Control qualité des données brutes, nettoyage des données Manipulation des fichiers FASTQ

Stéphanie Le Gras

DU Dijon

Objectifs

- Comprendre ce que sont les données brutes de séquençage haut débit (type Illumina)
- Comprendre comment elles sont obtenues
- Comprendre d'où peuvent provenir les biais du Séquençage Haut débit (SHD)
- Apprendre à préparer les données de SHD pour l'analyse secondaire des données
  - Vérifier la qualité des données et si nécessaire les nettoyer (enlever ce qui pourrait bruiter le signal i.e générer la détection de faux variants)

Plan

- Introduction
  - Rappel : séquençage
  - Exemple de contrôles qualités du séquençage
- Données brutes : Le format FastQ
- Qualité des données brutes
- Nettoyage des données brutes

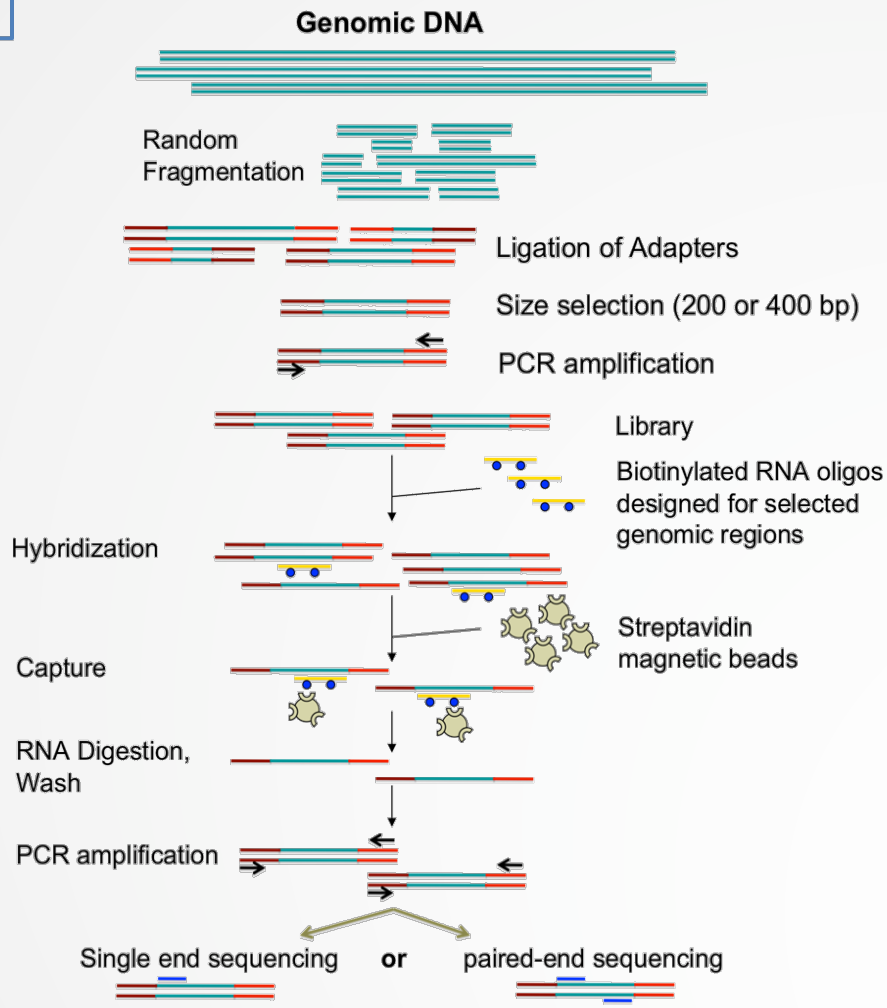
Rappel : Sequençage

Séquençage haut débit

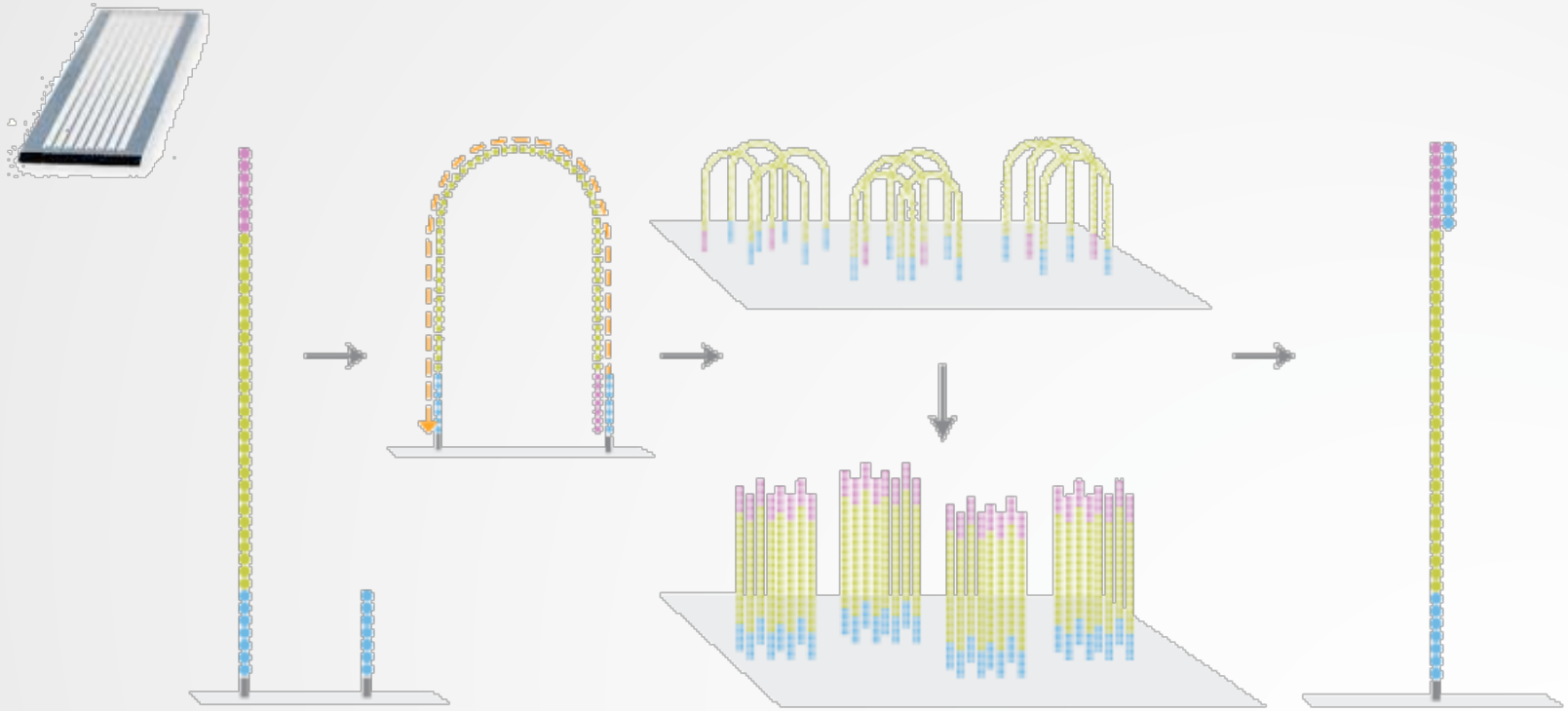
- 3 étapes principales
  - Préparation des librairies
  - Génération des clusters
  - Séquençage
- Analyse primaire

Préparation des librairies

Experiment

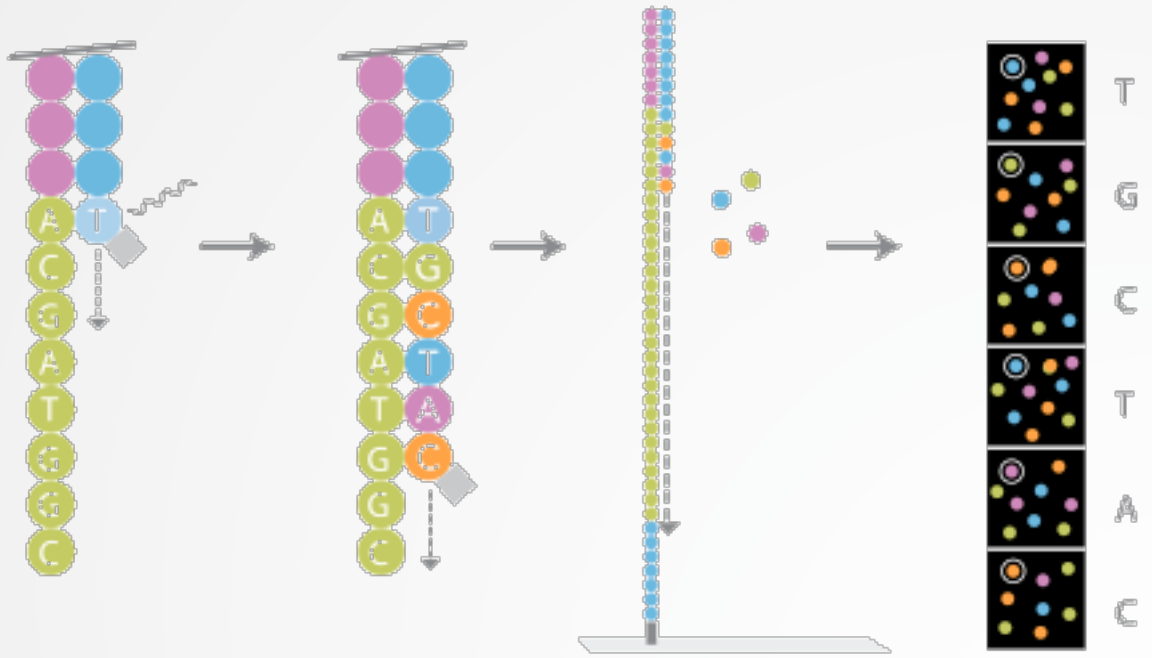


Génération des clusters



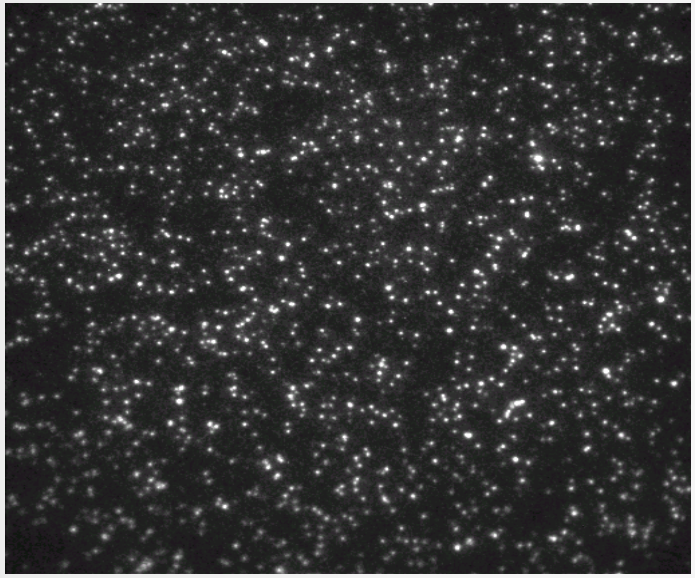
- Un cluster : ~1000 fois la même séquence d'ADN
- Nécessaire pour détecter la fluorescence pendant le séquençage

Séquençage



- Séquençage Illumina : Séquençage massivement parallèle

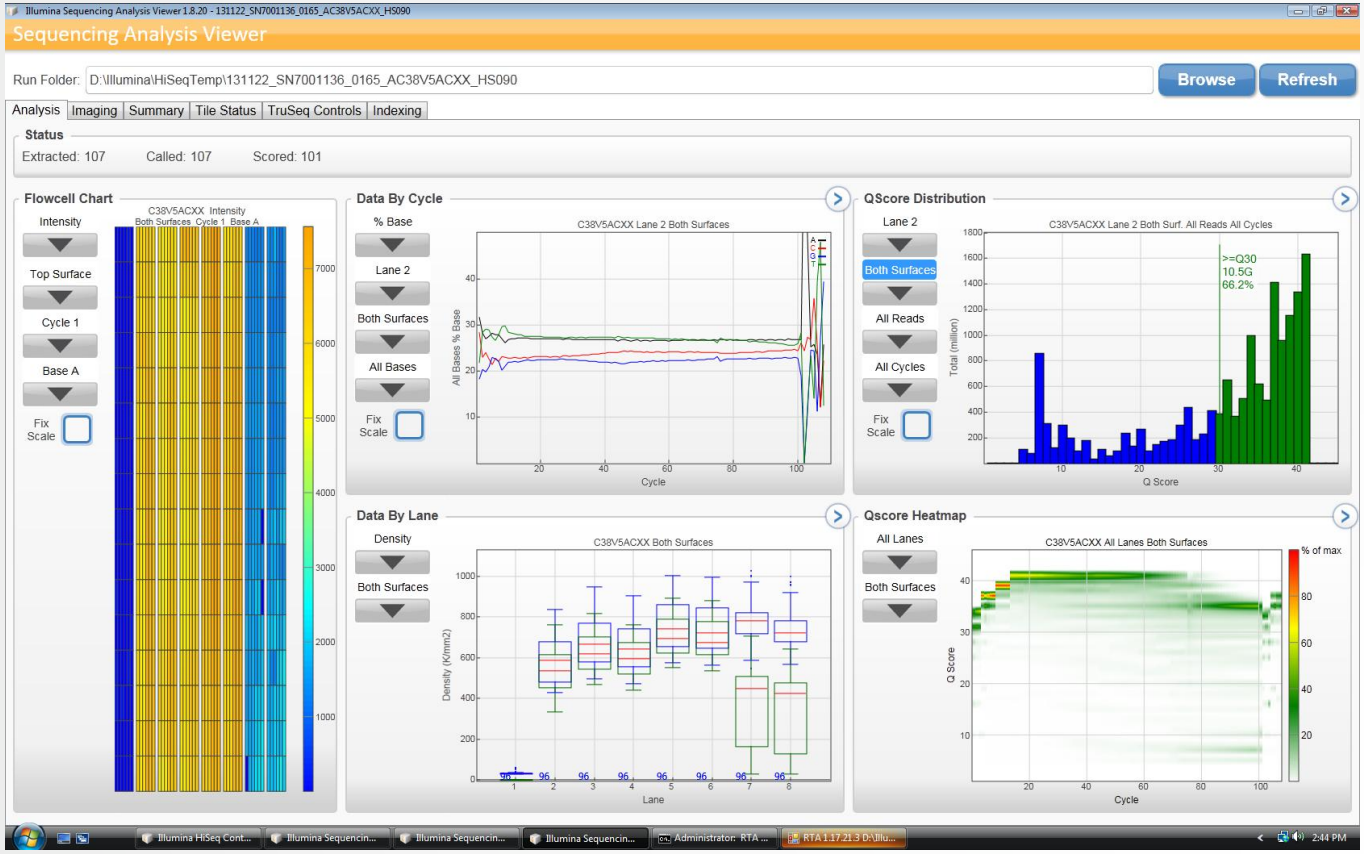
Analyse primaire



- Pipeline Illumina
  - Analyse d'image (extraction des intensités)
  - Appel de base
    - Identification des nucléotides
    - Calcul d'un score de qualité relatif à la probabilité d'erreur du nucléotide (0 <= Q <= 41)

QC pendant le séquençage

- L'analyse primaire est réalisée pendant le séquençage. On peut donc suivre en temps réel les statistiques du séquençage



QC pendant le séquençage

20/03/2017

Illumina Sequencing Analysis Viewer 1.8.20 - 131122_SN7001136_0165_AC38V5ACXX_HS090

Sequencing Analysis Viewer

Run Folder: D:\Illumina\HiSeqTemp\131122_SN7001136_0165_AC38V5ACXX_HS090 Browse Refresh

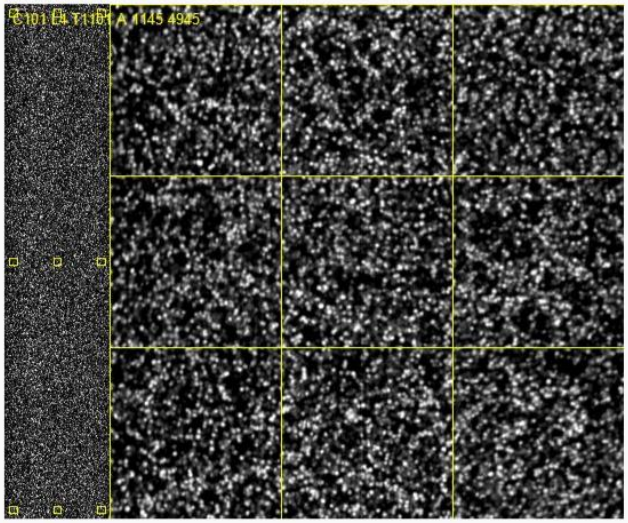
Analysis | Imaging | Summary | Tile Status | TruSeq Controls | Indexing

Cycle All | Lane All | Surface Both | Swath All | Section All

Index	Lane	Tile	Section	Cycle	Surface	Swath	Time	P90 A	P90 C	P90 G	P90 T	Error Rate	Corrected %A	Corrected %C
410	4	1101	1	86	Top	Left	11/26/201...	5174	8685	2508	6396		27.2	22.9
411	4	1101	1	87	Top	Left	11/26/201...	5165	8721	2435	6338		27.0	23.0
412	4	1101	1	88	Top	Left	11/26/201...	5169	8809	2498	6447		27.2	22.9
413	4	1101	1	89	Top	Left	11/26/201...	5096	8634	2418	6262		27.1	22.9
414	4	1101	1	90	Top	Left	11/26/201...	5085	8707	2377	6235		27.1	23.0
415	4	1101	1	91	Top	Left	11/26/201...	4995	8555	2370	6274		27.2	23.1
416	4	1101	1	92	Top	Left	11/26/201...	5022	8592	2357	6183		27.2	23.0
417	4	1101	1	93	Top	Left	11/26/201...	4985	8522	2371	6184		27.1	23.0
418	4	1101	1	94	Top	Left	11/26/201...	4913	8405	2328	6047		27.2	23.0
419	4	1101	1	95	Top	Left	11/26/201...	4935	8423	2379	6102		27.2	23.0
420	4	1101	1	96	Top	Left	11/26/201...	4899	8427	2465	6094		27.2	23.0
421	4	1101	1	97	Top	Left	11/26/201...	4879	8368	2354	6066		27.3	23.0
422	4	1101	1	98	Top	Left	11/26/201...	4828	8256	2331	6049		27.3	22.9
423	4	1101	1	99	Top	Left	11/26/201...	4774	8225	2324	5880		27.3	23.0
424	4	1101	1	100	Top	Left	11/27/201...	4768	8168	2313	5837		27.3	23.0
425	4	1101	1	101	Top	Left	11/27/201...	4636	8055	2254	5899		27.4	23.0
426	4	1101	1	102	Top	Left	11/27/201...	7494	10942	264	374		67.0	32.8
427	4	1101	1	103	Top	Left	11/27/201...	7272	10753	4021	7263		41.9	11.4
428	4	1101	1	104	Top	Left	11/27/201...	5049	9823	3543	4977		11.6	57.6
429	4	1101	1	105	Top	Left	11/27/201...	6638	9866	3277	6659		41.2	14.0
430	4	1101	1	106	Top	Left	11/27/201...	6465	9802	4007	2488		32.0	26.4
431	4	1101	1	107	Top	Left	11/27/201...	5710	9191	3140	5768		44.6	27.2
432	4	1101	1	108	Top	Left	11/27/201...	5195	9298	3220	7318		15.0	31.8
433	5	1101	1	1	Top	Left	11/23/201...	6843	9973	2229	7059		29.7	23.1
434	5	1101	1	2	Top	Left	11/23/201...	6520	9507	2015	6561		31.2	17.9
435	5	1101	1	3	Top	Left	11/23/201...	6587	9688	2624	6949		31.1	19.8
436	5	1101	1	4	Top	Left	11/23/201...	6533	9645	2922	7124		29.6	20.8

