Production des données à haut débit :
Contrôle qualité de données de séquençage

Valentin LOUX
Cédric MIDOUX
valentin.loux@inra.fr cedric.midoux@irstea.fr

31 janvier 2019

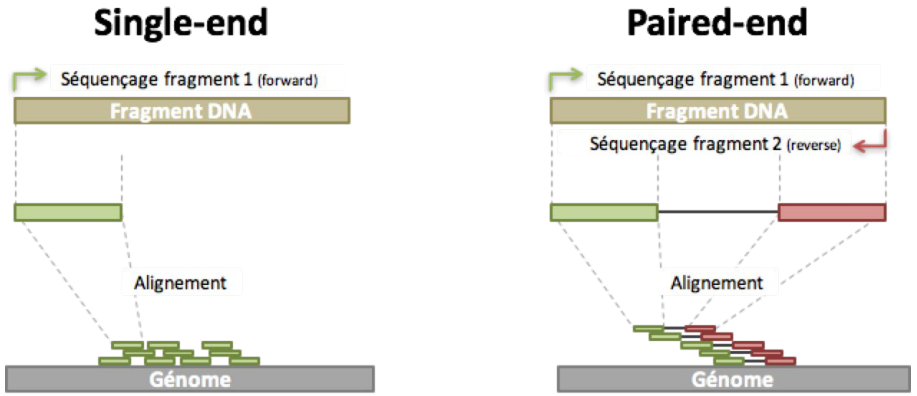
Quelques rappels

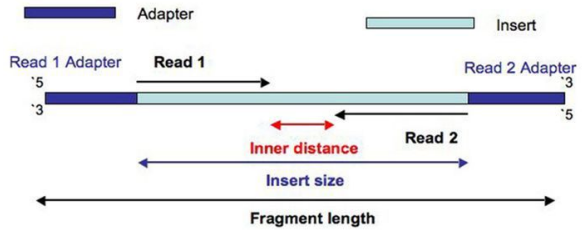
Vocabulaire

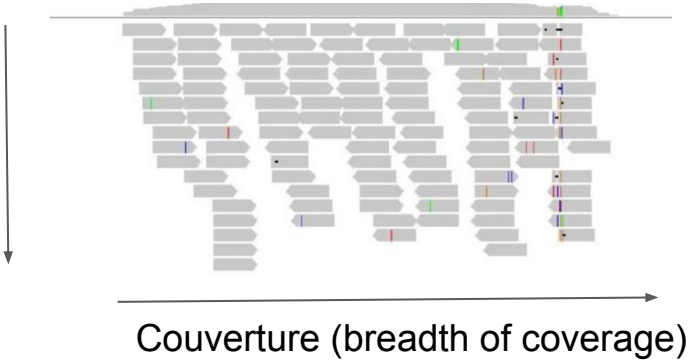
- **Read/Lecture** : morceau d'ADN séquencé
- Un fragment d'ADN = 1 ou plusieurs reads selon si le séquençage est **single end** ou **paired-end**
- **Insert** = taille du fragment
- **Profondeur = N * L / G**

  N= nbre de reads, L = longueur, G : taille genome haplotype

- **Couverture** = % du génome couvert



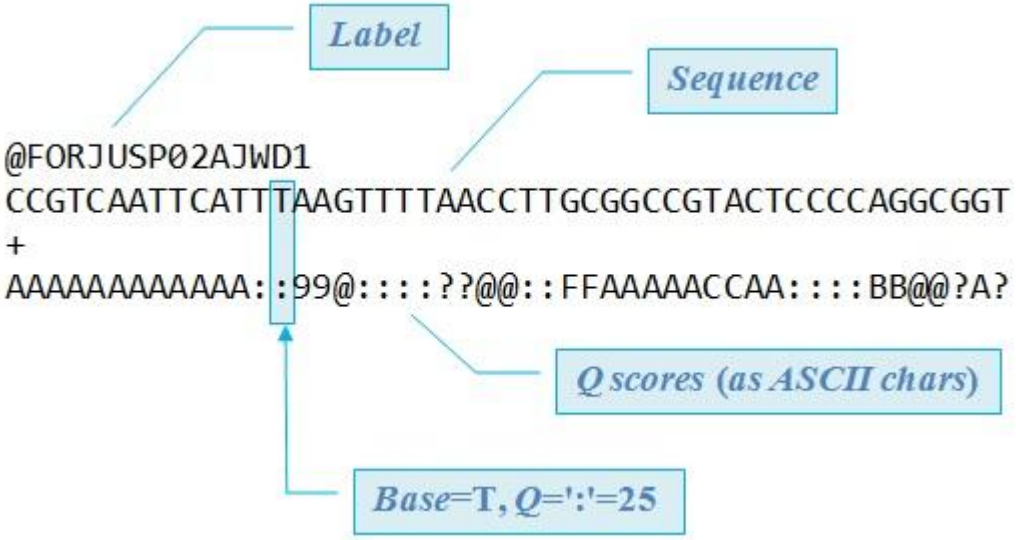




Le format de fichier FASTQ

- En bioinformatique, le format `.fastq` ou (`.fq`) est généralement utilisé pour stocker les séquences
- Fichier texte
- 1 lecture = 4 lignes
  - @ Label unique
  - Séquence
  - +
  - Score de qualité (phred)
- Compression avec gzip

  → `.fastq.gz`



Fastq : détails

```
@HWUSI-EAS1656:65:FC:8:1:14164:1001  2:N:0:ATTCCT
TGAACCAGCGGTTAATAAACAAAAATCGGAGGTTGAAGACTTTACCAAGCTGATTGATCGC
+
IIIHIIIIIIIIIIIIHIIIBIIHIIIHIFIIDIGGBDEHIHBIIIIIIIIIHBHDHEIG@
@HWUSI-EAS1656:65:FC:8:1:5620:1001  2:N:0:ATTCCT
CGGTAAAGCGAATACATTGTATTTGAAGCCCGTTCAACAAGATTTGCTGCAATATCATGAT
+
IIIHIIIIIIIIIIIIHIIIBIIHIIIHIFIIDIGGBDEHIHBIIIIIIIIIHBHDHEIG@
```

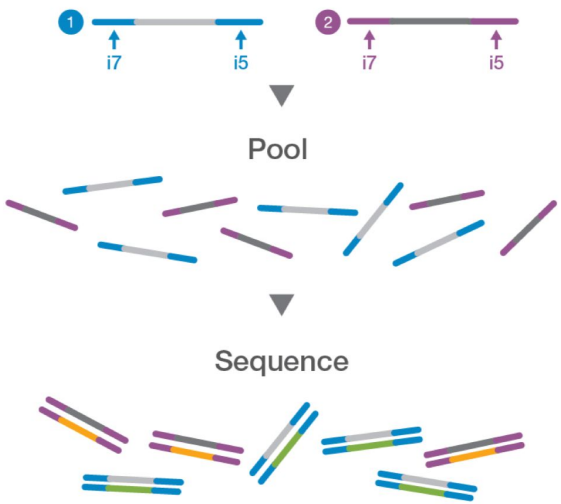
Exemple	Signification
HWUSI-EAS1656	nom de l'instrument (unique)
65	run number
FC	Flowcell Id
8	flowcell lane
1	numéro du "tile" dans la lane
14164	abscisse du cluster sur la tile
1001	ordonnée du cluster sur la tile
2	Read number (1|2 si pairé, 1 si single end)
N	is filtered (Y|N , Y si filtré)
0	Multiplexage
ATTCCT	Tag

