Formation à l'utilisation d'un séquenceur Oxford Nanopore

Session informatique

Paris •Mai 2024

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

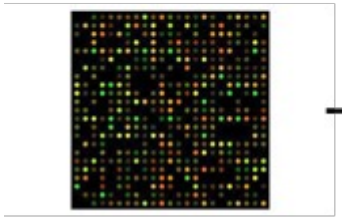
Stéphane Le Crom

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)

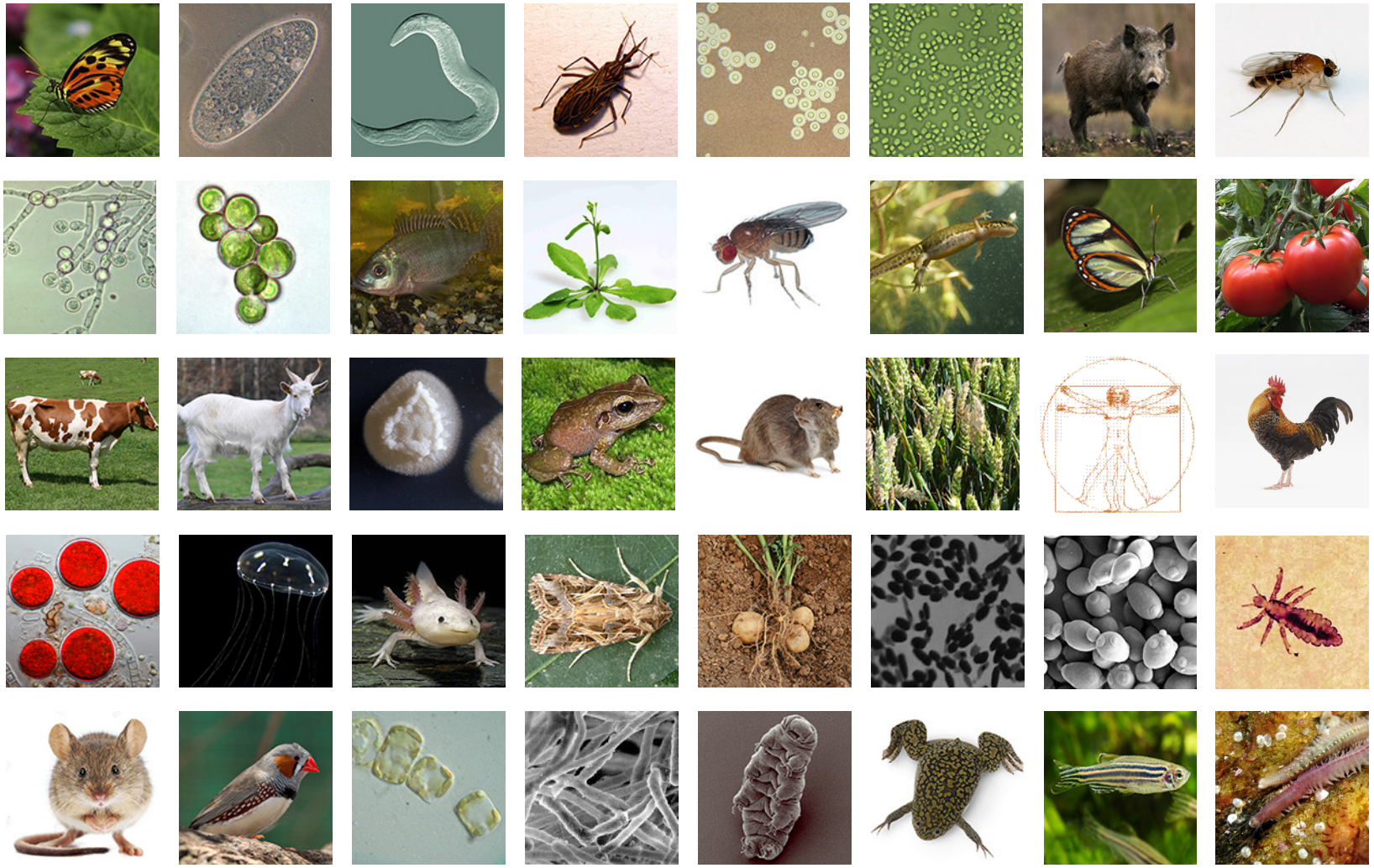
Illumina NextSeq 2000
2nd generation sequencer
(Since 2021)

Nanopore MinION
3rd generation sequencer
(Since 2016)

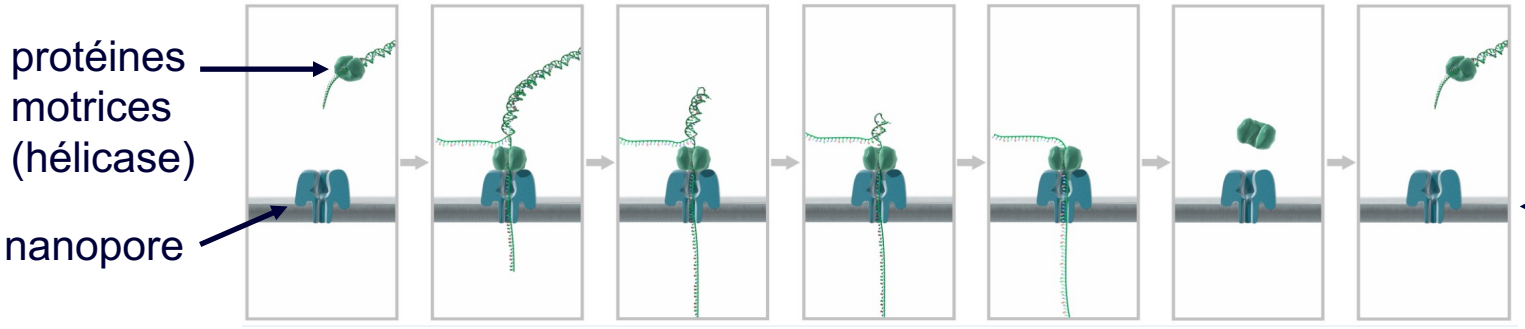
Nanopore PromethION
3rd generation sequencer
(Since 2022)

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

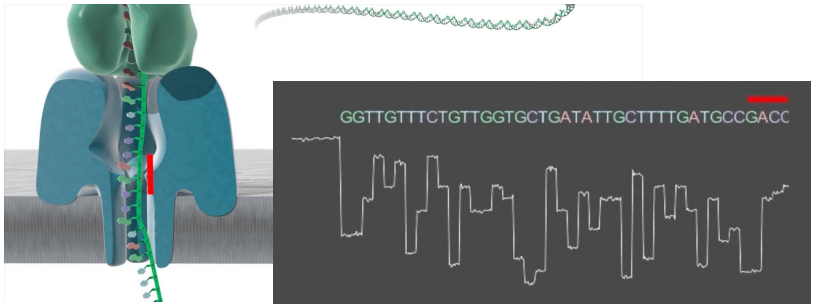
Spécialisée sur les organismes eucaryotes



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

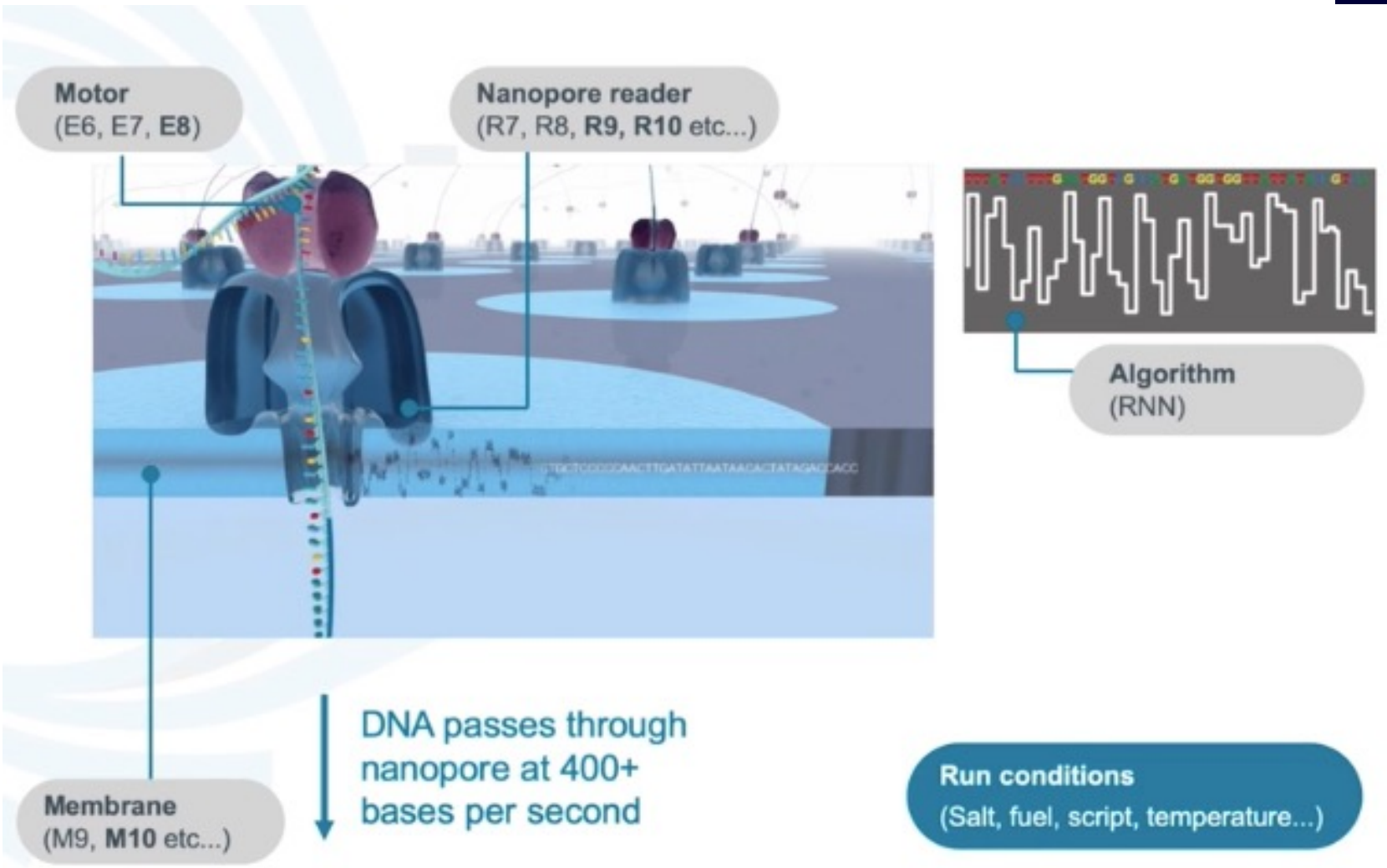
- Assemblage de génome simplifié
- ARN séquencés en pleine longueur

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…

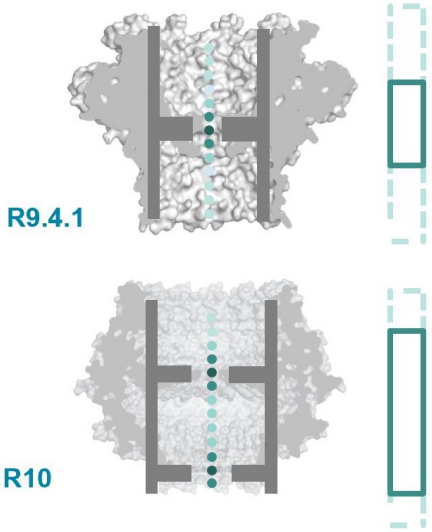
https://nanoporetech.com/

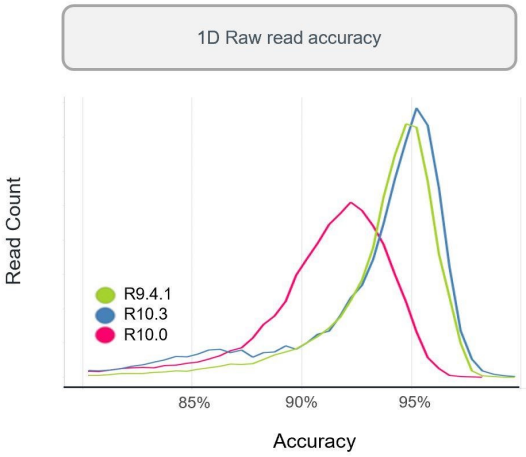
Les différents éléments d'un run

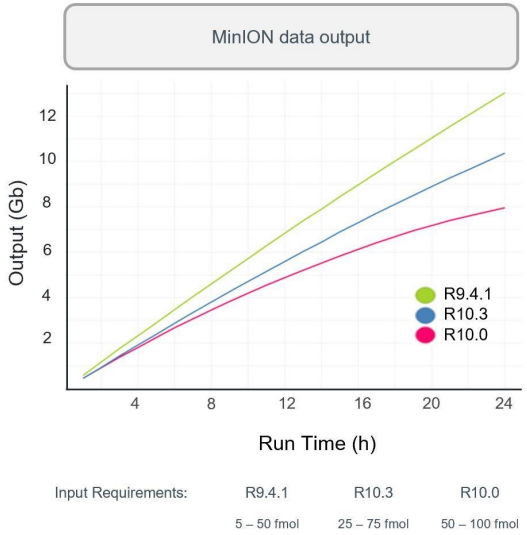


Pores types

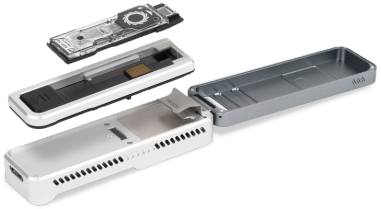
- There are **two types** of flowcell **pores** :
  - **R9.4.1** with one reader head (01/2017, 13 Gb/24h, 95 % accuracy, 5-50 fmol)
  - **R10.4.1** with two reader zones (03/2022, ?? Gb/24h, 99 % accuracy, ?? fmol)







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

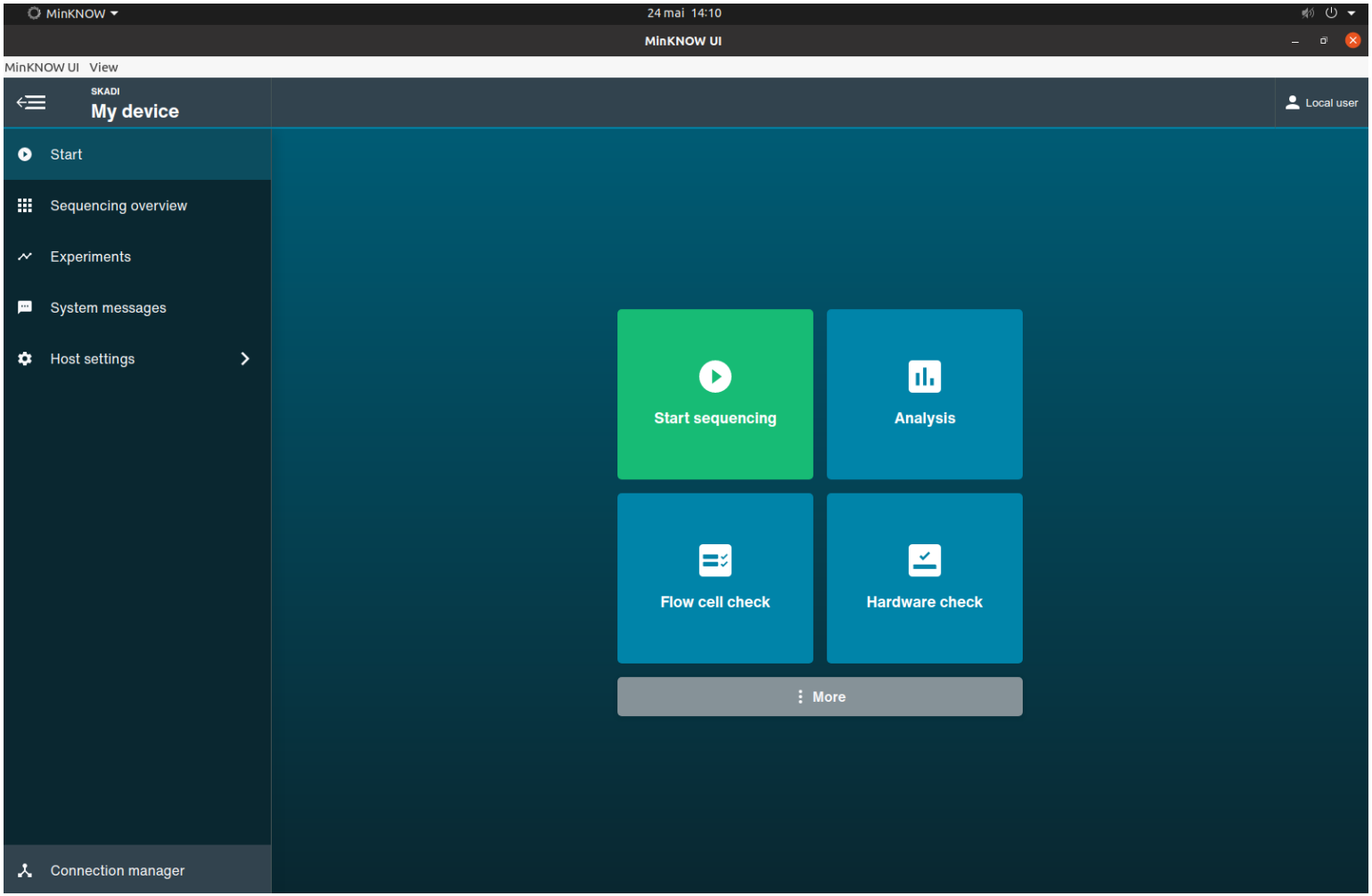
- Oxford Nanopore device are powered by a **standard Ubuntu Linux** system with additional packages :
  - **MinKNOW** suite
  - **Nvidia tools** for GPU computing
  - **Docker** for running containerized applications

- ONT uses **Ubuntu 20.04** for GridION and PromethION.

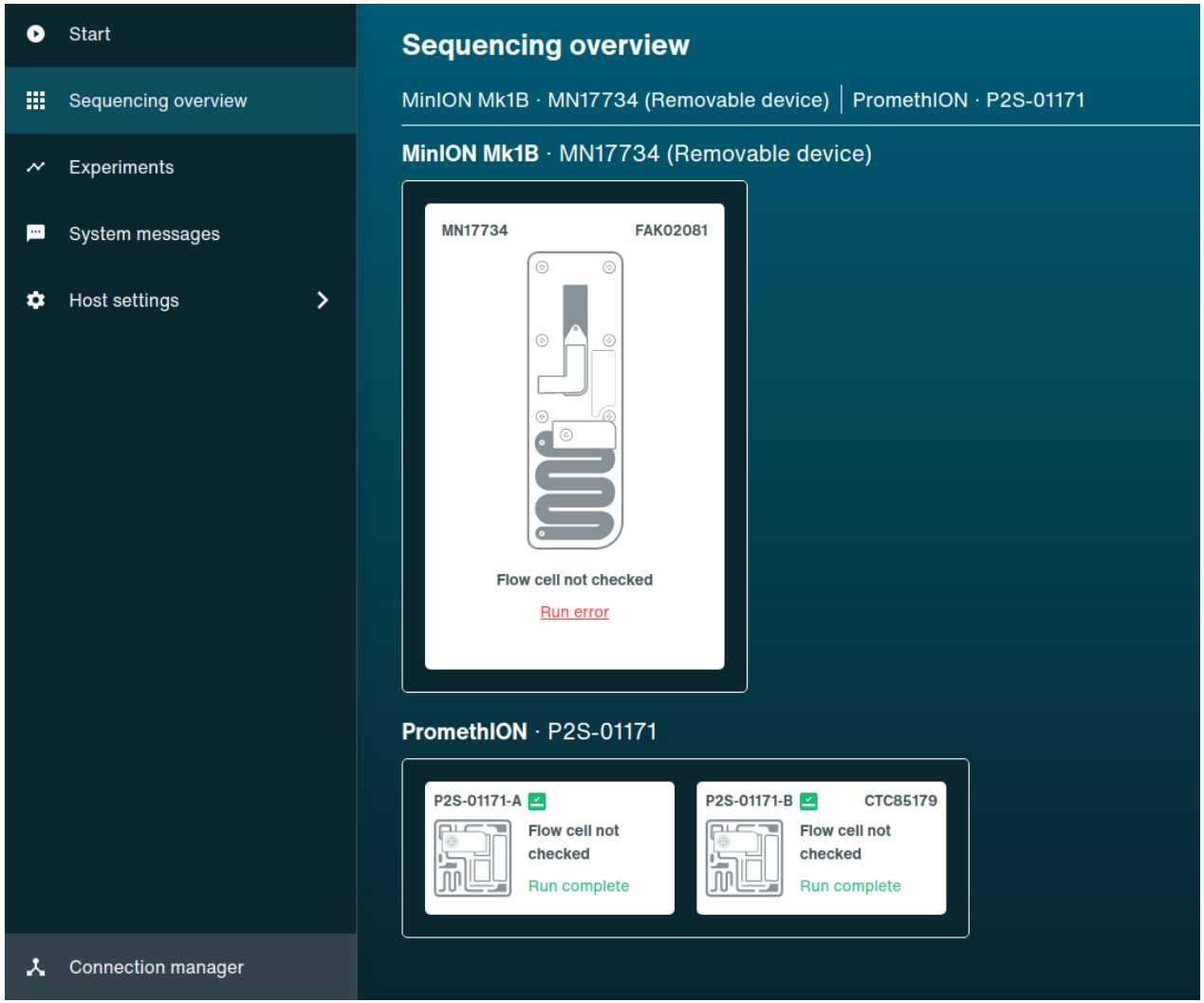
- ONT gives users full administrative rights, so you can modify the operating system as you want. However its at your own risks!

- All low level administration tasks are performed in **command line** through a **SSH** connection.