Rows=82844 | Disp=82844 | Sel=1 | Filter

A C G T



QC pendant le séquençage

20/03/2017

Illumina Sequencing Analysis Viewer 1.8.20 - 131122_SN7001136_0165_AC38V5ACXX_HS090

Sequencing Analysis Viewer

Run Folder: D:\Illumina\HiSeqTemp\131122_SN7001136_0165_AC38V5ACXX_HS090 Browse Refresh

Analysis | Imaging | Summary | Tile Status | TruSeq Controls | Indexing

Total	113.2	220.0	0.0	0.0	0.00	0.0	0.0	0.00	4290	106.0	84.6

Read 1

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
2	96	594 +/- 110	91.84 +/- 8.36	0.282 / 0.778	164.20	149.94	64.7	15.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	6335 +/- 494	79.3 +/- 10.6
3	96	679 +/- 116	92.21 +/- 2.24	0.228 / 0.198	187.69	172.40	94.1	17.2	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	6042 +/- 516	92.7 +/- 6.3
4	96	656 +/- 109	91.86 +/- 2.44	0.219 / 0.196	181.39	165.95	94.1	16.6	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	5884 +/- 485	94.5 +/- 6.0
5	96	756 +/- 117	93.19 +/- 2.18	0.212 / 0.195	208.89	194.05	94.6	19.4	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	6942 +/- 376	82.3 +/- 1.3
6	96	744 +/- 118	92.91 +/- 2.06	0.215 / 0.194	205.77	190.58	94.5	19.1	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	6617 +/- 446	85.6 +/- 4.2
7	96	779 +/- 91	48.66 +/- 24.60	0.498 / 0.148	215.32	103.72	66.2	10.4	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	1593 +/- 381	160.8 +/- 25.9
8	96	740 +/- 87	45.29 +/- 25.83	0.533 / 0.172	204.55	91.34	63.9	9.1	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	1641 +/- 332	161.8 +/- 33.0

Read 2 (I)

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
2	96	594 +/- 110	91.84 +/- 8.36	0.000 / 0.000	164.20	149.94	96.7	0.9	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	7159 +/- 357	0.0 +/- 0.0
3	96	679 +/- 116	92.21 +/- 2.24	0.000 / 0.000	187.69	172.40	91.4	1.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	7228 +/- 327	0.0 +/- 0.0
4	96	656 +/- 109	91.86 +/- 2.44	0.000 / 0.000	181.39	165.95	91.5	1.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	7289 +/- 373	0.0 +/- 0.0
5	96	756 +/- 117	93.19 +/- 2.18	0.000 / 0.000	208.89	194.05	69.9	1.2	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	1783 +/- 71	0.0 +/- 0.0
6	96	744 +/- 118	92.91 +/- 2.06	0.000 / 0.000	205.77	190.58	76.9	1.1	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	1821 +/- 226	0.0 +/- 0.0
7	96	779 +/- 91	48.66 +/- 24.60	0.000 / 0.000	215.32	103.72	54.6	0.6	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	4104 +/- 361	0.0 +/- 0.0
8	96	740 +/- 87	45.29 +/- 25.83	0.000 / 0.000	204.55	91.34	57.0	0.5	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	4122 +/- 369	0.0 +/- 0.0

Read 3

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
2	96	594 +/- 110	91.84 +/- 8.36	0.000 / 0.000	164.20	149.94	NaN	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0 +/- 0	0.0 +/- 0.0
3	96	679 +/- 116	92.21 +/- 2.24	0.000 / 0.000	187.69	172.40	NaN	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0 +/- 0	0.0 +/- 0.0
4	96	656 +/- 109	91.86 +/- 2.44	0.000 / 0.000	181.39	165.95	NaN	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0 +/- 0	0.0 +/- 0.0
5	96	756 +/- 117	93.19 +/- 2.18	0.000 / 0.000	208.89	194.05	NaN	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0 +/- 0	0.0 +/- 0.0
6	96	744 +/- 118	92.91 +/- 2.06	0.000 / 0.000	205.77	190.58	NaN	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0 +/- 0	0.0 +/- 0.0
7	96	779 +/- 91	48.66 +/- 24.60	0.000 / 0.000	215.32	103.72	NaN	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0 +/- 0	0.0 +/- 0.0
8	96	740 +/- 87	45.29 +/- 25.83	0.000 / 0.000	204.55	91.34	NaN	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0 +/- 0	0.0 +/- 0.0

Copy to Clipboard... Generate IVC Plots...

Comment obtenir des données de SHD

- En produisant vos propres données de séquençage
  - Centre National de Séquençage/Génotypage
  - Plateforme technologique
  - Compagnie privée
- En utilisant des données publiques
  - SRA : NCBI Sequence Read Archive
  - ENA : EMBL/EBI European Nucleotide Archive

Données brutes :
Le format Fastq

Le format FastQ

- Extension *.fastq
- Fichier texte : peut être ouvert avec un simple éditeur de texte (! taille)
- Contient des séquences nucléotidiques + valeurs de qualité (fasta + Qualité)
- Aucune information relative à un génome

```
1  @HWI-ST1136:117:HS055:3:1101:1134:2244 1:N:0:GCCAAT
2  GCCGCGCCGAGCCGGGCCCGTGGCCCGCCGGTCCCCGTCCCGGGGGTTGG
3  +
4  @CCFFFDFHHHHHJJIGIJJJGGICHEBB<@6?=BBB2<@DD6@BB5<@D
5  @HWI-ST1136:117:HS055:3:1101:1250:2246 1:N:0:GCCAAT
6  CCTCCAGCAACTTCCTGATGGTTCTGTGAGGCCGGGAACACAGTGTAGTT
7  +
8  CC@FFFDDFDHHHJJJJJIJIEHIEGGHHIIGIJGIEEHGHD@?FBGIHI
9  @HWI-ST1136:117:HS055:3:1101:1159:2247 1:N:0:GCCAAT
10 ATCAAACAAAAGCCAGAGAACCCACAACCGCATTGCATAGTCACAGGTGT
11 +
12 @@@FDDDEFGFFGIIGIJIICGGGHIGHIGHIGIHIJJGGJIIIIIECHG
13 @HWI-ST1136:117:HS055:3:1101:1432:2218 1:N:0:GCCAAT
14 CCAGNCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATCA
15 +
16 CCCF#2ADHHGDHIJIHIJIIIIIIGIGEHIGIIGGGHIEF;CGGCHGGI
```

Identifiant
Séquence
Qualité

Signification de l'identifiant

- @HWI-ST1136:117:HS055:3:1101:1134:2244 1:N:0:GCCAAT
  - HWI-ST1136 : Nom du séquenceur
  - 117 : identifiant du run
  - HS055 : identifiant de la flowcell
  - 3 : numéro de ligne
  - 1101 : numéro du tile
  - 1134 : coordonnée X
  - 2244 : coordonnée Y

  - 1 : Numéro de la paire (1 ou 2)
  - N : booléen indiquant le passage du filtre qualité
    - Y : La séquence est de mauvaise qualité
    - N : la séquence a passé le filtre de qualité
  - 0 : 0 lorsque aucun des bit contrôles n'est activé, sinon c'est un nombre
  - GCCAAT : Index de la librairie (en cas de multiplexage)

Exemple de données pairées

LCD-01_1_ATCACG_L007_R1_045.fastq.gz

```
@HWI-ST1136:163:HS087:7:2310:17264:70630 1:N:0:ATCACG
GTTAGAGGCCAAGGTACAGTGGCCTGTCTTTGTAAATGTGCCTTTATGT
+
CCCFFFFFHHHHHJFHHIJHIJIJJIIJJJJIHHIJIIGIIJIJJJJJI
@HWI-ST1136:163:HS087:7:2310:17415:70636 1:N:0:ATCACG
TGGAGCCTTGGTAACTTTTTGTAGTTTGTATGCGTTTTTGTTTGGTCTC
+
BCCFFFFFHHHHHIJJJJJJJJHIJJJJJIJJJJJIIJJJJJIJJJJFGIJ
@HWI-ST1136:163:HS087:7:2310:17337:70637 1:N:0:ATCACG
CTGTTACCCCTCCATTCAGGGGTATGAAGAAGGGCTTCACCTGTAGTTC
+
@CCFFFFFHHHHHJJIJJJJJJJBFHIJJIJJIIIIJIJIJJHHGHIJJJ
```

LCD-01_1_ATCACG_L007_R2_045.fastq.gz

```
@HWI-ST1136:163:HS087:7:2310:17264:70630 2:N:0:ATCACG
TAATTTTTTGCATCCTGAAAAGTGTTGGAAGTTGGGTTTTCATAGTCAA
+
CCCFFFFFHHHHHJJJIJJJJJICHGIIJIJJIIIJIJFHIIJHIJJJIJJ
@HWI-ST1136:163:HS087:7:2310:17415:70636 2:N:0:ATCACG
TGTTCATATGTATGAGATAGATTTGAAAAATCTACTAATTTTTAAAATC
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@HWI-ST1136:163:HS087:7:2310:17337:70637 2:N:0:ATCACG
TCCTGACATCAAGCACACTGCTTCTGCATCTATGTGGCACCTAAAACAA
+
CCCFFFFFHHHHHJJIIJJJJJJJJJFIIIJJJJJIJIJIIJIJJJJJJJ
```

- Convention : Les lectures sens 1 et sens 2 du même cluster sont à la même ligne entre les deux fichiers (R1 et R2)

L'encodage de la qualité

- Score de qualité = Score Phred
- Score de qualité donné par le séquenceur
- 1 symbole ASCII = 1 valeur de qualité
- ASCII : Norme de codage de caractère en informatique
- Score Phred (Sanger) : ASCII – 33
  - 0 <= p <= 41

- Score Phred = -10 log10 p
- p : probabilité d'avoir une erreur de séquençage

Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
32	20	040		Space	64	40	100	@	@
33	21	041	!	!	65	41	101	A	A
34	22	042	"	"	66	42	102	B	B
35	23	043	#	#	67	43	103	C	C
36	24	044	$	$	68	44	104	D	D
37	25	045	%	%	69	45	105	E	E
38	26	046	&	&	70	46	106	F	F
39	27	047	'	'	71	47	107	G	G
40	28	050	(	(	72	48	110	H	H
41	29	051	)	)	73	49	111	I	I
42	2A	052	*	*	74	4A	112	J	J
43	2B	053	+	+	75	4B	113	K	K
44	2C	054	,	,	76	4C	114	L	L
45	2D	055	-	-	77	4D	115	M	M
46	2E	056	.	.	78	4E	116	N	N
47	2F	057	/	/	79	4F	117	O	O
48	30	060	0	0	80	50	120	P	P
49	31	061	1	1	81	51	121	Q	Q
50	32	062	2	2	82	52	122	R	R
51	33	063	3	3	83	53	123	S	S
52	34	064	4	4	84	54	124	T	T
53	35	065	5	5	85	55	125	U	U
54	36	066	6	6	86	56	126	V	V
55	37	067	7	7	87	57	127	W	W
56	38	070	8	8	88	58	130	X	X
57	39	071	9	9	89	59	131	Y	Y
58	3A	072	:	:	90	5A	132	Z	Z
59	3B	073	;	;	91	5B	133	[	[
60	3C	074	<	<	92	5C	134	\	\
61	3D	075	=	=	93	5D	135	]	]
62	3E	076	>	>	94	5E	136	^	^
63	3F	077	?	?	95	5F	137	_	_

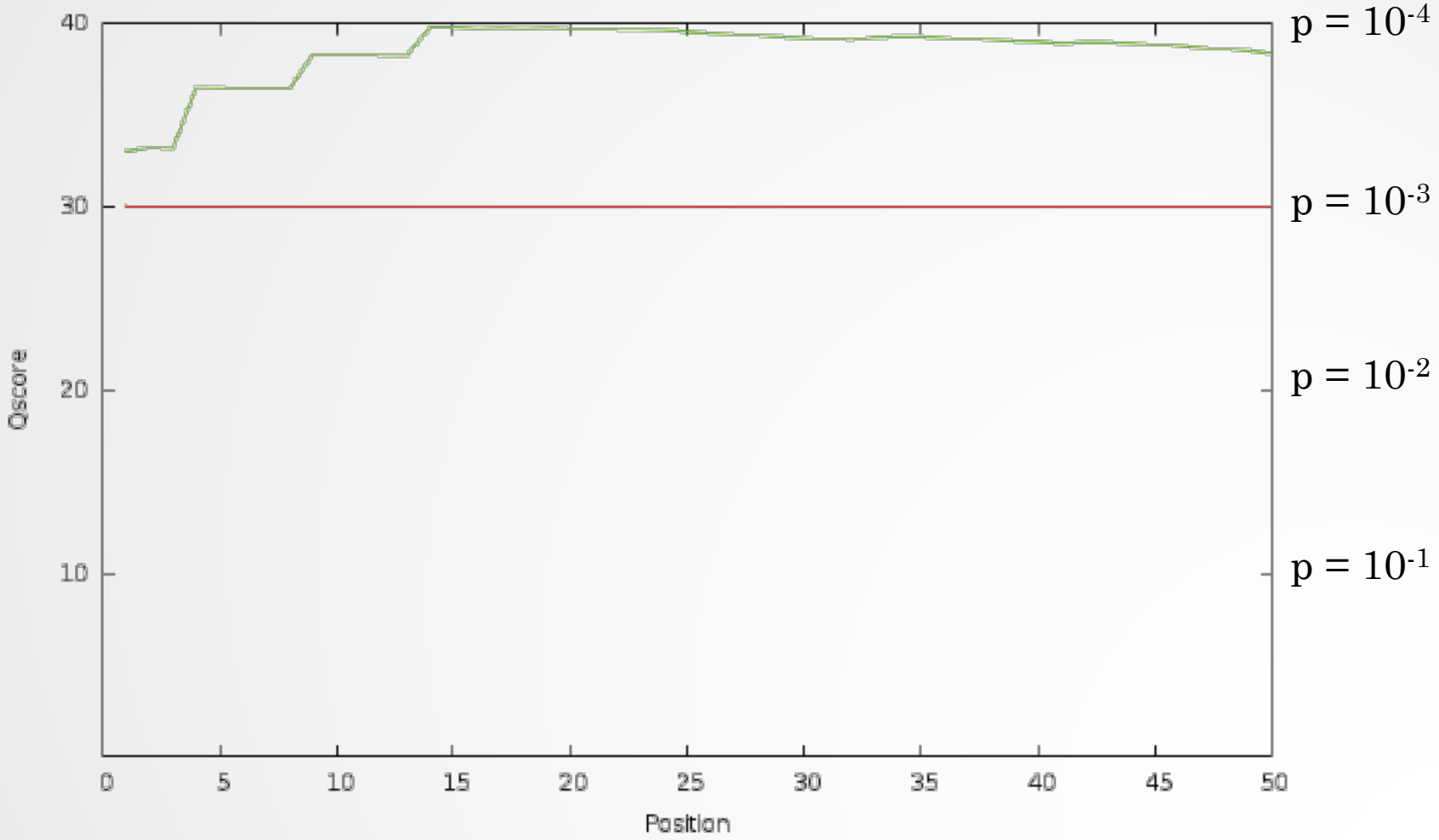
Exemple

```
@HWI-ST1136:117:HS055:3:1101:1134:2244 1:N:0:GCCAAT
GCCGCGCCGAGCCGGGCCCGTGGCCCGCCGGTCCCCGTCCCGGGGGTTGG
+
@CCFFFDFHHHHHJJIGIJJJGGICHEBB<@6?=BBB2<@DD6@BB5<@D
```

- 1er nucléotide : G
- Qualité associée : @
- Partie Pratique : Déterminez la valeur de qualité associée
  - Score Phred = 64 – 33 = 31
  - $-10 \log_{10} p = 31$
  - $p = 10^{(-31/10)} = 7{,}9 \times 10^{-3}$

Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
32	20	040		Space	64	40	100	@	@
33	21	041	!	!	65	41	101	A	A
34	22	042	"	"	66	42	102	B	B
35	23	043	#	#	67	43	103	C	C
36	24	044	$	$	68	44	104	D	D
37	25	045	%	%	69	45	105	E	E
38	26	046	&	&	70	46	106	F	F
39	27	047	'	'	71	47	107	G	G
40	28	050	(	(	72	48	110	H	H
41	29	051	)	)	73	49	111	I	I
42	2A	052	*	*	74	4A	112	J	J
43	2B	053	+	+	75	4B	113	K	K
44	2C	054	,	,	76	4C	114	L	L
45	2D	055	-	-	77	4D	115	M	M
46	2E	056	.	.	78	4E	116	N	N
47	2F	057	/	/	79	4F	117	O	O
48	30	060	0	0	80	50	120	P	P
49	31	061	1	1	81	51	121	Q	Q
50	32	062	2	2	82	52	122	R	R
51	33	063	3	3	83	53	123	S	S
52	34	064	4	4	84	54	124	T	T
53	35	065	5	5	85	55	125	U	U
54	36	066	6	6	86	56	126	V	V
55	37	067	7	7	87	57	127	W	W
56	38	070	8	8	88	58	130	X	X
57	39	071	9	9	89	59	131	Y	Y
58	3A	072	:	:	90	5A	132	Z	Z
59	3B	073	;	;	91	5B	133	[	[
60	3C	074	<	<	92	5C	134	\	\
61	3D	075	=	=	93	5D	135	]	]
62	3E	076	>	>	94	5E	136	^	^
63	3F	077	?	?	95	5F	137	_	_