Multiplexage / Indexes

Le multiplexage permet de passer **plusieurs échantillons** par run (diminution des coûts)

Illlumina : Kits permettant de passer de **4 à 396** échantillons suivant le type de librairie et de machines. Généralement (MiSeq) : **96 échantillons** maximum

Chaque librairie est identifié par une (ou deux) séquence de **7 nucléotides** , ce qui permet la **séparation** des échantillons après run



https://emea.illumina.com/science/education/minimizing-index-hopping.html?langsel=/fr/

En cas de séquences d'index trop proches, possibilité d' **Index Hoping qui aboutit à des mélanges d'échantillons**

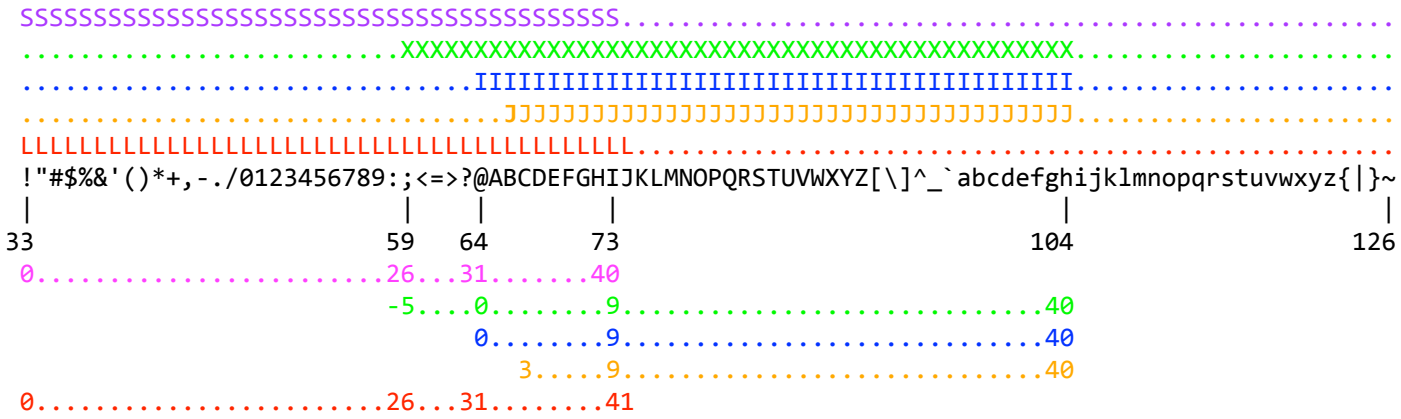
Score de qualité : Phred score

- La précision de lecture de chaque base est estimé par la machine
- Q = -10 log(P)
- P : probabilité d'erreur
- Ce score est encodé en ASCII
- Il existe plusieurs encodages suivant les technologies et générations de séquenceurs

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Encodage qualité dans FastQ



S : Sanger

X : Solexa

I : Illumina 1.3+

J : Illumina 1.5+

L : Illumina 1.8+

TP : récupération des données

- Les séquences publiées sont mises à disposition sur SRA (Sequence Reads Archive) du NCBI (ou l'équivalent à l'EMBL European Nucleotide Archive)
- On peut les télécharger facilement avec `fastq-dump` du SRA Toolkit
- `fastq-dump -h`

  Attention, par défaut , fichier fastq **unique** (interleaved) **non compréssé** !

Exercice 1

Récupérer un jeu de séquences publique

1. Se connecter au cluster de l'IFB
2. Activer le module `conda`
3. Activer l'environnement conda

   `dubii2019_m4_production_donnees`

4. Repérer le numéro d'accession correspondant à ce DOI 10.1128/MRA.01052-18 (genome de *E. coli K12*)
5. Télécharger, avec `fastq-dump`, les fichiers fastq correspondant sous forme de **reads pairés séparés et gzippés**

Attention à l'organisation de vos repertoires

Exercice 1

Récupérer un jeu de séquences publique

```
ssh login@core.cluster.france-bioinformatique.fr

module load conda

. activate dubii2019_m4_production_donnees

fastq-dump  --split-files --gzip SRR8082143
```

Et sans outil externes ?

https://www.ebi.ac.uk/ena/data/view/SRR8082143&display=html

puis `wget`

Contrôle qualité

Objectifs du Contrôle Qualité

Les séquences sont-elles conformes au niveau de prestation attendu ?

- Taille ?
- Nombre de séquences ?
- Bonne qualité ?
- Présence résiduelle d’adaptateurs ? D’indexes ?

- Les séquences peuvent-elle convenir pour répondre aux questions biologiques posées et dans quelles limites ?
- Présentent-elles des biais inattendus ?

Biais techniques à prendre en compte lors de l’analyse

fastQC

FastQC : **Quality Control for FastQ files** :

- **Contrôle qualité** des séquences générées.
- Utilisable via une interface graphique ou en **ligne de commande.**
- Rapport de contrôle de qualité qui peut permettre de repérer des problèmes venant du **séquenceur ou des librairies**.
- **Rapport HTML** construit à partir de différents modules d'analyse spécialisés.
-

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

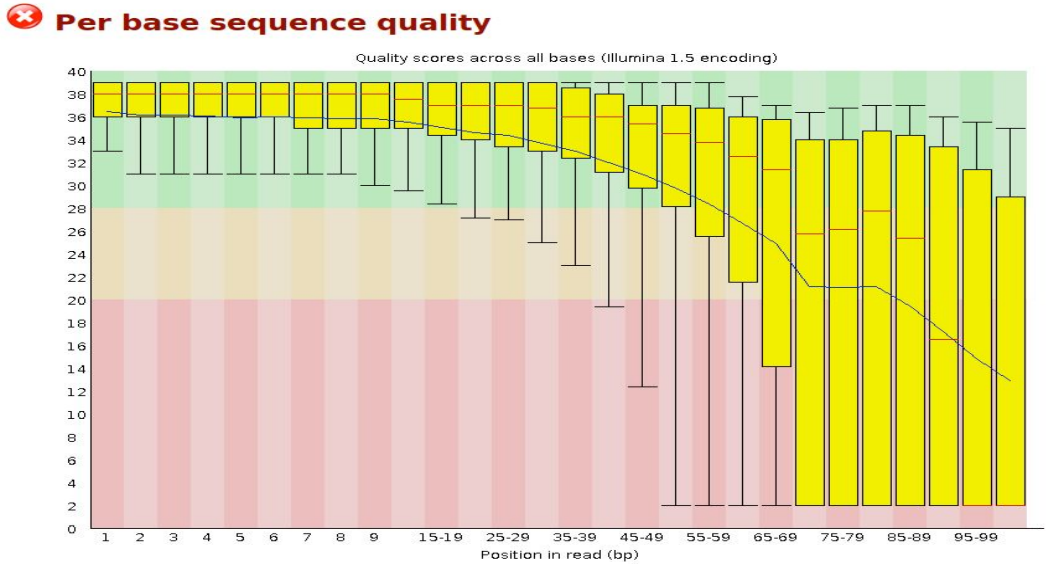
fastQC - Statistiques

Basic Statistics

Measure	Value
Filename	B02EWABXX_R1_s_8_sequence.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	134247743
Filtered Sequences	0
Sequence length	101
%GC	45

