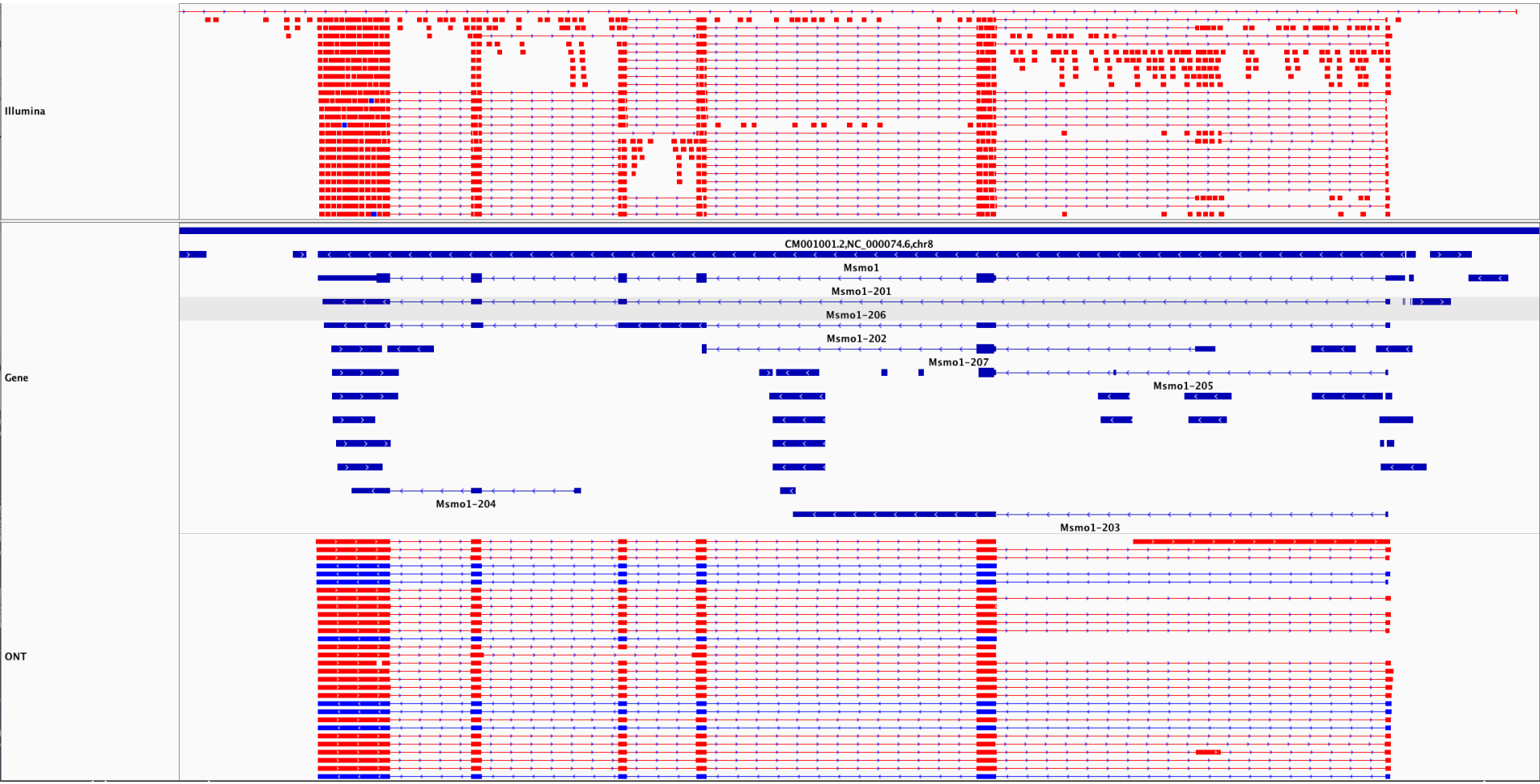
```
6b109bc4-8bbd-408e-9089-
7e4b530ce5e6    0       3       141738285       60      152S16M1D33M3I8M2D19M1I24M1D7M1D45M2I1M1I66M1D7M4992
N48M1602N37M1D79M2D8M3I25M2I35M4I36M1I4M1D49M1I5M2I4M1D67M3I1M1D2M1I5M1I13M2D17M1I28M1D9M1D64M2D5M1D26M1I12M
117S    *       0       0       GGTATGCTTCGTTCGGTT…GAAGTCA    ,,,,%%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 alignement du format sam

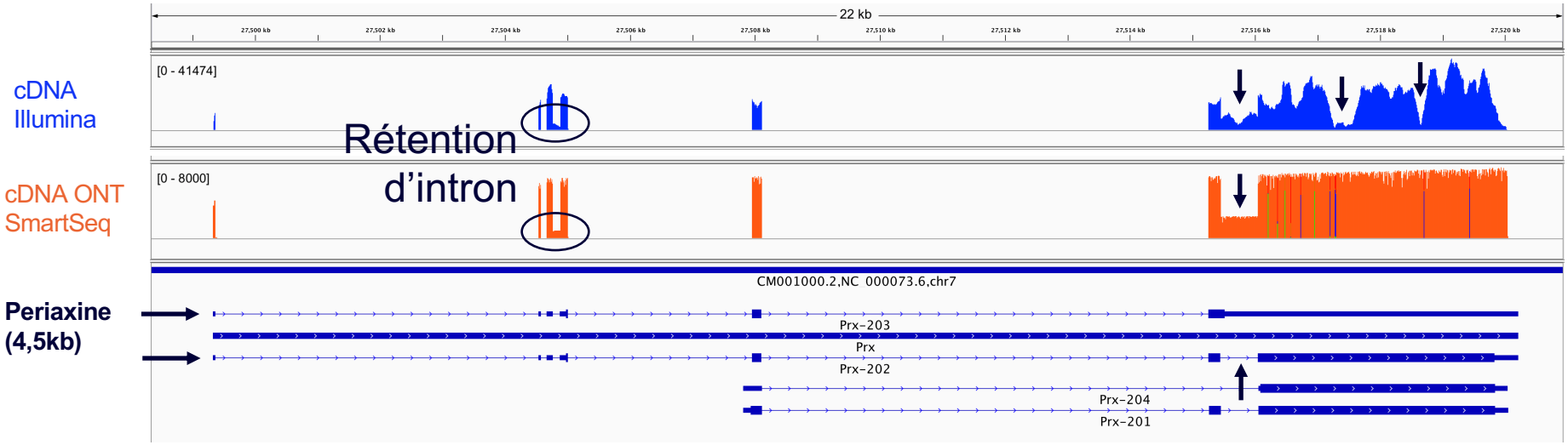
Chaque alignement est décrit par une ligne tabulée
Certains champs sont obligatoires :

QNAME	Id Lecture	6b109bc4-8bbd-408e-9089-7e4b530ce5e6
FLAG	016 sens / 16 reverse	0
RNAME	Ref Sequence	3
POS	Start alignement	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		GGTATGCTTCGTTCGGTT…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 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