Exemple : Graphe de qualité moyenne



Q30 = proportion de nucléotides ayant une qualité supérieure à 30

Attention à la version de l'encodage des qualités (Illumina)

S - Sanger and Illumina 1.8+,	Offset 33,	Quality range (0, 40)	(!,I)
X - Solexa,	Offset 64,	Quality range (-5, 40)	(;,h)
I - Illumina 1.3+	Offset 64,	Quality range (0, 40)	(@,h)
J - Illumina 1.5+	Offset 64,	Quality range (3, 40)	(B,h)
L - Illumina 1.8+	Offset 33,	Quality range (0, 41)	(!,J)

Nos données TESTS

Syndrome Bardet-Biedl

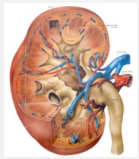
- Redin et al., 2012
- Genetique
  - Autosomique recessive
  - hautement hétérogène : 16 gènes BBS (274 exons, ~45kb)
  - Rare ~1/100000 - ~1/150000
- Phenotype

Retinopathy

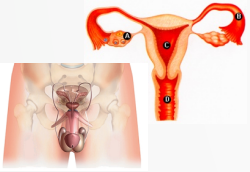
Polydactyly

Cognitive defects

Obesity



Renal anomalies

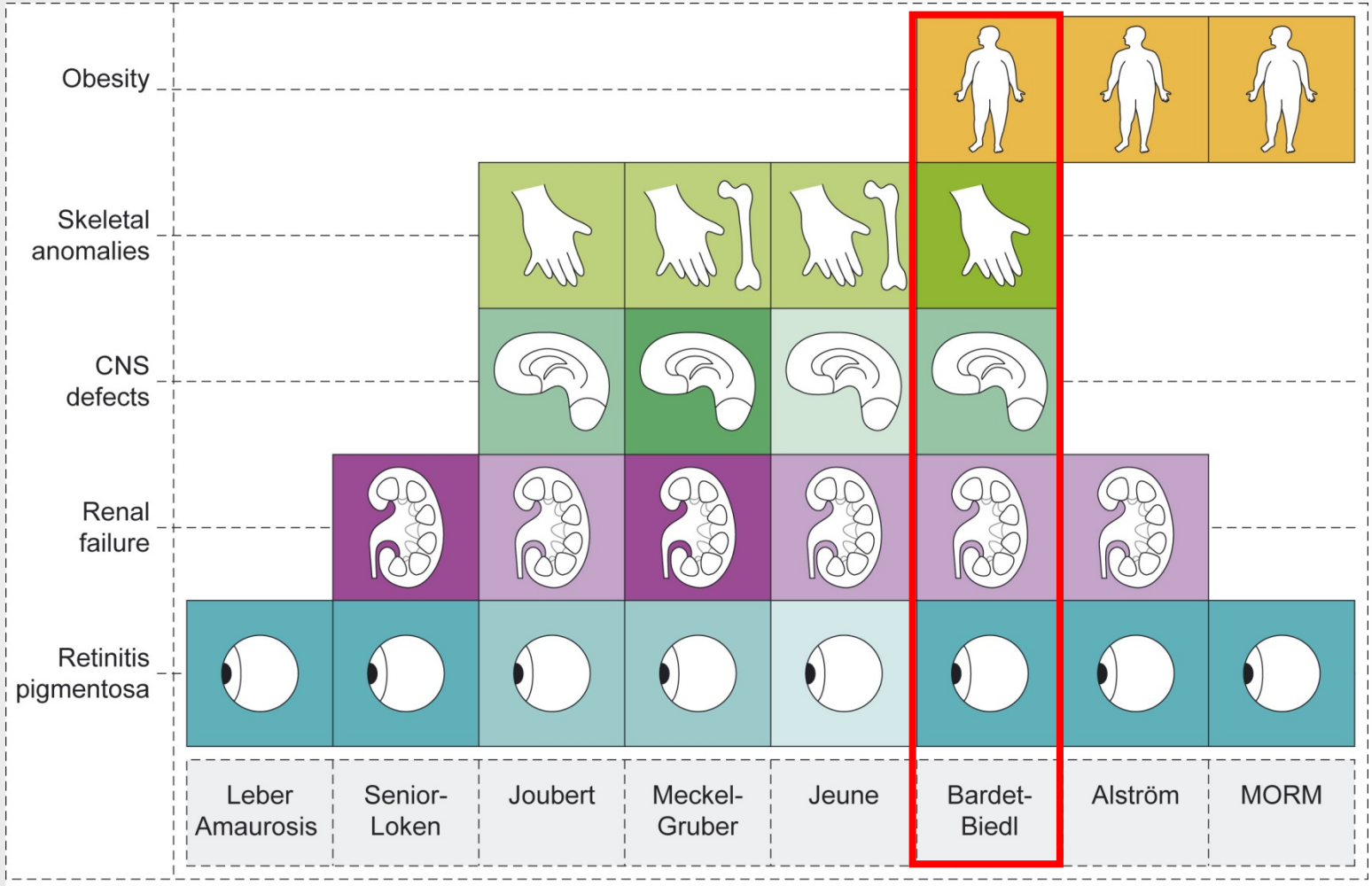


Hypogonadism

Main Features	Minor features
Retinal dystrophy	Myopia, cataract, astigmatism, strabism
Postaxial Polydactyly	Syndactyly, Brachydactyly
Intellectual disability, Developmental delay	Hearing defects, Smell defects
Renal dysfunction	Diabetes, glucidic intolerance
Hypogonadism/ Hydrometrocolpos	Cardiopathy, liver fibrosis
Hypertension	Ataxias

Beales et al 1999

Ciliopathies



Diagnostic BBS

- Séquençage Sanger exhaustif
  - Couteux
  - Beaucoup de gènes impliqués
- Screening des mutation récurrentes et des gènes fréquemment mutés (BBS1, BBS10, BBS12) combinés à de l'alignement hétérozygote
- Screening systématique et automatique de tous les gènes BBS
  - Capture + NGS

Design expérimental

- Design de la capture (à la carte): exons de 30 genes (16 gènes BBS + 14 gènes d'autres ciliopathies)
- 52 patients:
  - Cohort de preuve de principe: 14 patients dont les mutation sont connues (identifiées en Sanger)
  - 1 cohort: 38 patients avec mutation inconnue
- Le patient provient d'une autre cohorte analysée après la validation de la preuve de principe et après les bons résultats sur la première cohorte

Qualité des données brutes

Partie pratique n°1

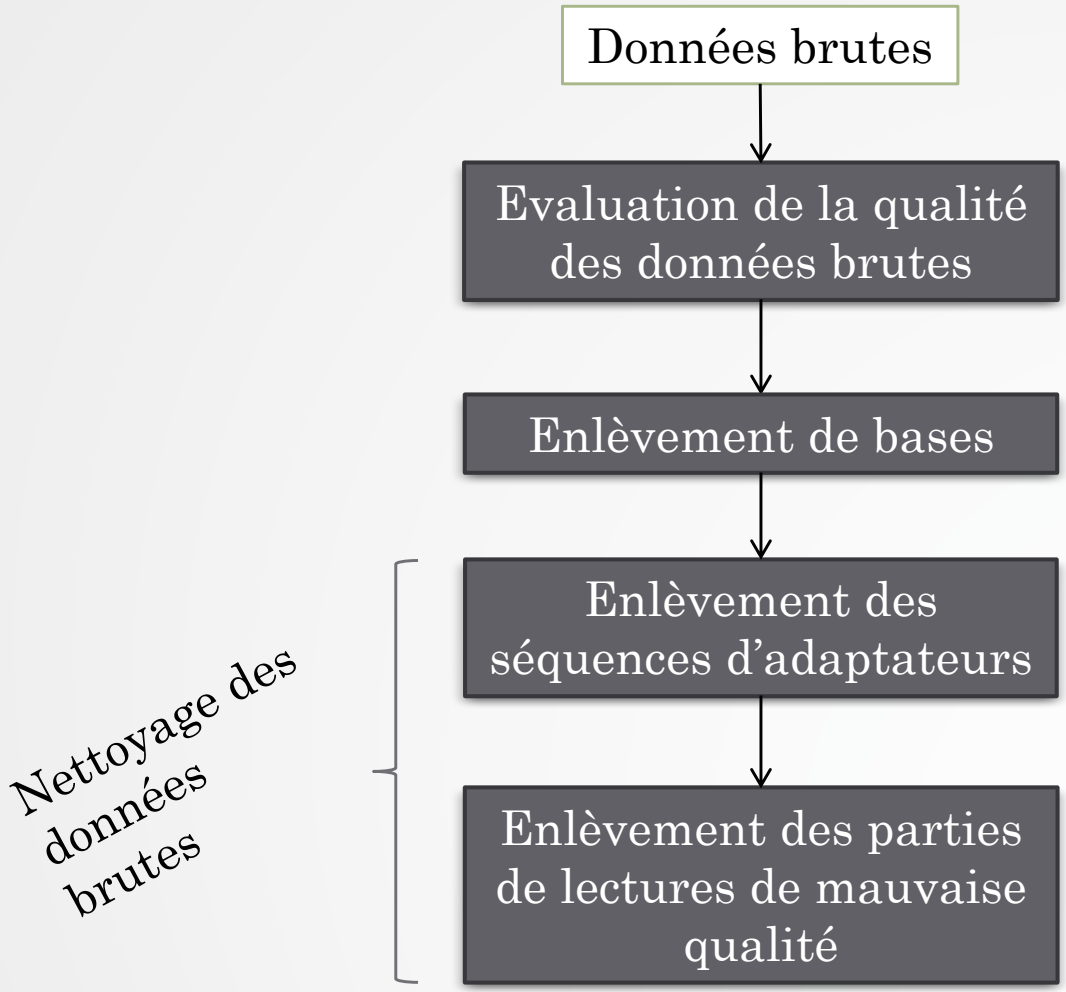
- Wiki: http://genomeast.igbmc.fr/wiki/
- Choisir: Training
- Choisir DNA-seq/DU dijon
- Choisir: Quality control of raw sequencing data, data cleaning, FASTQ file handling
- Réaliser la création de l'environnement de travail
- Remplacer [your login] par votre login!

Partie pratique n°2

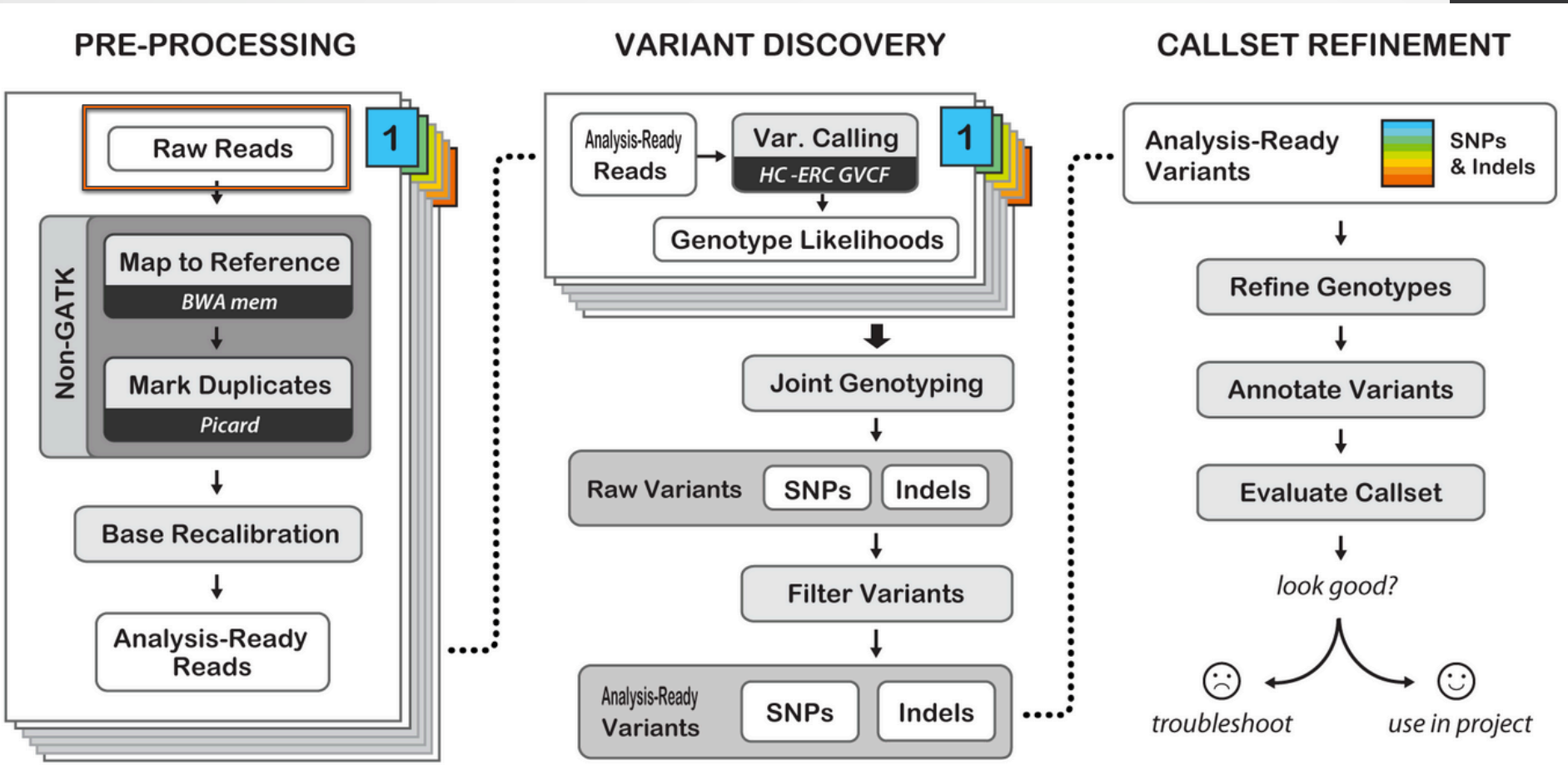
Gestion des données brutes de séquençage

- Il existe plusieurs outils développés pour la gestion des données brutes issues du séquenceur :
  - Evaluer la qualité des données
  - Corriger les problèmes de qualité
  - Manipuler les fichiers (transformation de formats).
- Toujours penser à lire les spécifications pour être sûr que l'outil fait ce que vous souhaitez (Attention aux surprises!)

Processus

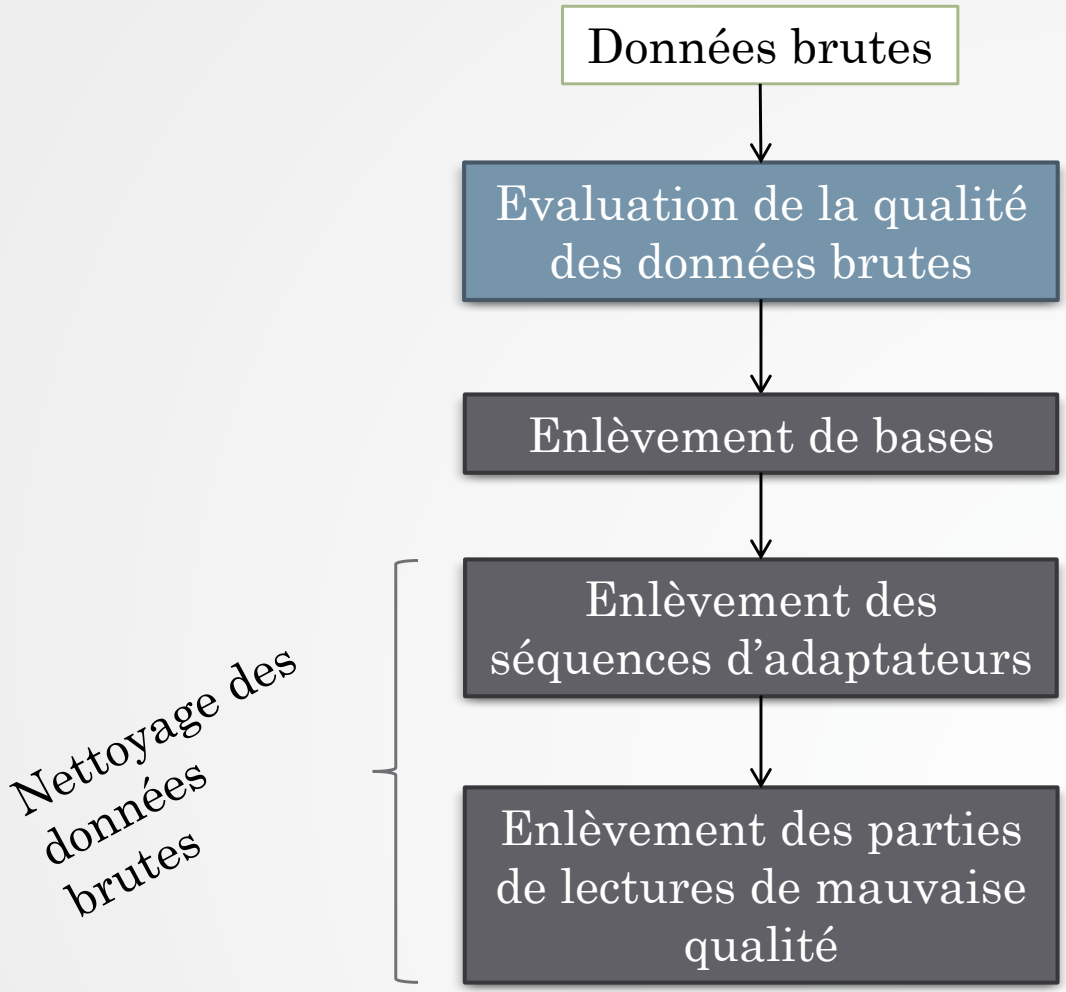

Données brutes
Evaluation de la qualité des données brutes
Enlèvement de bases
Enlèvement des séquences d'adaptateurs
Enlèvement des parties de lectures de mauvaise qualité
Nettoyage des données brutes