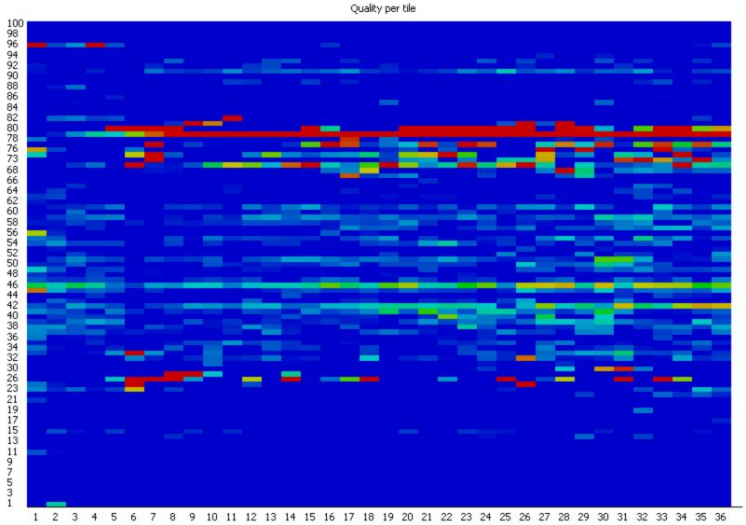
Filename: Nom du fichier que vous avez analysé.
File type: Nucleotide space ou color space, converti ensuite en nucleotide space.
Encoding: Format d'encodage de la qualité.
Total Sequences: Nombre total de séquences analysées.
Filtered Sequences: Lorsque le mode "Casava" est utilisé, il s'agit du nombre de séquences supprimées et non analysées.
Sequence Length: Taille de la plus courte et la plus longue séquence. Si toutes les séquences ont la même taille, seule une taille est donnée.
%GC: Le GC% total.

fastQC - Qualité par base



Ligne rouge	qualité médiane
Barre jaune	inter-quartile (25%-75%)
Moustaches	premier et dernier déciles
Ligne bleue	qualité moyenne

fastQC - Qualité par tuile



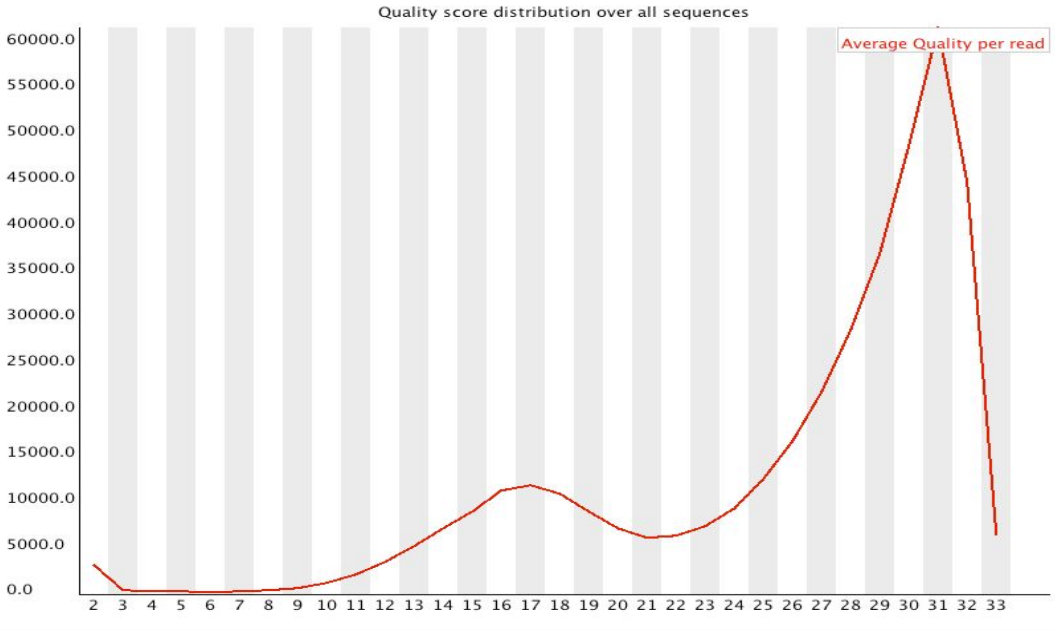
Une tuile (~position) a t'elle un comportement atypique ?

La couleur indique la **déviation** par rapport à la qualité moyenne du run.

Couleurs froides <= qualité moyenne au cycle

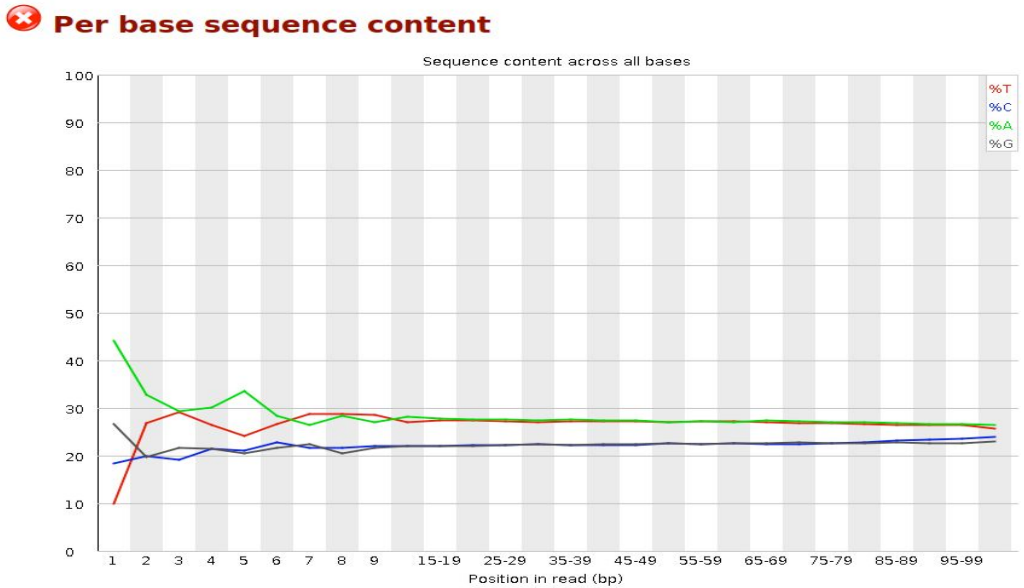
Couleurs chaudes >= qualité moyenne au cycle