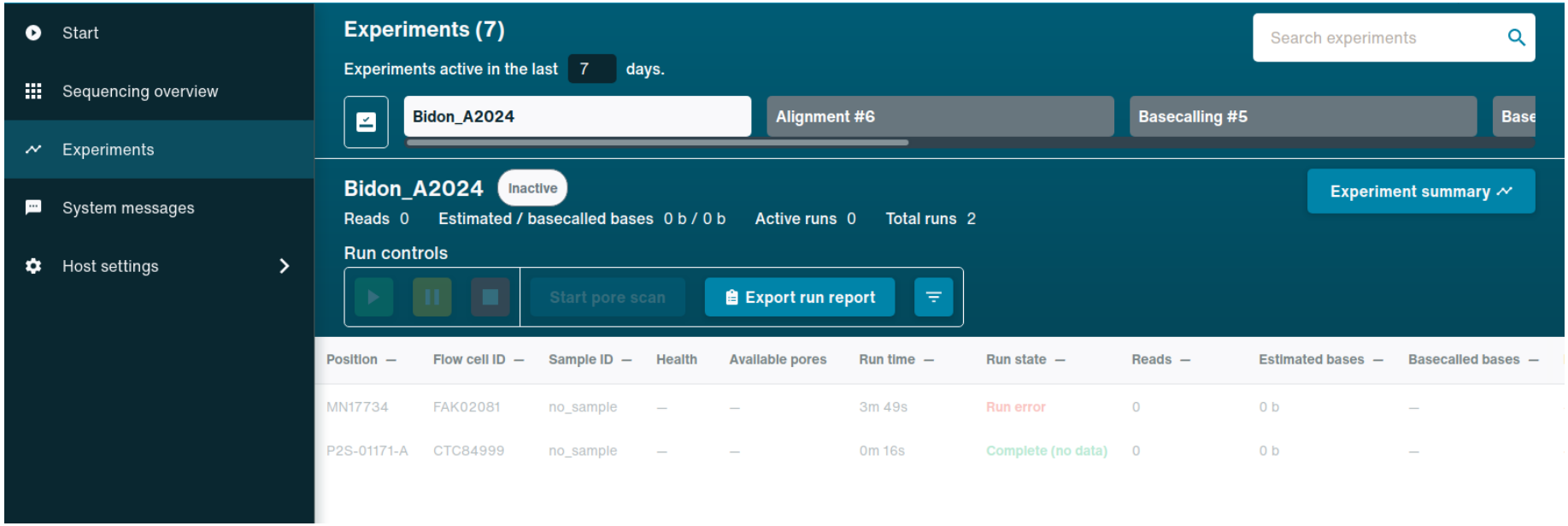
Utilisation de MinKNOW



Section Sequencing overview

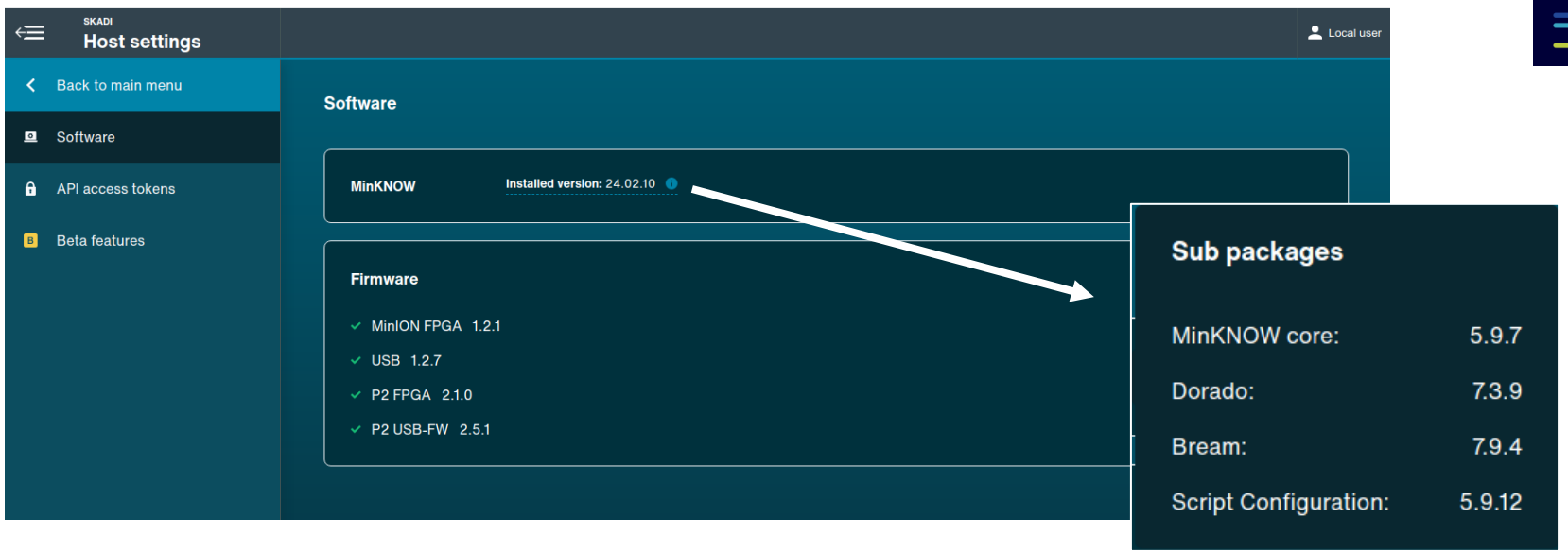


Section Experiment



- **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**

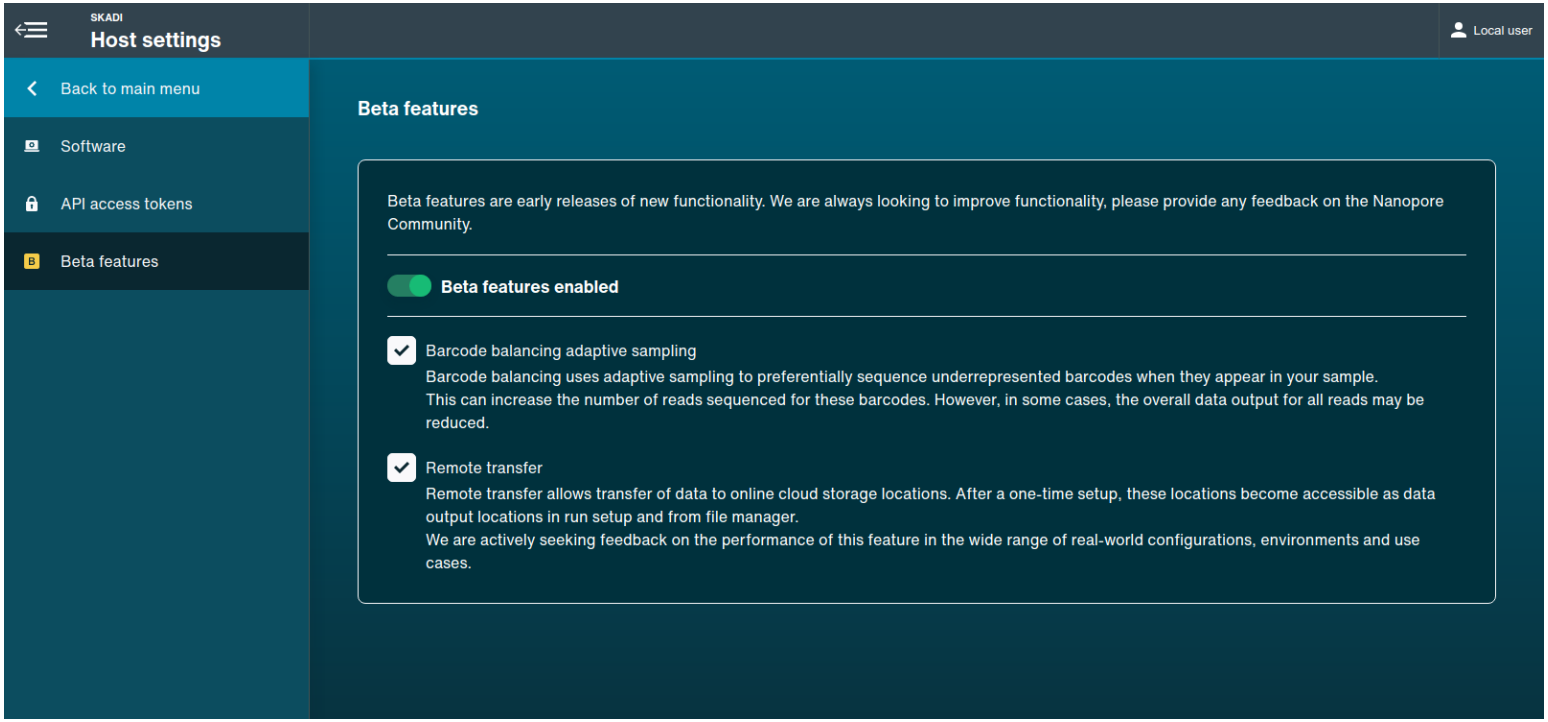
Section Software : Host settings



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

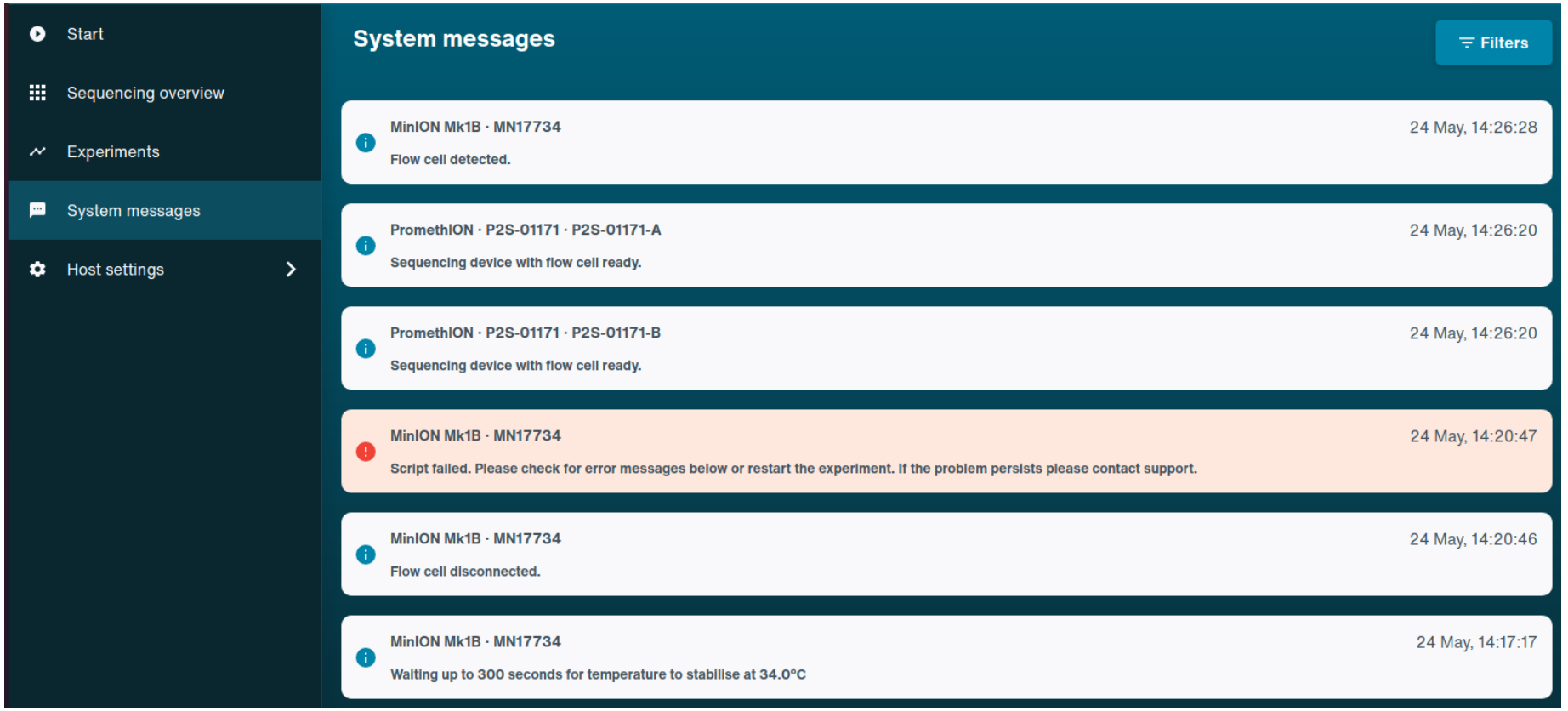
- La gestion du séquençage : MinKNOW
  - Dorado : Basecalling
  - Bream : FC check
- La communication séquenceur-FC-PC

Section Software : Beta features

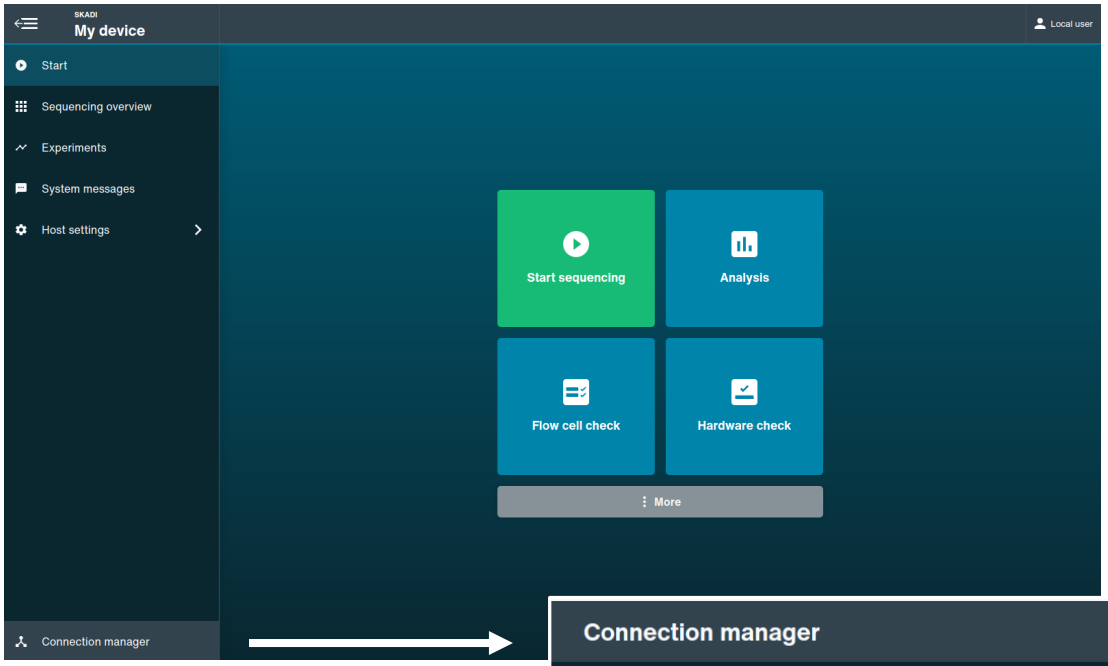


- Autorise MinKNOW a proposer ses nouveautés en cours de validation

Section Système message

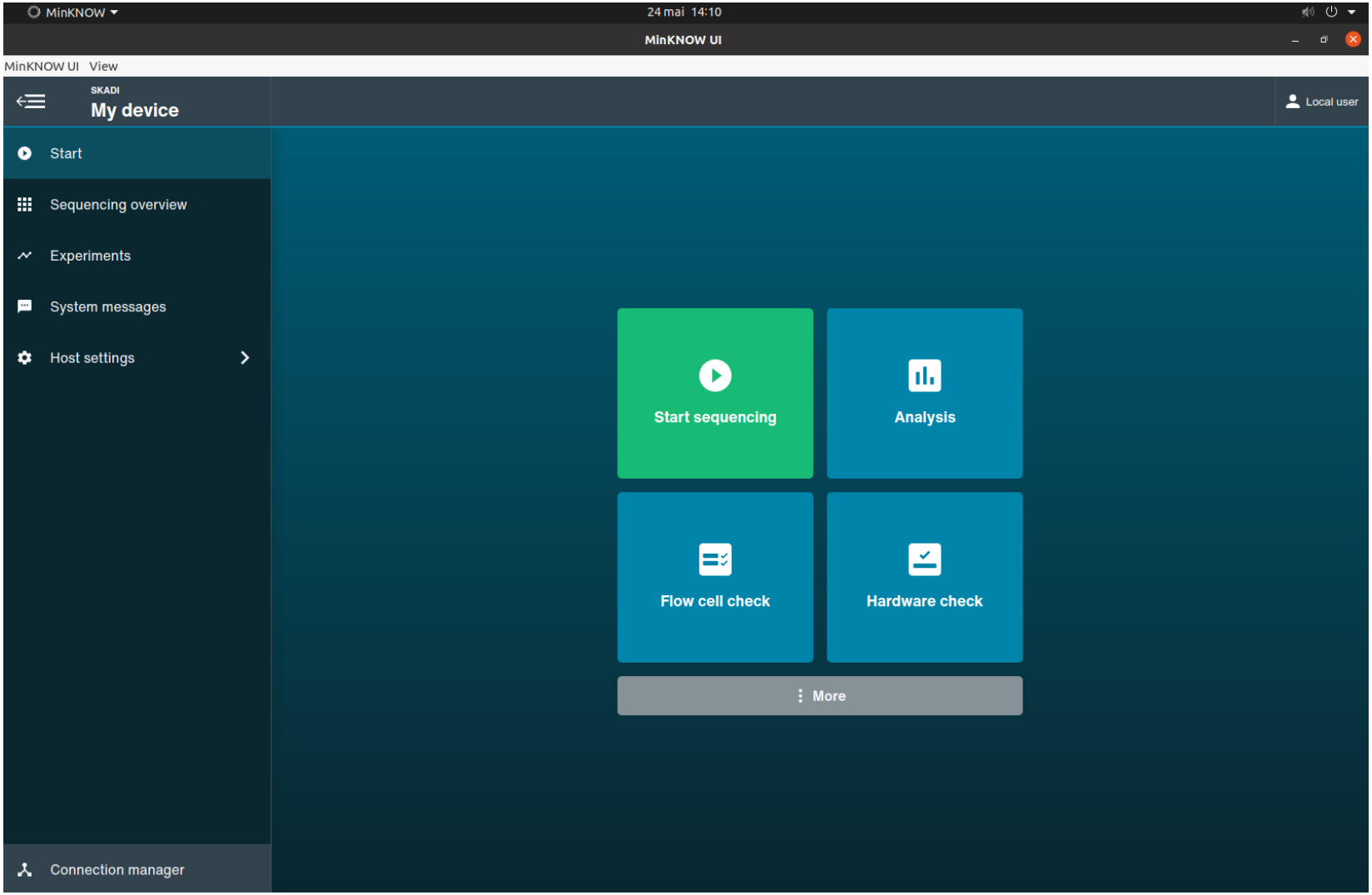


Section Connection manager

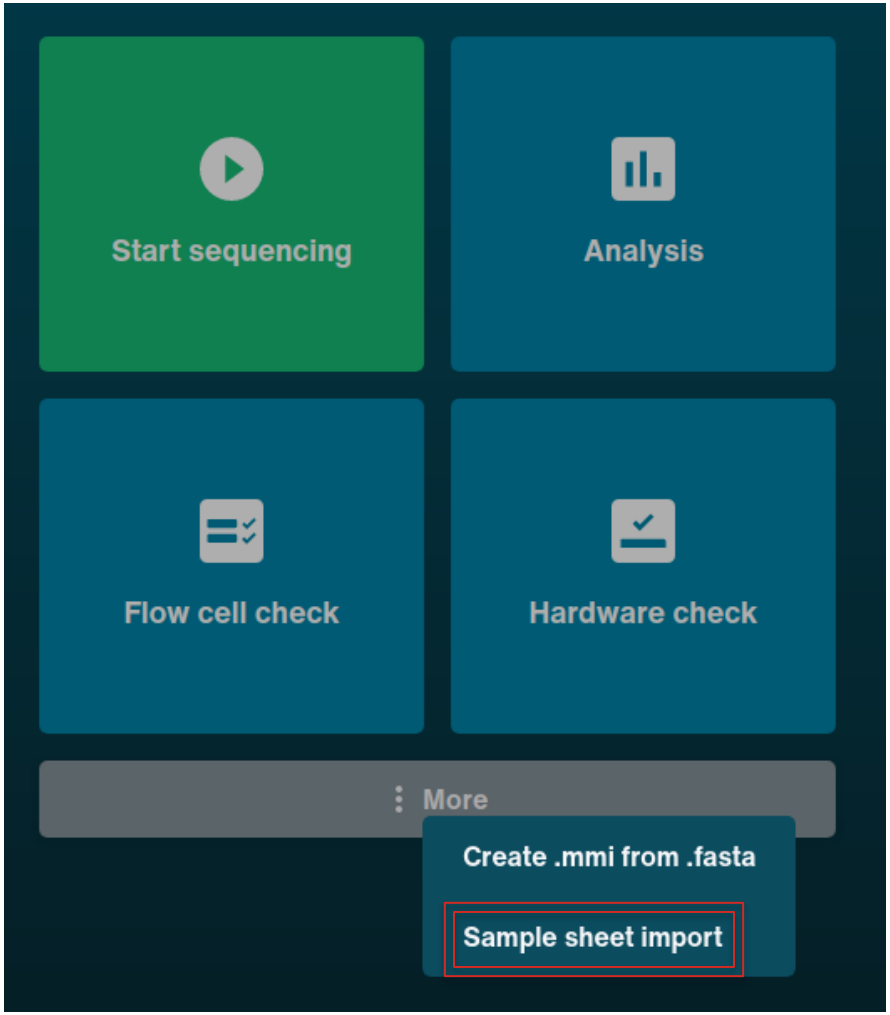


Une liste des séquenceurs disponibles sur le PC et leur état

Section Start

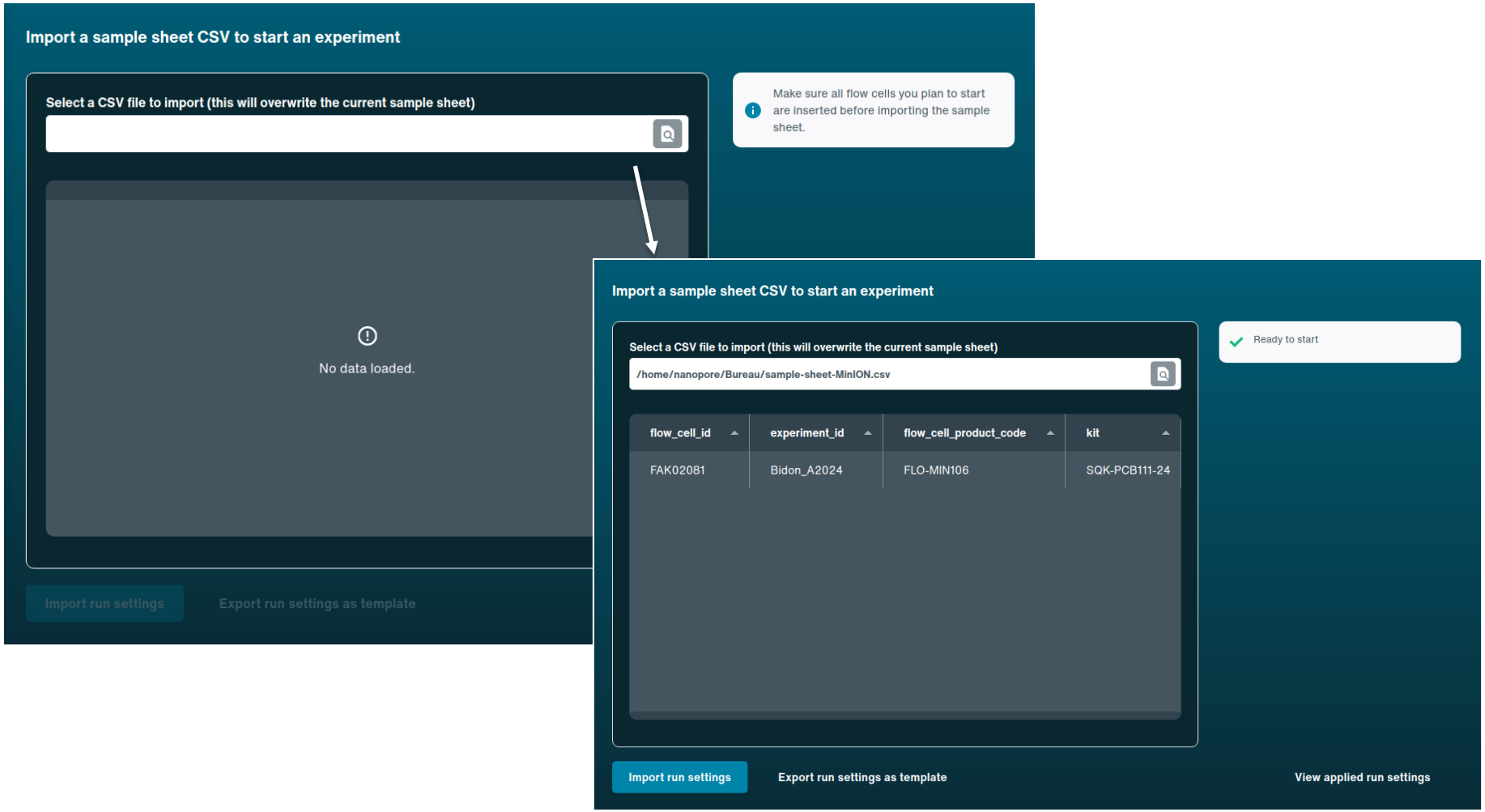


Demarrage du run avec une samplesheet et un fichier settings (format json)



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

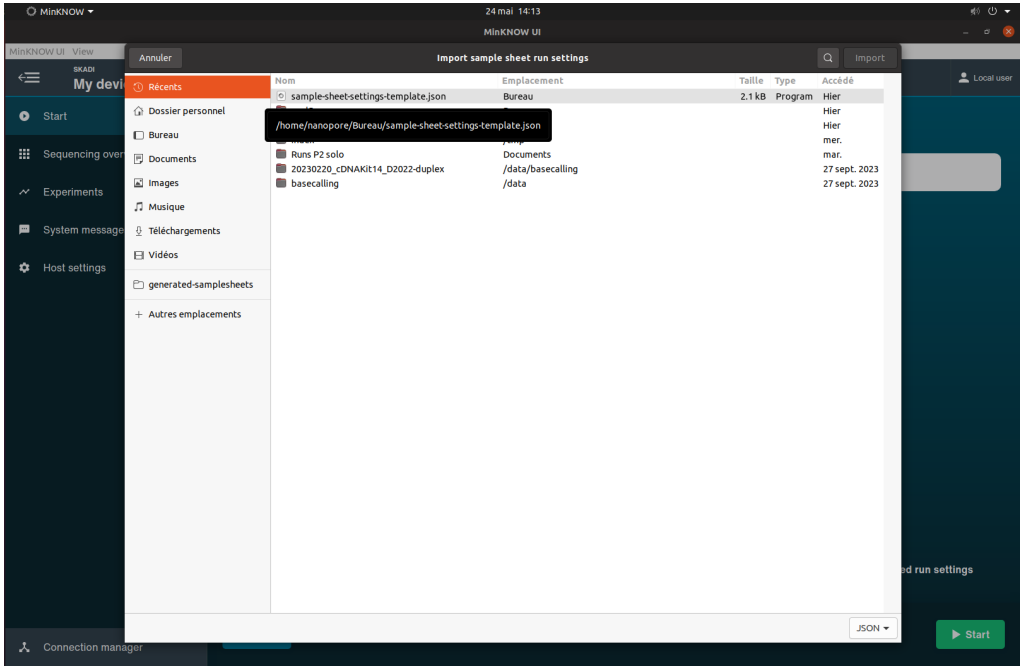


Définition des feuilles de route : Samplesheet

	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

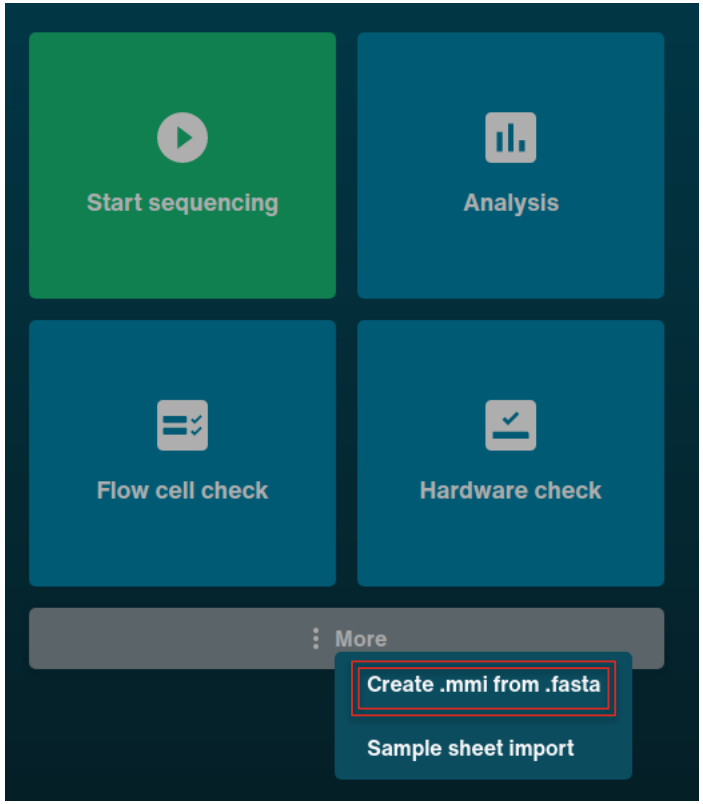
- Permet de refaire à l'identique un run déjà défini sans avoir à redéfinir manuellement toutes les étapes de MinKNOW
  - Fichier json : texte avec une architecture particulière



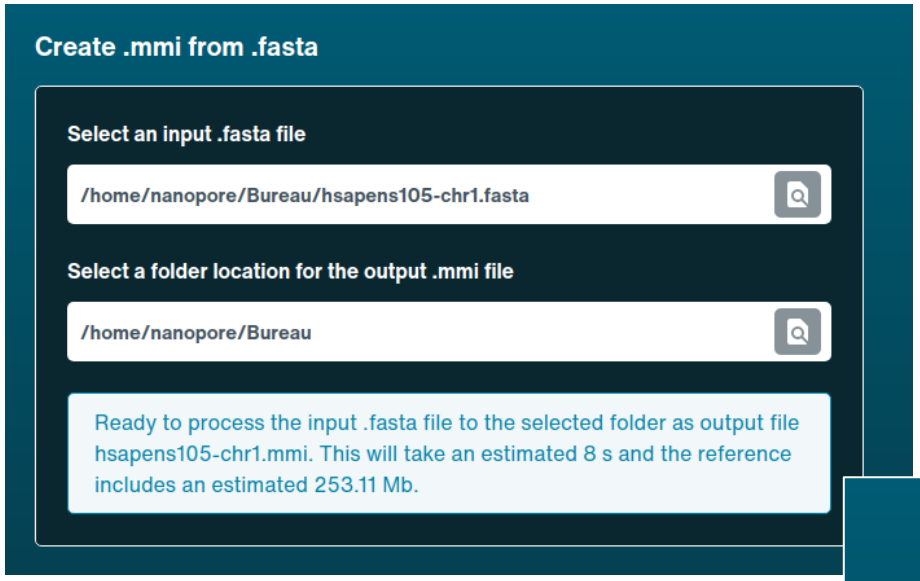
=> Voir le fichier samplesheet et le fichier json dans les documents de la formation

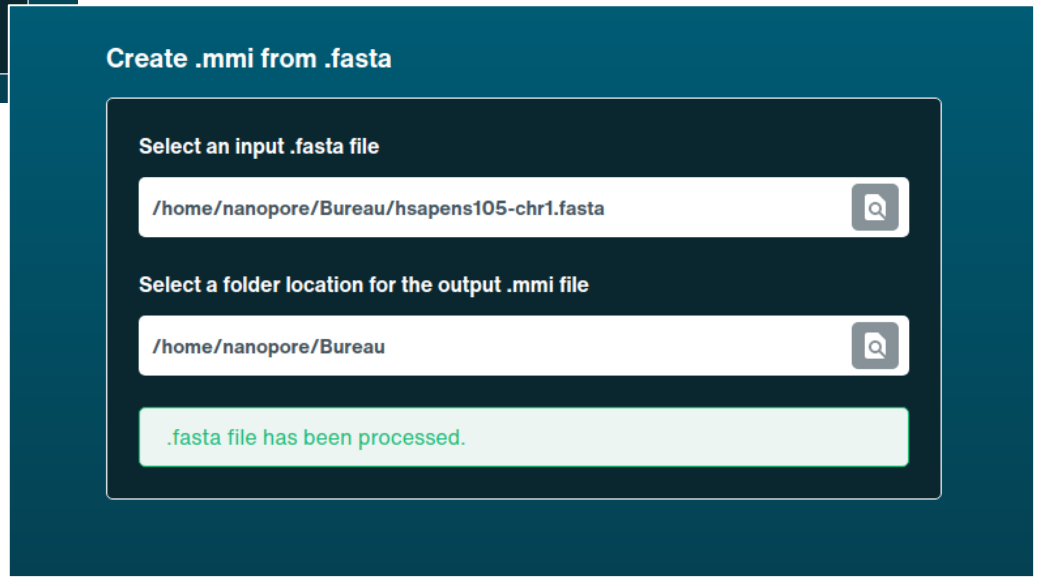
Création de l'index du génome

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



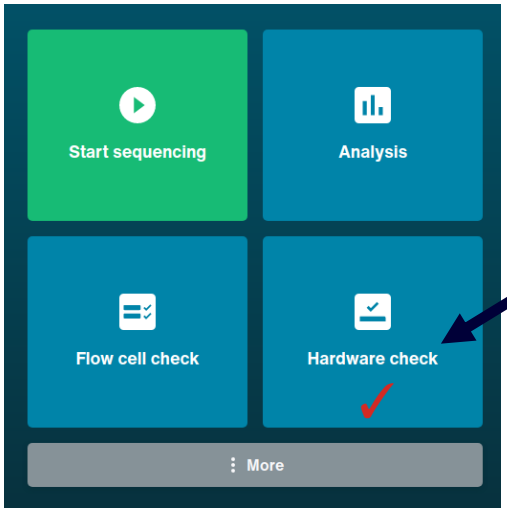
Création de l'index du génome

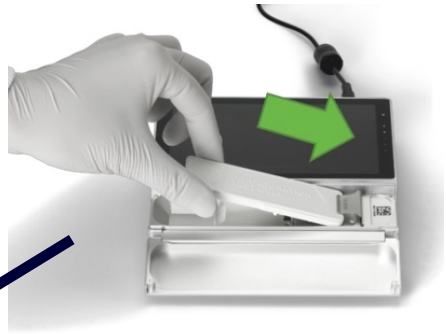


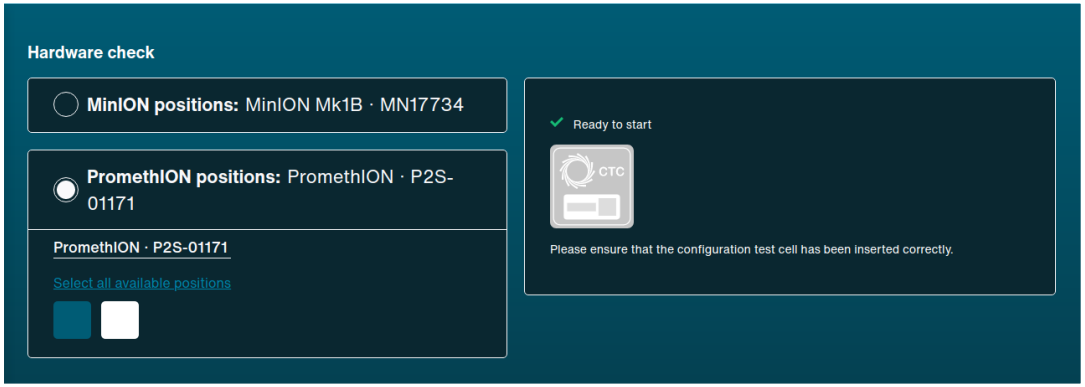


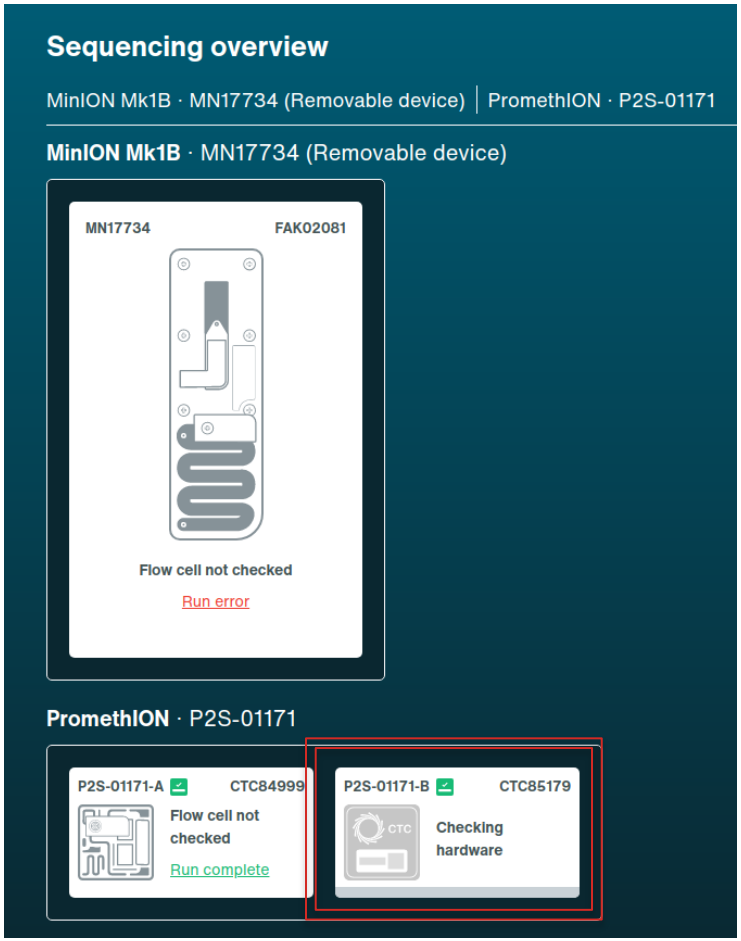
Vérification initiale du séquenceur

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



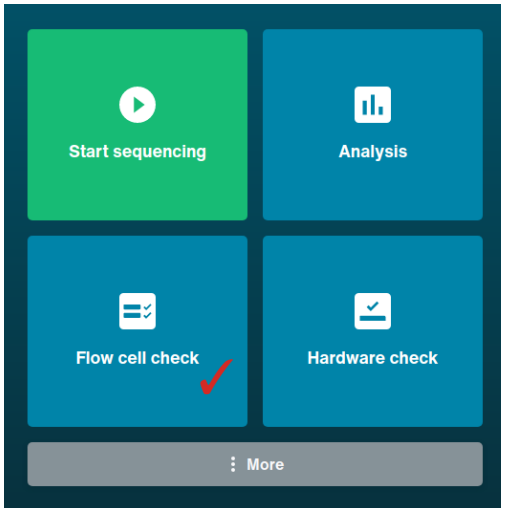






Vérification initiale de la flowcell avant de lancer le run

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 :