Processus



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

Processus

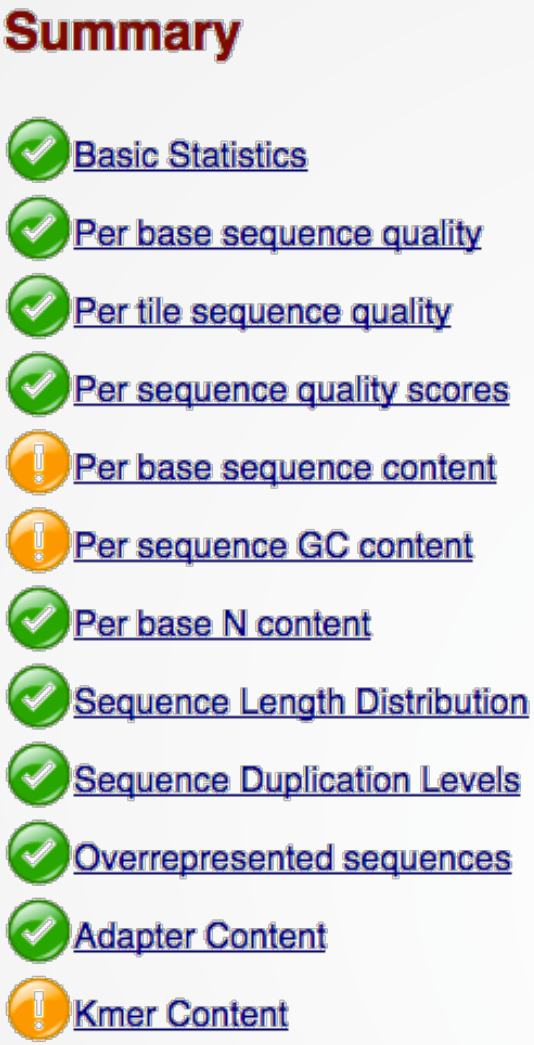

Données brutes
Evaluation de la qualité des données brutes
Enlèvement de bases
Enlèvement des séquences d'adaptateurs
Enlèvement des parties de lectures de mauvaise qualité
Nettoyage des données brutes

Evaluation de la qualité des données brutes

- Outils : FastQC, SolexaQA, Fastx-toolkit, NGS QC toolkit…
- FastQC (Babraham Institute)
  - Import de fichiers BAM, SAM, FastQ (tous les encodages de qualité sont supportés)
  - Lancement en ligne de commande ou via une interface
  - Fournit un rapport sur la qualité des données
  - Permet d'évaluer les problèmes
  - Rapport contient des tableaux et des graphes
  - HTML
  - Fonctionne sur des fichiers compressés
  - Estimation sur un échantillon du fichier d'entrée pour accélérer le temps de calcul

Partie pratique n°3

FastQC

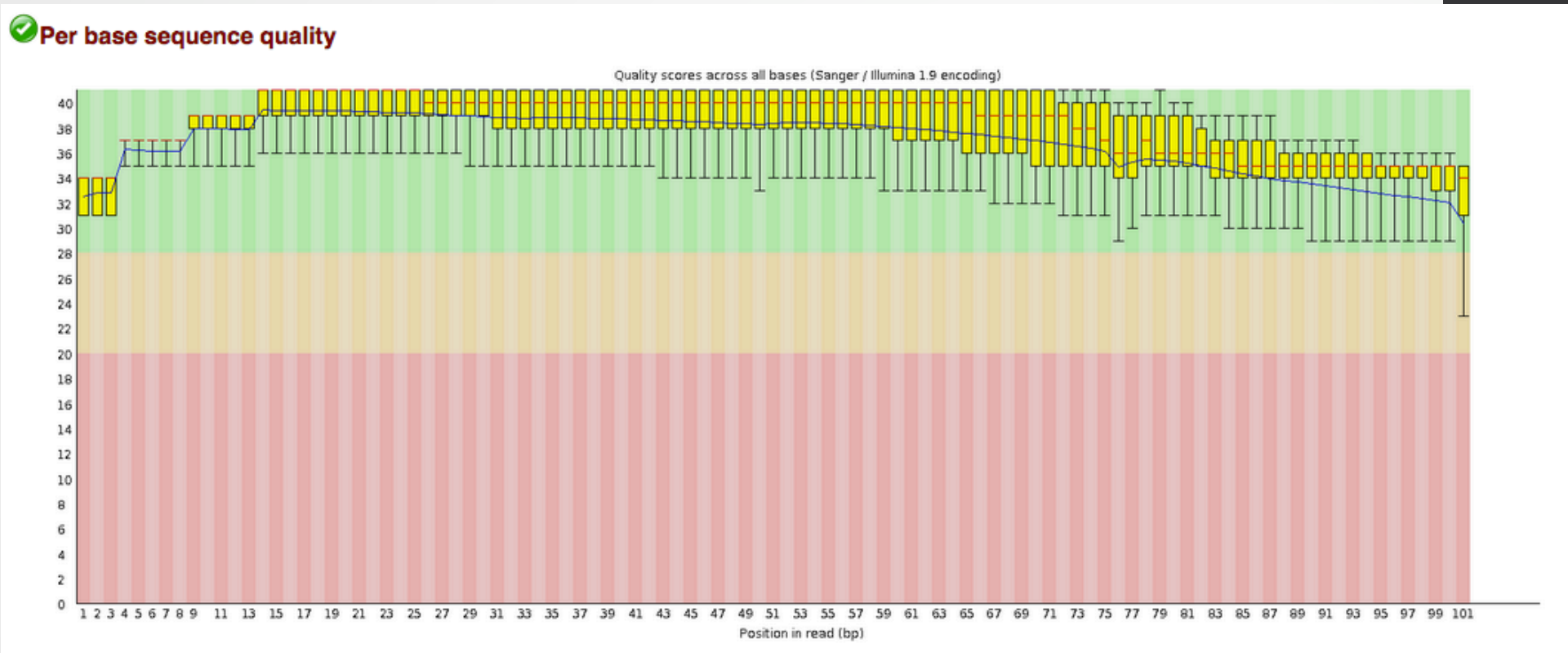

Summary
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
Kmer Content

FastQC

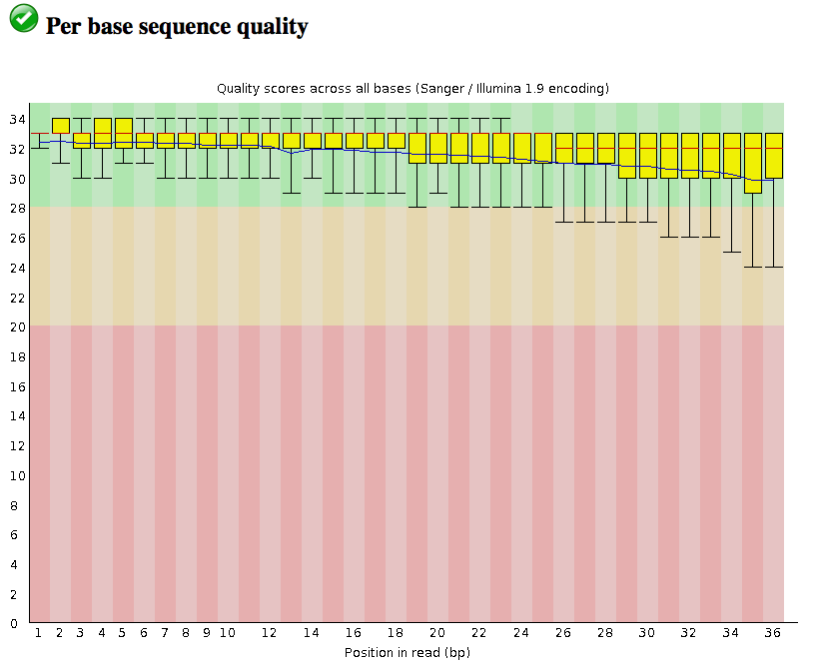
Basic Statistics

Measure	Value
Filename	CRN-107_11-R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	280508
Sequences flagged as poor quality	0
Sequence length	101
%GC	38