fastQC - Qualité moyenne des séquences



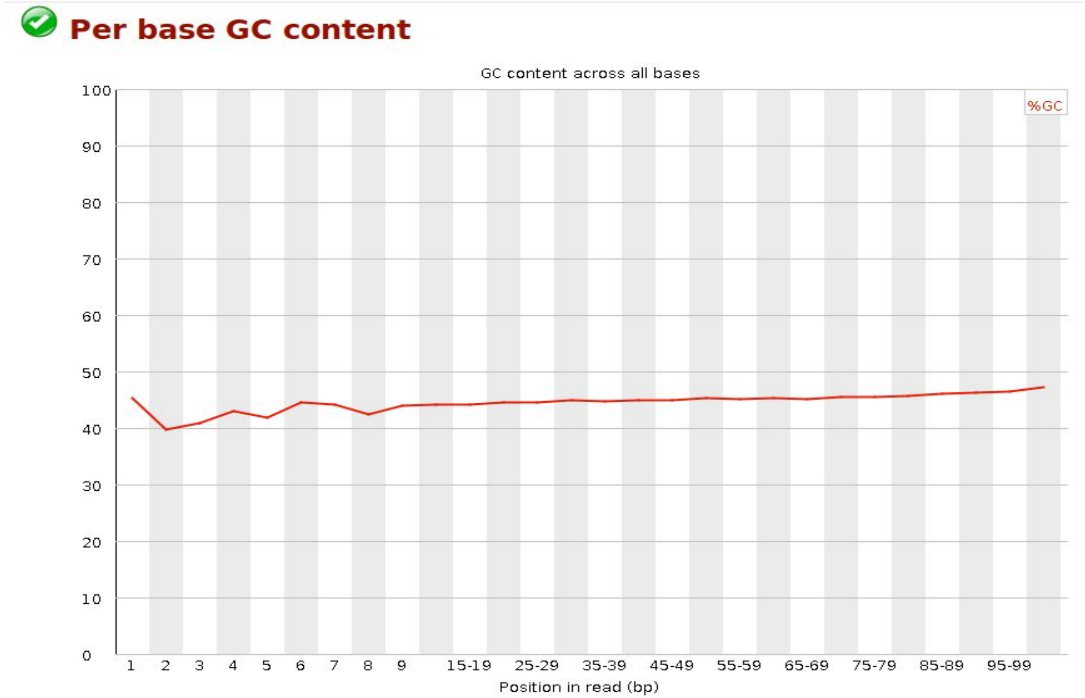
Un sous-ensemble de lectures a t'il une **qualité faible** ? Ce peut être dû à une mauvaise acquisition du signal

fastQC - contenu par base



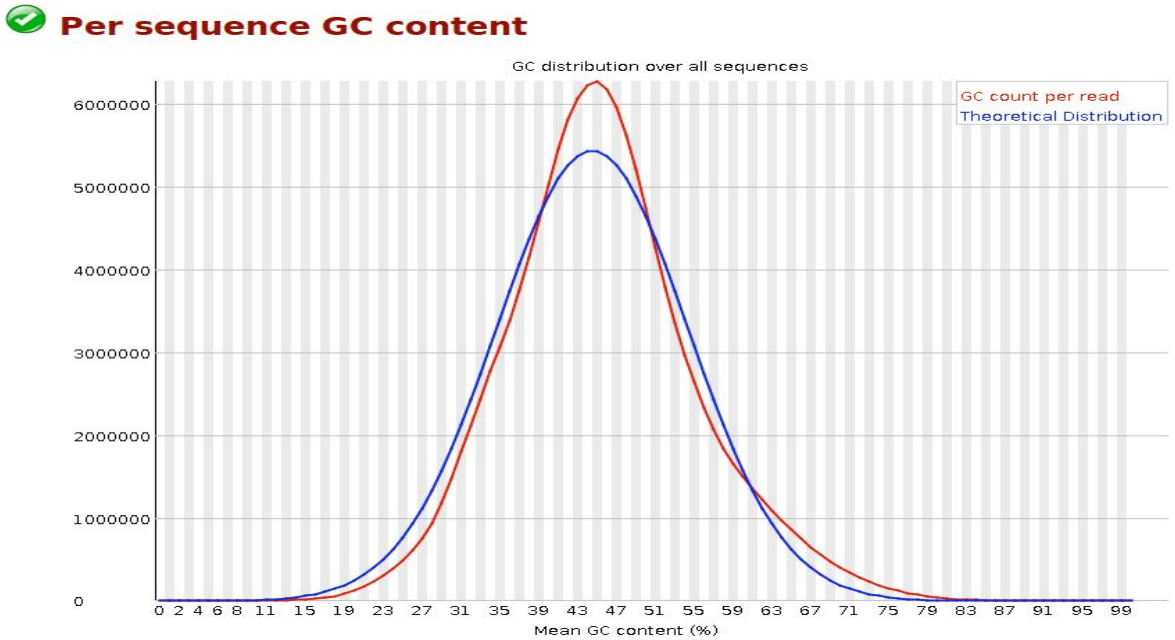
Proportion des quatre bases sur chaque position des lectures.
La proportion relative des bases doit suivre celle du génome.
Un biais important indique en général une séquence **sur-représentée** qui contamine votre librairie.