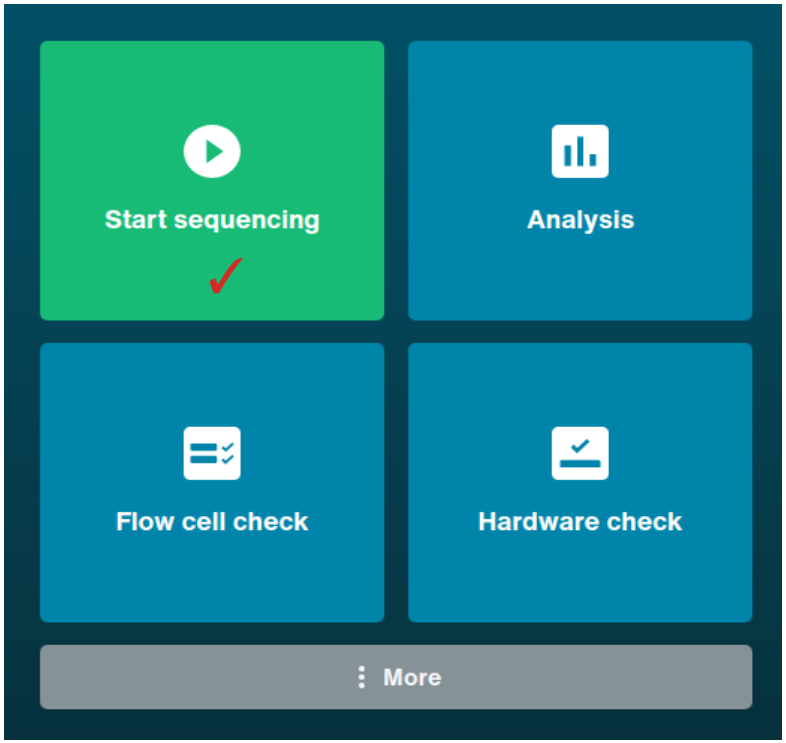
- 50 dans le cas d'une flowcell Flongle
- 800 dans le cas d'une flowcell MinION
- 5000 dans le cas d'une flowcell PromethION

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

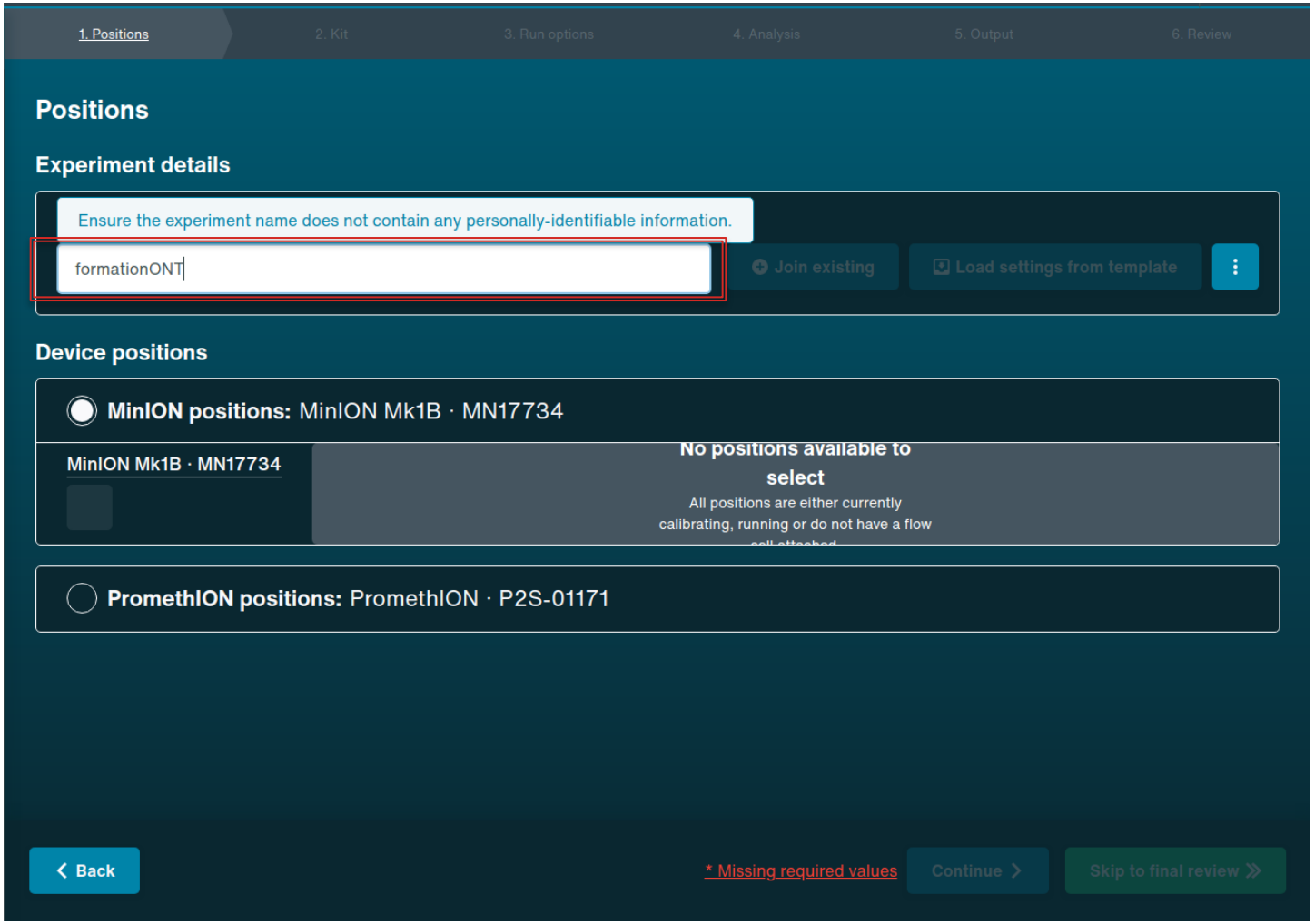
Paramétrage et lancement d'un run

Il faut définir

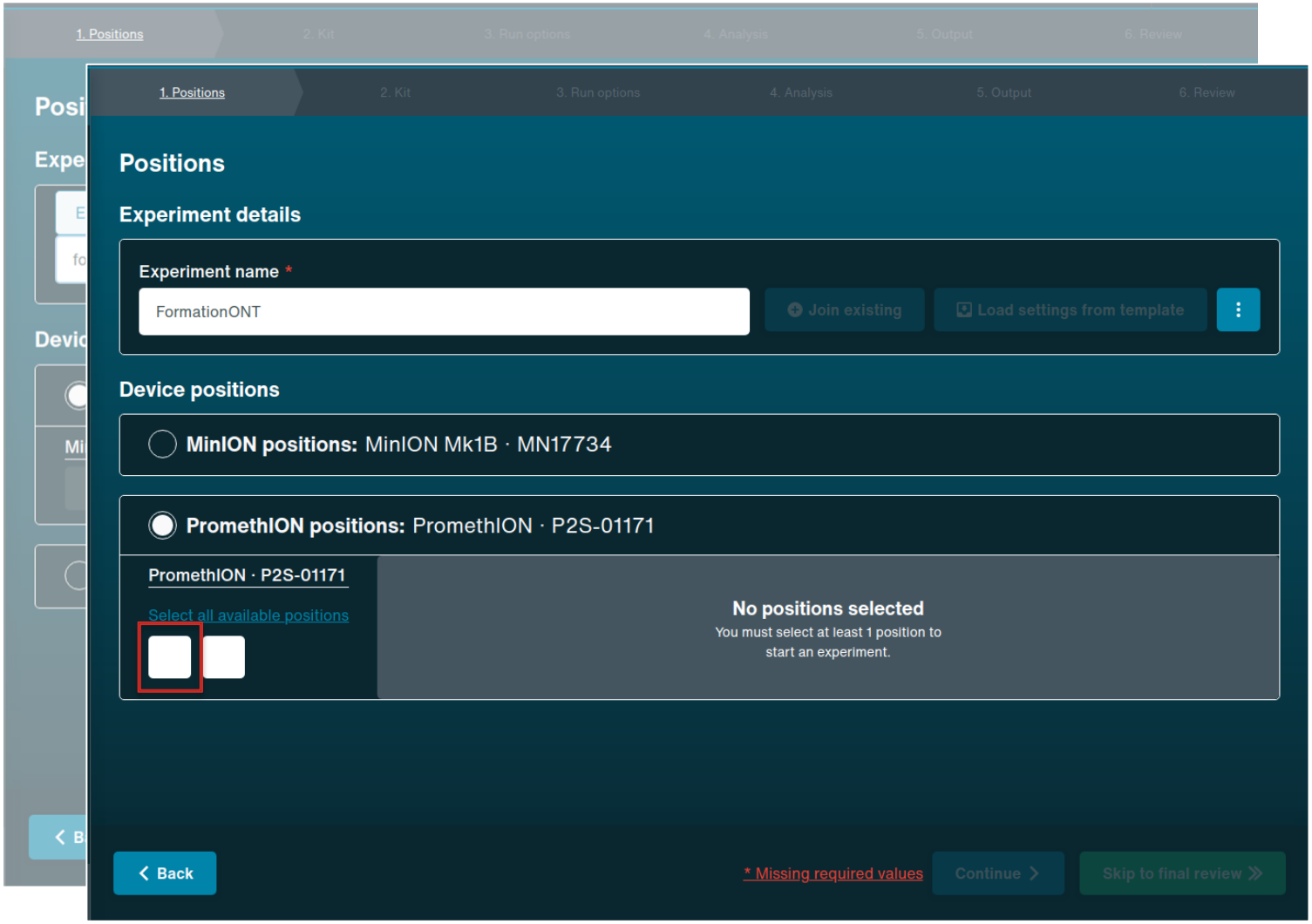
- Votre expérience
- Le kit utilisé
- Le basecalling (à la volée ou après de run dans la section Analysis)
- Les format de sortie de vos données
- L'alignement (à la volée ou après) et par conséquent, les séquences références