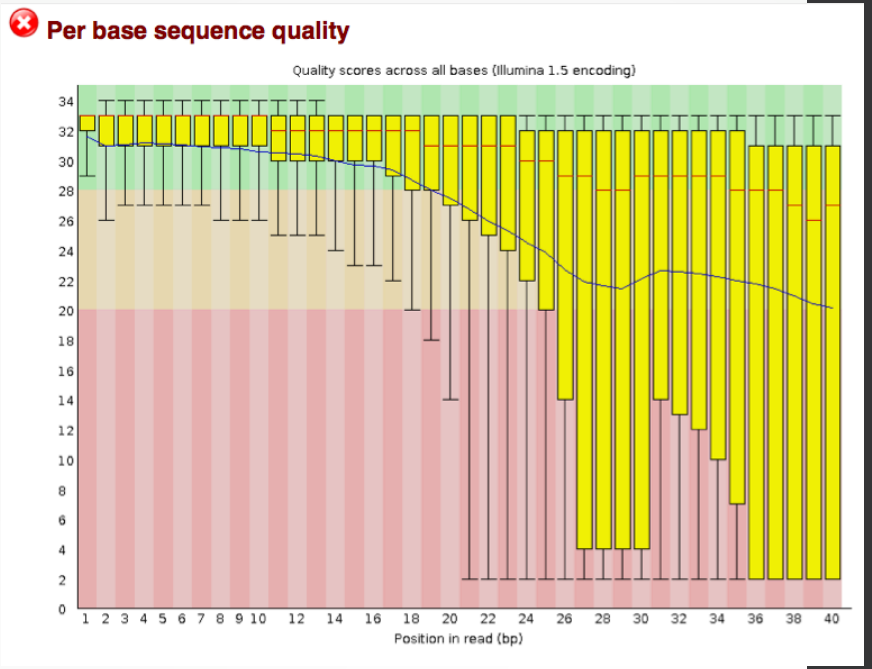
FastQC


Per base sequence quality
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

FastQC



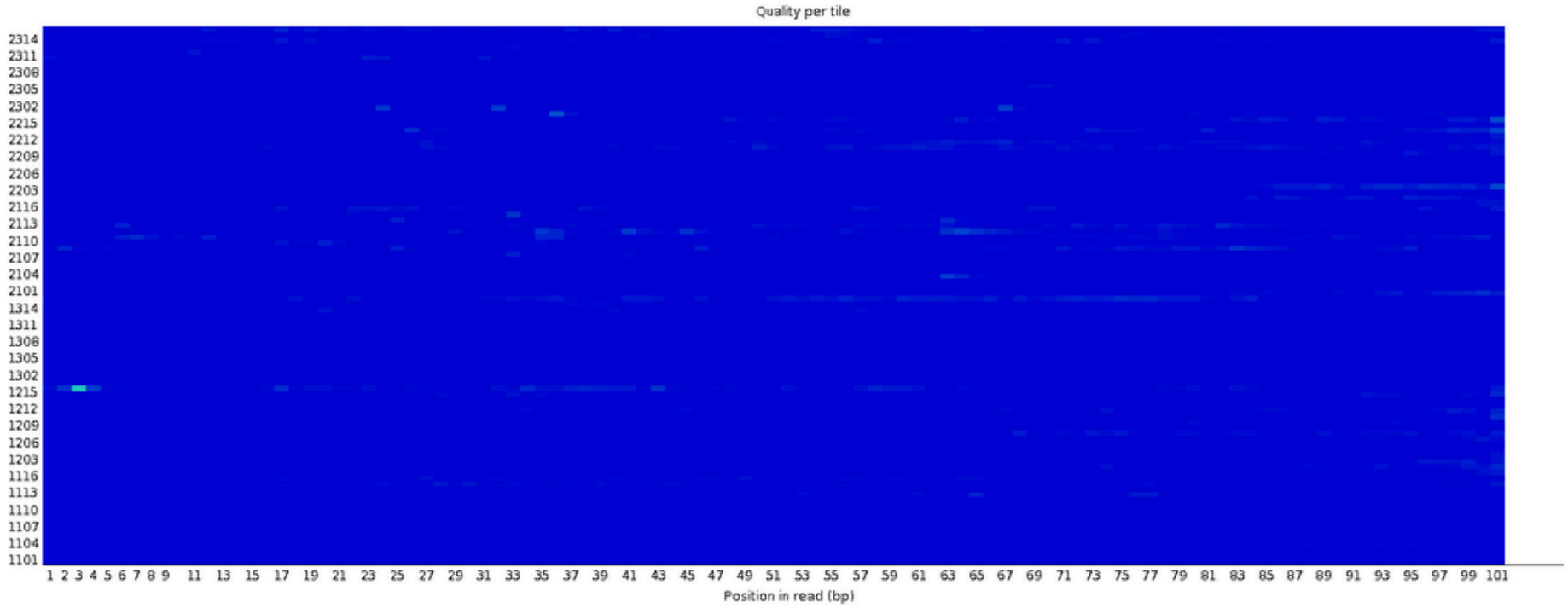
Données de bonne qualité



Données de mauvaise qualité

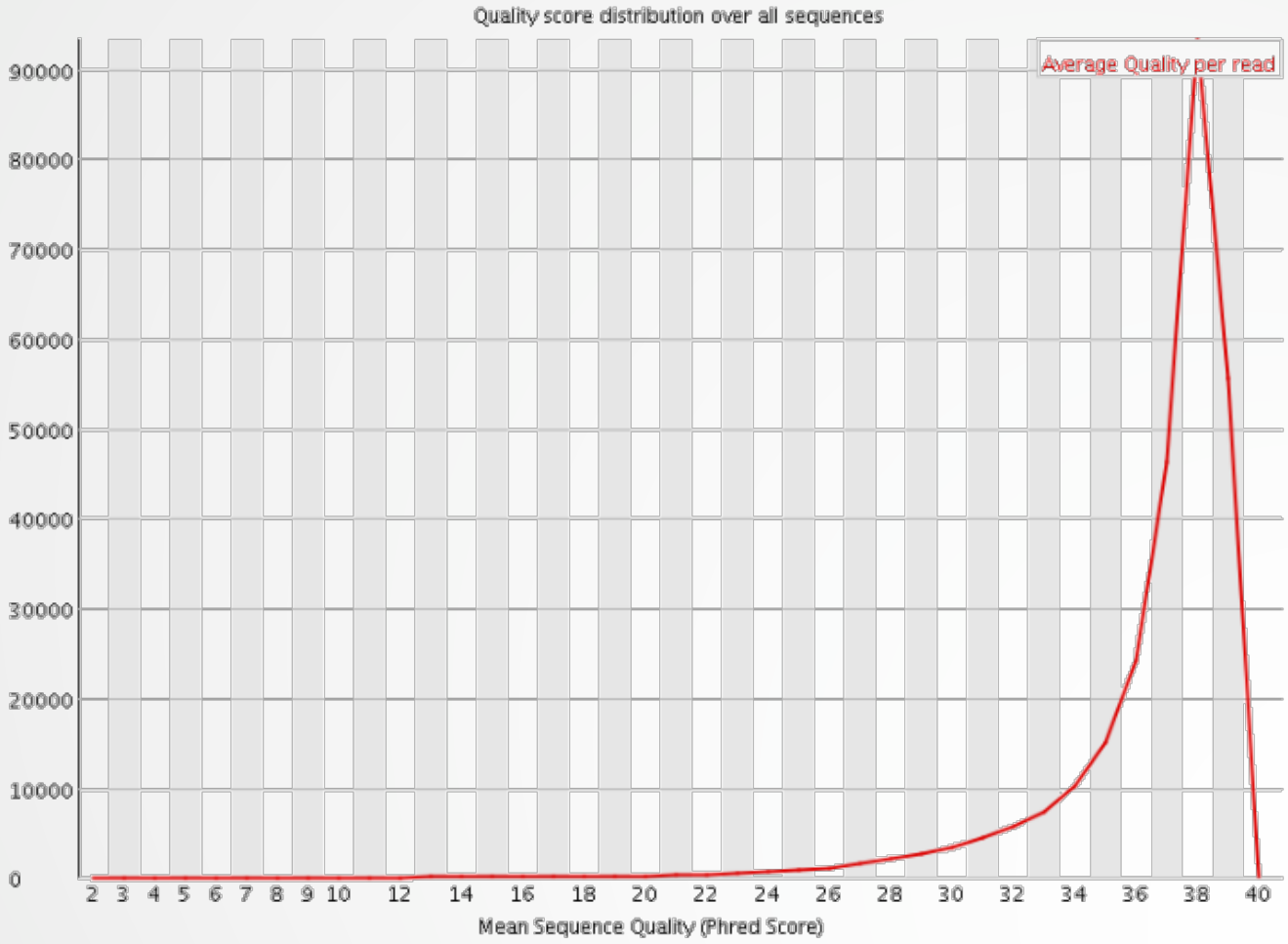
FastQC

Per tile sequence quality

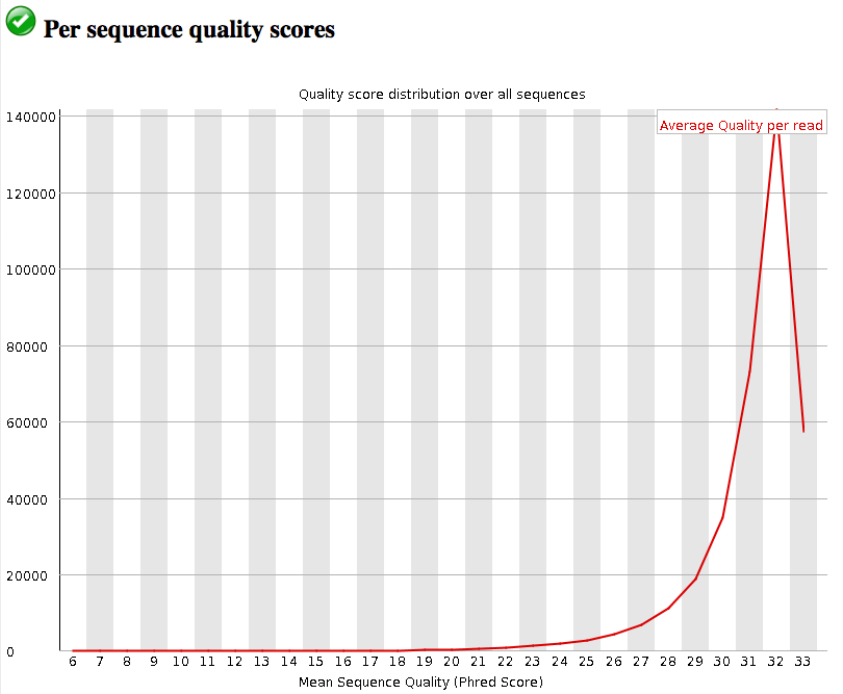


FastQC

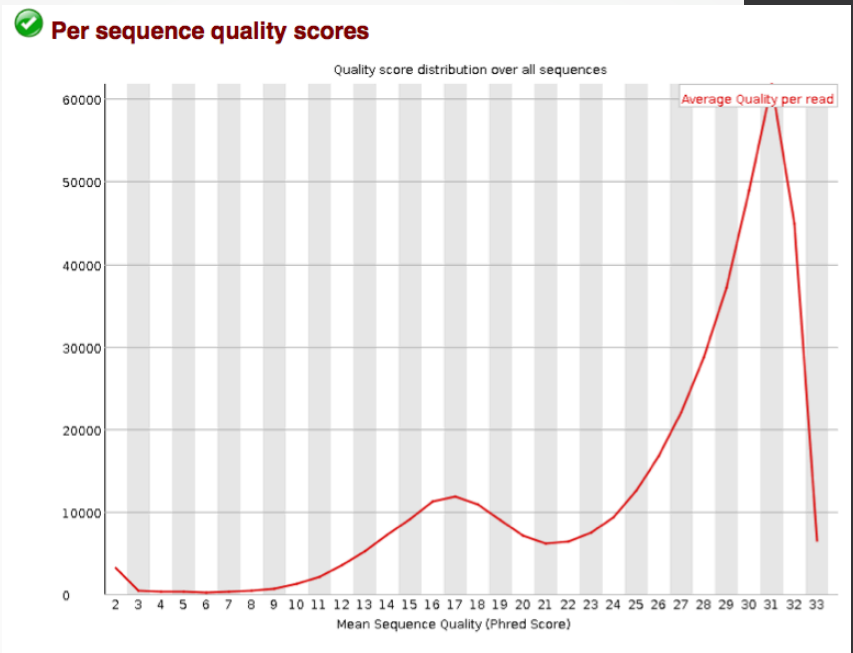
Per sequence quality scores



FastQC

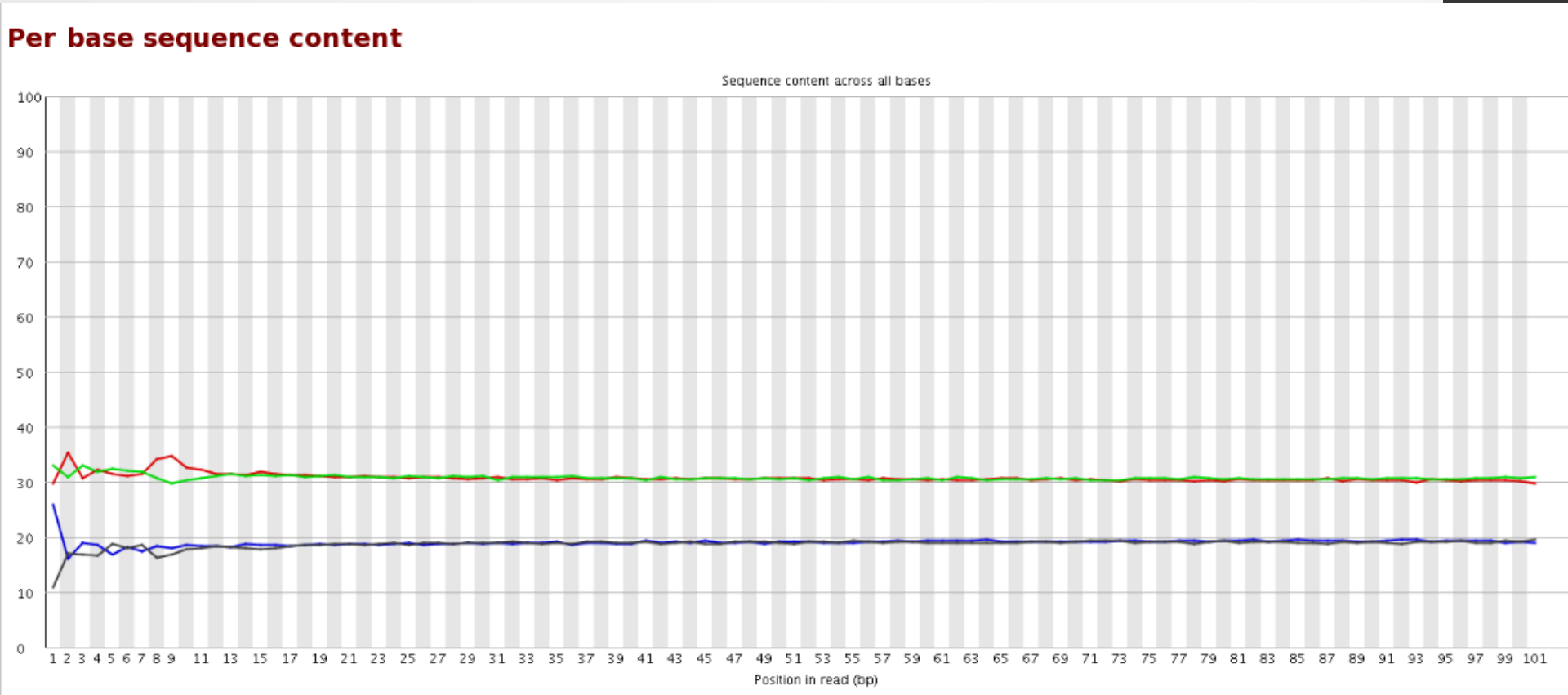


Données de bonne qualité

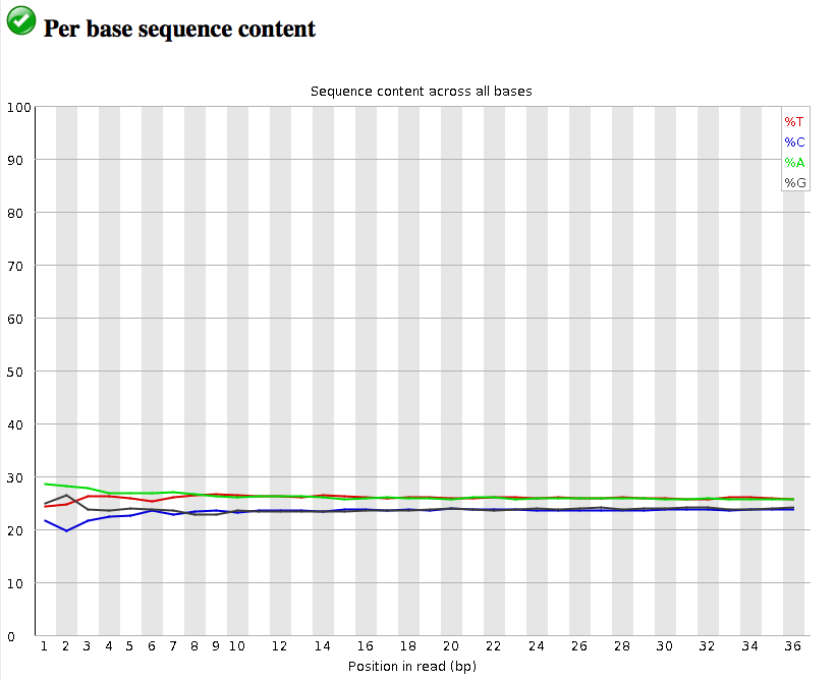


Données de mauvaise qualité

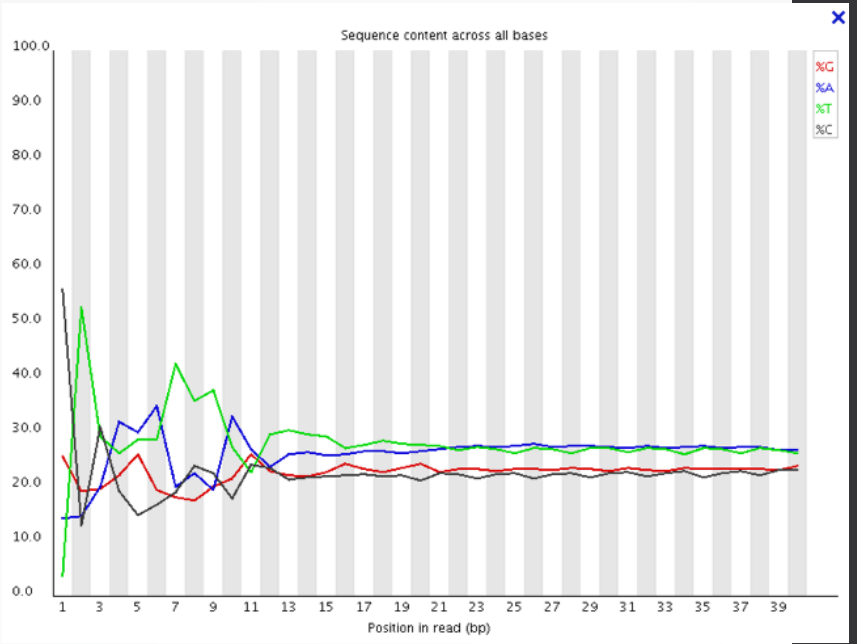
FastQC


Per base sequence content
Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99 101
Position in read (bp)

FastQC

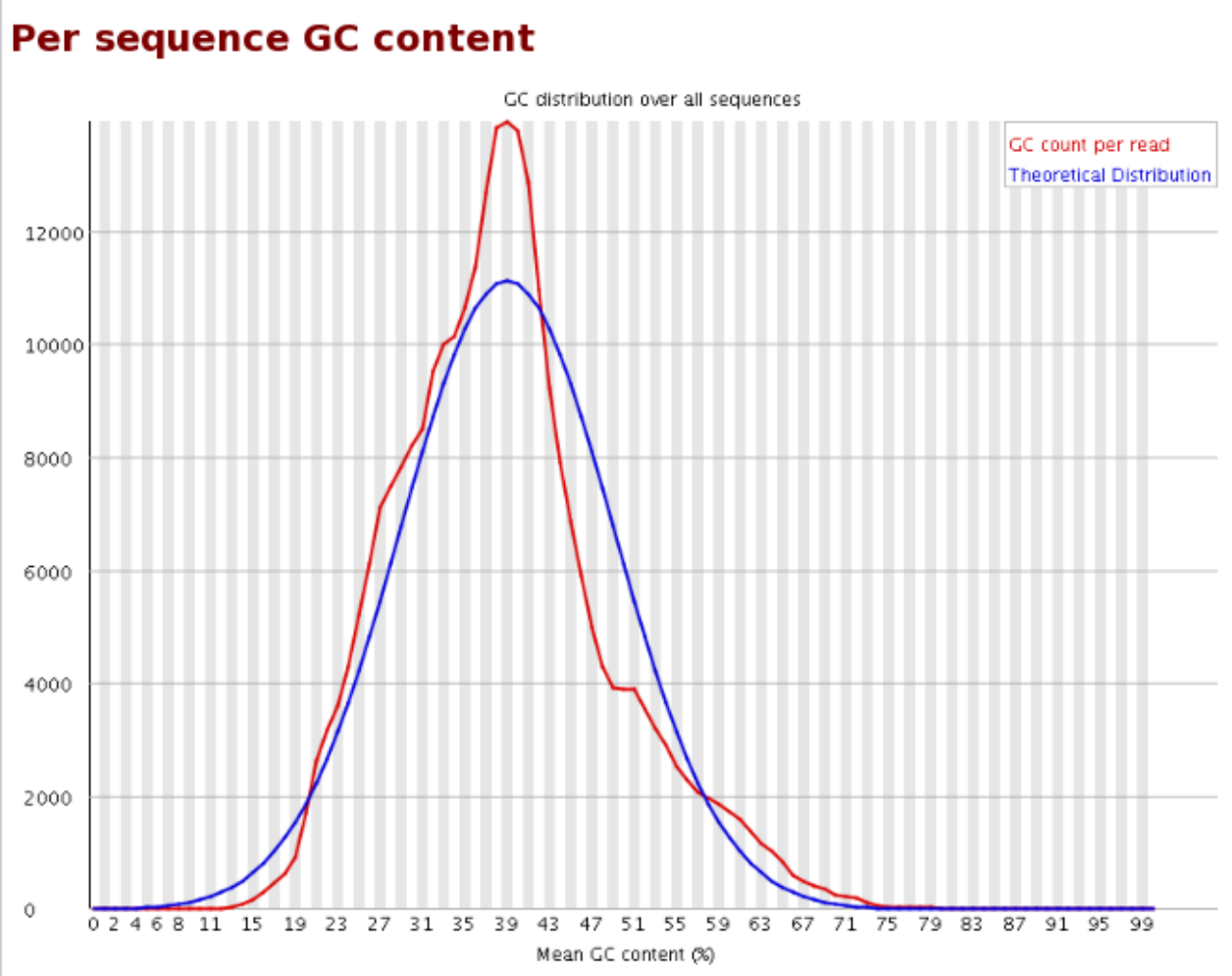


Données de bonne qualité

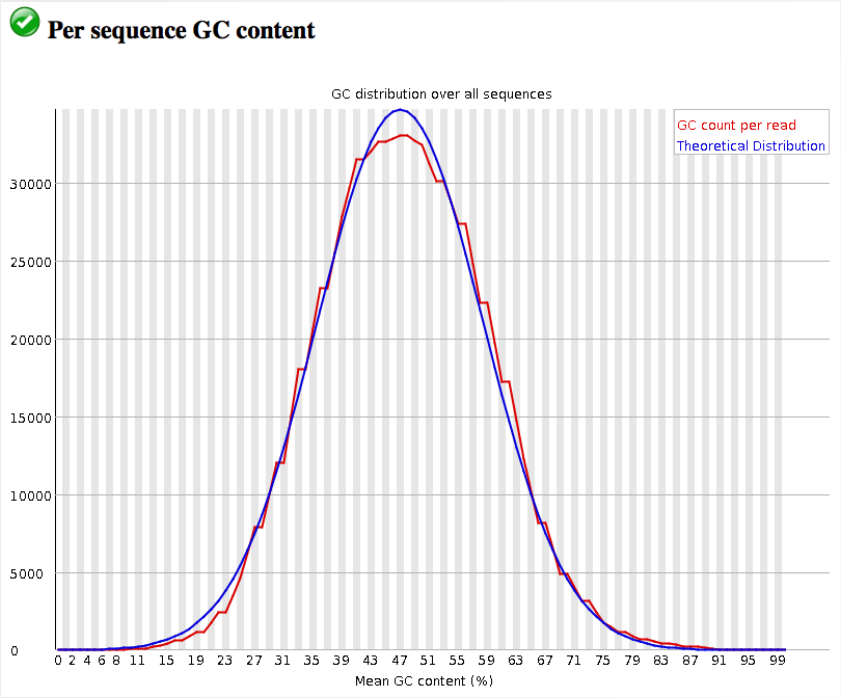


Données de mauvaise qualité

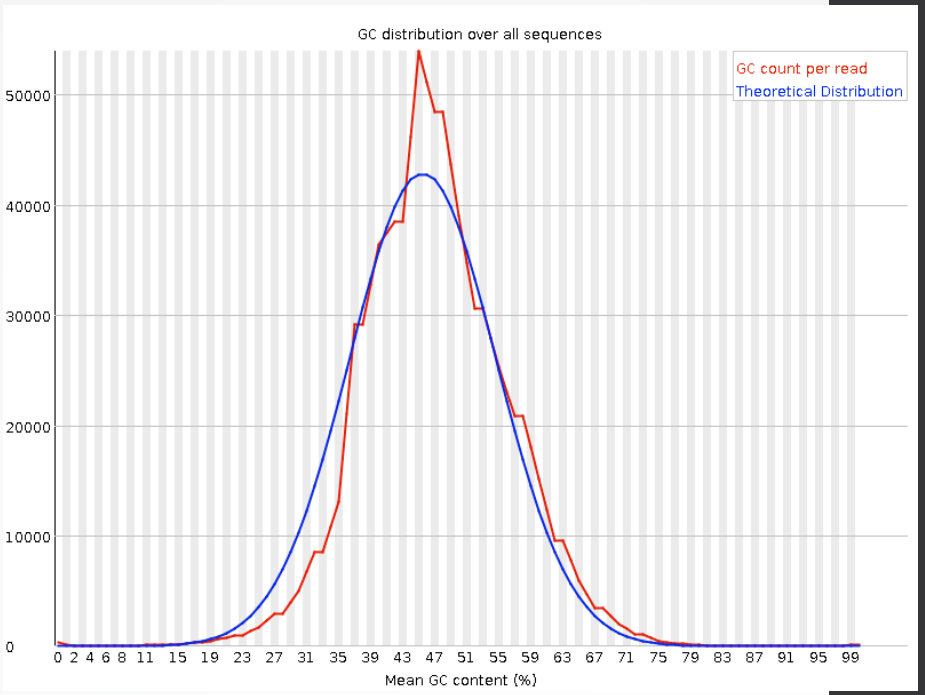
FastQC


Per sequence GC content
GC distribution over all sequences
GC count per read
Theoretical Distribution
12000
10000
8000
6000
4000
2000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