fastQC - GC% par base



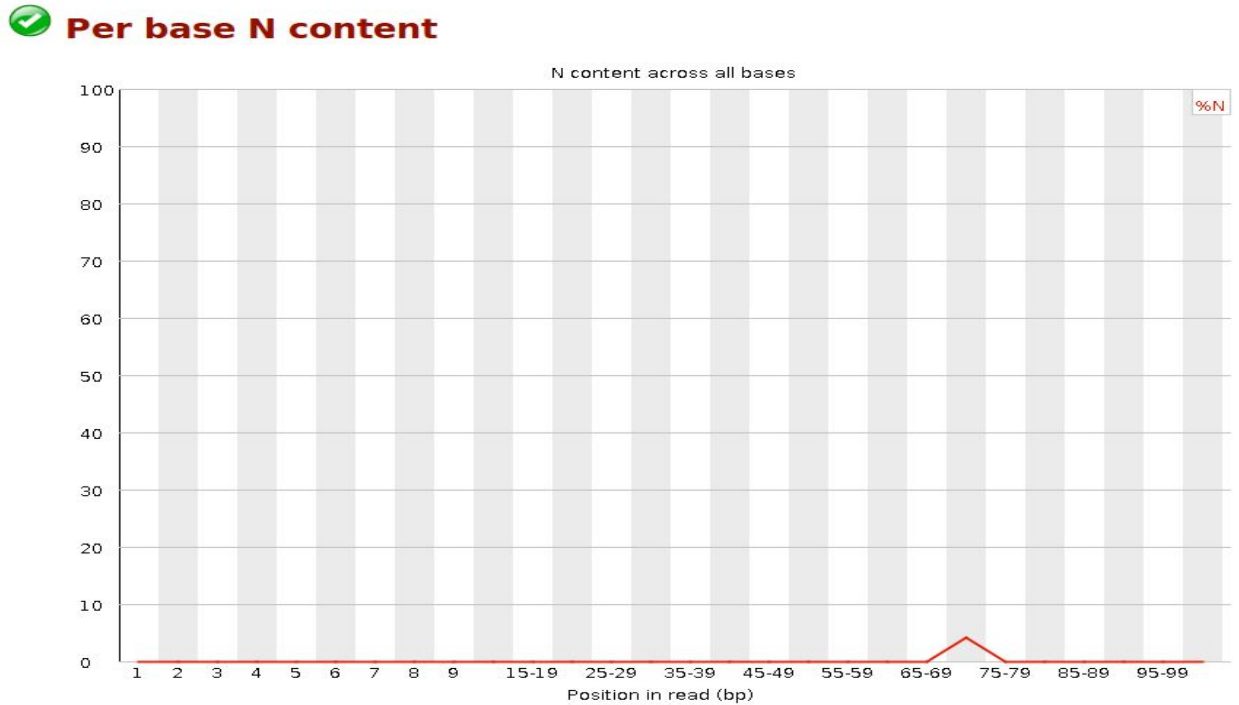
Une librairie réellement aléatoire donne une courbe avec **peu de fluctuations**.
Le GC% doit refléter celui du génome.

fastQC - GC% par séquence



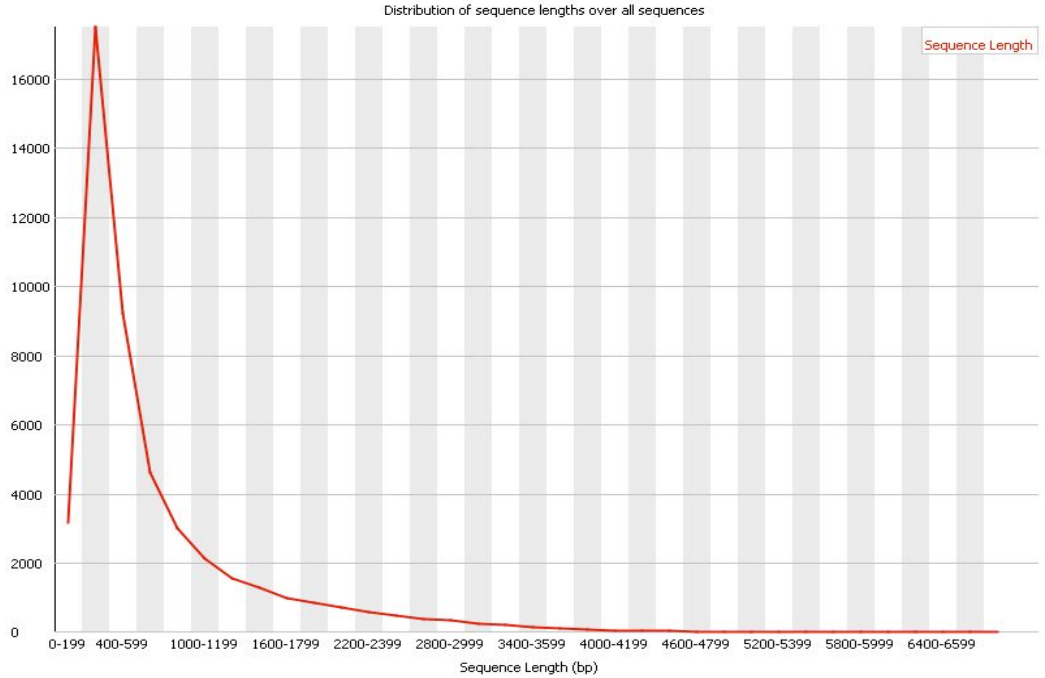
La courbe donne le **contenu en GC** calculé sur l'intégralité des lectures, et donne une comparaison avec la modélisation d'une distribution de GC selon une loi normale.

fastQC - %N par séquence



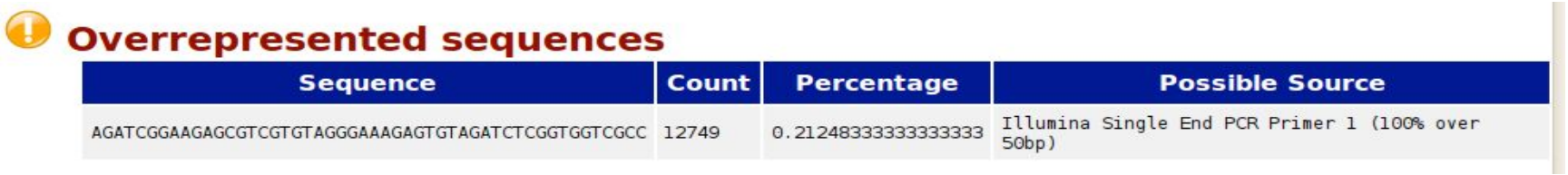
Pourcentage de N le long des lectures. Peut être un signe d'un problème sur un cycle particulier ?

fastQC - taille des séquences



En Illumina, **toutes les séquences d'un run** font la même taille.

fastQC - Séquences sur-représentées

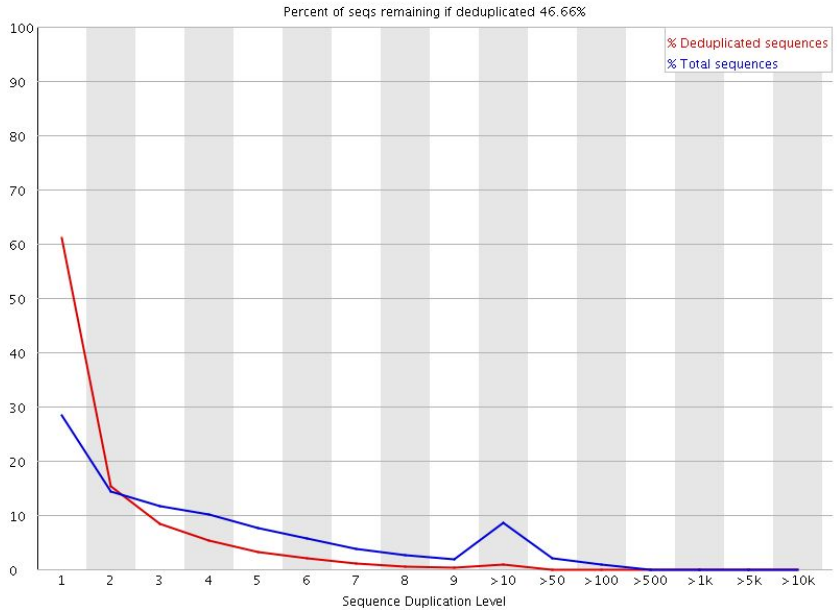


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCC	12749	0.21248333333333333	Illumina Single End PCR Primer 1 (100% over 50bp)