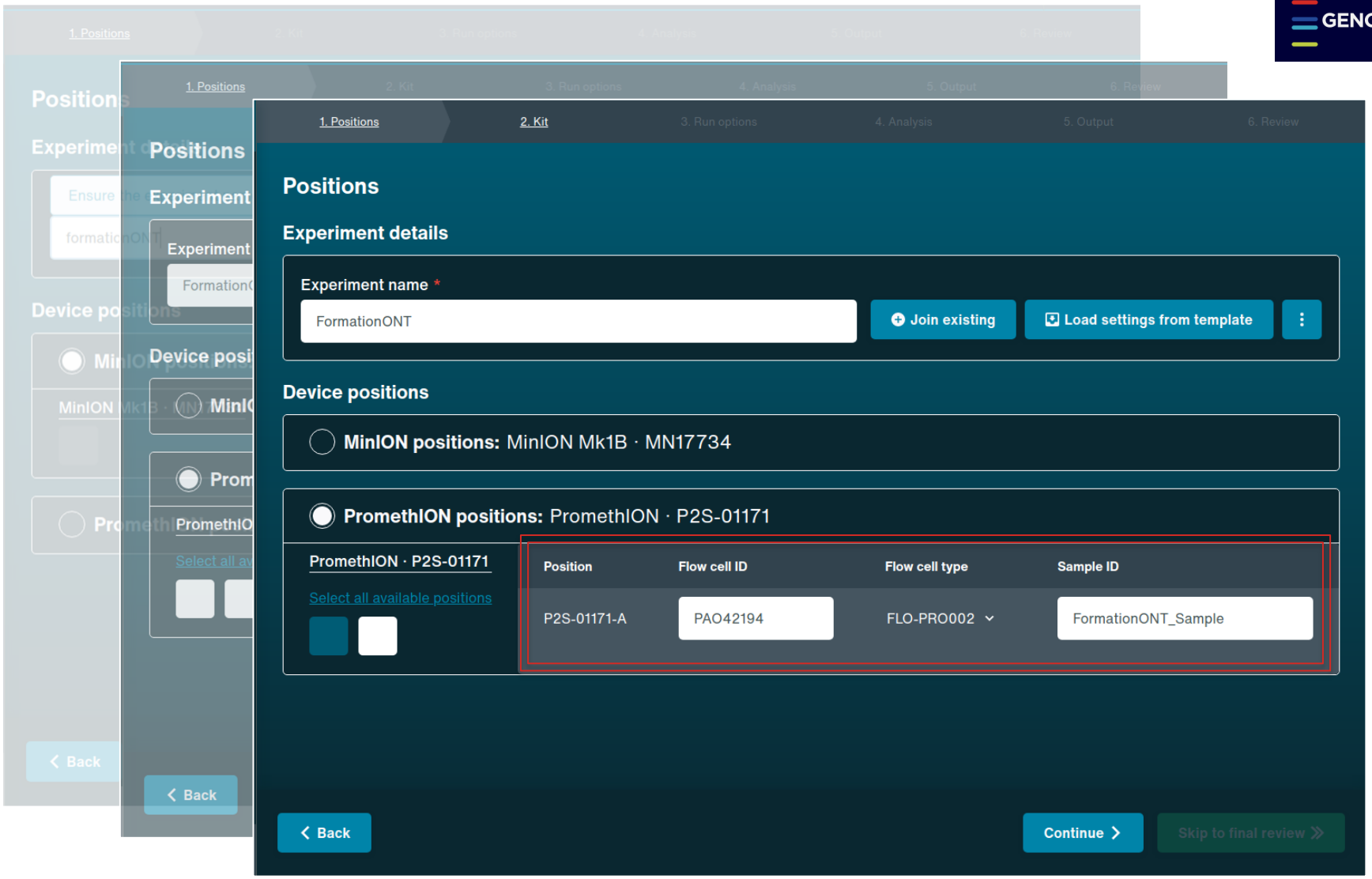
Définition de l'expérience



Définition de l'expérience

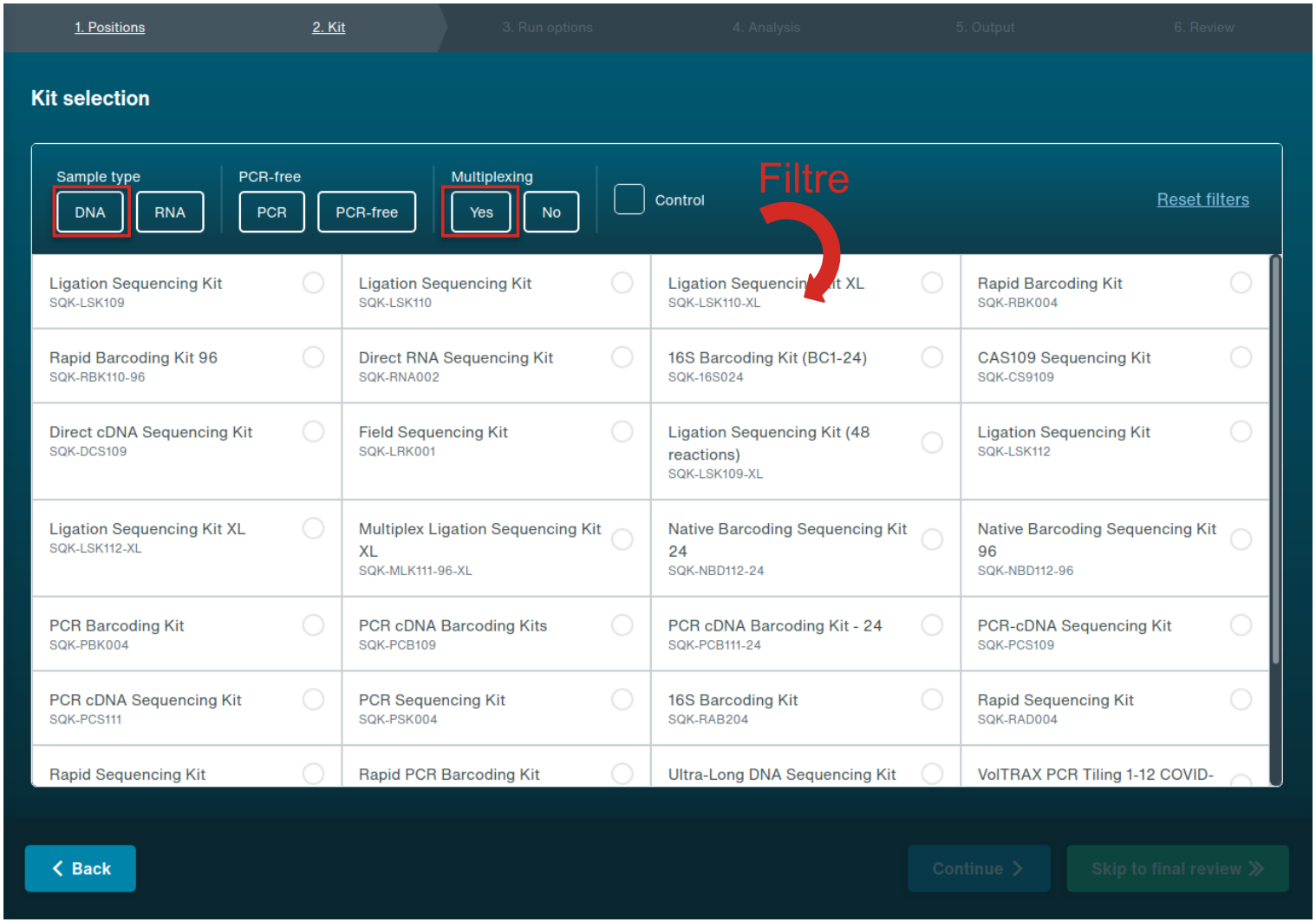


Définition de l'expérience



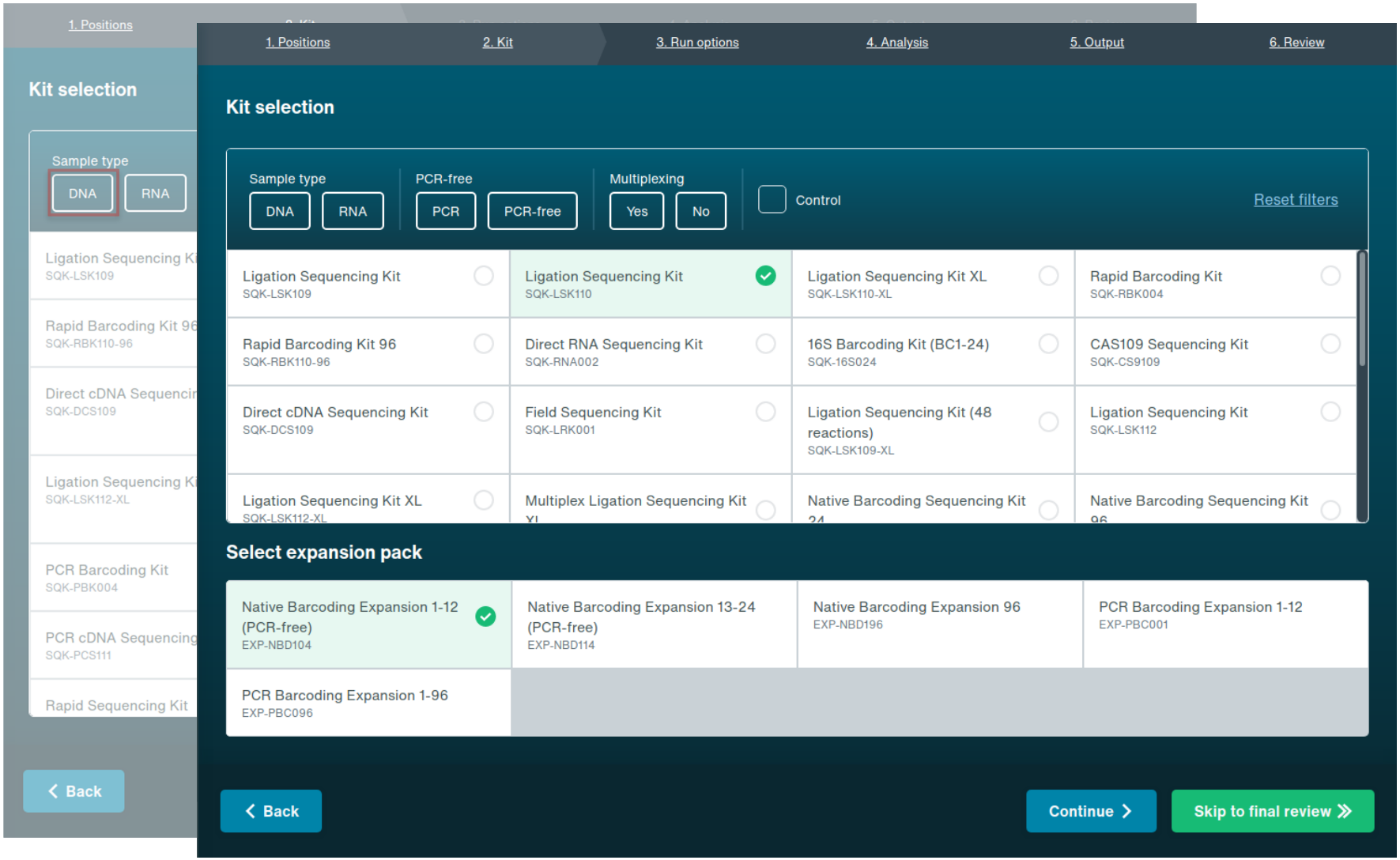
Choix du kit de séquençage

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

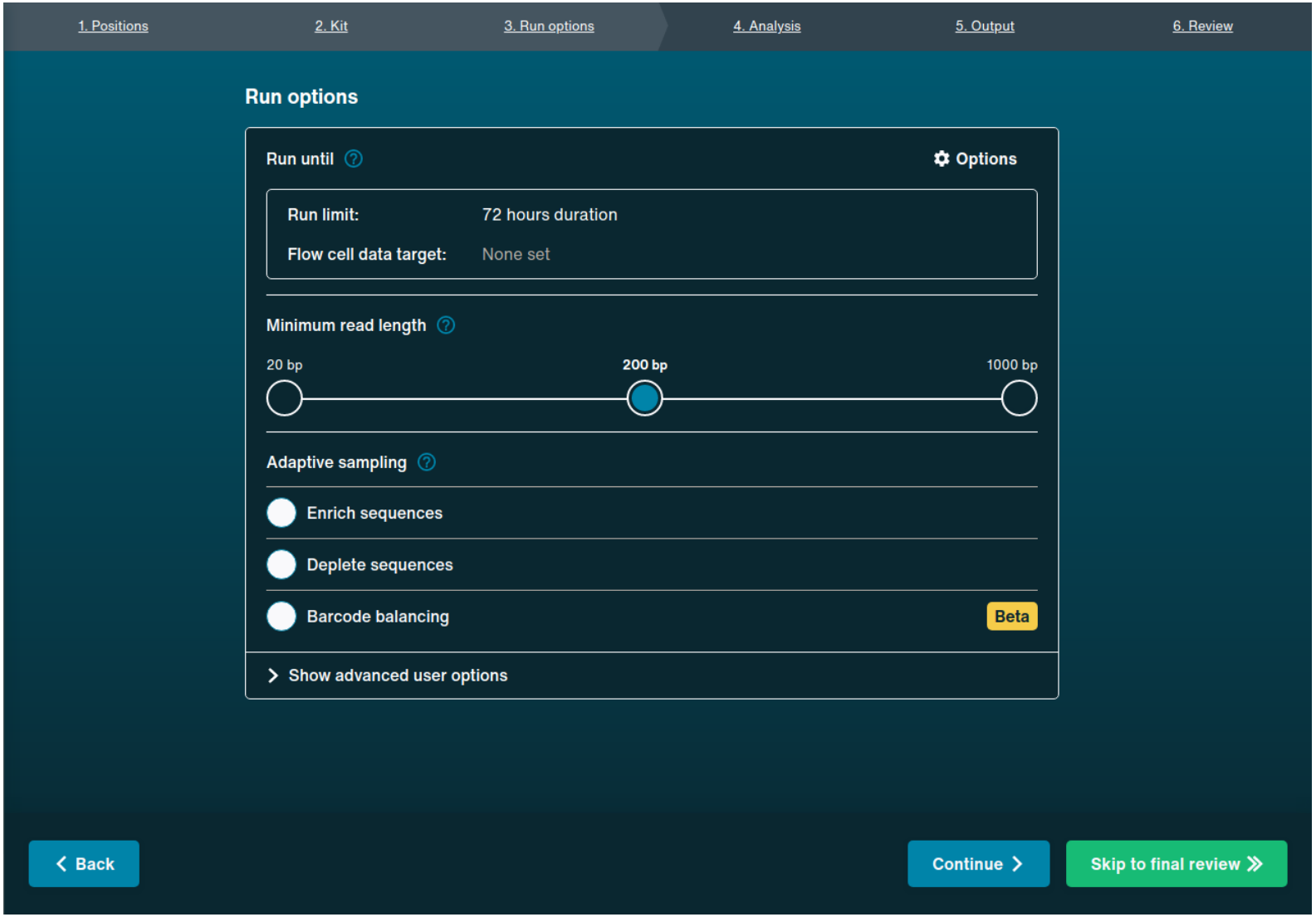


Choix du kit de séquençage

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

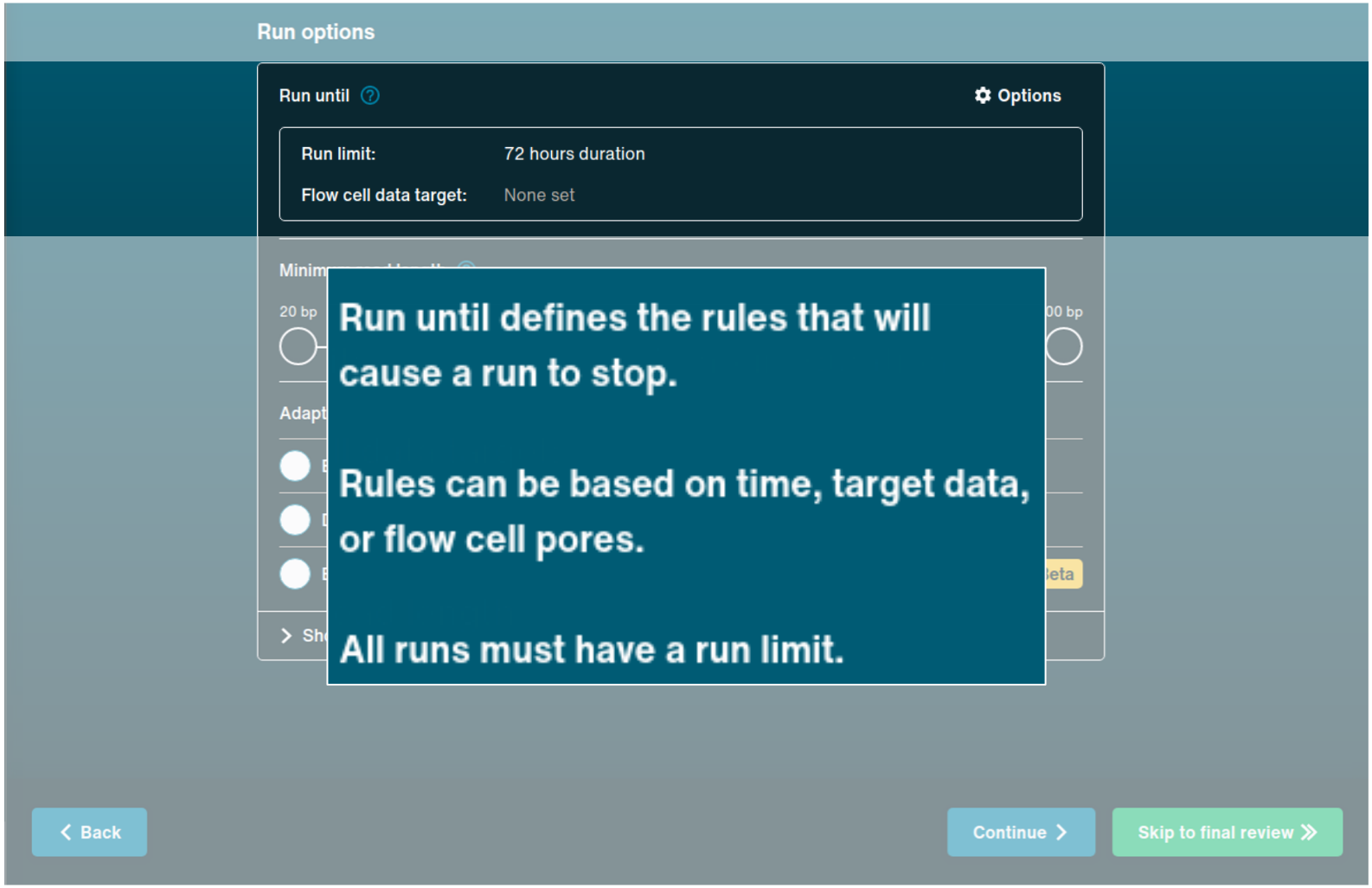


Configuration des options du run

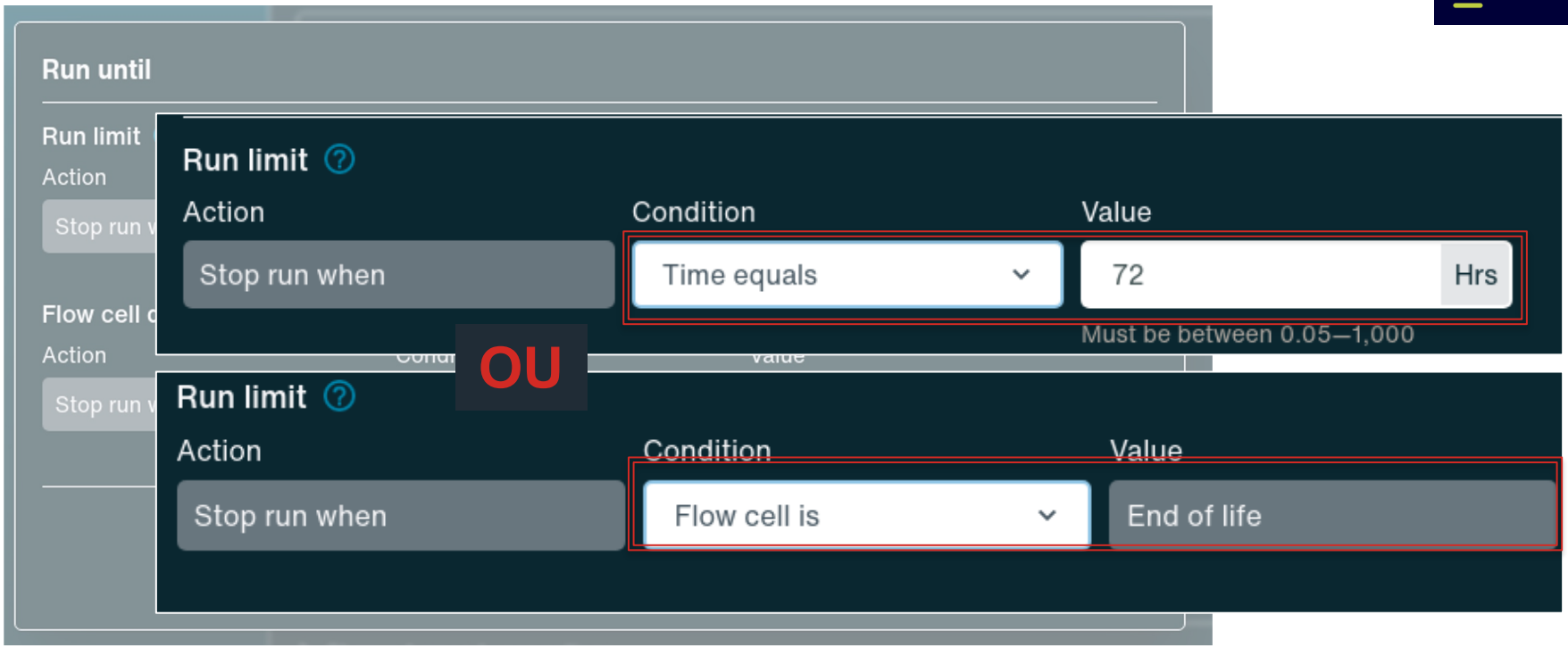


Les conditions d'arrêt normal du run :

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

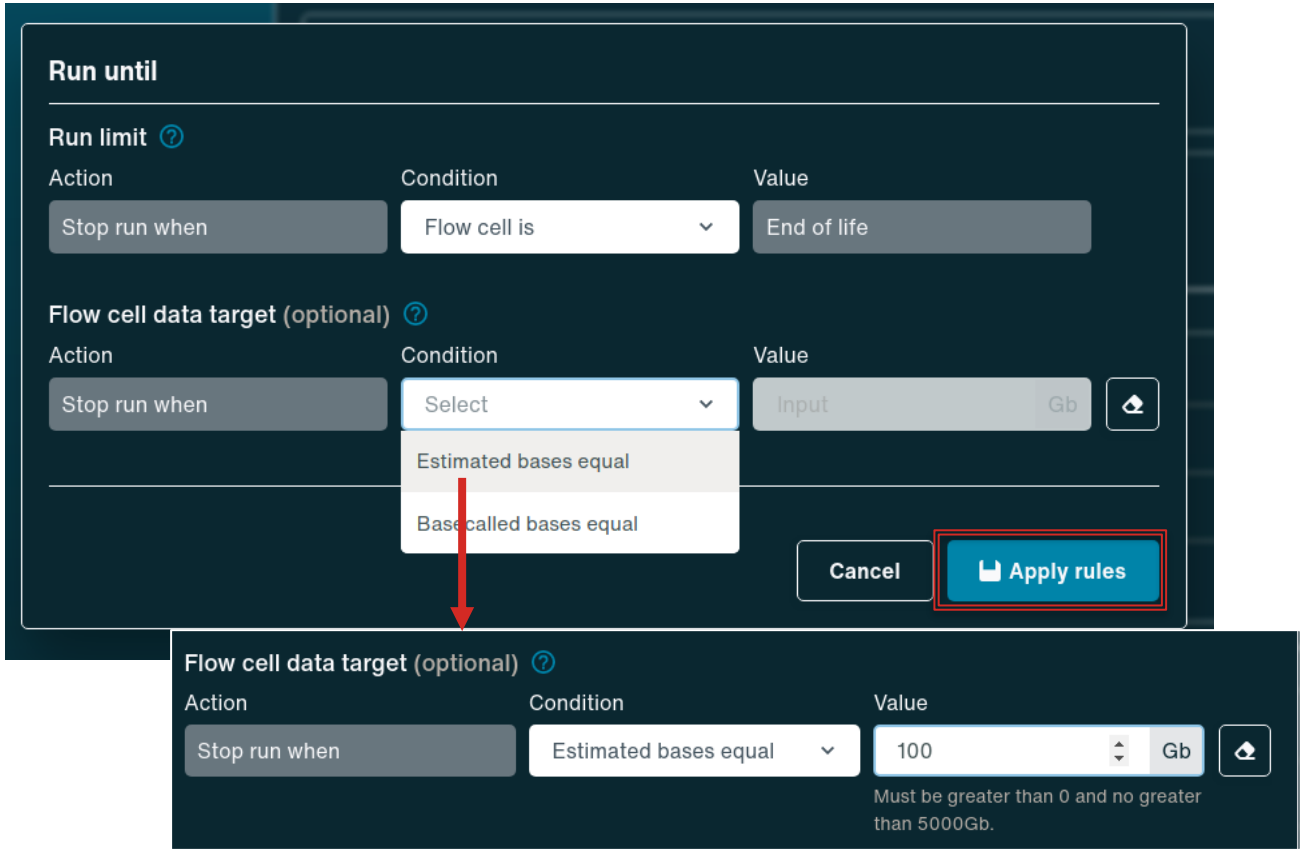


Les conditions d'arrêt normal du run : La durée

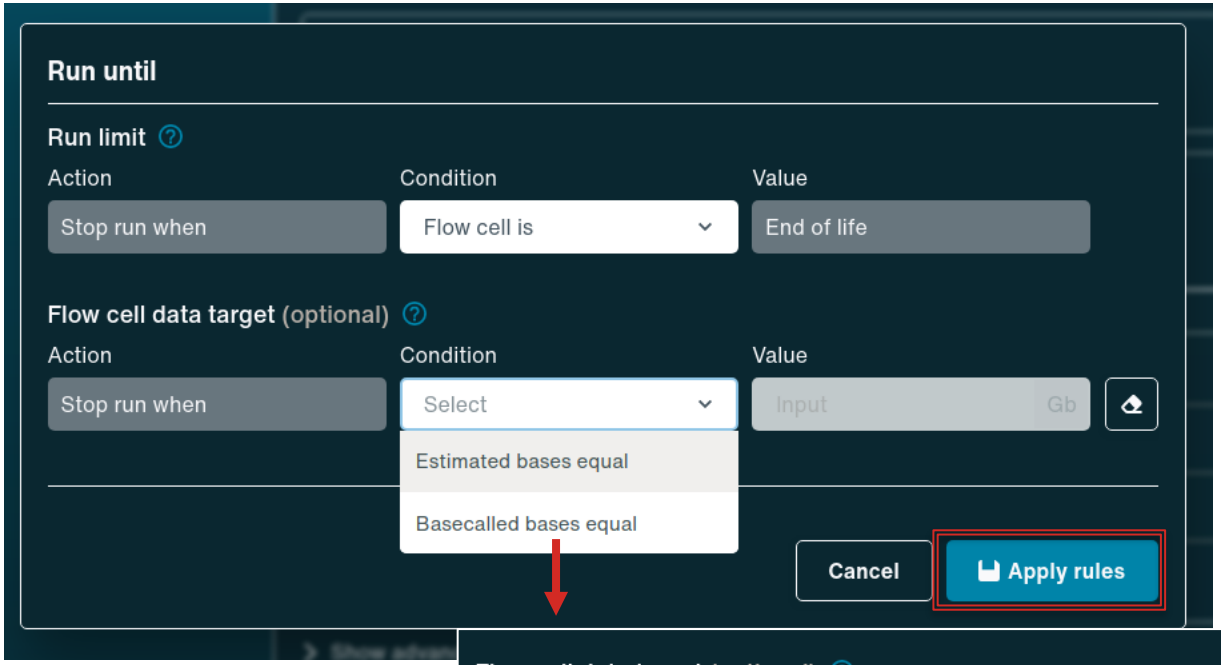


- En RNASeq, les runs durent 72h ou jusqu'à épuisement de la FC

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

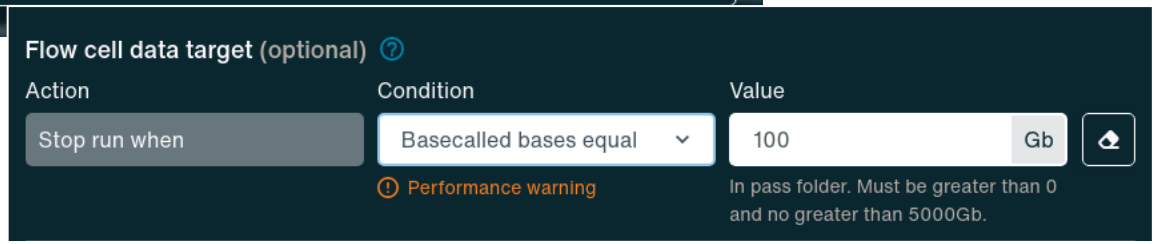


Les conditions d'arrêt normal du run : Le volume de données produit

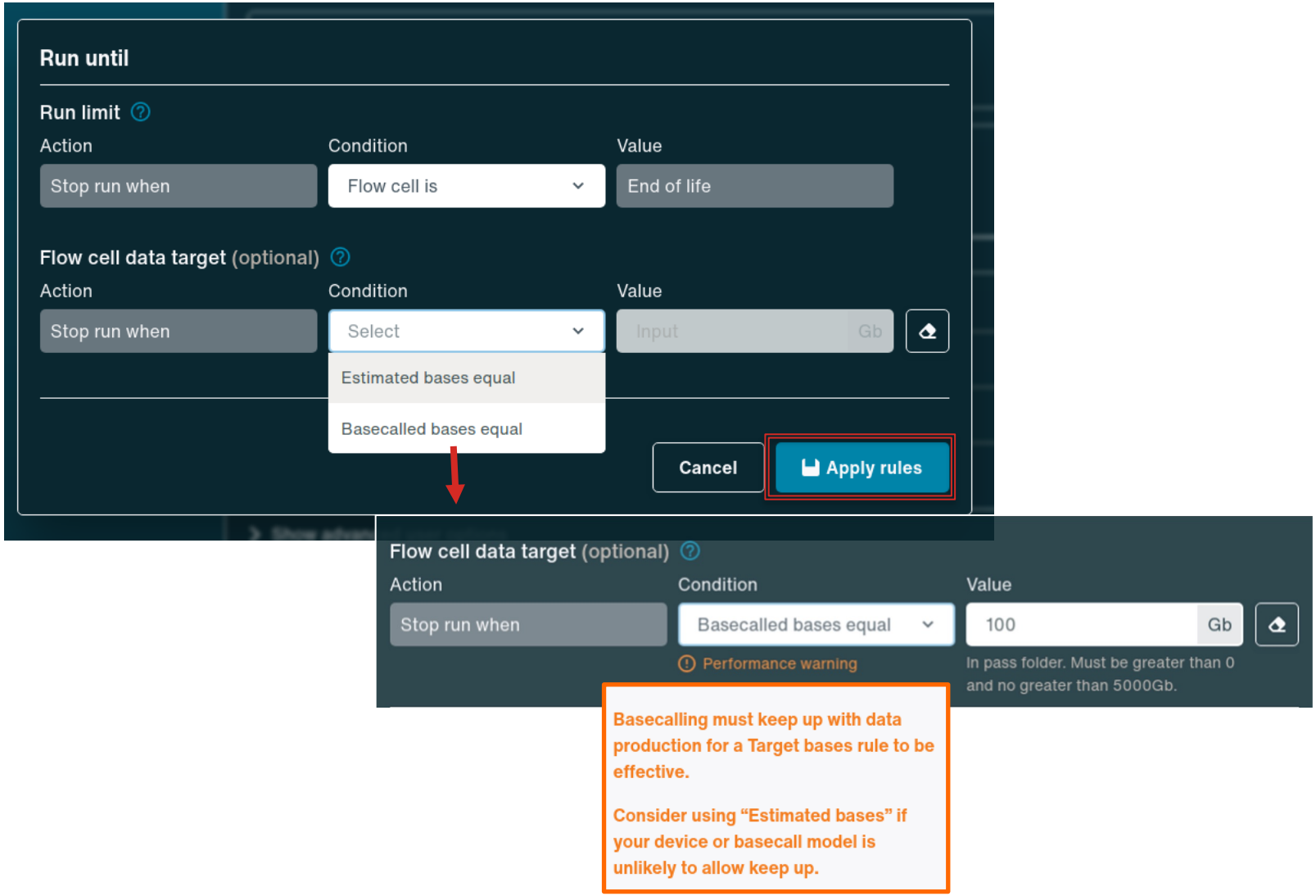


A target base rule will stop the run once the defined amount of estimated or basecalled bases are detected.

Targets are set per flow cell.

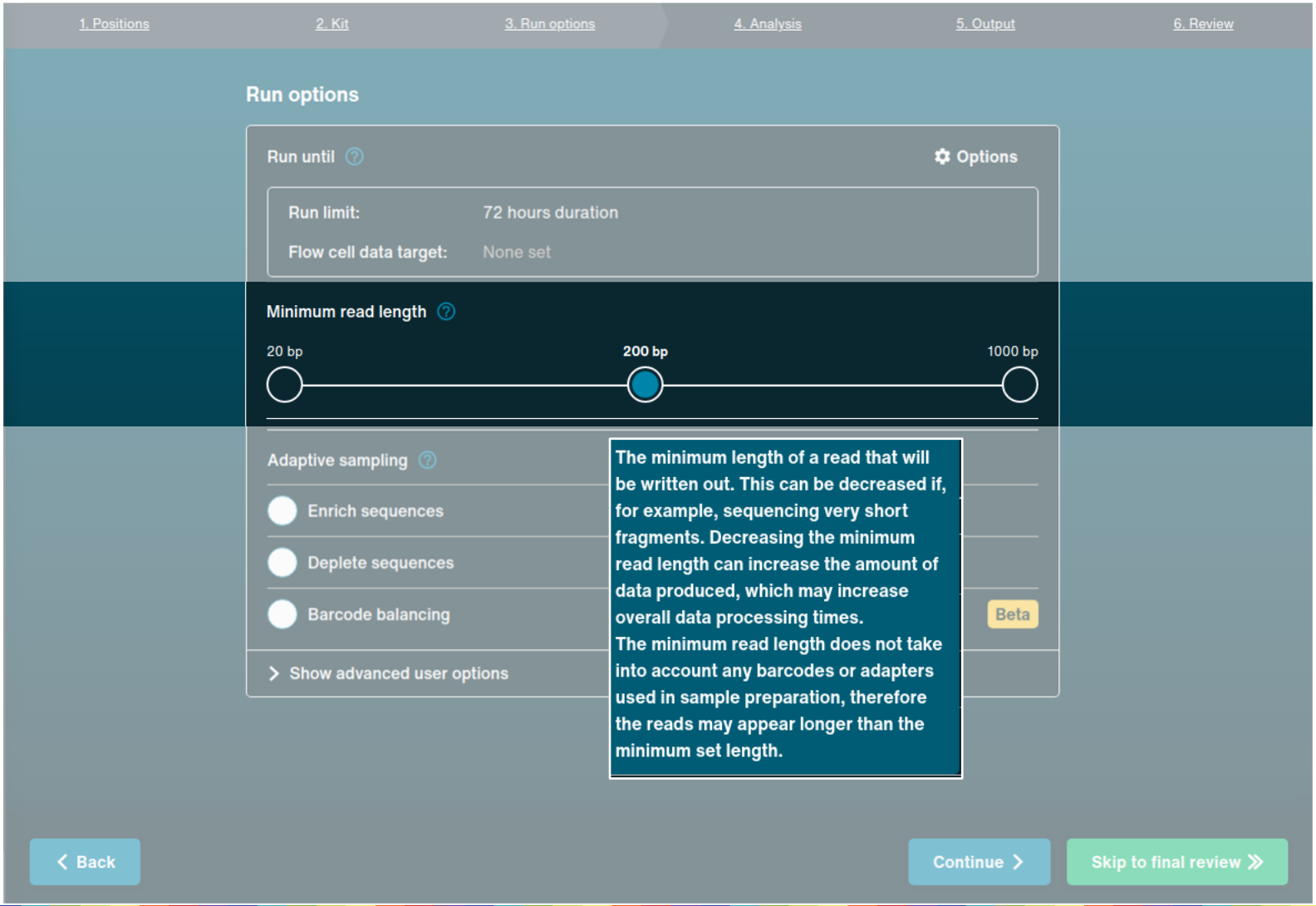


Les conditions d'arrêt normal du run : Le volume de données produit



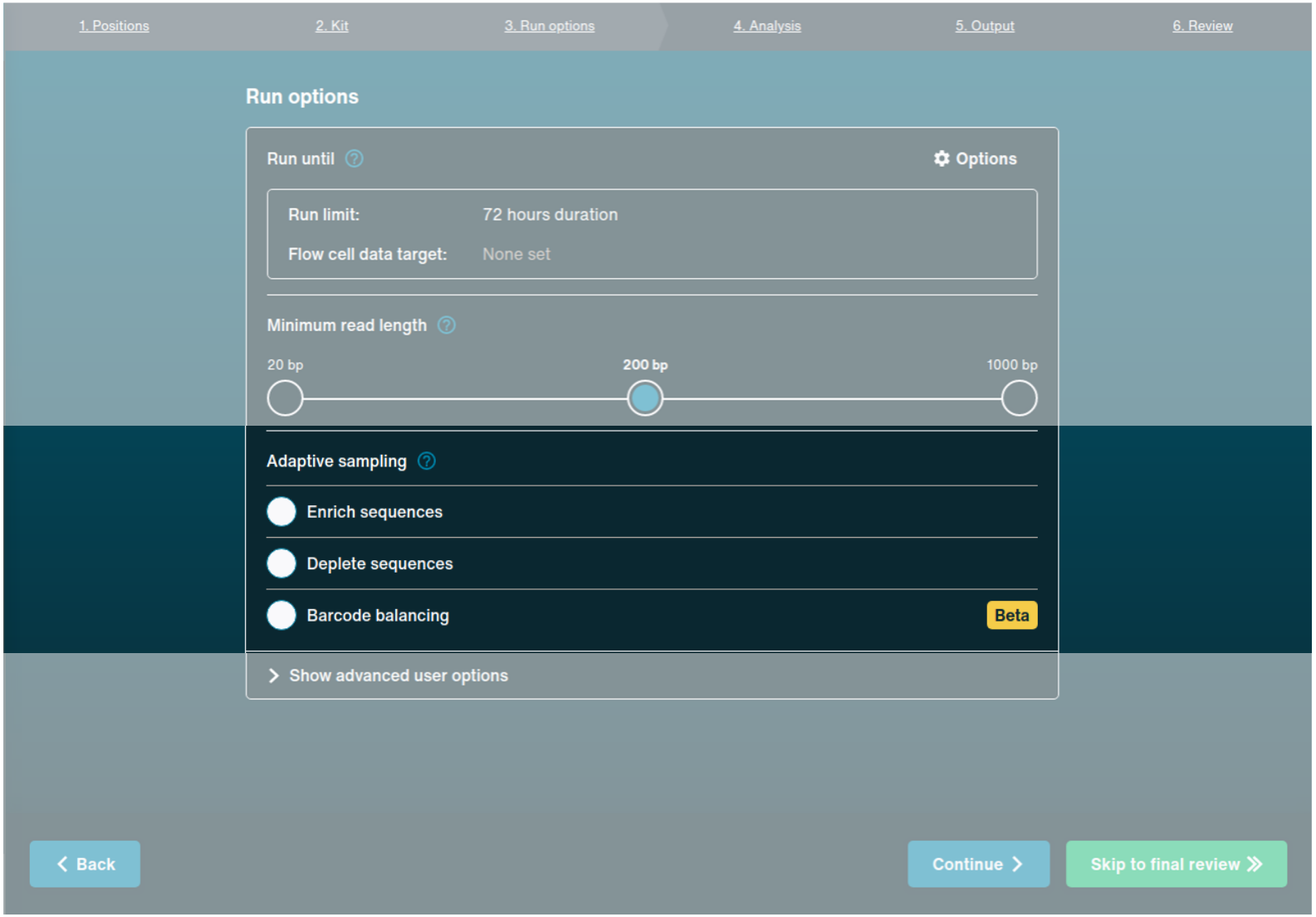
Filtre sur les données produites en préambule

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



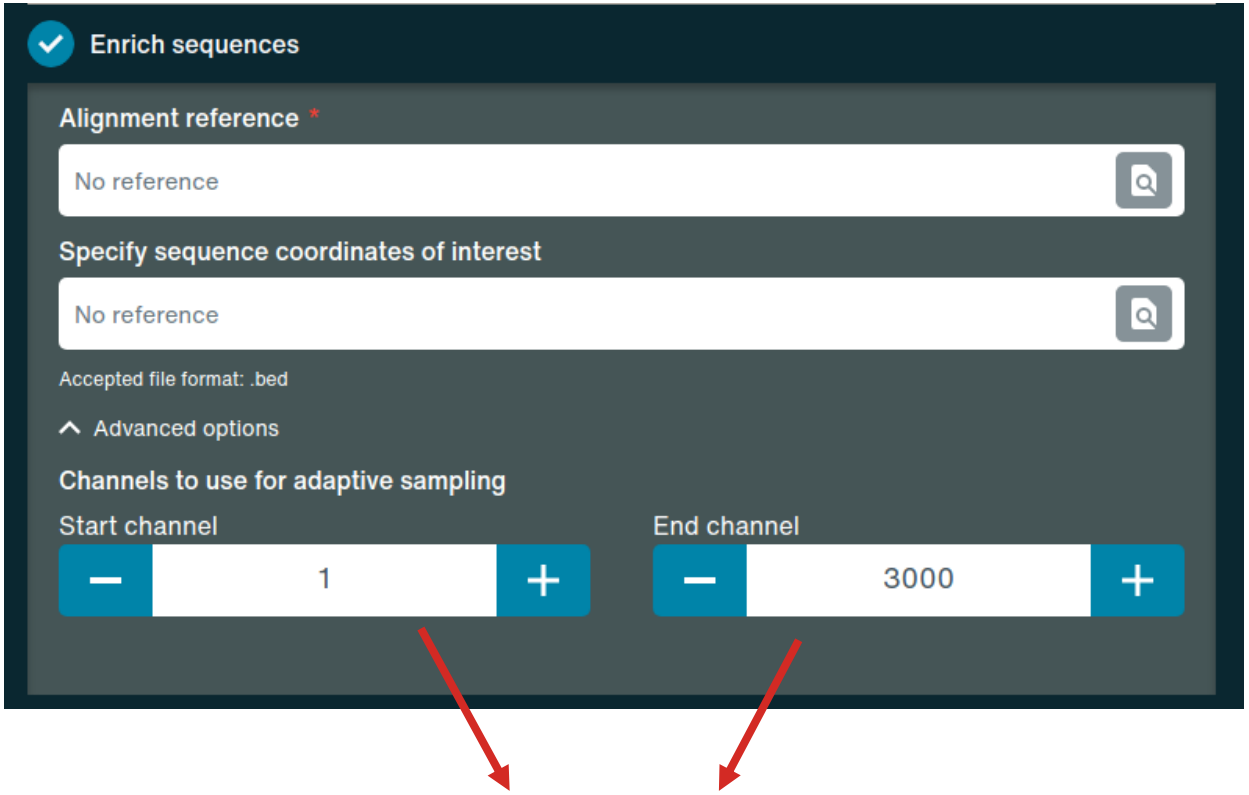
Filtre sur les données produites en préambule