FastQC



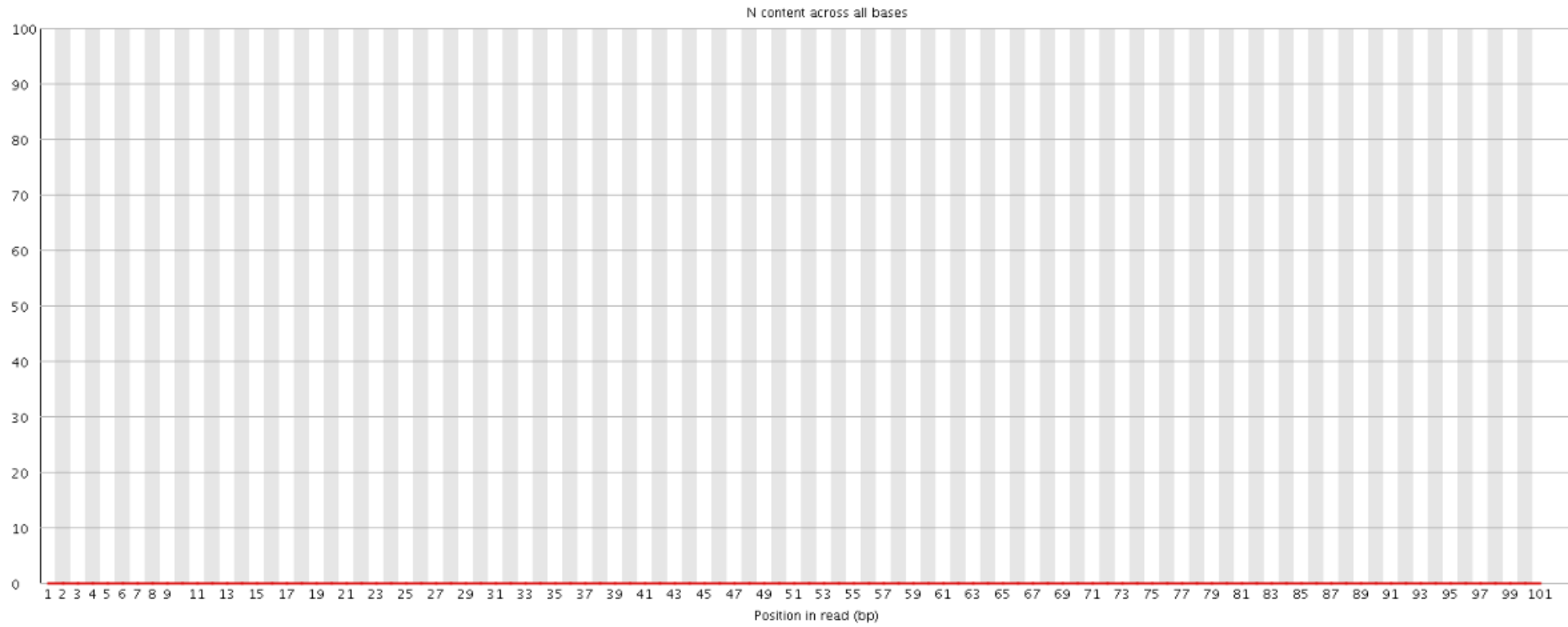
Données de bonne qualité



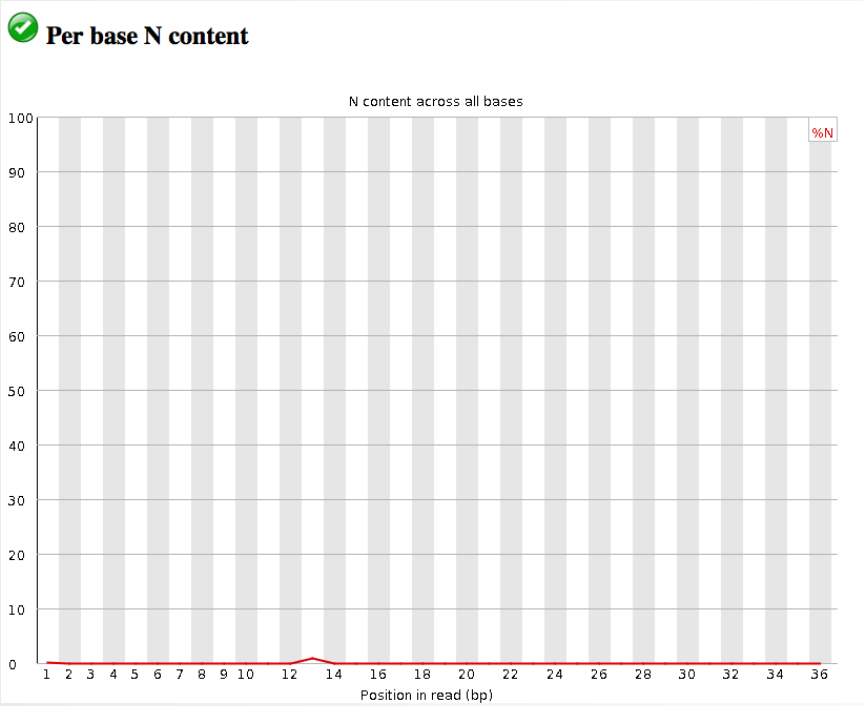
Données de mauvaise qualité

FastQC

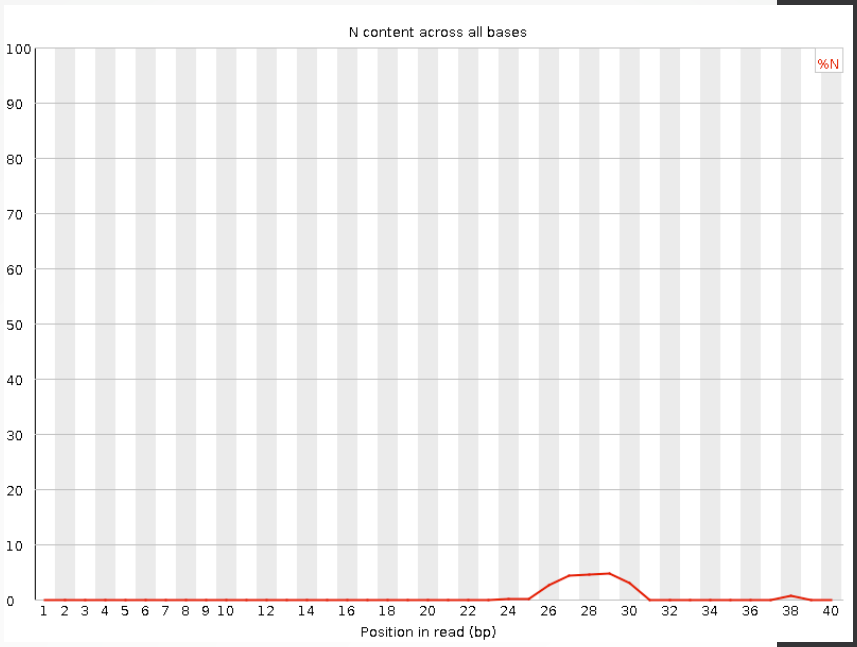
Per base N content



FastQC

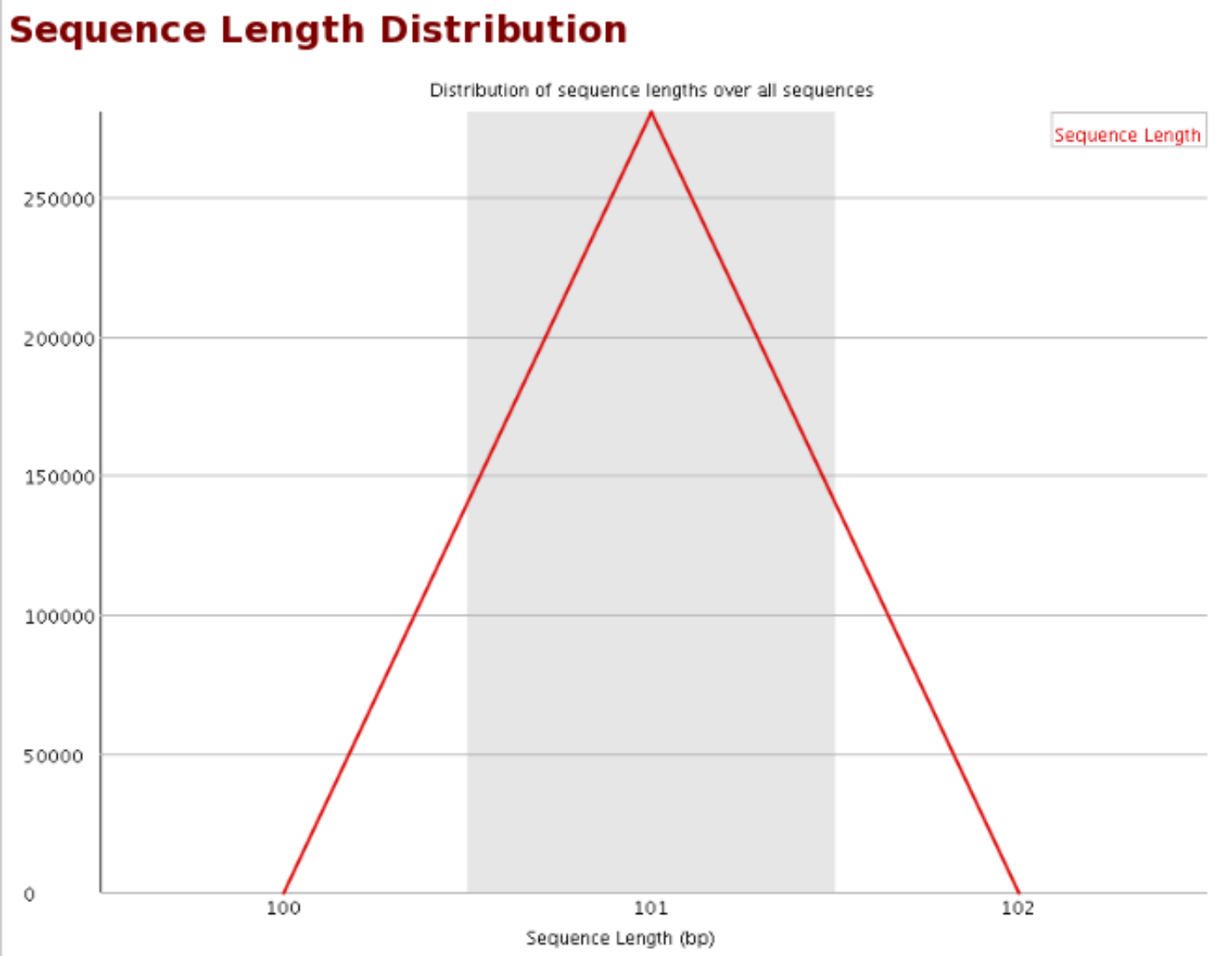


Données de bonne qualité



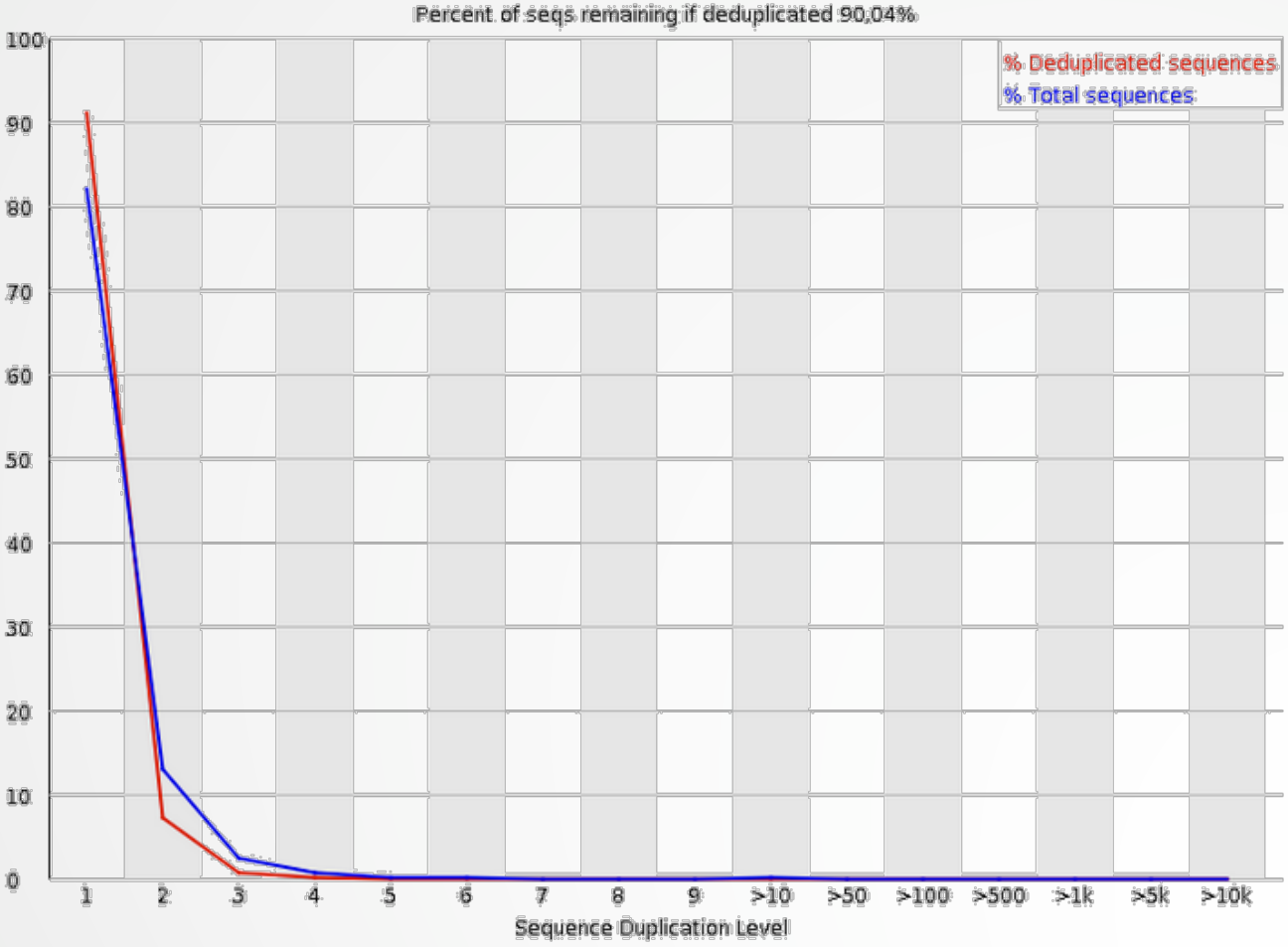
Données de mauvaise qualité

FastQC


Sequence Length Distribution
Distribution of sequence lengths over all sequences
Sequence Length
250000
200000
150000
100000
50000
0
100
101
102
Sequence Length (bp)