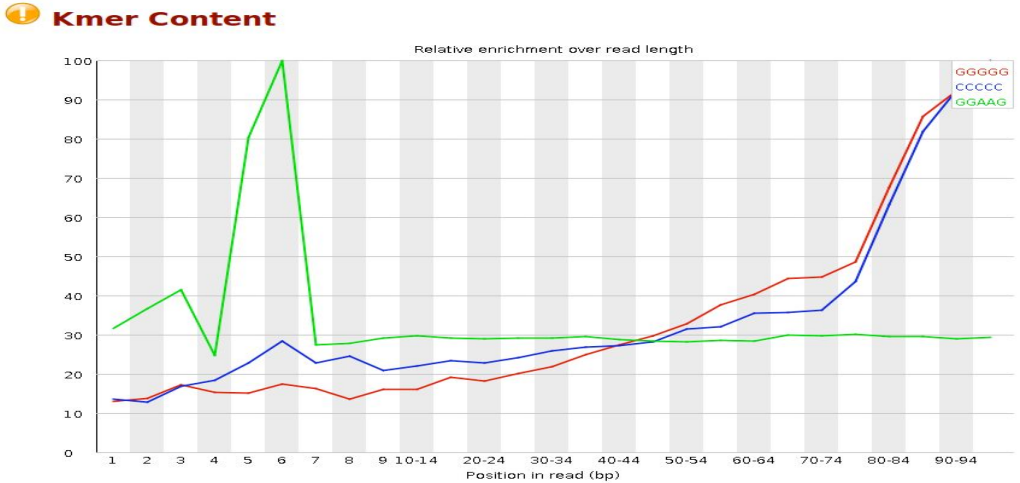
Séquences exactes (parmis les 100 000 premières lectures) représentant **plus de 0,1%** du total des lectures.
Ces séquences sont également comparées à une **base de données d'amorces**.

fastQC - Séquences dupliquées



Le graphique donne le nombre relatif de séquences (**50 bp sur les 10 000 premières séquences**) dupliquées dans l'ensemble du jeu de données.
En bleu, le pourcentage de séquence dupliquées
En rouge, le pourcentage une fois les séquences dé-dupliquées
Un faible niveau de duplication peut indiquer un grande profondeur de séquençage, un haut niveau un problème lors de la préparation de la banque (amplification …)

fastQC - Analyse en k-mers

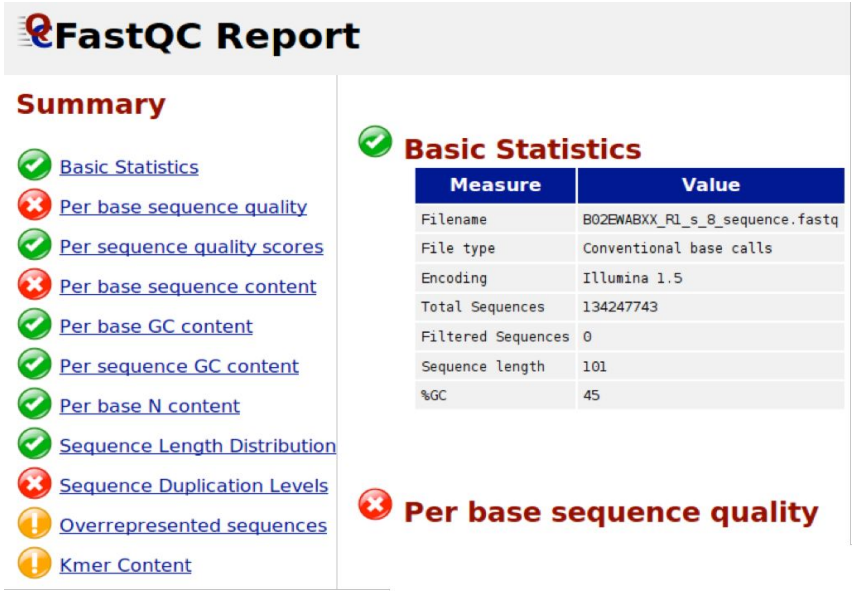


Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
GGGGG	26794195	3.5409238	9.029529	95-97
CCCCC	18119795	2.396203	6.178688	95-97
GGAAG	17810150	1.5521879	5.0448737	6

Le graphique donne les **7-mers sur-représentés**, ainsi que leur position dans les séquences.

fastQC - Rapport synthétique

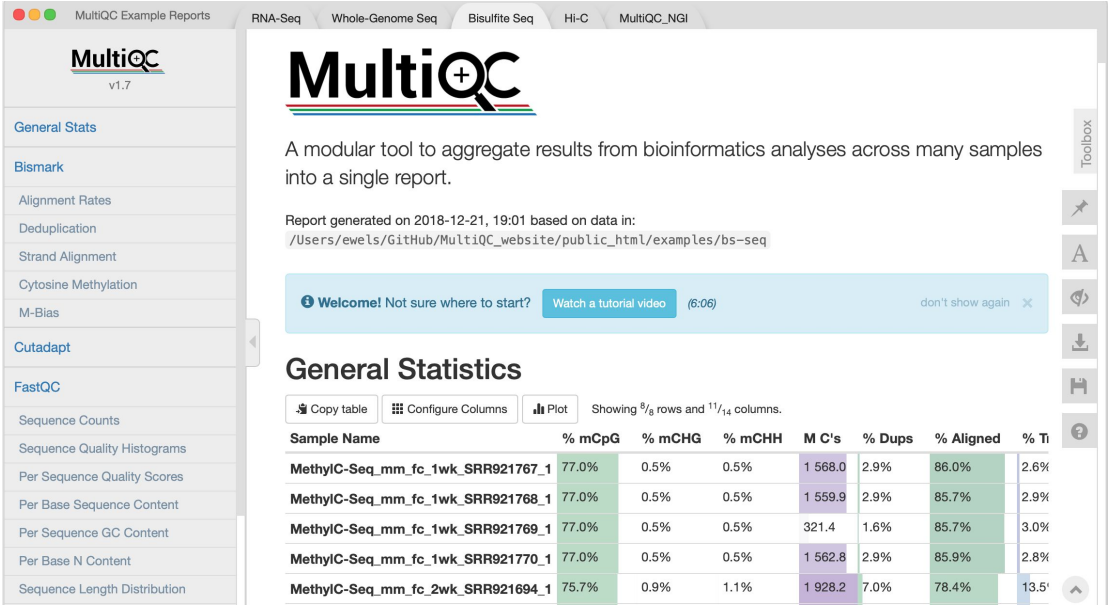
Vert	normal
Orange	légèrement anormal
Rouge	inattendu