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



Enrichissement

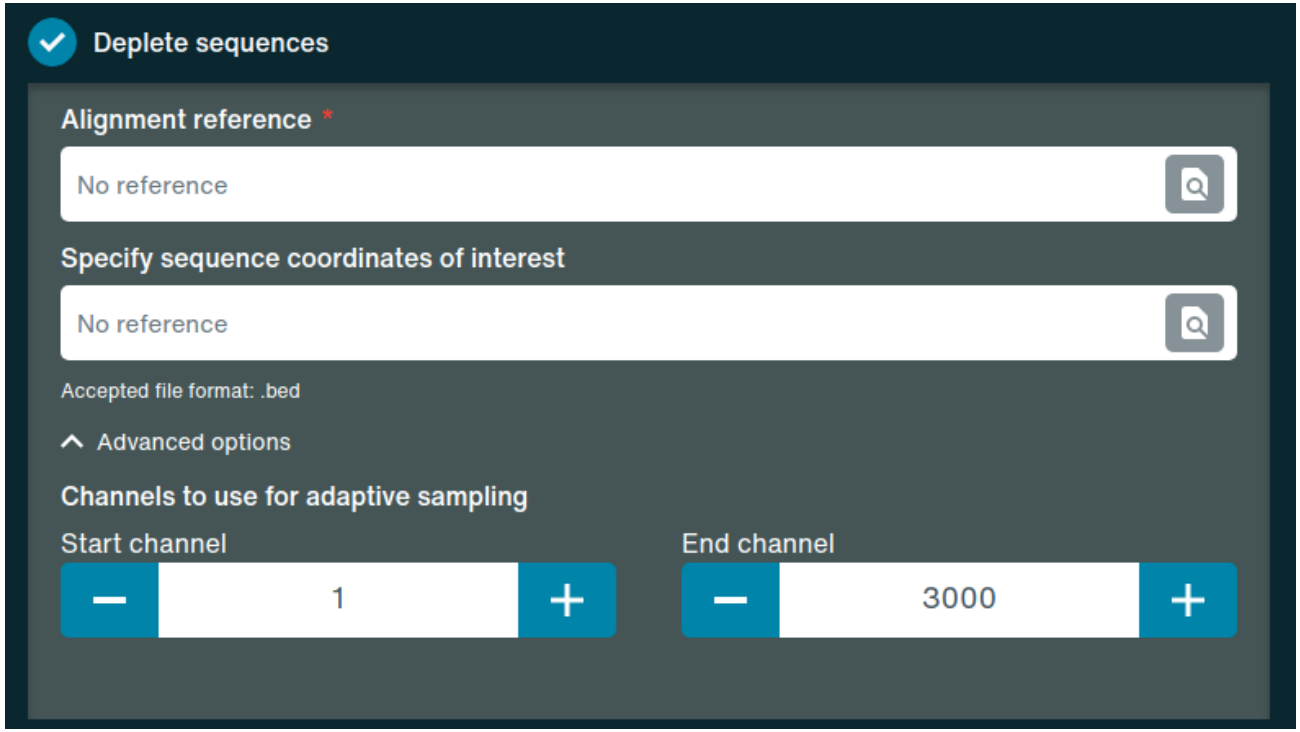
- Catalogue des séquences que l'on souhaite garder à fournir



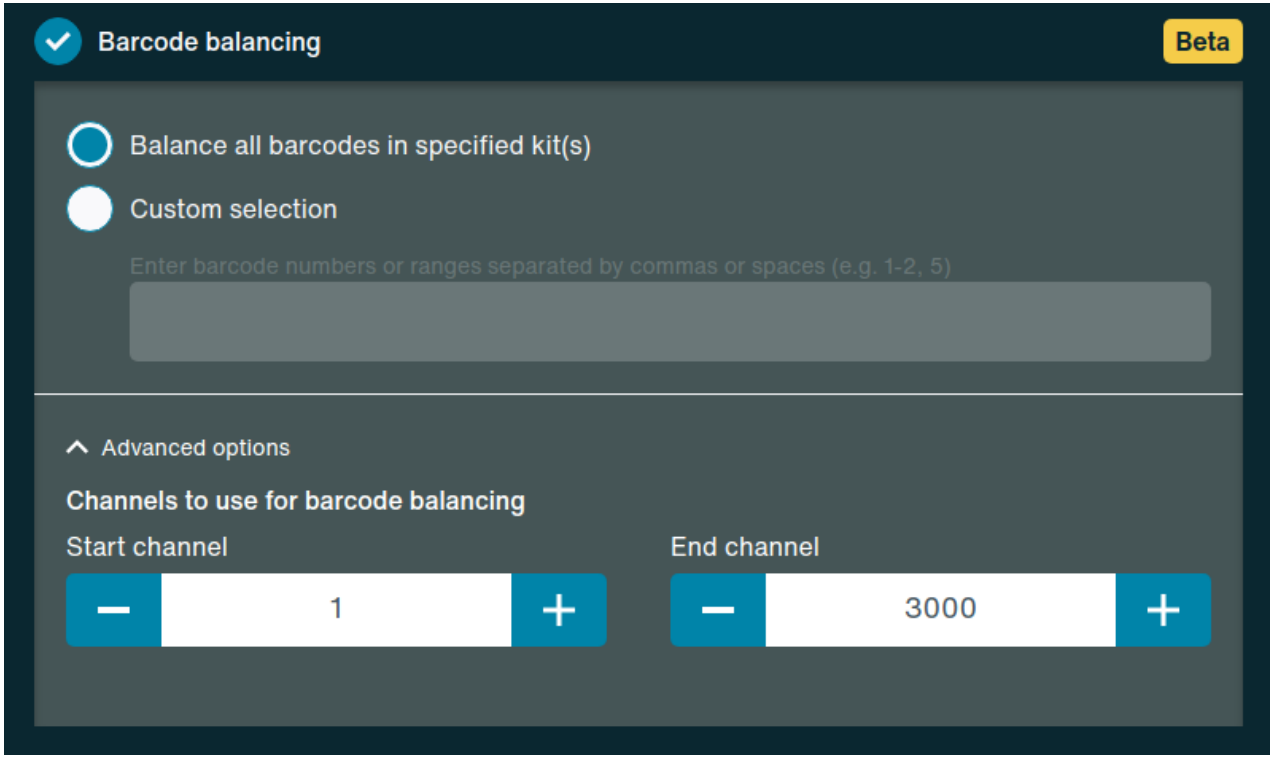
Il est possible de mobiliser seulement une partie de la FC dans cette tache

Appauvrissement

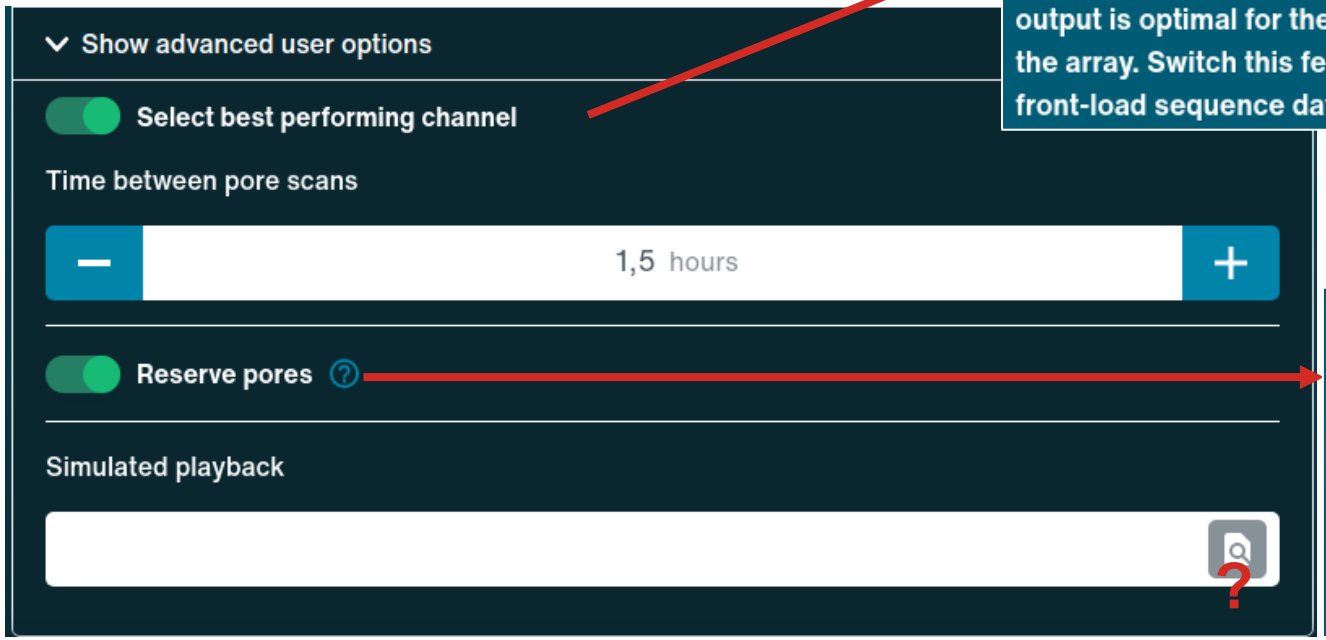
- Catalogue des séquences que l'on ne souhaite pas garder à fournir



Equilibrage des codes barres utilisés en cours de séquençage



Paramètres avancés

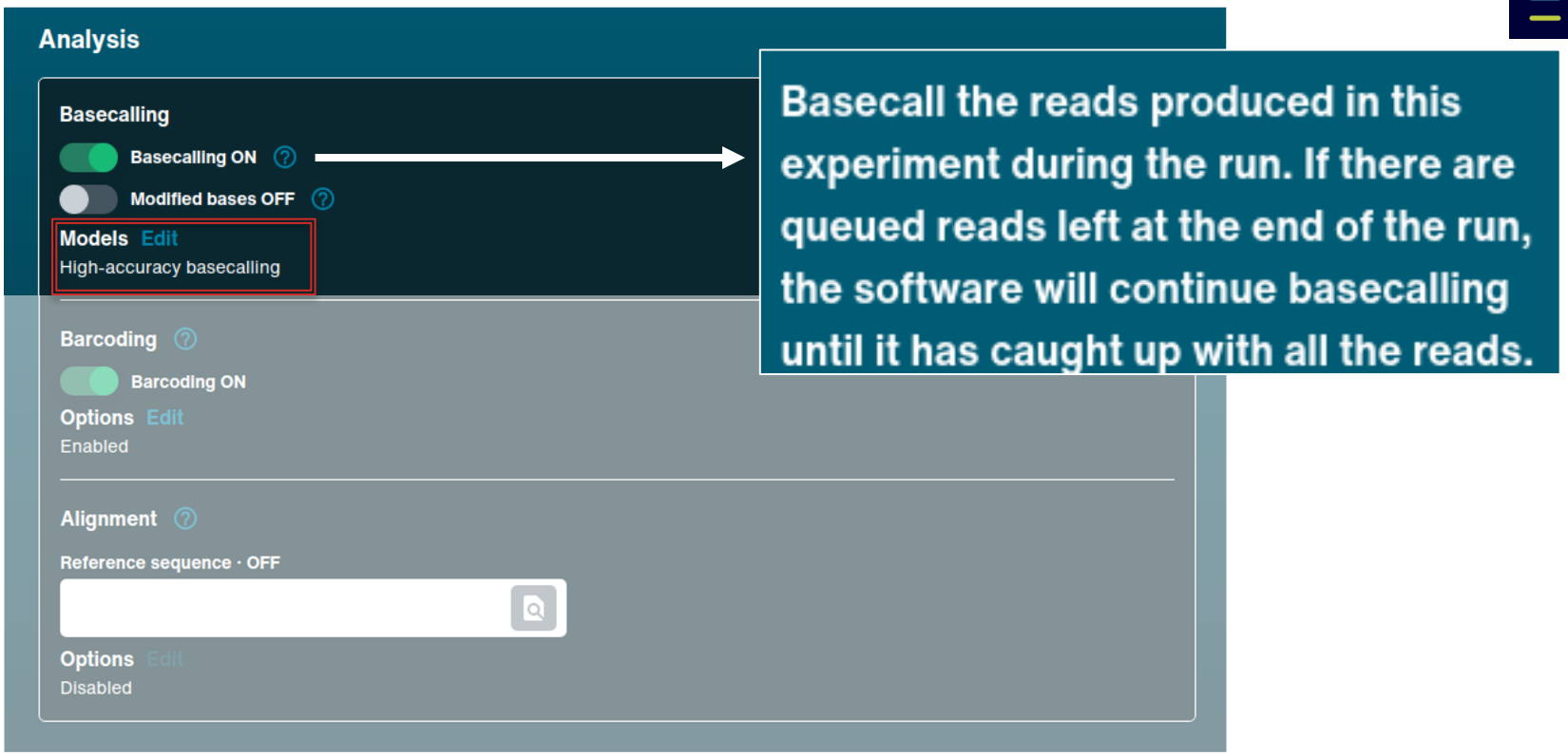


Delay using some of the pores until the voltage has dropped later in the run, such that other pores can catch up. This can be used when e.g. wanting to save pores until their accuracy and output is optimal for the voltage across the array. Switch this feature off to fully front-load sequence data acquisition.

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

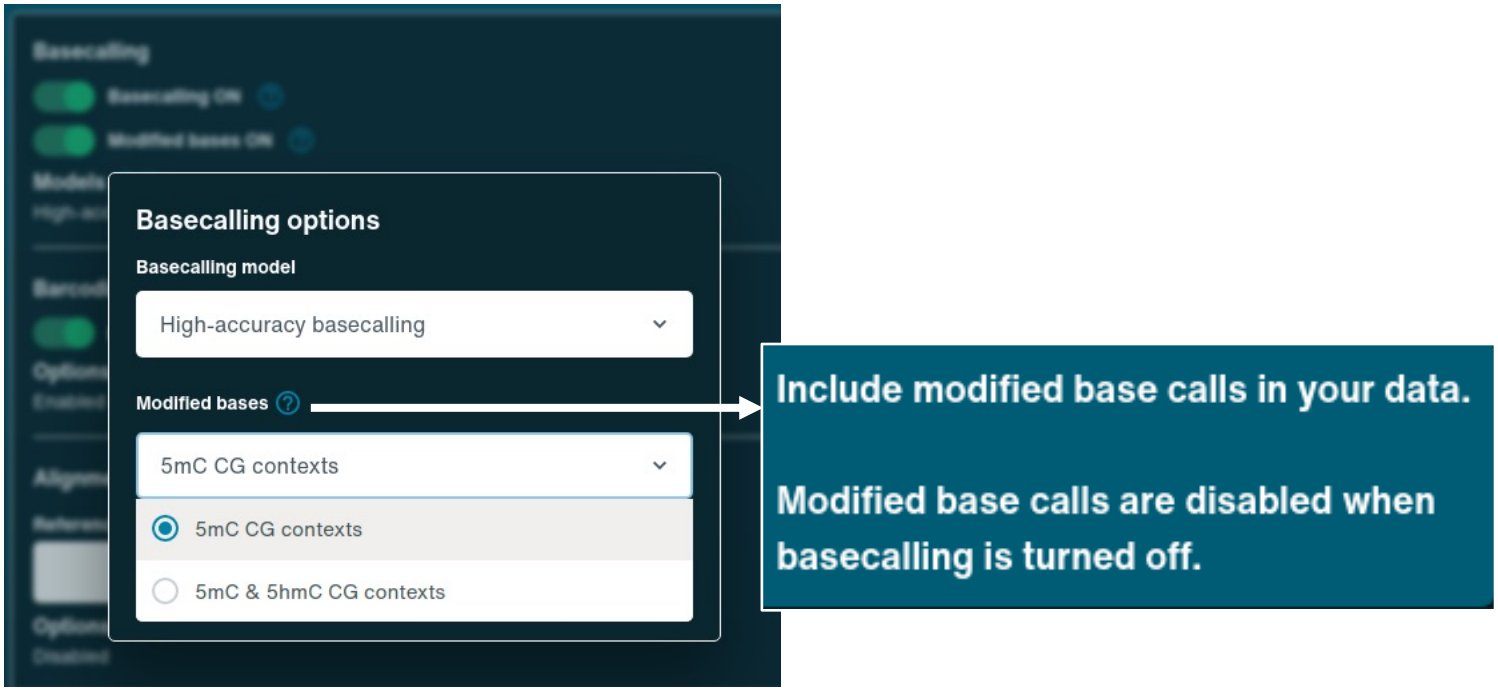


Trois possibilités de basecalling :

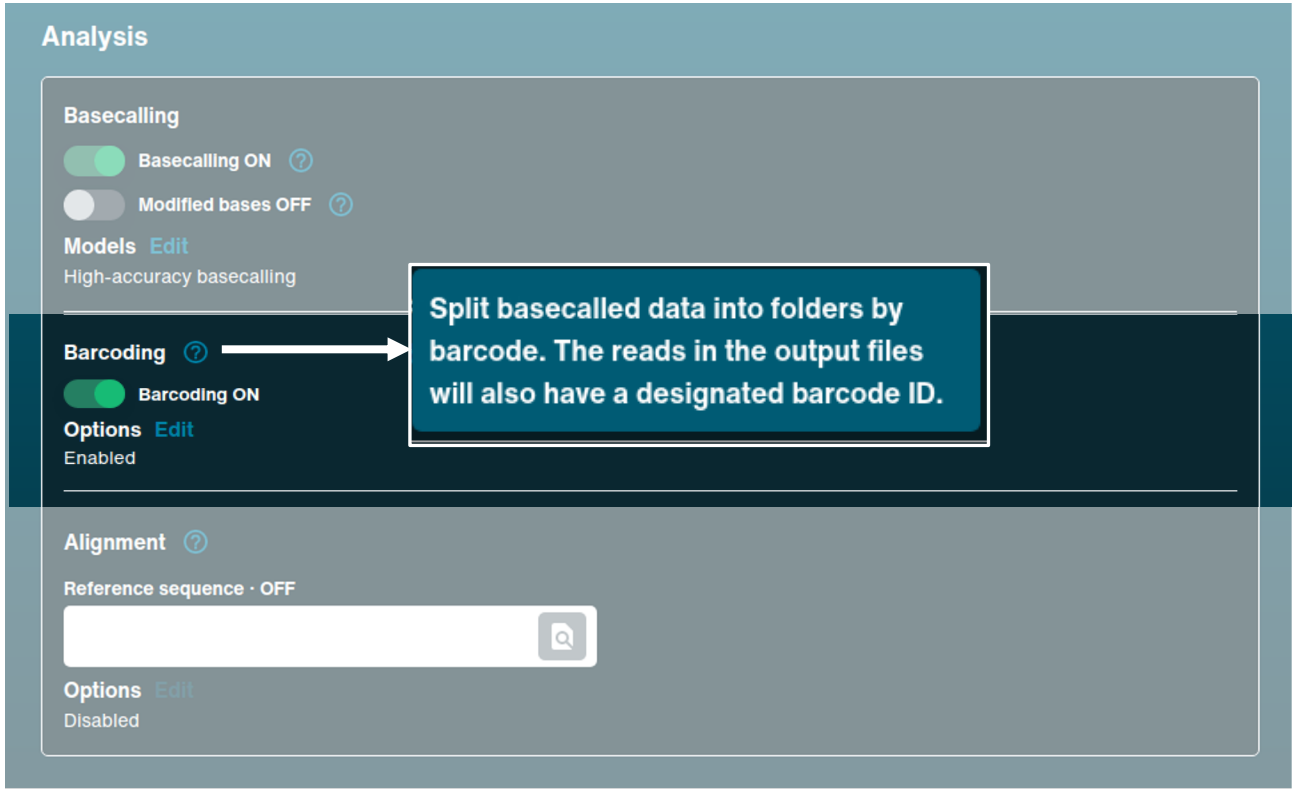
- Fast : Pratique pour le diagnostique parce rapide
- High-accuracy : Plus long mais moins d'erreur
- Super-accurate : Encore plus long mais moins d'erreur

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

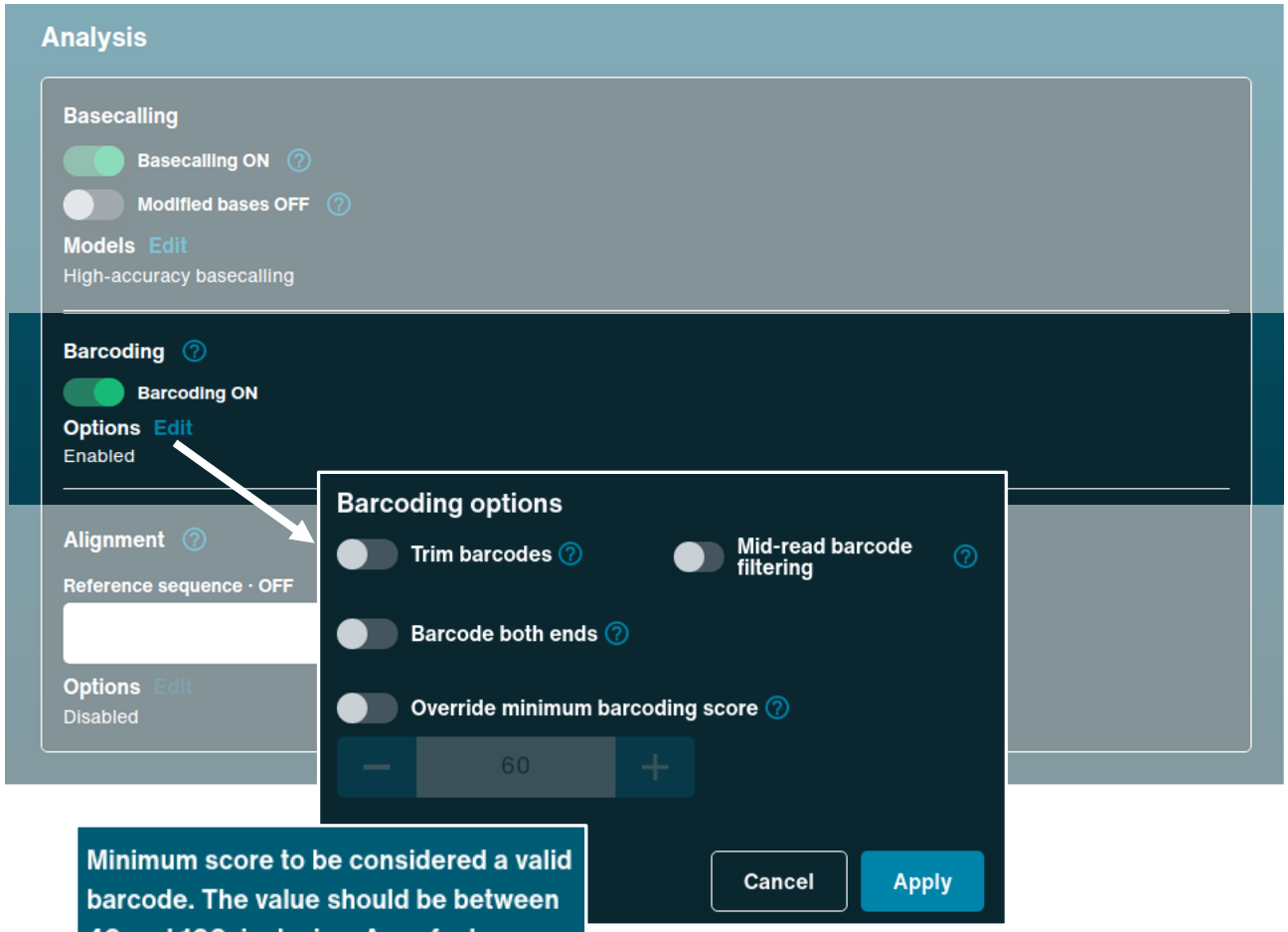
- Passage du signal électrique au fichier séquence
- Dictionnaires de bases possibles incluent certaines bases modifiées
  - Attention car le catalogue dépend de la FC et du protocole de fabrication de banque



Gestion des codes barres



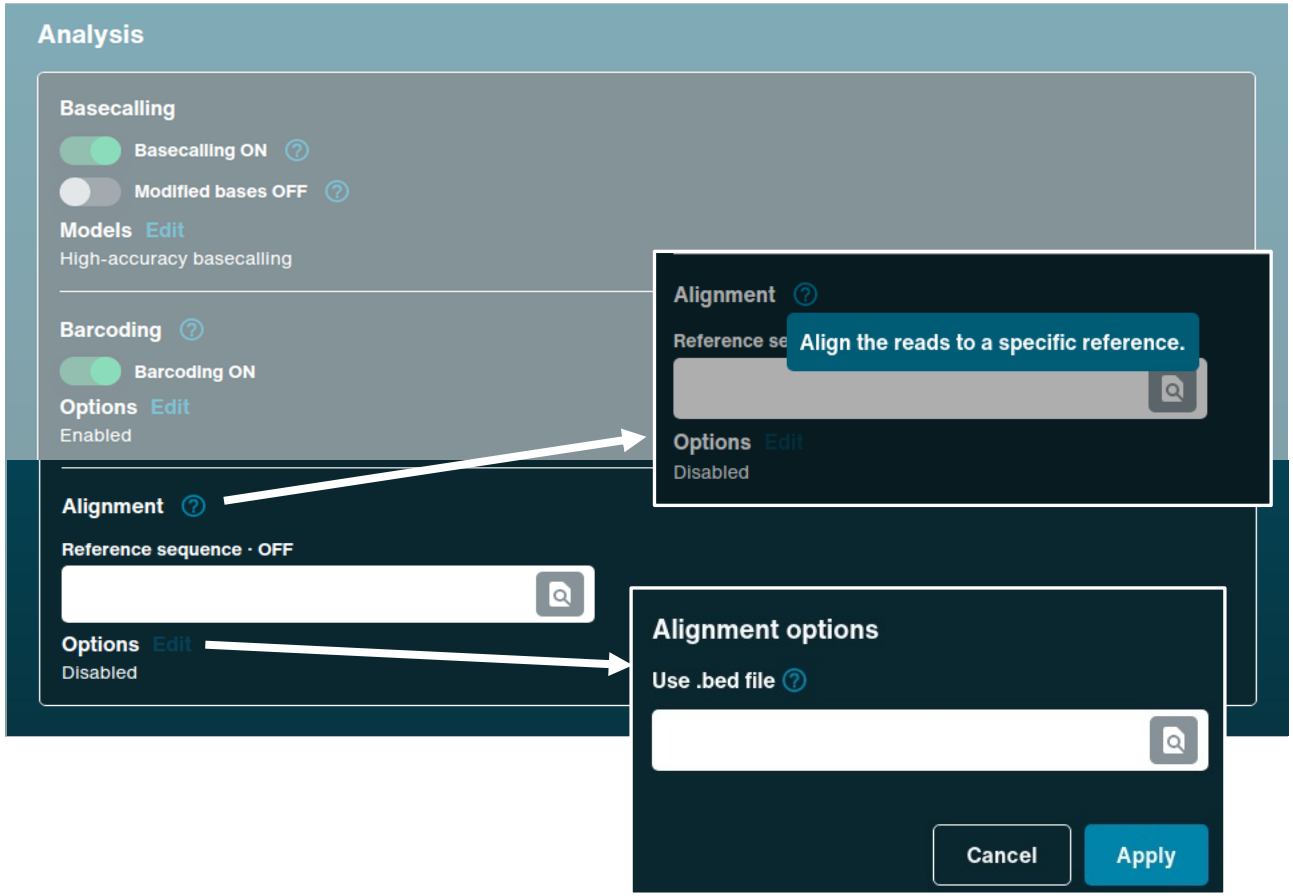
Gestion des codes barres : le triming



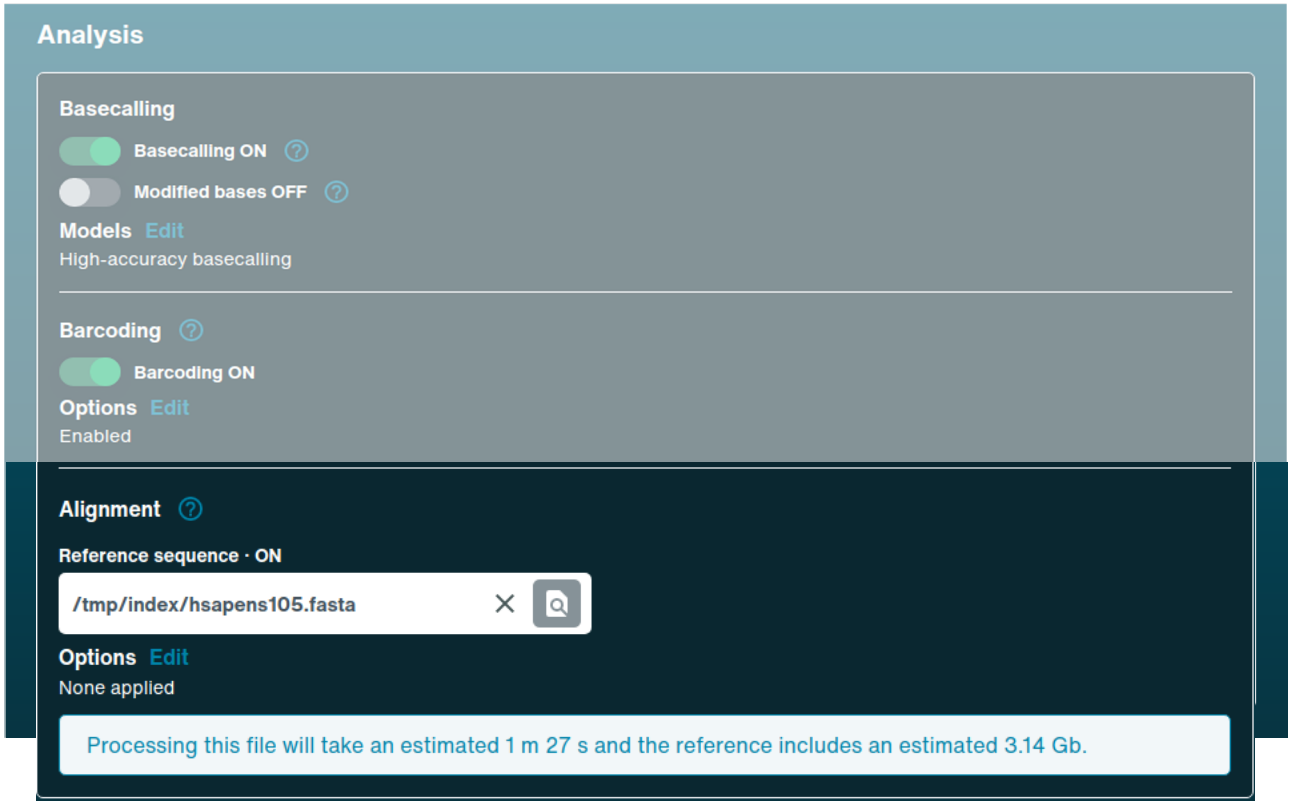
Minimum score to be considered a valid barcode. The value should be between 40 and 100, inclusive. A perfect barcode match will score 100, but lower scores will be assigned to imperfect matches based on the number of errors. A higher score threshold increases specificity at the expense of more unclassified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