FastQC

Sequence Duplication Levels

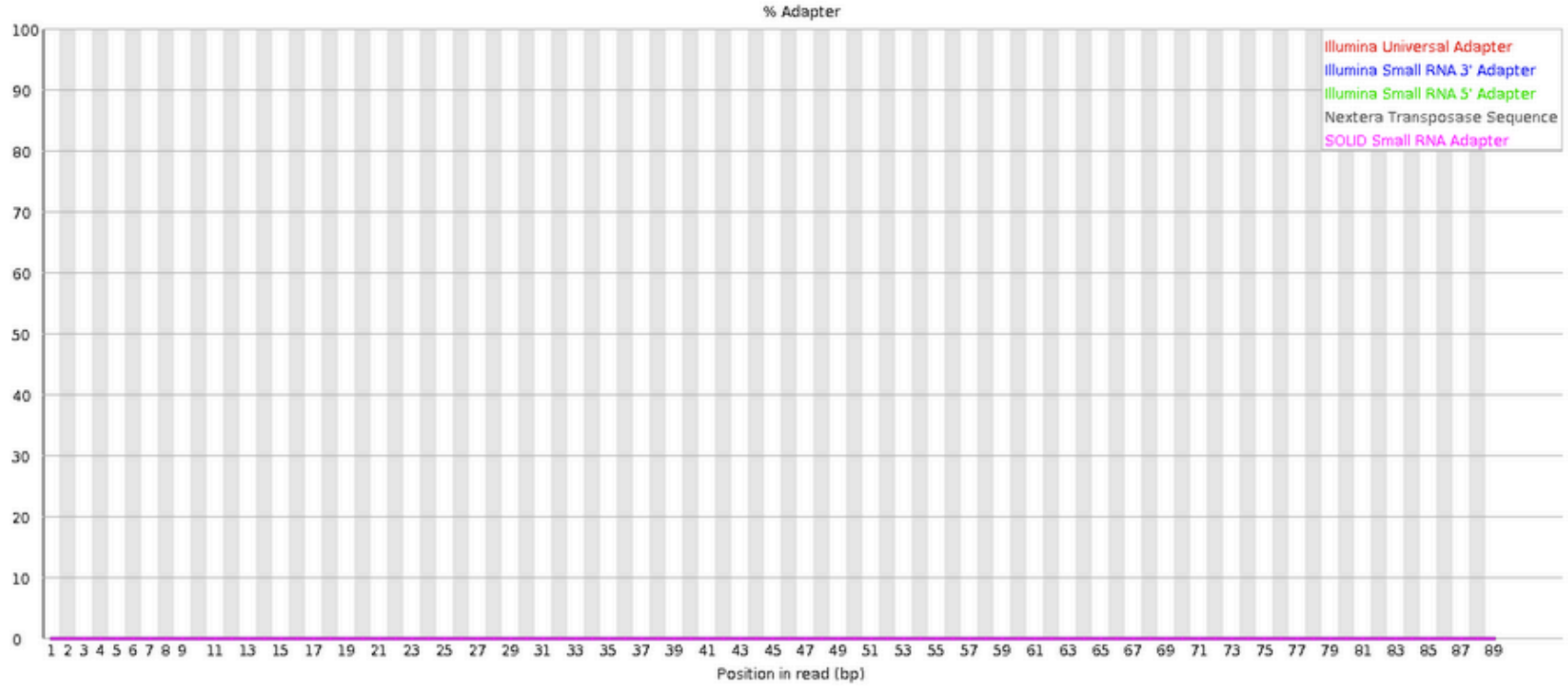


FastQC

Overrepresented sequences

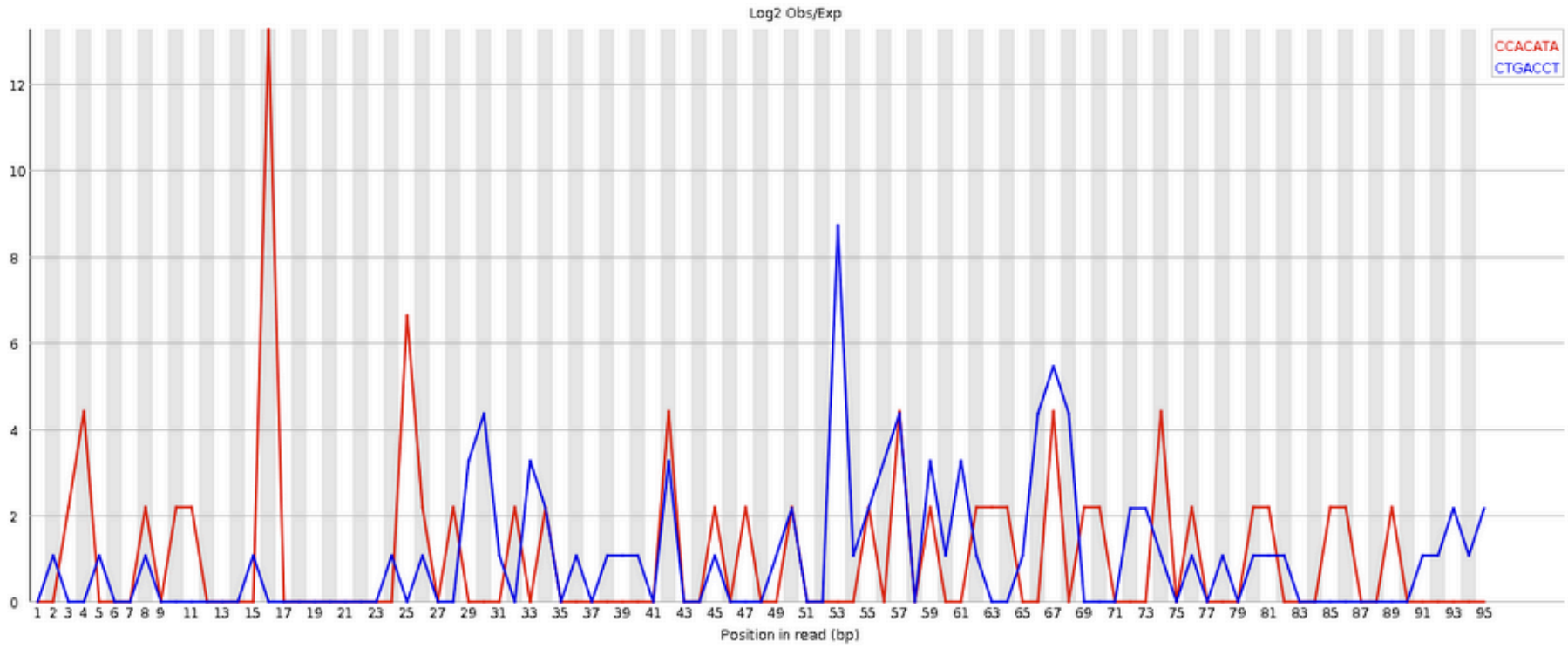
No overrepresented sequences

Adapter Content



FastQC

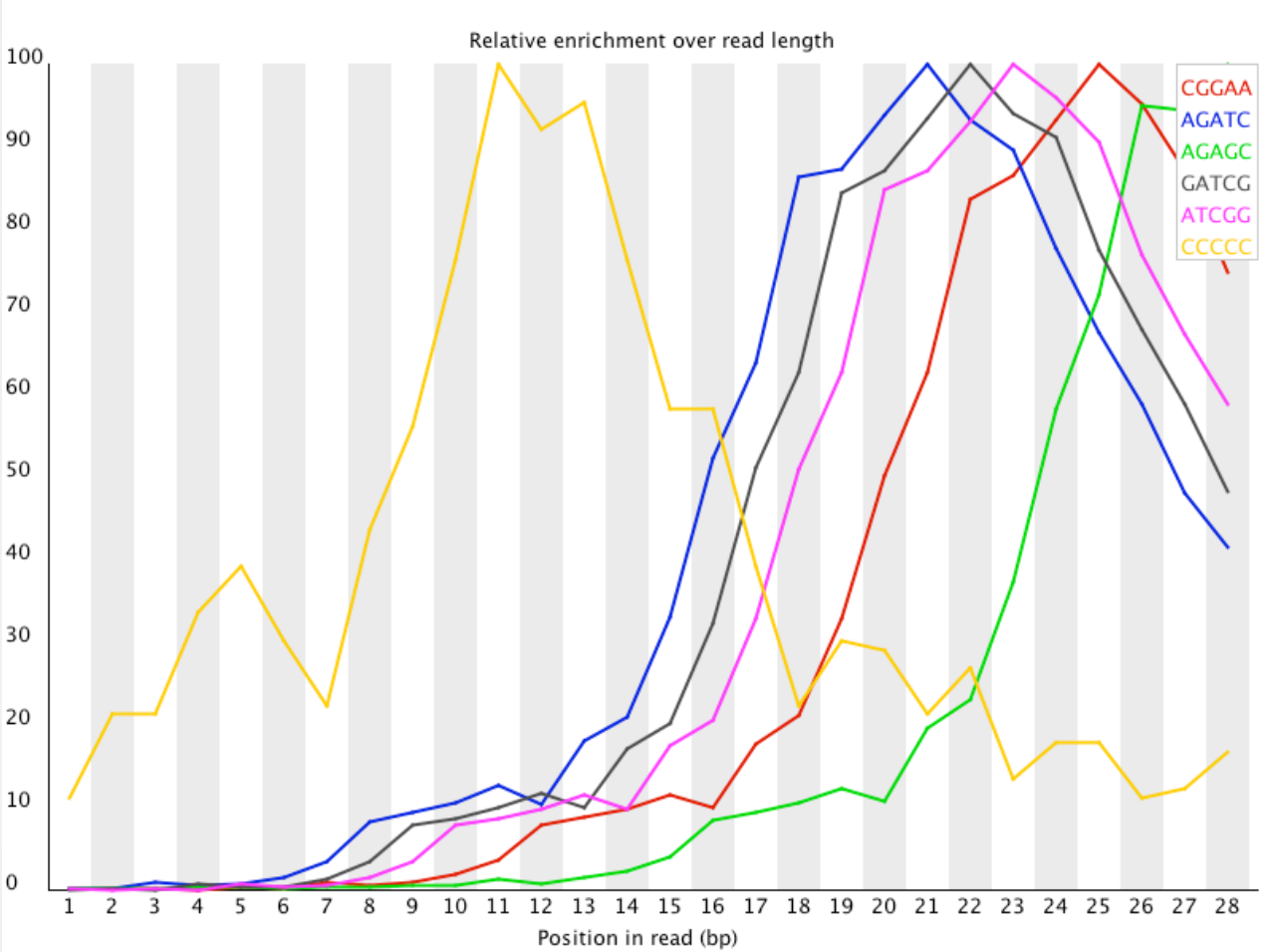
Kmer Content



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CCACATA	215	0.008498441	13.255788	16
CTGACCT	435	0.008957403	8.735616	53

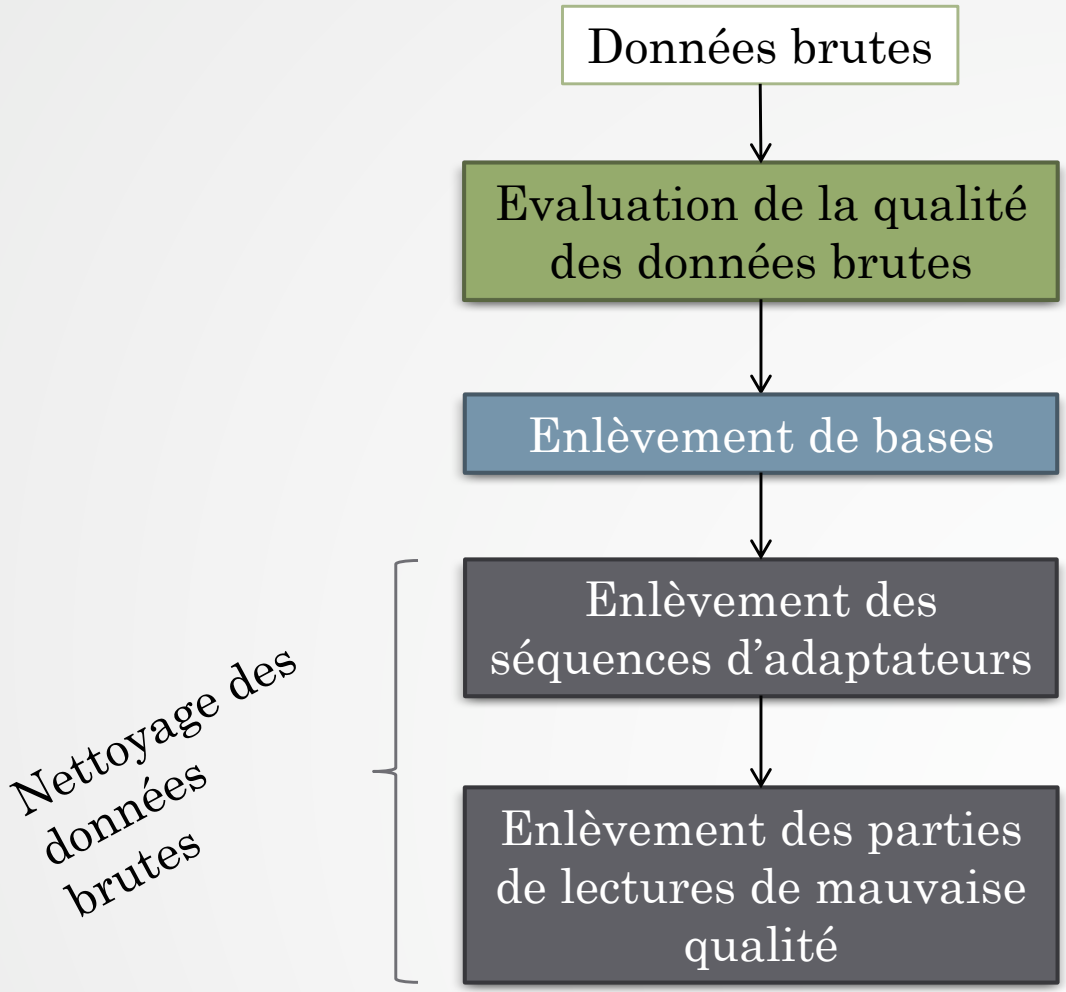
FastQC

Données biaisées



Nettoyage des données brutes

Processus

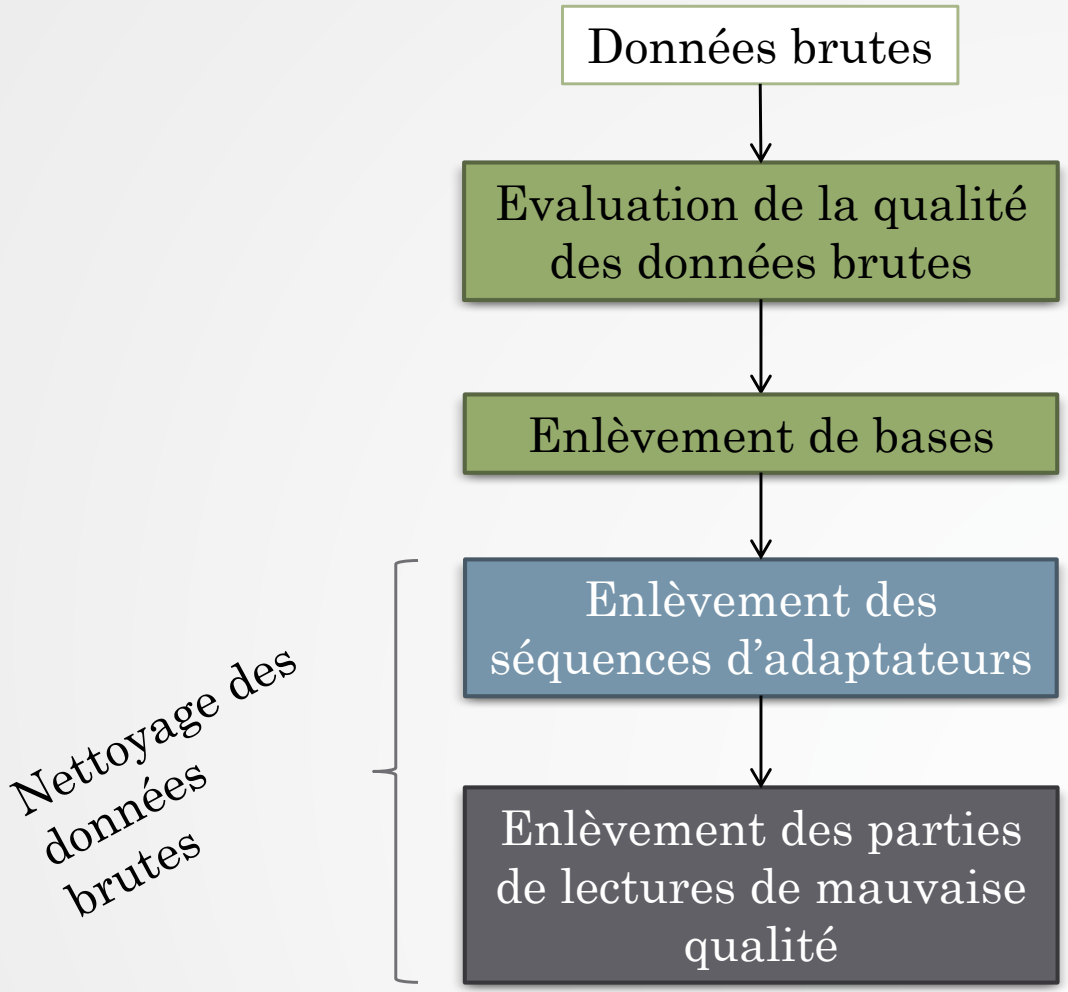

Données brutes
Evaluation de la qualité des données brutes
Enlèvement de bases
Enlèvement des séquences d'adaptateurs
Enlèvement des parties de lectures de mauvaise qualité
Nettoyage des données brutes

Enlèvement de la dernière base

- La taille des lectures attendue est 2x100 et non pas 2x101
- Lorsque l'on séquence, nous séquençons toujours une base de plus car les bases n+1 sont utilisées pour calculer les statistiques des bases à la position n
- La dernière base doit être enlevée

Partie pratique n°4

Processus


Données brutes
Evaluation de la qualité des données brutes
Enlèvement de bases
Enlèvement des séquences d'adaptateurs
Enlèvement des parties de lectures de mauvaise qualité
Nettoyage des données brutes

Elimination des séquences contaminantes

- Quel type de contamination?
  - Adaptateurs
  - Primer de séquençage
  - Autres…
- Pourquoi ces contaminants?
  - Les fragments d'ADN séquencés sont plus petits que la taille des lectures
  - Des dimers d'adaptateurs se sont formés lors de la préparation de la librairies.
- Pourquoi les enlever?
  - Ces séquences non génomiques peuvent poser un problème lors de l'alignement.

Elimination des séquences contaminantes

- A quoi dois-je faire attention ?
  - Certains outils n'enlèvent la séquence d'adaptateur que si les lectures contiennent exactement la séquence d'adaptateur (pas de gestion des erreurs de séquençage).
  - Attention aux données pairées! On ne peut pas enlever une lecture d'un sens sans enlever la lecture de l'autre sens. Il faut donc analyser les deux fichiers fastq en même temps.
  - Certains outils ne fonctionnent pas sur des données pairées
- Outils: ClipReads (GATK), fastx-toolkit, homerTools, Trimmomatic, Cutadapt…

Partie pratique n°5

Comprendre les adaptateurs

Where

Trufseq Universal Adaptor:

5´AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT3´

--> reverse complementary

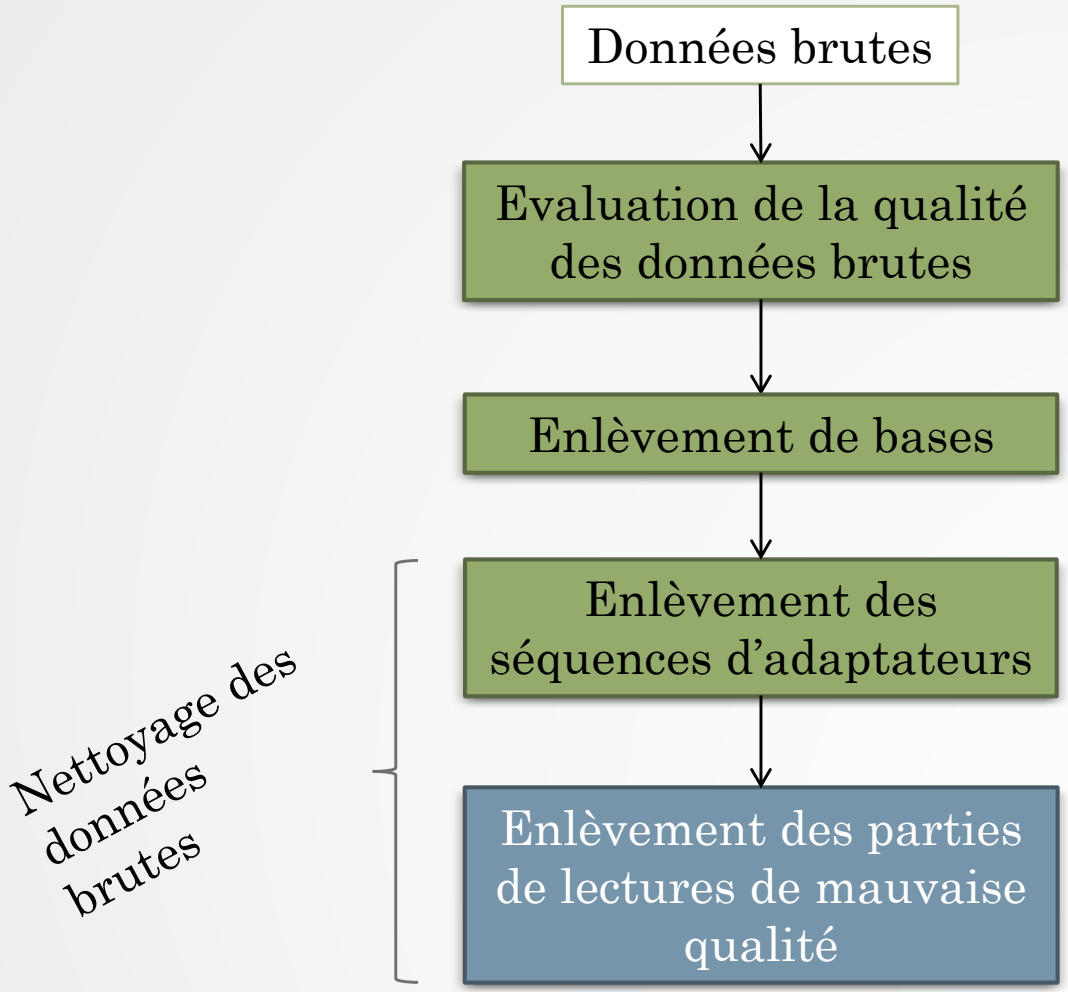
5´AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT3´

TruSeq Indexed Adapter:

5´GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNN-ATCTCGTATGCCGTCTTCTGCTTG3´

http://tucf-genomics.tufts.edu/documents/protocols/TUCF_Understanding_Illumina_TruSeq_Adapters.pdf

Processus


Données brutes
Evaluation de la qualité des données brutes
Enlèvement de bases
Enlèvement des séquences d'adaptateurs
Enlèvement des parties de lectures de mauvaise qualité
Nettoyage des données brutes

Elimination des parties de lectures de mauvaise qualité

- Pourquoi est ce que la fin des lectures est de moins bonne qualité?
  - Problème de chimie
- Quelle conséquence?
  - Les suites de nucléotides de mauvaise qualité à la fin des lectures peuvent induire des variants détectés à tord lors de la détection des variants.
- Comment corriger le problème?
  - Enlever les nucléotides de mauvaise qualité
  - Attention aux données pairées!
- Outil : Fastqx toolkit, SolexaQA…