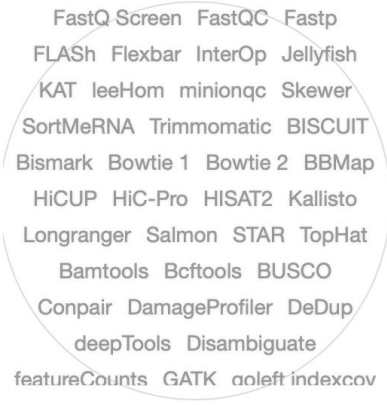


Les évaluations supposent une **distribution homogène des séquences.** Ceci peut ne pas correspondre à votre plan expérimental (e.g. RNA-Seq).

multiQC

- `FastQC` produit un rapport par librairie séquencée (deux si paired-end)
- `MultiQC` permet d'**agréger ces rapports**
- **Agrégation** de rapports de 72 autres outils





MultiQC: Summarize analysis results for multiple tools and samples in a single report
Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller
Bioinformatics (2016)
doi: 10.1093/bioinformatics/btw354
PMID: 27312411

TP : QC d’un jeu de données Illumina

Appliquer au jeu de données téléchargé précédemment les étapes de contrôle qualité.

Exercice 2

Analyses fastqc

1. Donner le taille des fichiers, le nombre de lignes (**sans les décompresser**)
2. Observer le début des fichiers téléchargés (**sans les décompresser**)
3. Réaliser un `fastqc` sur les séquences téléchargées
4. Regrouper les analyses avec `multiqc`
5. Commenter les **métriques**

Exercice 2

Analyses fastqc

```
ls -alh SRR8082143_?.fastq/gz
```

```
zcat SRR8082143_1.fastq.gz|wc
-l
```

```
zcat
SRR8082143_1.fastq.gz|head
```

```
fastqc SRR8082143_?.fastq/gz
```

```
multiqc .
```

Improved genome sequencing using an engineered transposase (fig 1)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5240201/

Interlude :
Quelles questions poser à son prestataire avant tout séquençage ?

Nettoyage qualité - Preprocess

Preprocess

- Après avoir étudié les reads brut, on cherchera ne conserver que ceux de qualité suffisante

- Objectifs :
    - Retirer les adapteurs présents
    - Filtrer les reads de mauvaise qualité / trop court

Retirer les adapteurs présents – `cutadapt`

- Points d'attention particulier:
  - Que faire des reads sans adapteurs ?
  - Comment gérer des paires de reads ?
  - La séquence (et l'orientation) des adapteurs
  - La tolérance au missmatch
- `cutadapt --help`
- `cutadapt --length <INT> -o <OUT_FILE> <IN_FILE>`
- `cutadapt -a <ADAPTER> [options] [-o <OUT_FILE>] <IN_FILE>`
- `cutadapt -a <ADAPTER> -g <ADAPTER> --match-read-wildcards --discard-untrimmed -o <OUT_FILE_R1> -p <OUT_FILE_R2> <IN_FILE_R1> <IN_FILE_R2>`

Retirer les adapteurs présents – `trimmomatic`

- Points d'attention particulier:
  - Que faire des reads sans adapteurs ?
  - Comment gérer des paires de reads ?
  - La séquence (et l'orientation) des adapteurs
  - La tolérance au missmatch
- `trimmomatic -h`
- `trimmomatic SE <inputFile> <outputFile> <trimmer1>...`
- `trimmomatic PE [-validatePairs] [-basein <inputBase> | <inputFile1> <inputFile2>] [-baseout <outputBase> | <outputFile1P> <outputFile1U> <outputFile2P> <outputFile2U>] <trimmer1>...`

Filtrer les reads – `sickle`

- Points d'attention particuliers:
  - Métriques des filtres (Qualité ? Longueur ?)
  - Comment gérer des paires de reads ?
- `sickle se --help` ou `sickle pe -help`
- `sickle se [options] -f <fastq sequence file> -t <quality type> -o <trimmed fastq file>`
- `sickle se -q 20 -l 50 -n -g -t <solexa | illumina | sanger> -f <INPUT.fastq> -o <OUTPUT.fastq.gz>`
- `sickle pe -q 20 -l 50 -n -g -t <solexa | illumina | sanger> -f <INPUT_R1.fastq> -r <INPUT_R2.fastq> -o <OUTPUT_R1.fastq.gz> -p <OUTPUT_R2.fastq.gz> -s <OUTPUT_single.fastq.gz> # separate files`
- `sickle pe [options] -c <INTPUT_interleaved.fastq> -t <quality> -M <OUTPUT_interleaved.fastq> # one file with interleaved reads`

Exercice 3

Preprocessing

1. Réaliser le nettoyage qualité des données
   - primer1 : CTGTCTCTTATACACATCT
   - primer2 : AGATGTGTATAAGAGACAG
   - length > 50bp
   - quality > 20
2. Analyser ses effets grâce à un contrôle qualité après nettoyage

Préprocess “tout-en-un” – `fastp`

- Permet d’enchainer les étapes suivantes :
  - Contrôle qualité
  - Trimming des adapteurs (sequences les plus courantes incluses)
  - Filtre qualité
  - Contrôles qualité post-preprocess
- `fastp –help`
- `fastp -i <in.fq> -o <out.fq>`
- `fastp -i <in_R1.fq.gz> -I <in_R2.fq.gz> -o <out_R1.fq.gz> -O <out_R2.fq.gz>`
- `fastp -i <in_R1.fq> -I <in_R2.fq> -o <out_R1.fq> -O <out_R2.fq> -l <LENGHT> -q <QUALITY> --thread <THREADS> --html <REPORT.html>`

Exercice 4

fastp

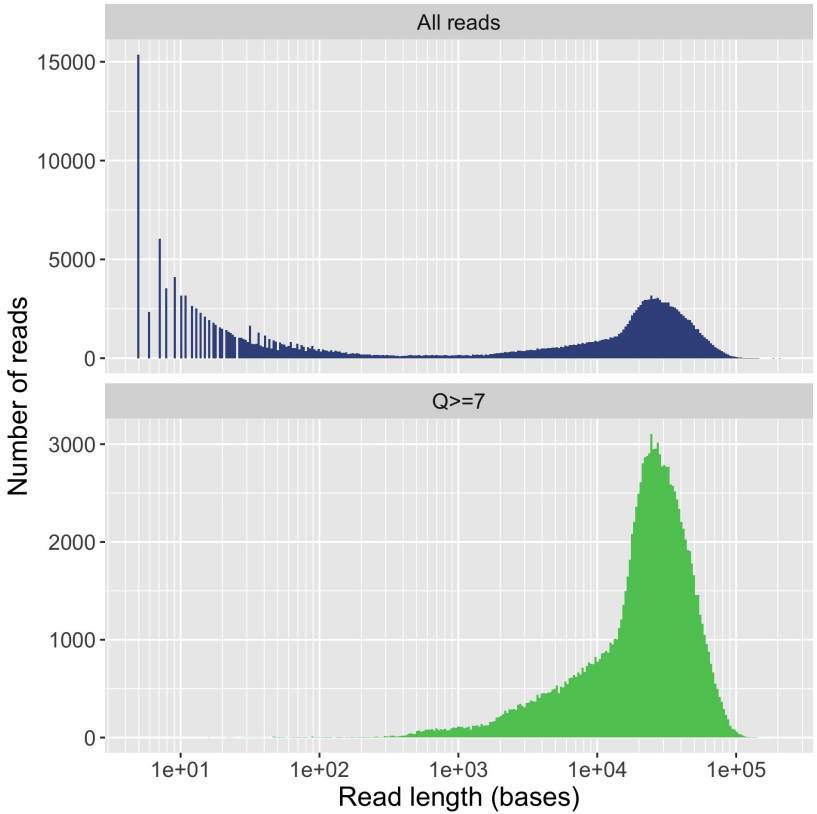
1. Réaliser le nettoyage qualité des données avec `fastp`
2. Commentez le rapport html