Alignement des séquences en simultané

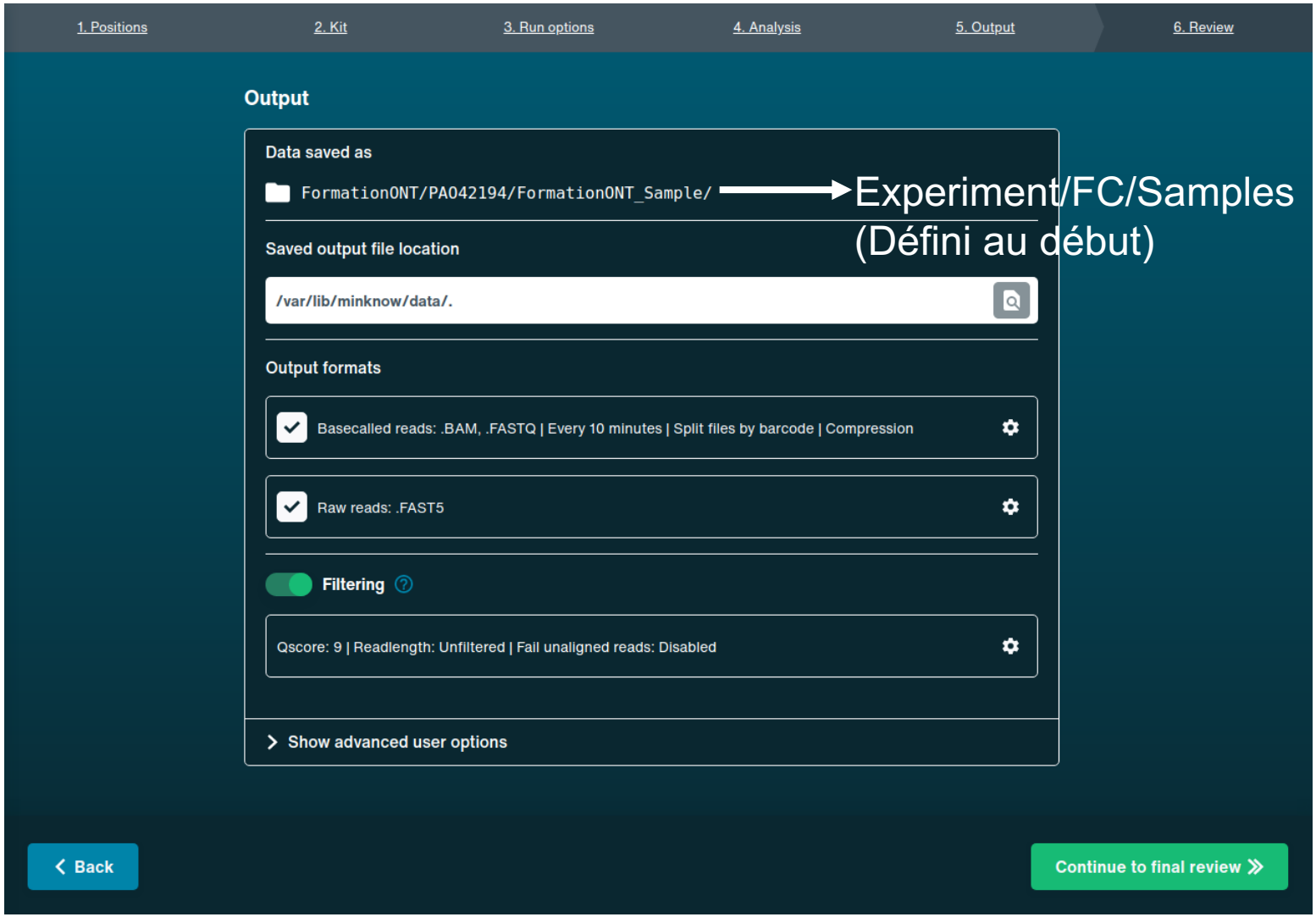
- Peut être fait à postériori
- Peut avoir un intérêt pour du séquençage de procaryotes, des séquençages courts…
- Réalisé avec Minimap2
  - Index à faire en début de paramétrage (fichier .mmi) ou à charger en lieu d'un fasta



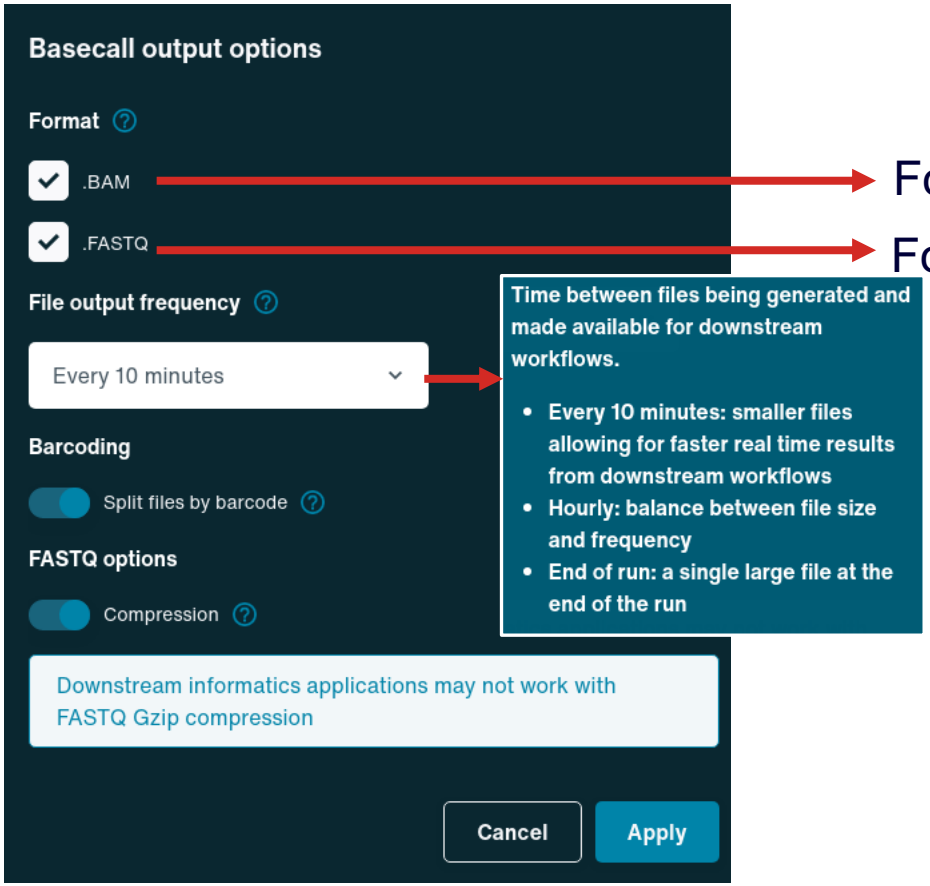
Alignement des séquences en simultané



Les fichiers de sortie



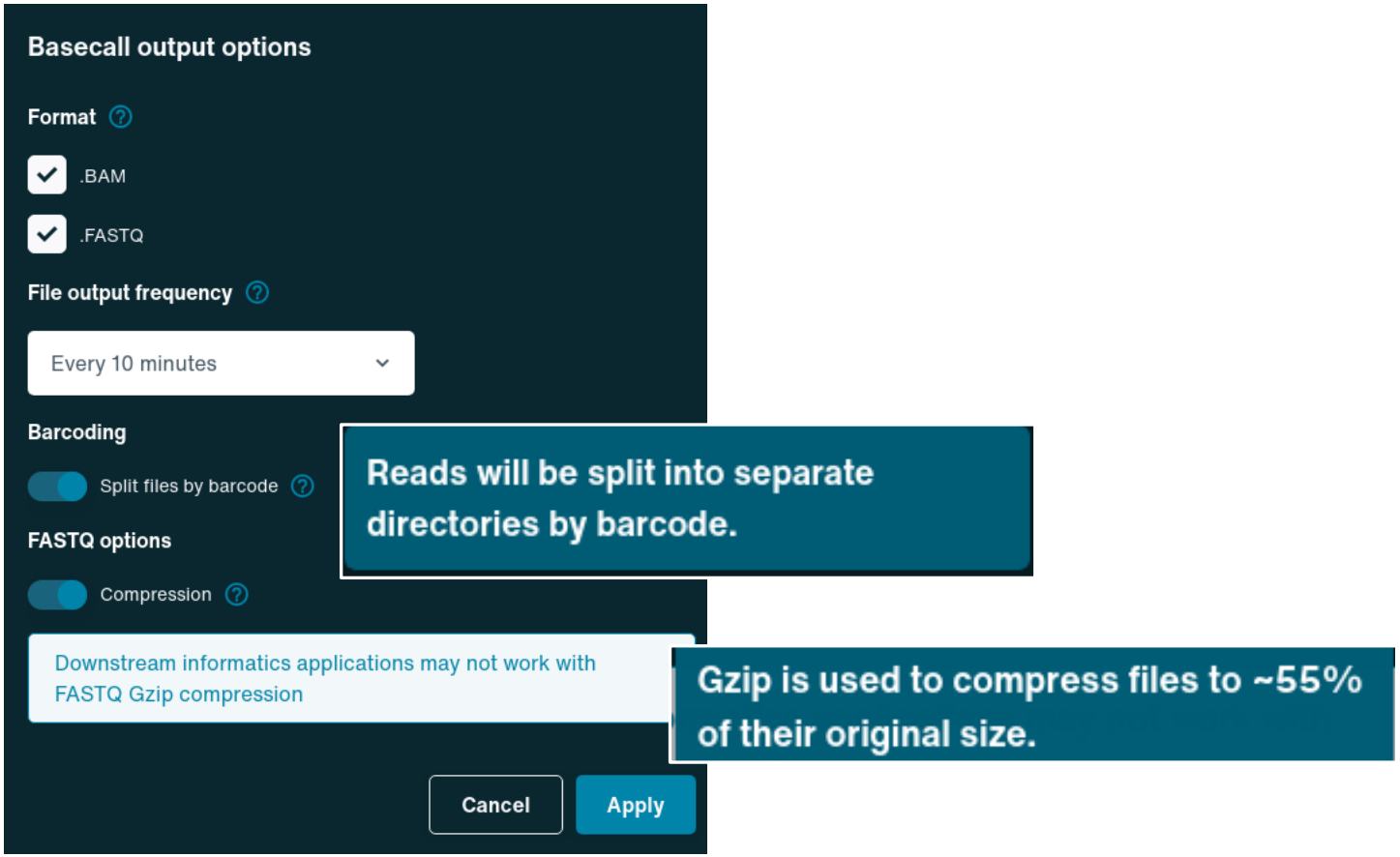
Les fichiers de sortie :
Les formats des données après le basecalling



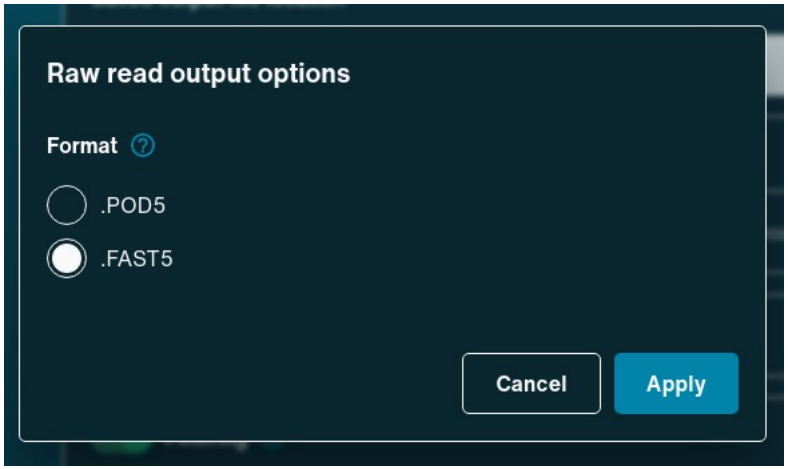
Format binaire

Format non binaire sur 4 lignes

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



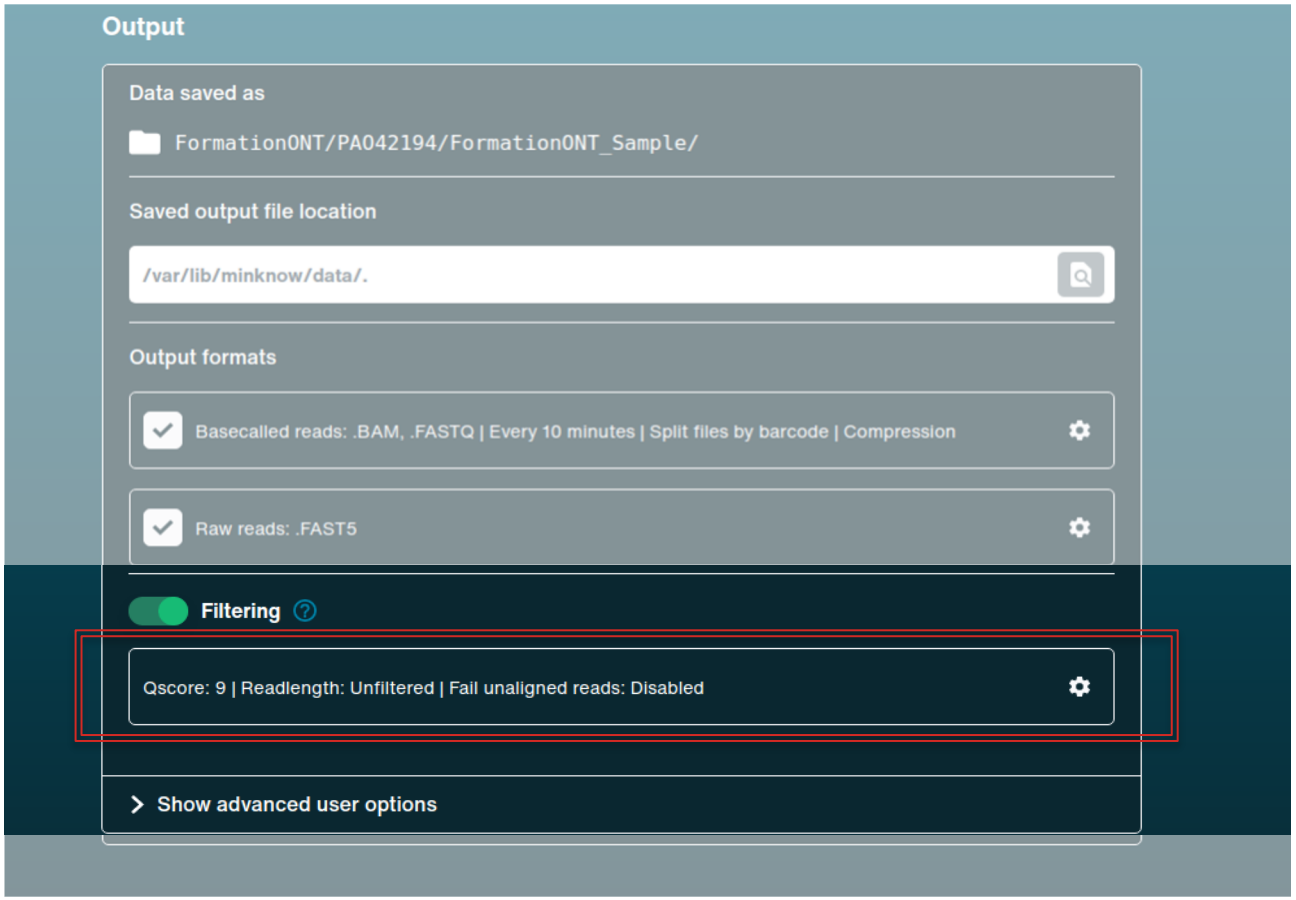
Les fichiers de sortie : Les formats des données brutes



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

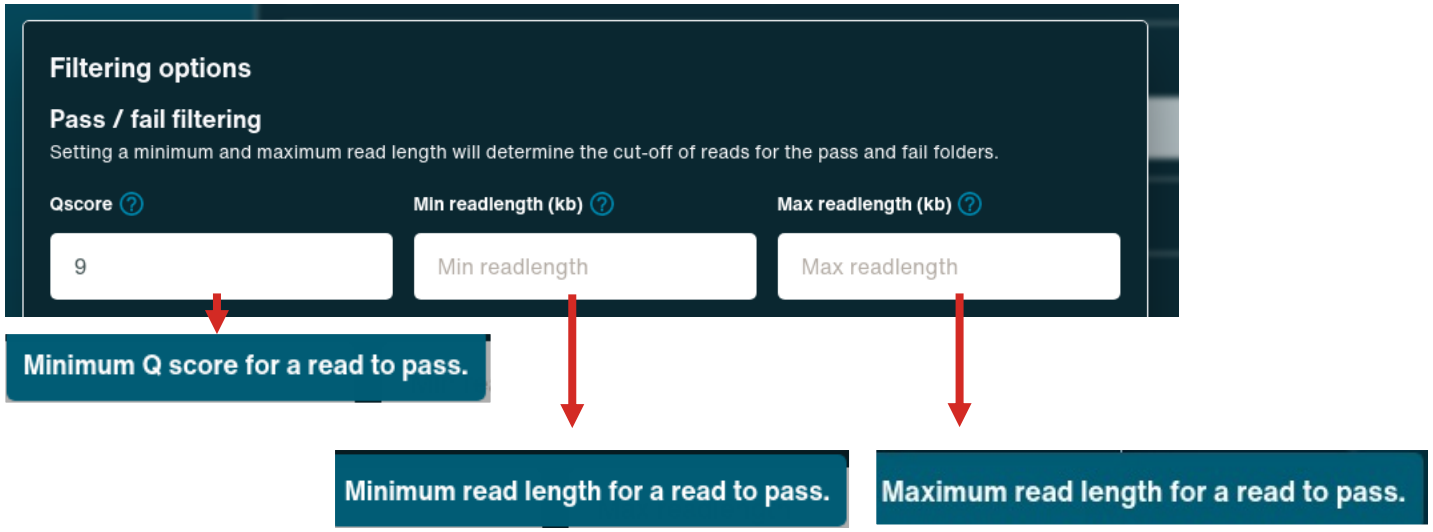
Les fichiers de sortie : Les filtres appliqués



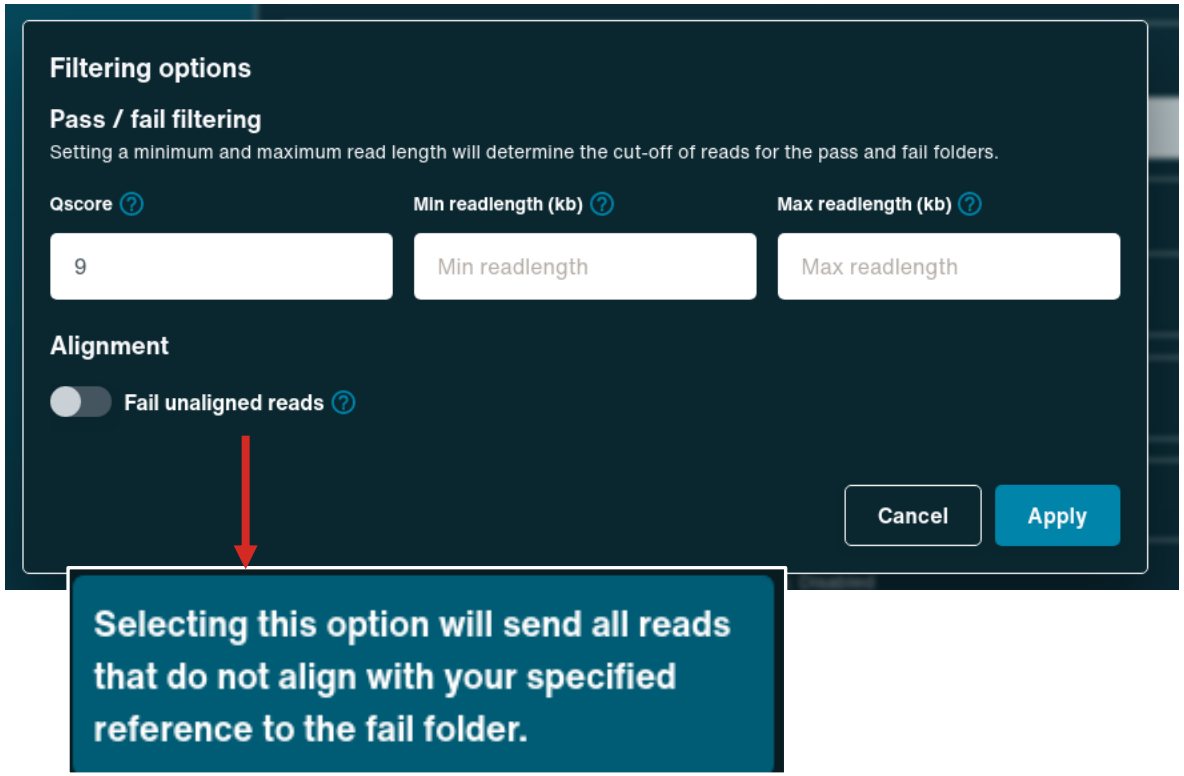
Par défaut avec le modèle high-accuracy :

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

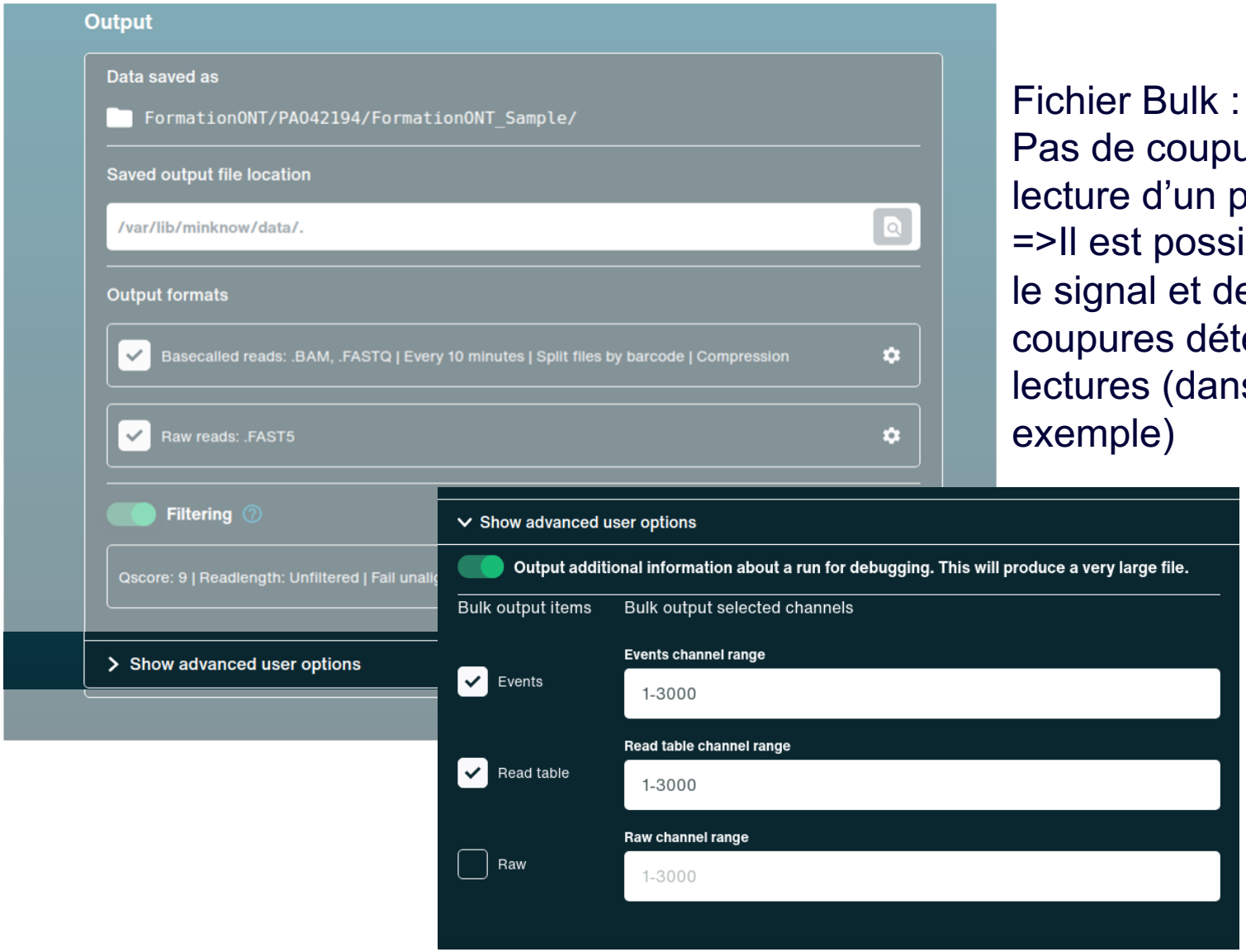
Les fichiers de sortie :
Les filtres appliqués sur les séquences brutes