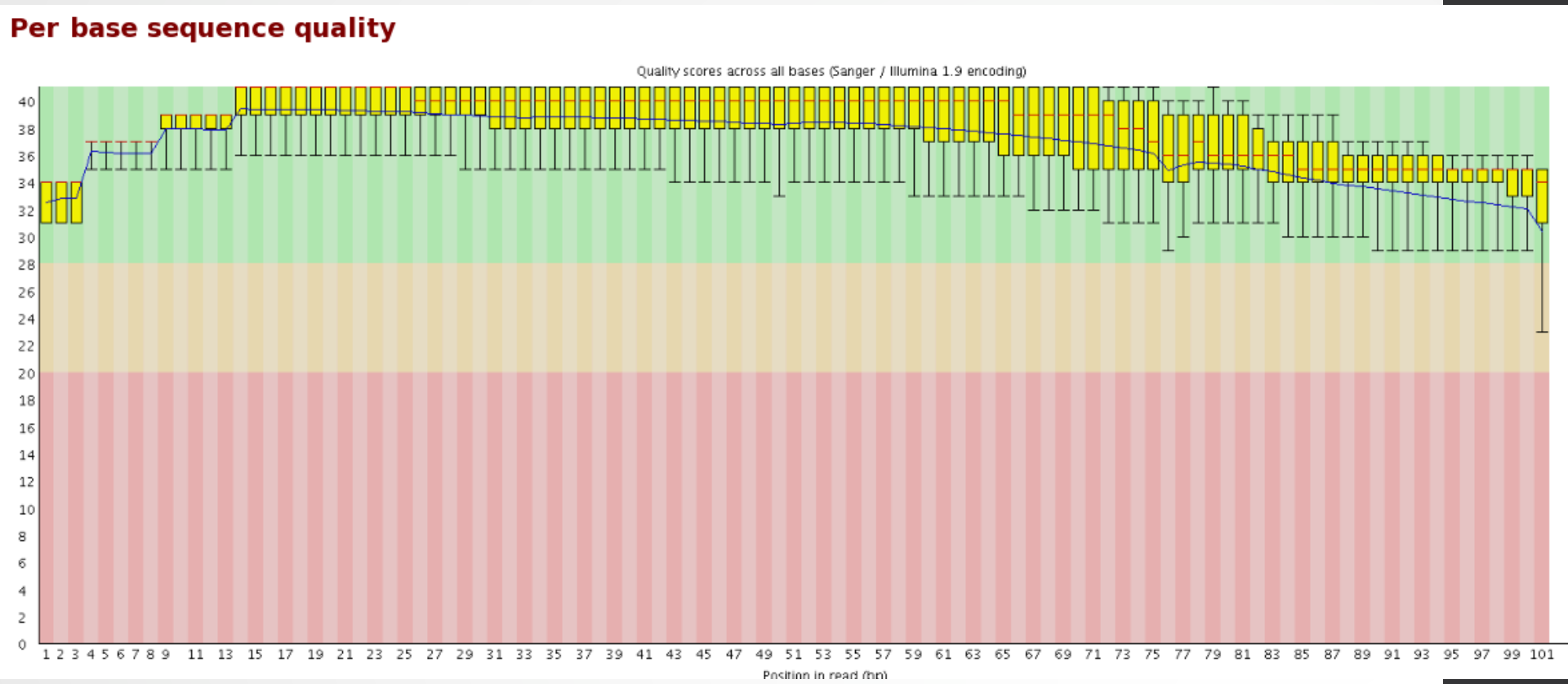
Partie pratique n°6

Partie pratique n°7

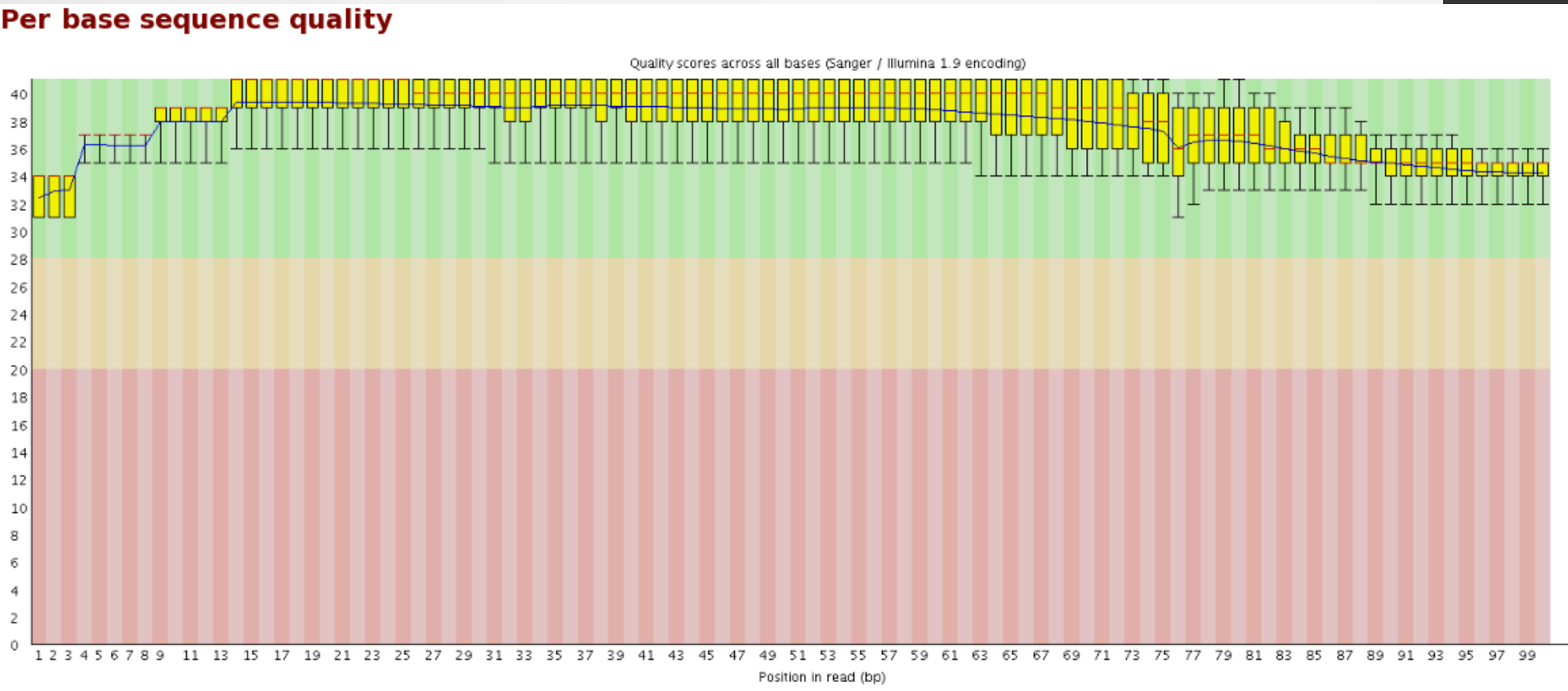
Partie pratique n°8

Partie pratique n°9

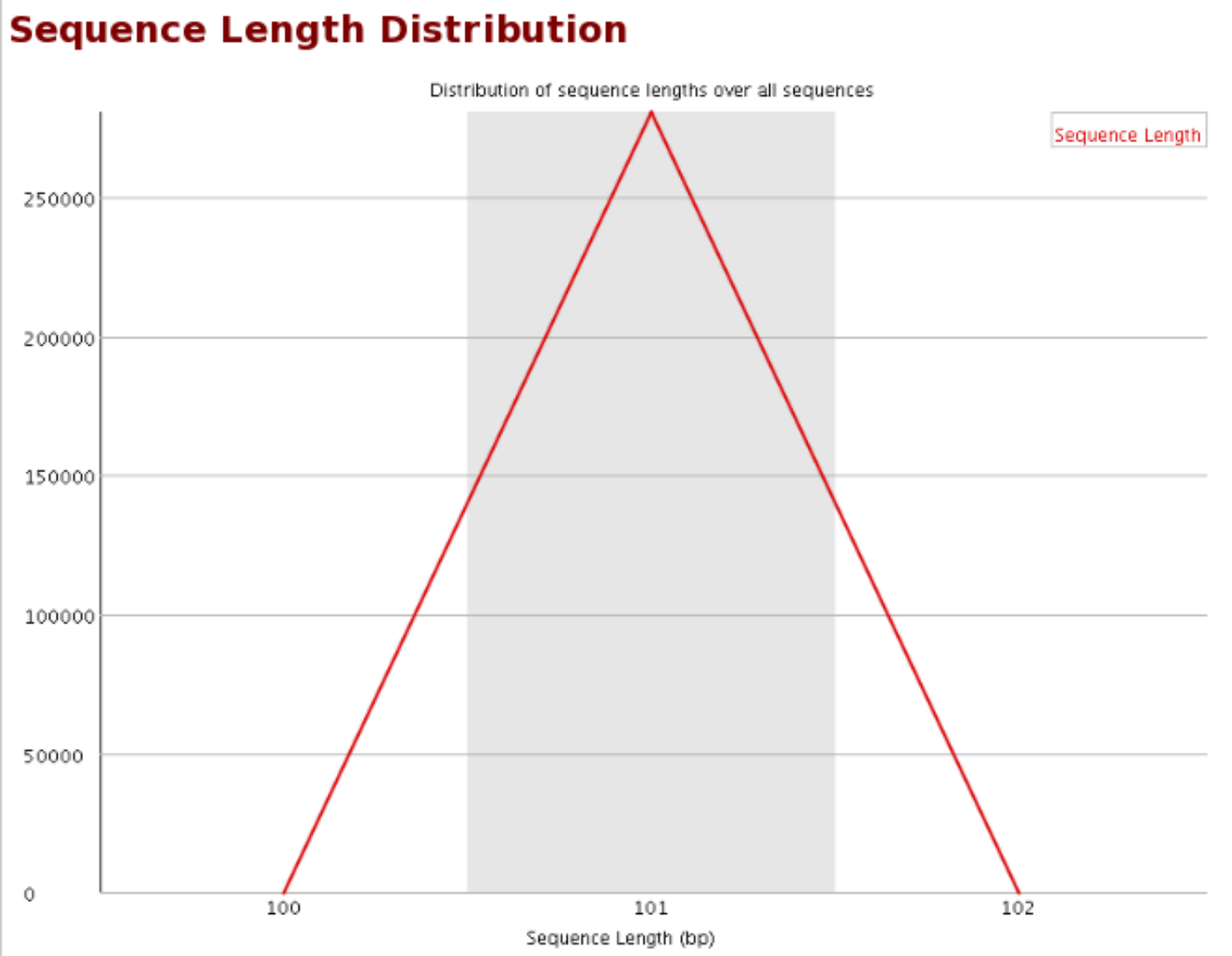
FastQC (avant)



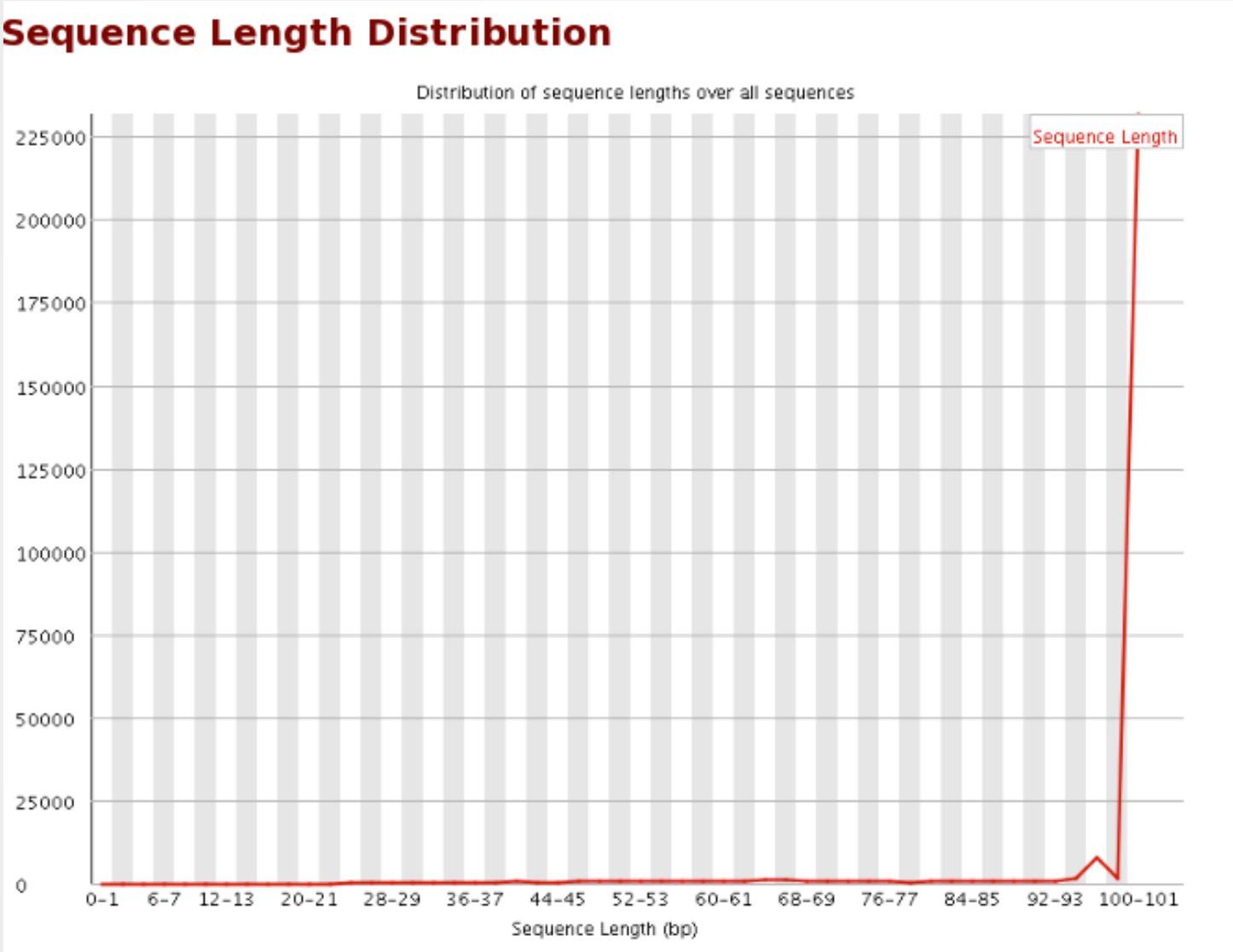
FastQC (après)



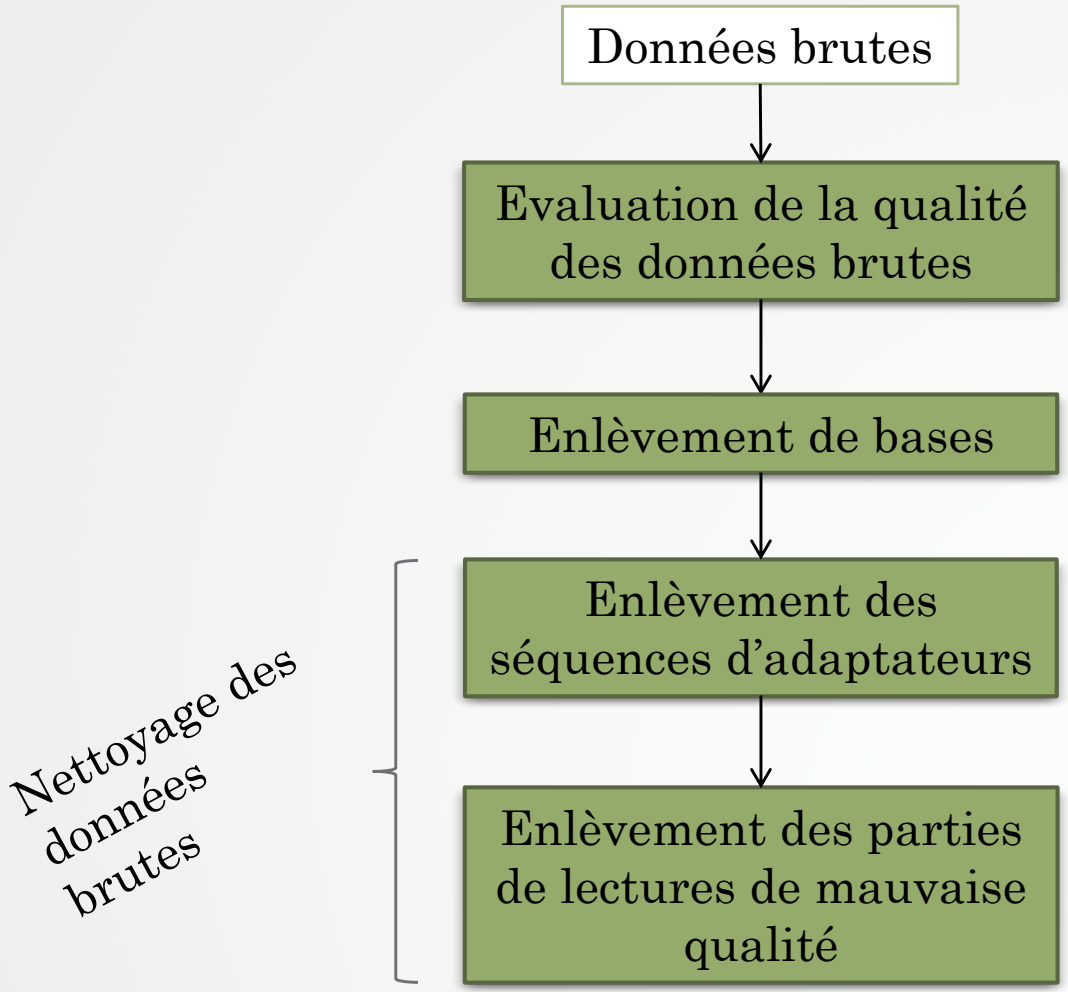
FastQC (avant)



FastQC (après)



Processus


Données brutes
Evaluation de la qualité des données brutes
Enlèvement de bases
Enlèvement des séquences d'adaptateurs
Enlèvement des parties de lectures de mauvaise qualité
Nettoyage des données brutes

Références

- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- Murray P. Cox, Daniel A. Peterson, and Patrick J. Biggs. SolexaQA: at-a-glance quality assessment of illumina second-generation sequencing data. BMC Bioinformatics , 11(1):485, September 2010. PMID:20875133.
- Cutadapt (http://code.google.com/p/cutadapt/)
- Fastx-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)