5. Spécificités des *long reads*

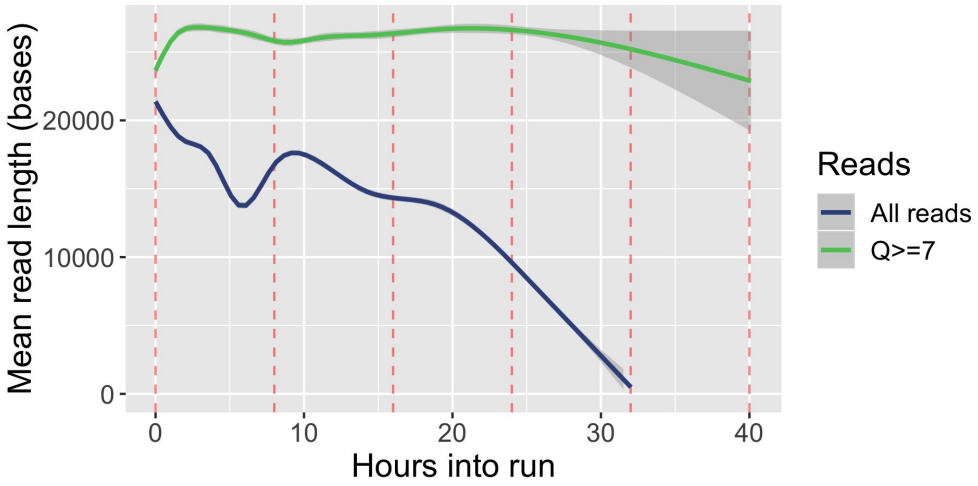
Spécificités des *long reads* (PacBio, Nanopore, …)

- Les outils de contrôle qualité traditionnel ne sont pas dimensionnés pour les *long reads*
- Des métriques sont regroupées dans `sequencing_summary.txt` lors du base-calling
- Visualisation grâce à des scripts R tel que `MinIONQC.R`

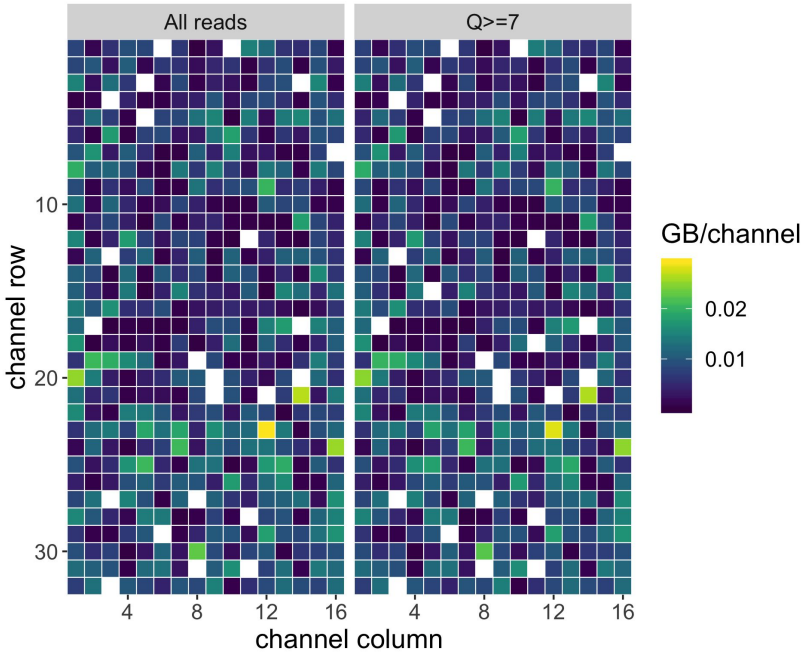
- Pour plus d'info voir le GitHub du projet



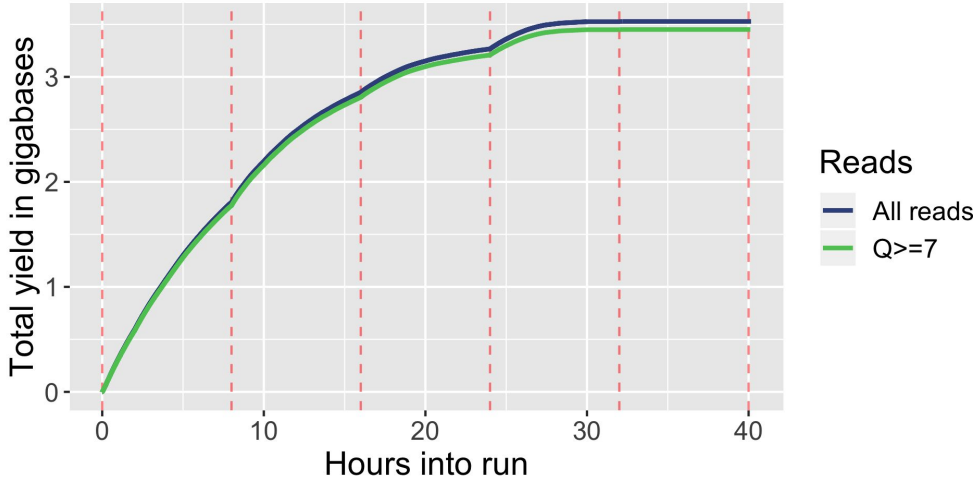
length histogram



length by hours



gb per channel



yield over time

Conclusion

- QC : outils de stats descriptives permettant une évaluation **rapide** de la qualité intrinsèque du séquençage.

- Pre-processing : étape indispensable de toute analyse de données.
- Pour les reads courts :
  - filtrage adaptateur + nettoyage
- Pour les reads longs:
  - pre-processing généralement fait de façon conjointe avec l’assemblage ou le mapping

Références

- SRA Toolkit : Site - GitHub
- fastqc : Andrews, 2010 - Site
- multiqc : Ewels, 2016 - Site - GitHub
- cutadapt : Martin, 2011 - Site - GitHub
- sickle : Joshi, 2011 - GitHub
- fastp : Chen, 2018 - GitHub