Les fichiers de sortie :
Les filtres appliqués sur les alignements

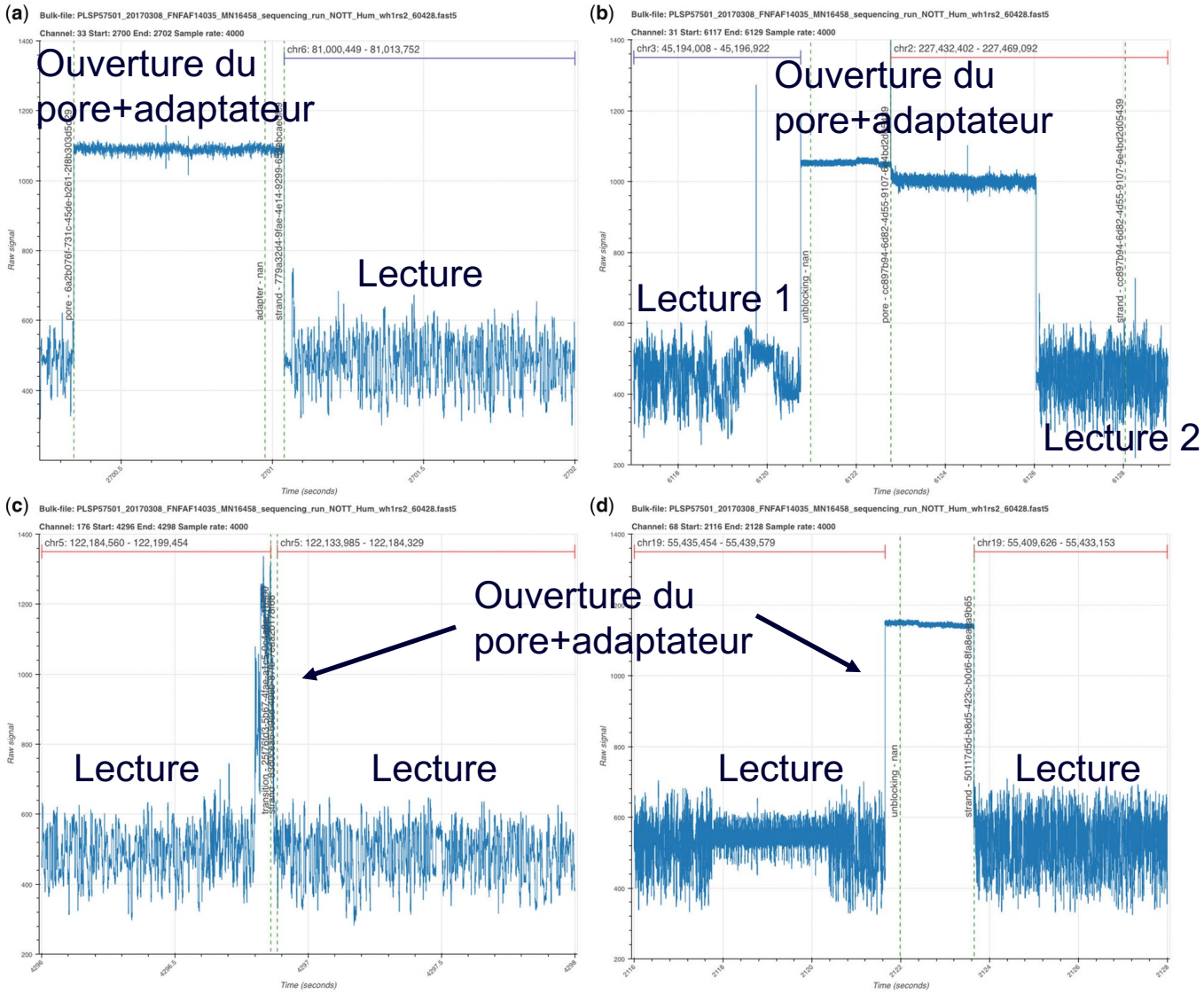


Les fichiers de sortie : Les options avancés



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



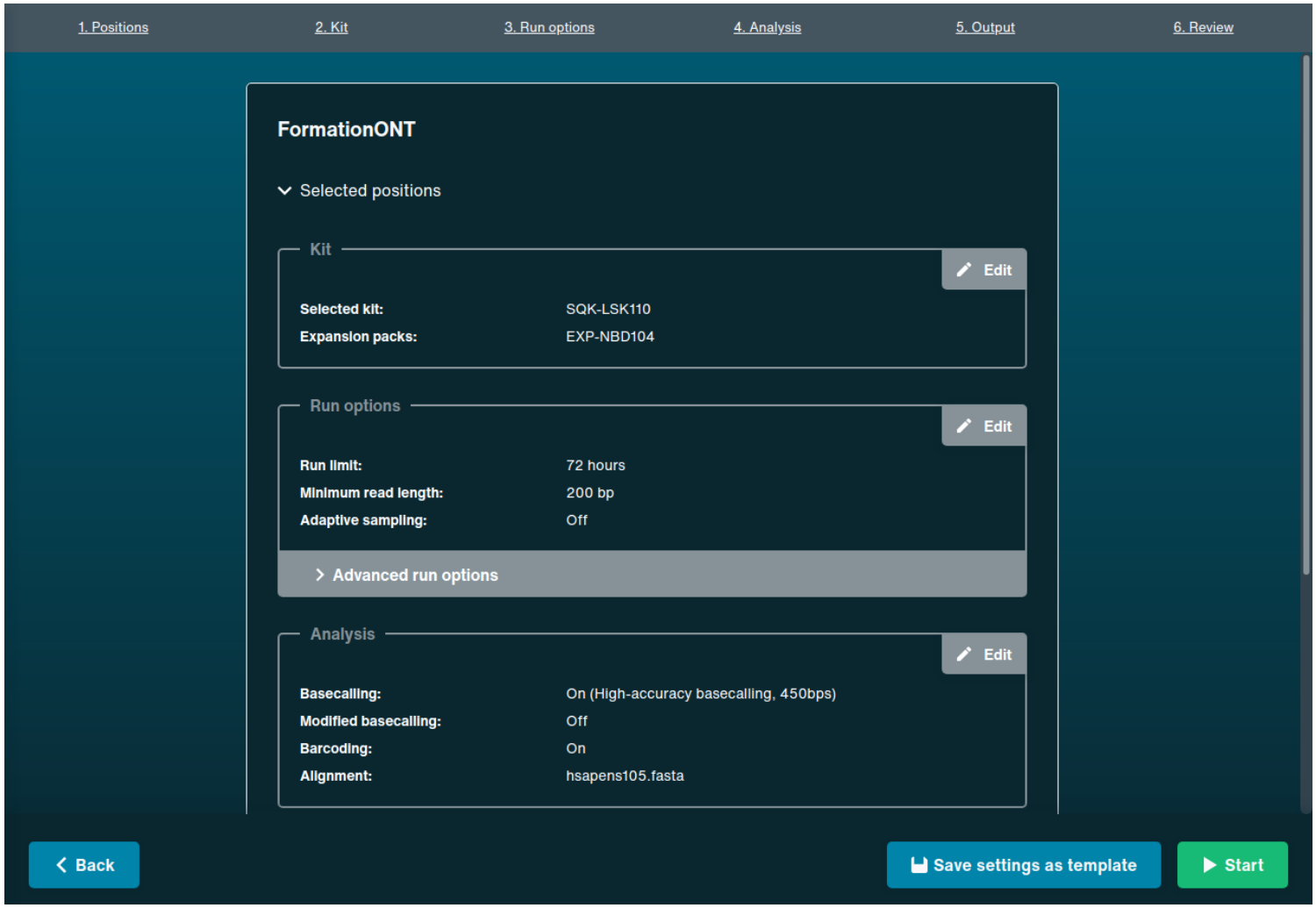
Les lectures adjacentes s'alignent à des positions non adjacentes
=>Le découpage de MinKNOW est justifié

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

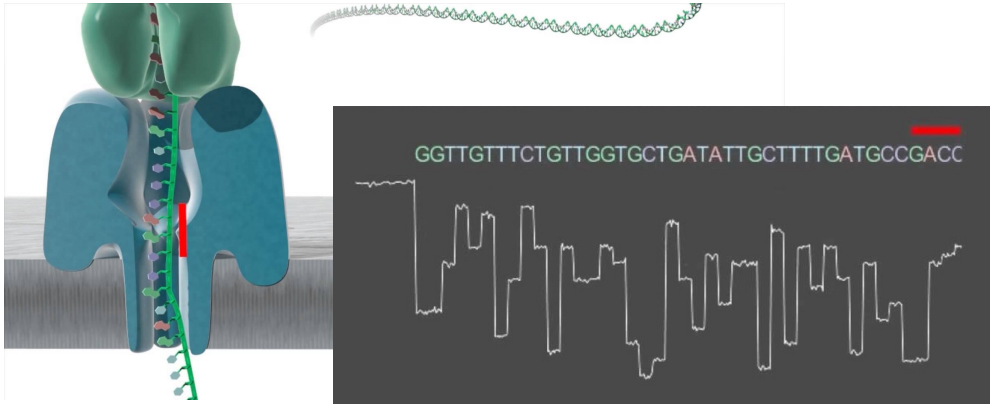
Récapitulatif avant lancement du run

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



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

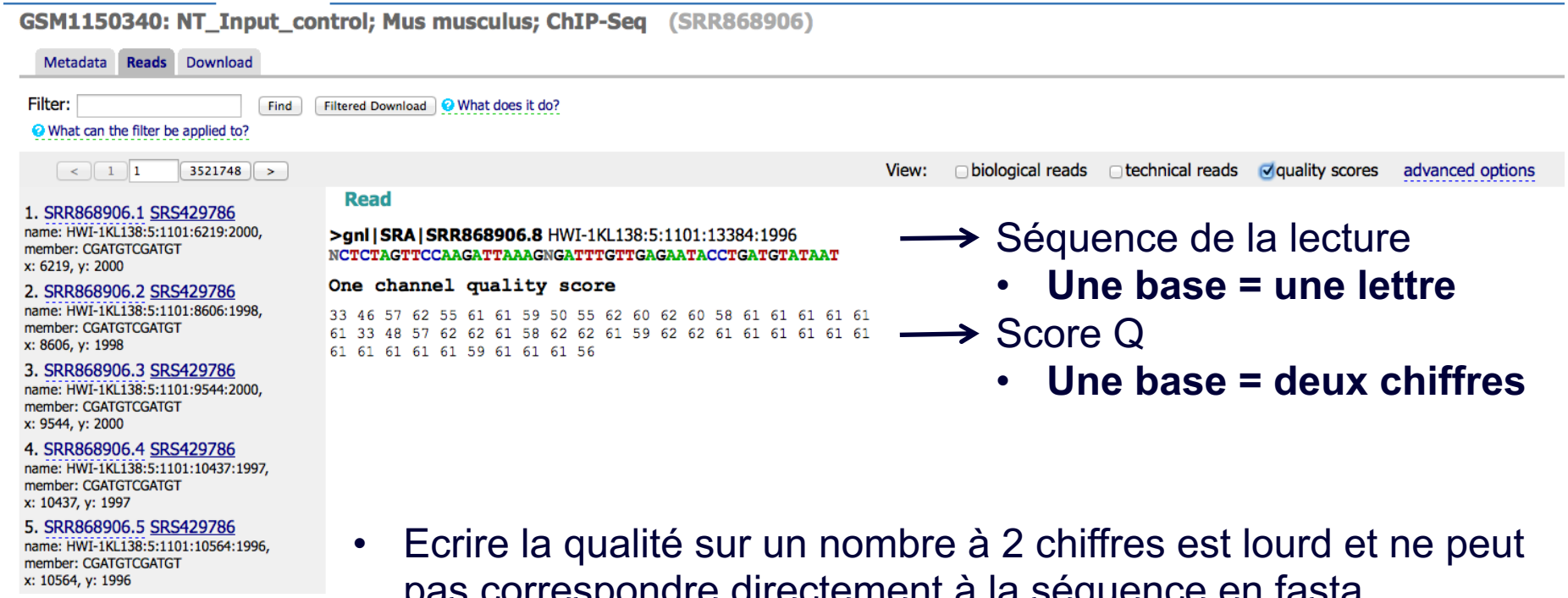
Probabilité d'identifier une base par erreur

Les 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 %

- 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 ?



→ 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 »

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]

Début de codage Phred = Phred+33

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,)*('%$$
```

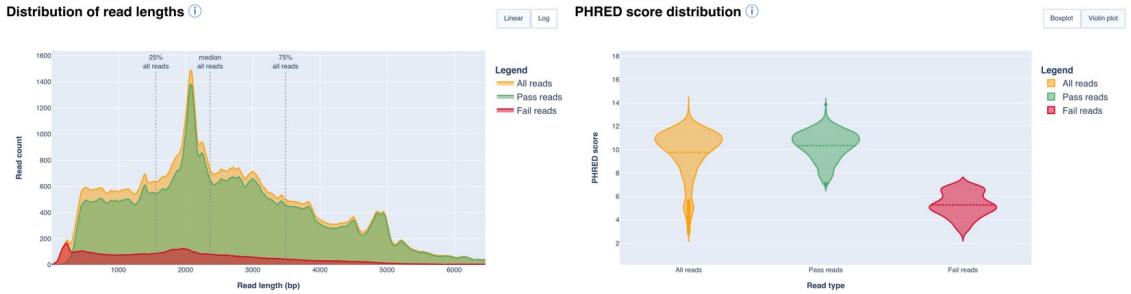
Les outils pour manipuler les fichiers FASTQ

- **Les outils de QC**
  - **Nanopack** : https://github.com/wdecoster/nanopack
  - **ToulligQC** : https://github.com/GenomiqueENS/toulligQC

ToulligQC pypi package 2.6 downloads 45k python 3.8 License GPL v3 Follow @Genomique_ENS

ToulligQC is dedicated to the QC analyses of Oxford Nanopore runs. This software is written in Python and developped by the GenomiqueENS core facility of the Institute of Biology of the Ecole Normale Superieure (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 ne 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

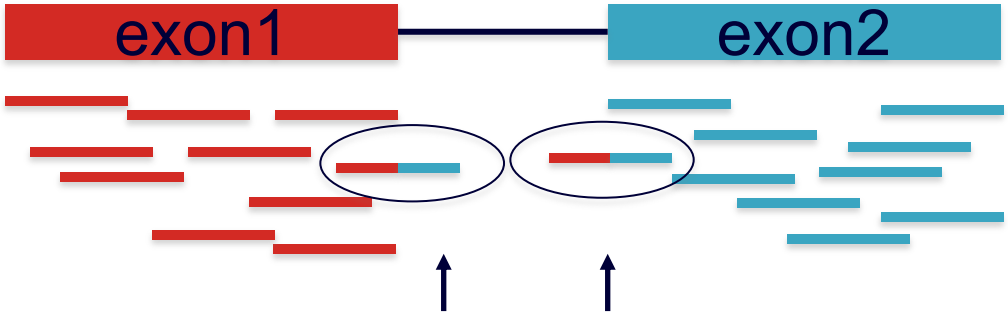
L'alignement est idéalement fait sur le génome

- 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 :

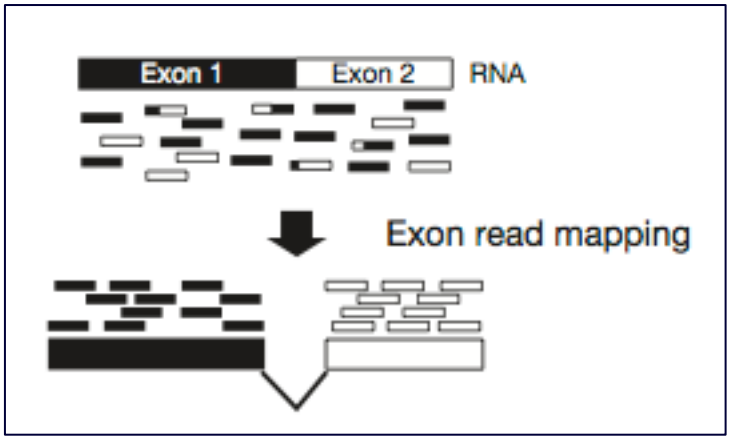


Alignement au niveau de la jonction exon-exon problématique car pas de gaps possibles

BWA

Aligneur développé pour l'alignement de séquences génomiques

- **Pas d'alignement sur les jonctions prévu**



Nat Methods 2011 Garber

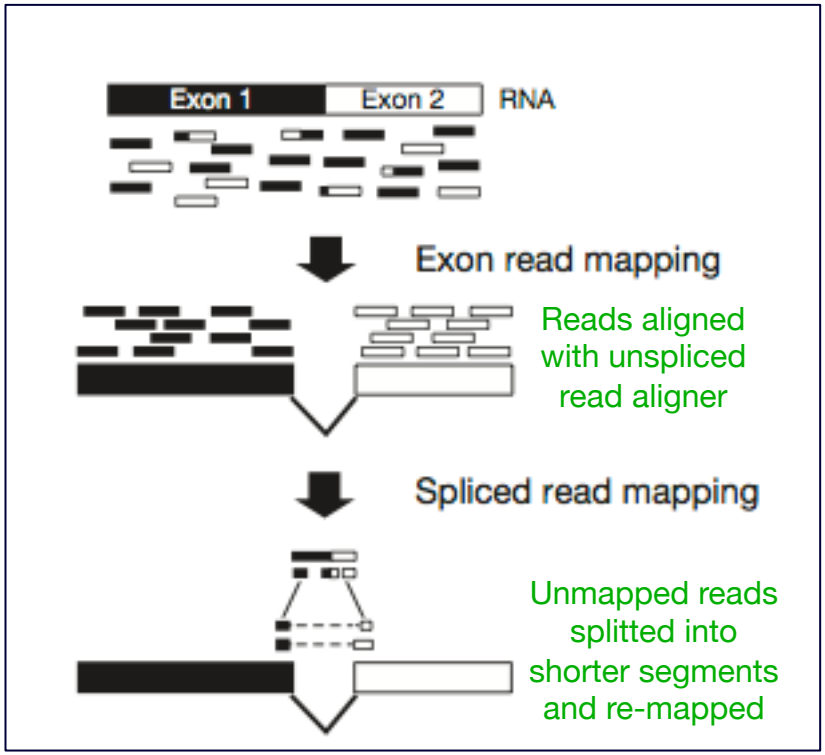
S'ils sont utilisés en RNASeq

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

TopHat, GSNAP…

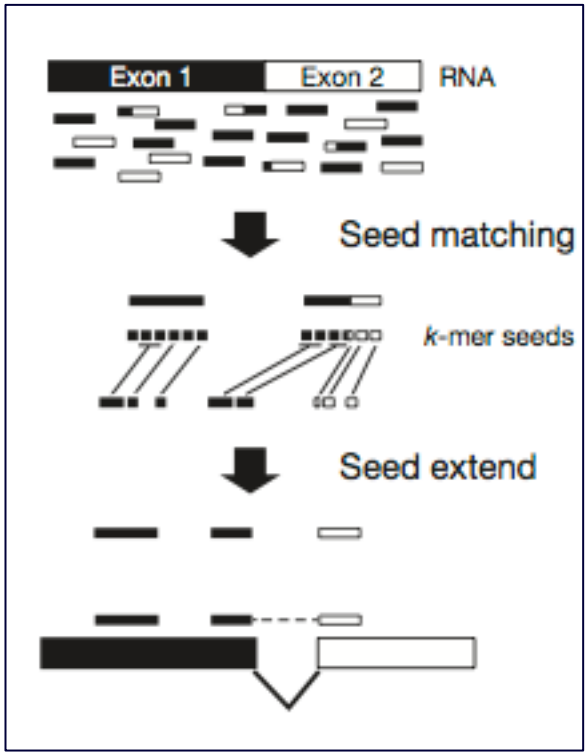
Aligneurs développés pour l'alignement d'ARNm

- **Les jonctions sont prises en comptes**



Stratégie TopHat : exon first

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



Stratégie GSNAP : 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

Alignement des cDNA :

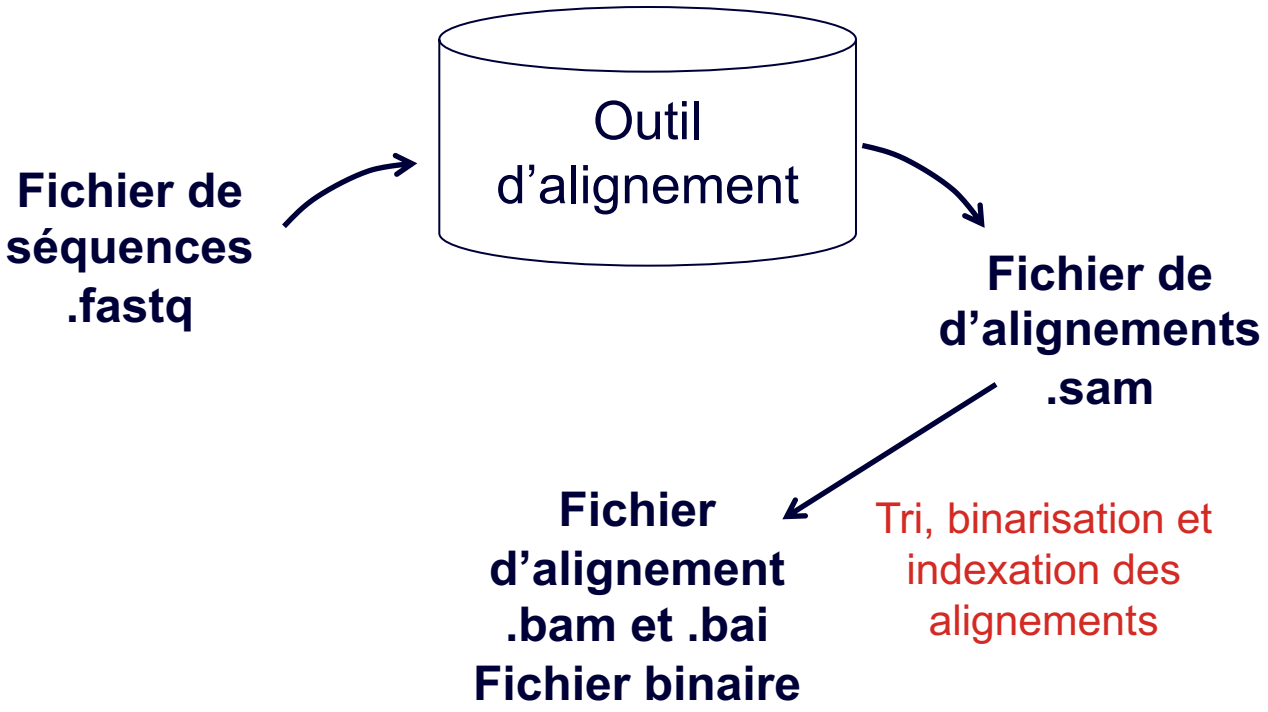
```
minimap2 -ax splice ref.fa nanopore-cdna.fa > aln.sam
```

Alignement des ARN :

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

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	152S16M1D33M3I8M2D19M1I24M1D7M1D45M2I1M1I66M1D7M4992N48M1602N37M1D79M2D8M3I25M2I35M4I36M1I4M1D49M1I5M2I4M1D67M3I1M1D2M1I5M1I13M2D17M1I28M1D9M1D64M2D5M1D26M1I12M117S
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 format BAM pour stocker les données non alignées (format Binary Alignment Map)

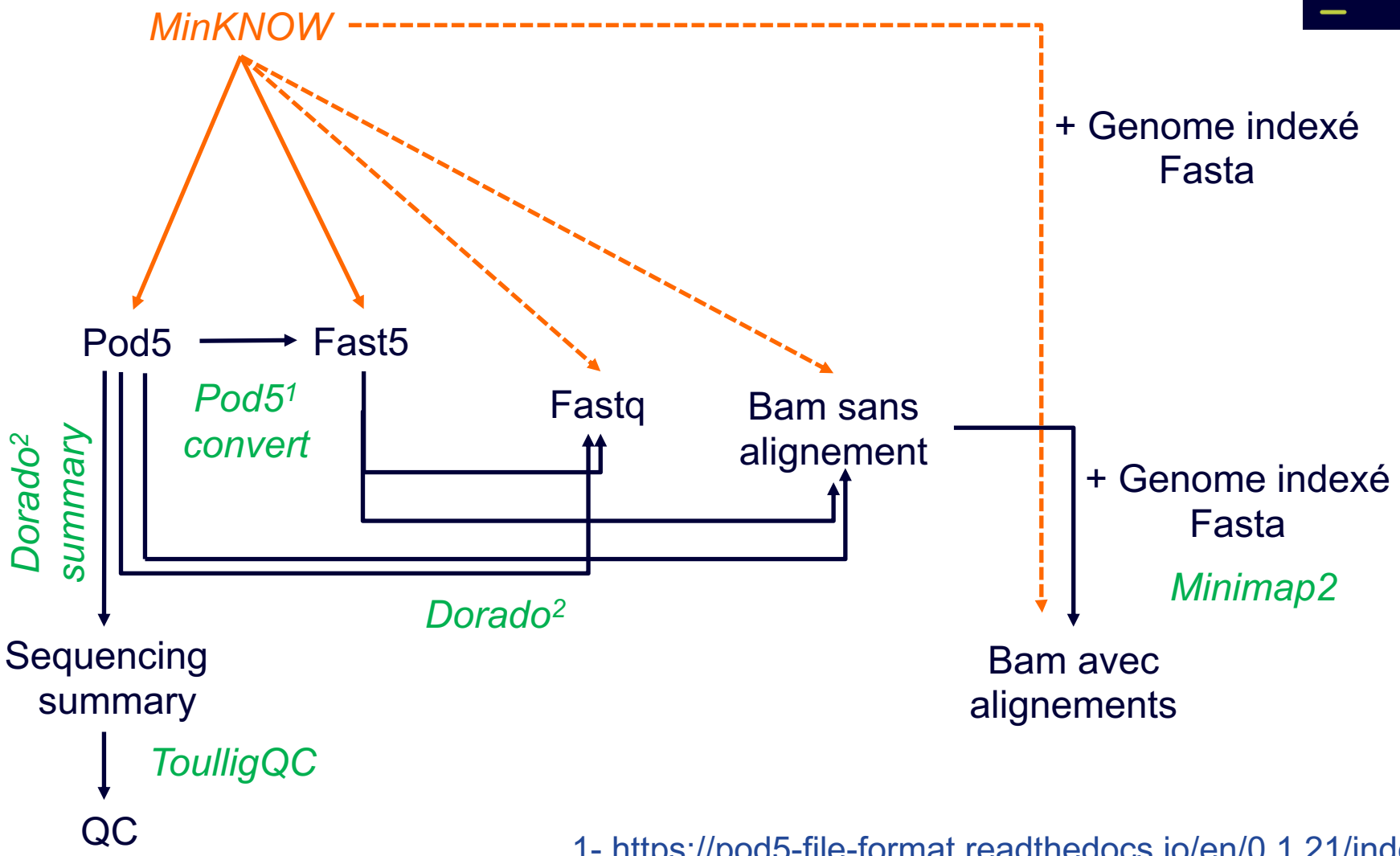
Les fichiers BAM « non alignés » peuvent être utilisés pour stocker les fichiers FASTQ.

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)

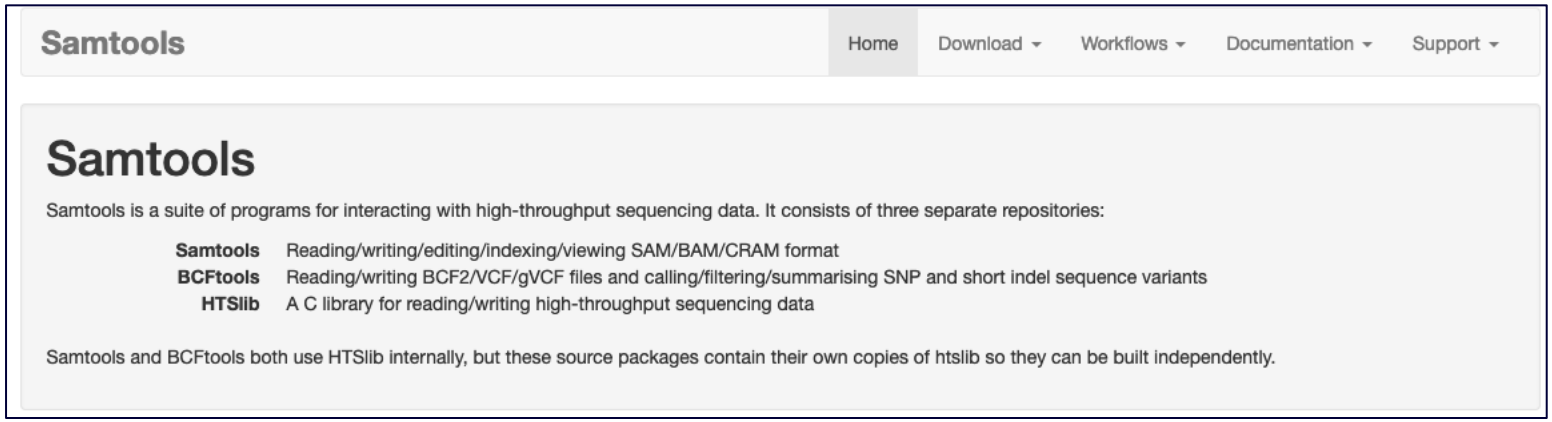
Avec et sans MinKNOW



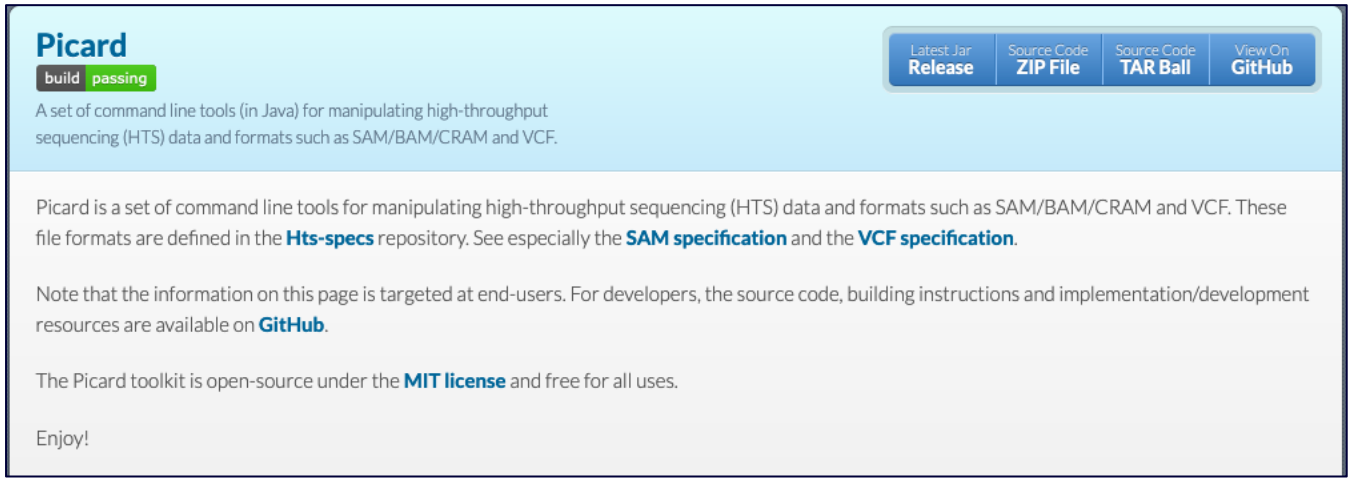
1- https://pod5-file-format.readthedocs.io/en/0.1.21/index.html
2- https://github.com/nanoporetech/dorado

Les outils pour manipuler les fichiers SAM/BAM

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



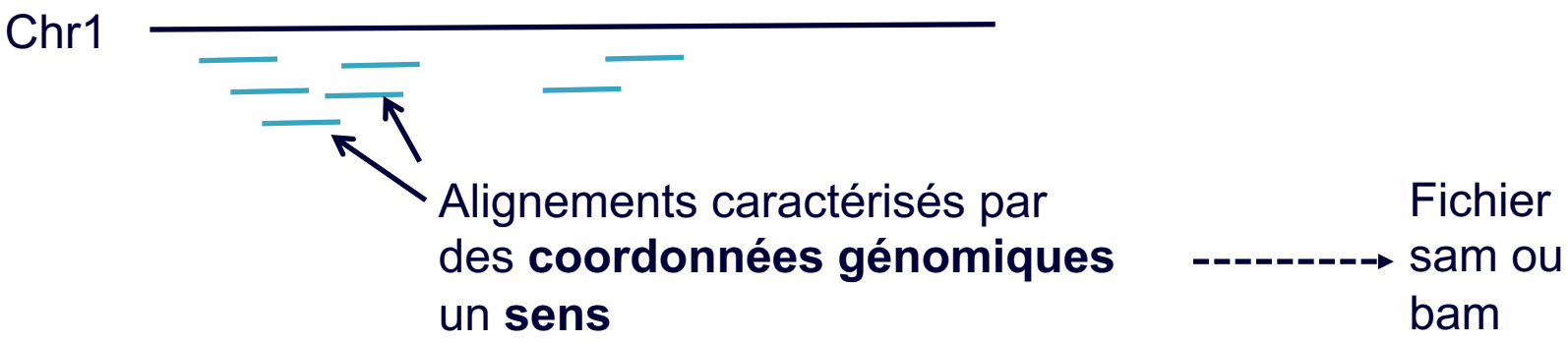
- Picard : https://broadinstitute.github.io/picard/



De l'alignement au nom de gène

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

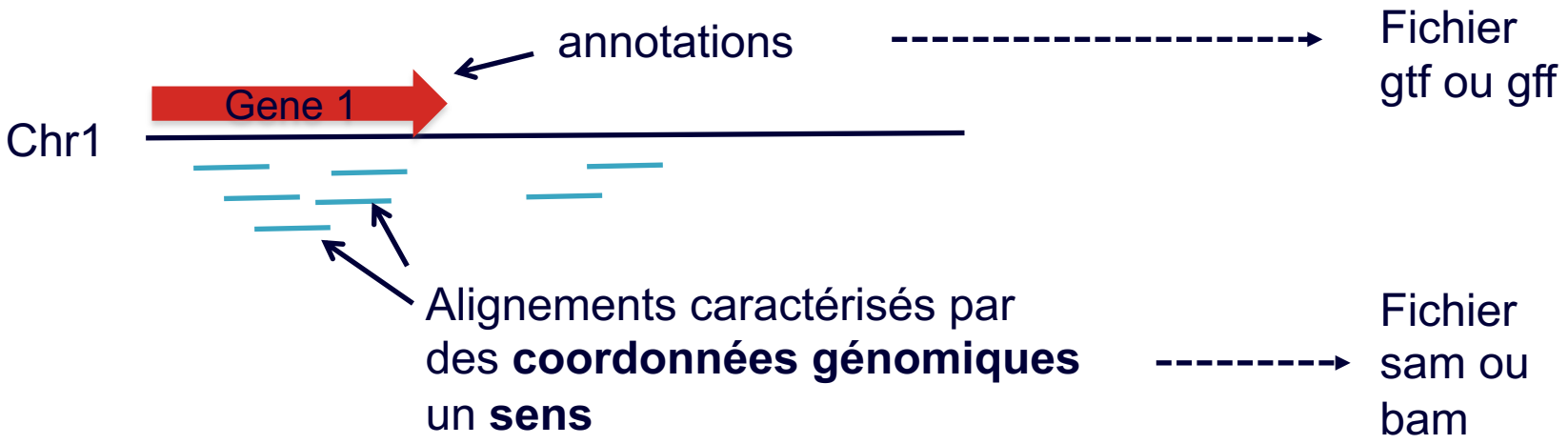
- **Il contient simplement des coordonnées**



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**

Ensembl

Génomes et transcriptomes

Annotations

Show 10 entries | Show/hide columns | Filter

★	Species	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** *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM
Y	**Mouse** *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM
Y	**Zebrafish** *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	-	-	BAM
	Alpaca *Vicugna pacos*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-
	Anole lizard *Anolis carolinensis*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	BAM
	Armadillo *Dasypus novemcinctus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	BAM
	Bushbaby *Otolemur garnettii*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-
	C.intestinalis *Ciona intestinalis*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-
	C.savignyi *Ciona savignyi*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-
	Caenorhabditis elegans *Caenorhabditis elegans*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-

Showing 1 to 10 of 66 entries | << < 1 2 3 4 5 > >>

Description du format gtf

Coordonnées de l'élément sur le génome

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	.	-	.

Le seqid doit correspondre au >seqid du fichier fasta correspondant

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	"OTTMUSG00000026353";
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

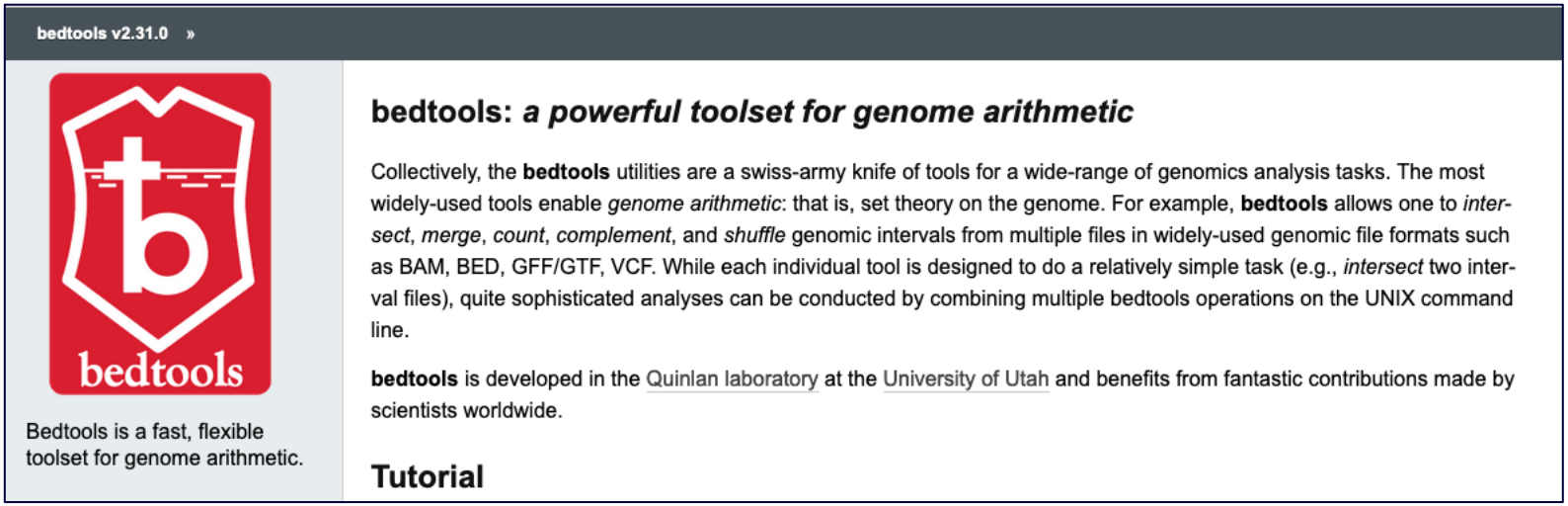
Les fichiers d'annotation au format BED (Browser extensible Data)

- 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

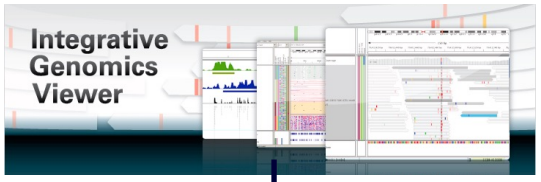


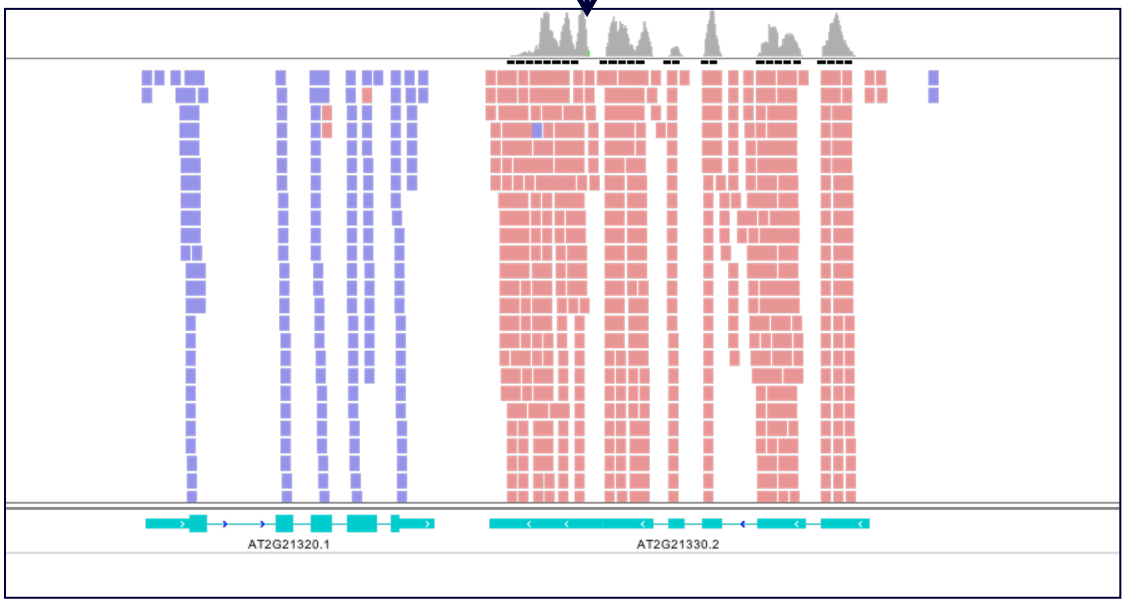
Visualiser ses données
Integrative Genomics Viewer (IGV)

Annotations.gtf ou .bed

Genome.fasta

Fichiers .bam et .bai





couverture

alignements

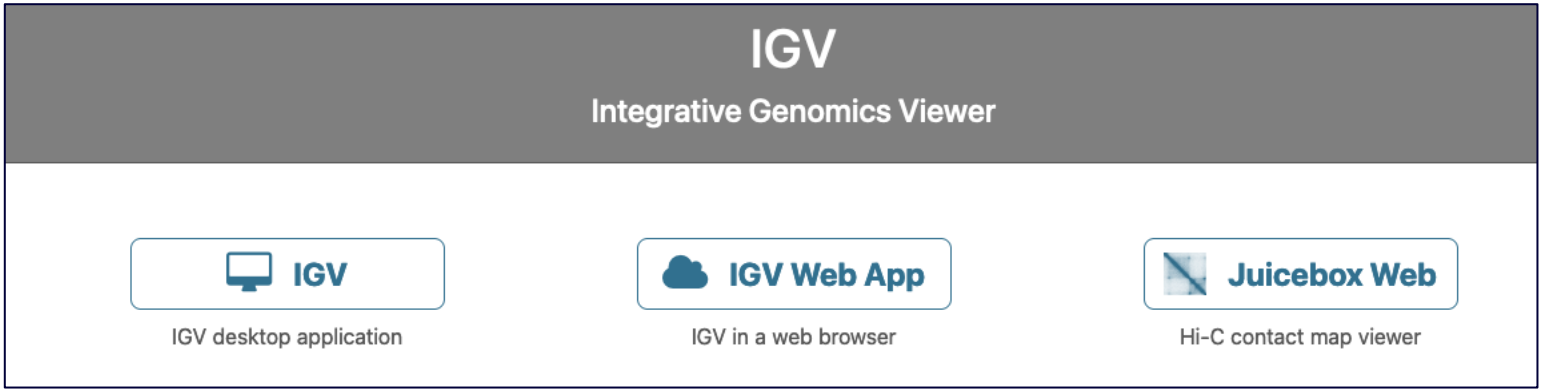
Pistes

Séquences et annotations

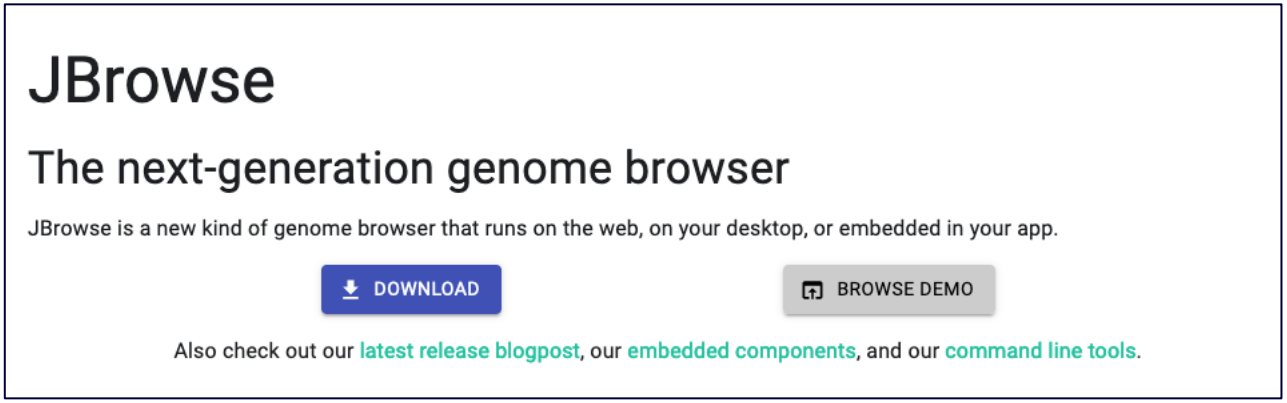
http://www.broadinstitute.org/igv/home

Les outils pour visualiser les données après alignement

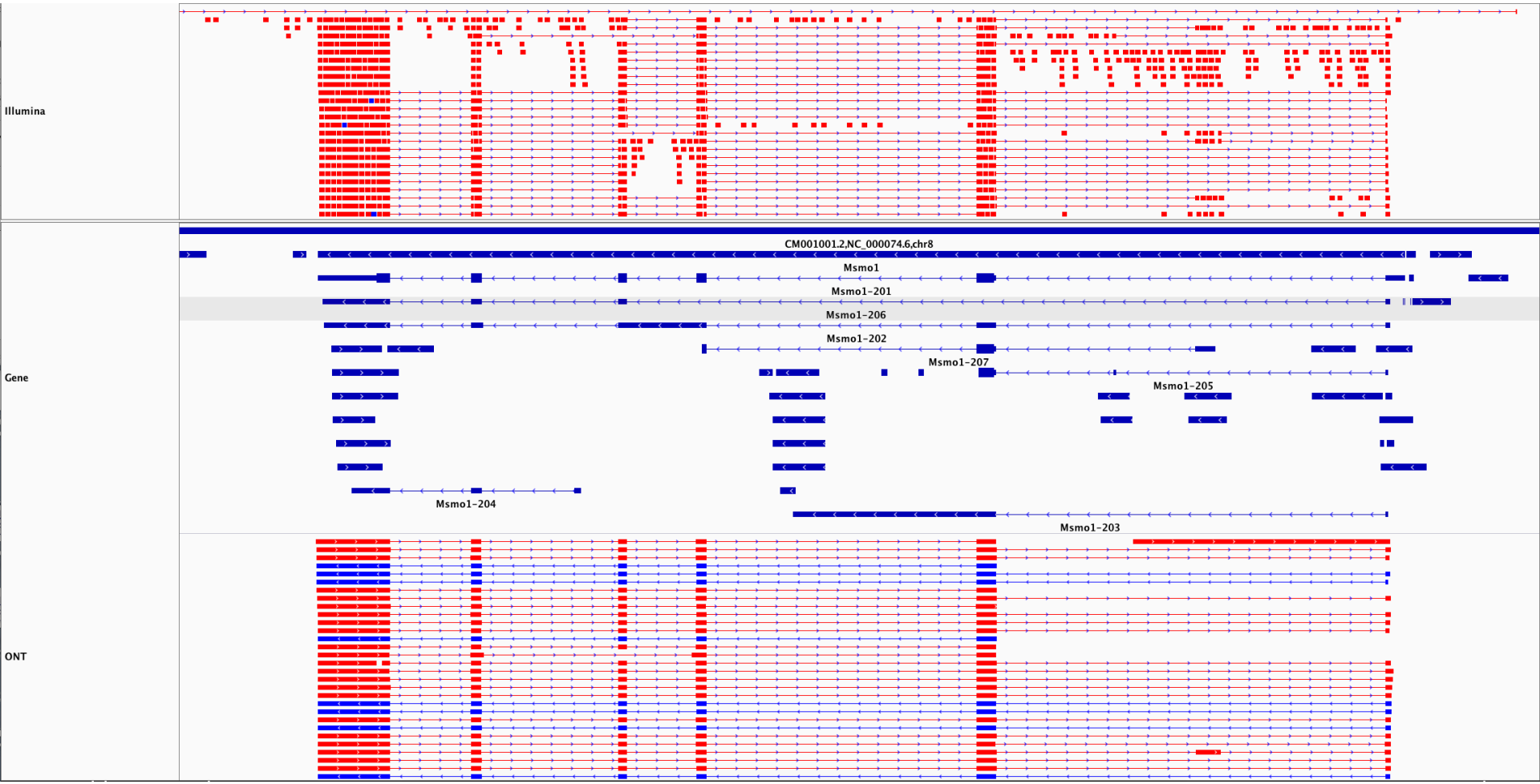
- **Integrative Genome Viewer** : https://igv.org/



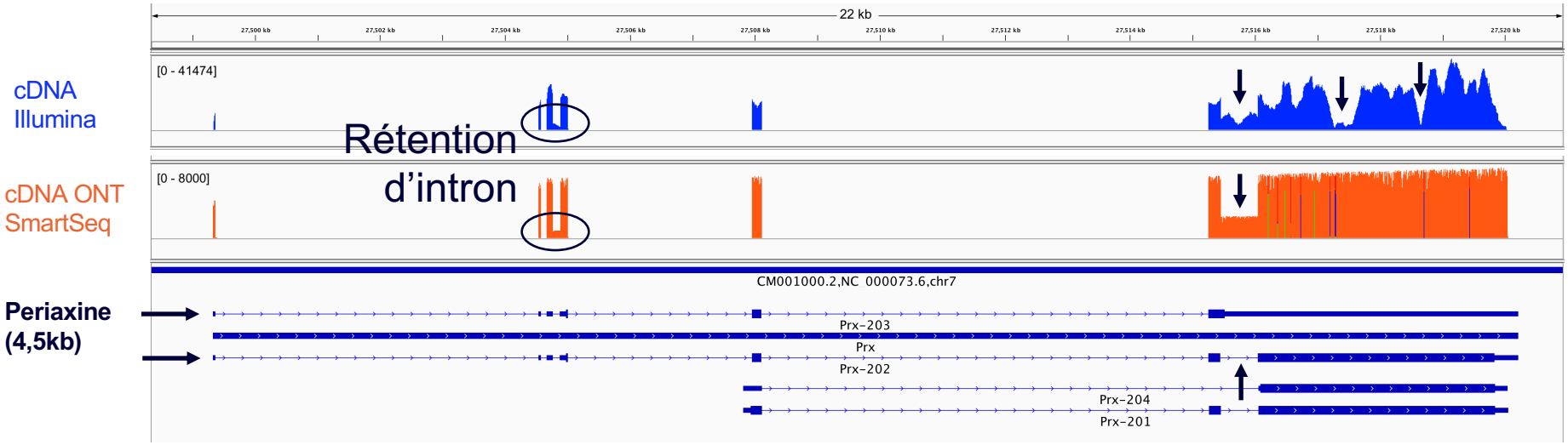
- **JBrowse2** : https://jbrowse.org/jb2/



Le séquençage de transcrits sur nanopores
Exemples de lectures alignées Illumina / ONT



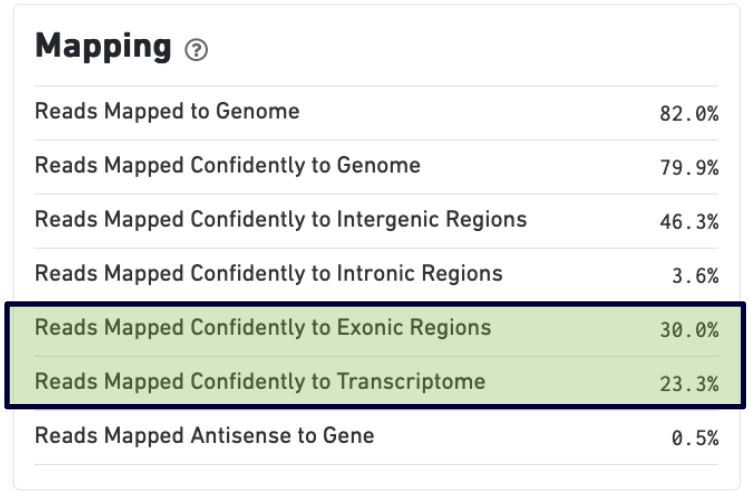
Le séquençage ONT permet une couverture homogène le long du transcrit



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

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 ?

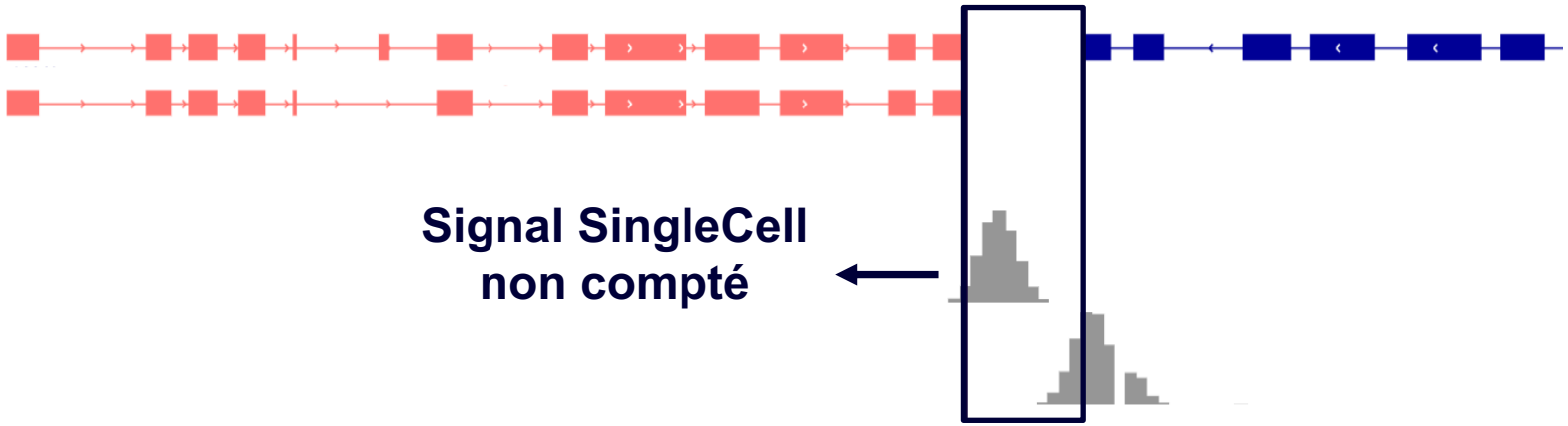
Le séquençage ONT et ré-annotation du génome
Application en SingleCell

Mapping

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…

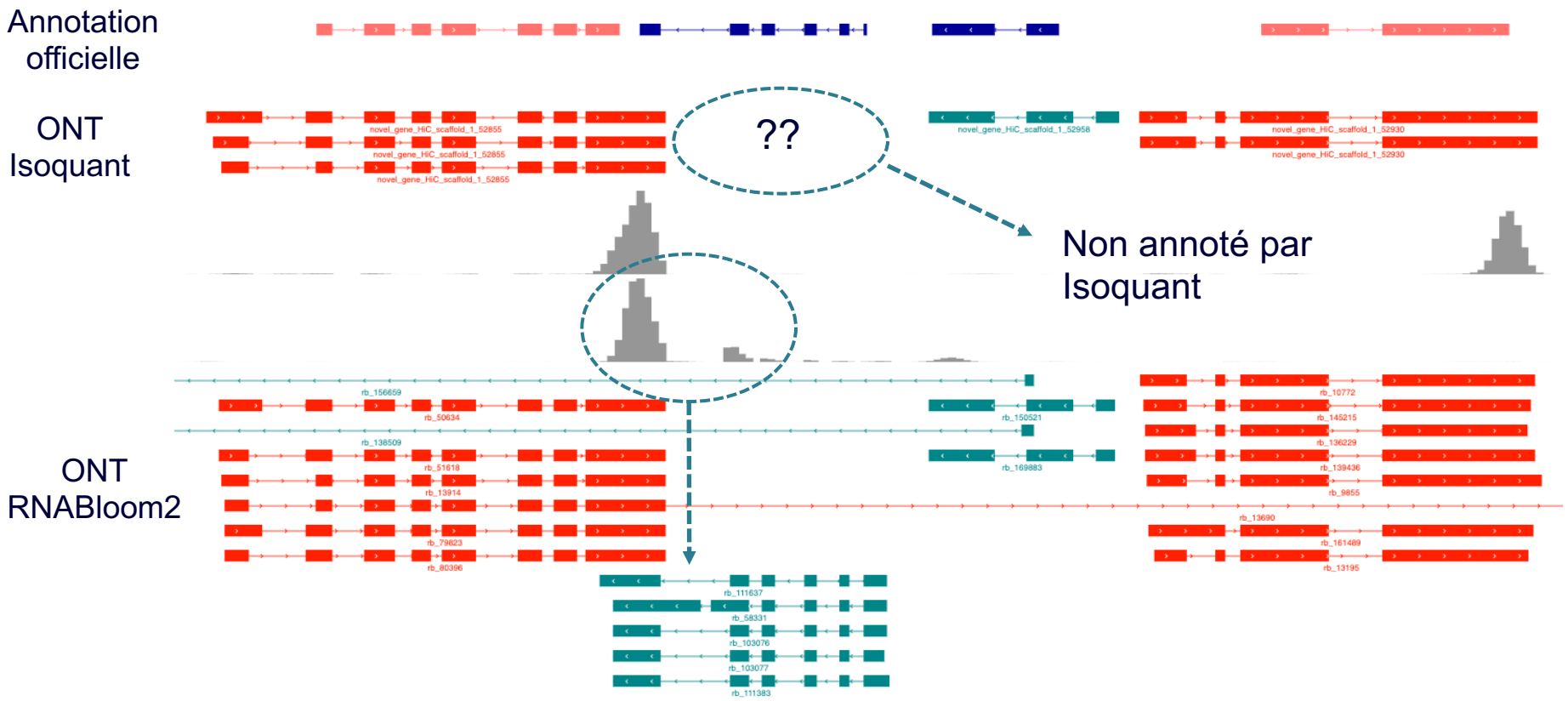
Annotation officielle



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



- **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