Food and Agriculture Organization of the United Nations

Module 1

Contrôle qualité basique des lectures brutes

Mushal Allam

Institut national des maladies transmissibles, Afrique du Sud

Clause de non-responsabilité

La traduction de cette présentation de l'anglais vers le français a été effectuée par la FAO en collaboration avec le Centre de référence de la FAO pour la résistance aux antimicrobiens au Danemark, à savoir l'Institut national de l'alimentation de l'Université technique du Danemark (DTU), dans le cadre du Projet mondial pour la sécurité sanitaire de la FAO sur la surveillance des maladies infectieuses et les systèmes d'alerte rapide financé par the United States Agency for International Development (USAID) (Agence des États-Unis pour le développement international).

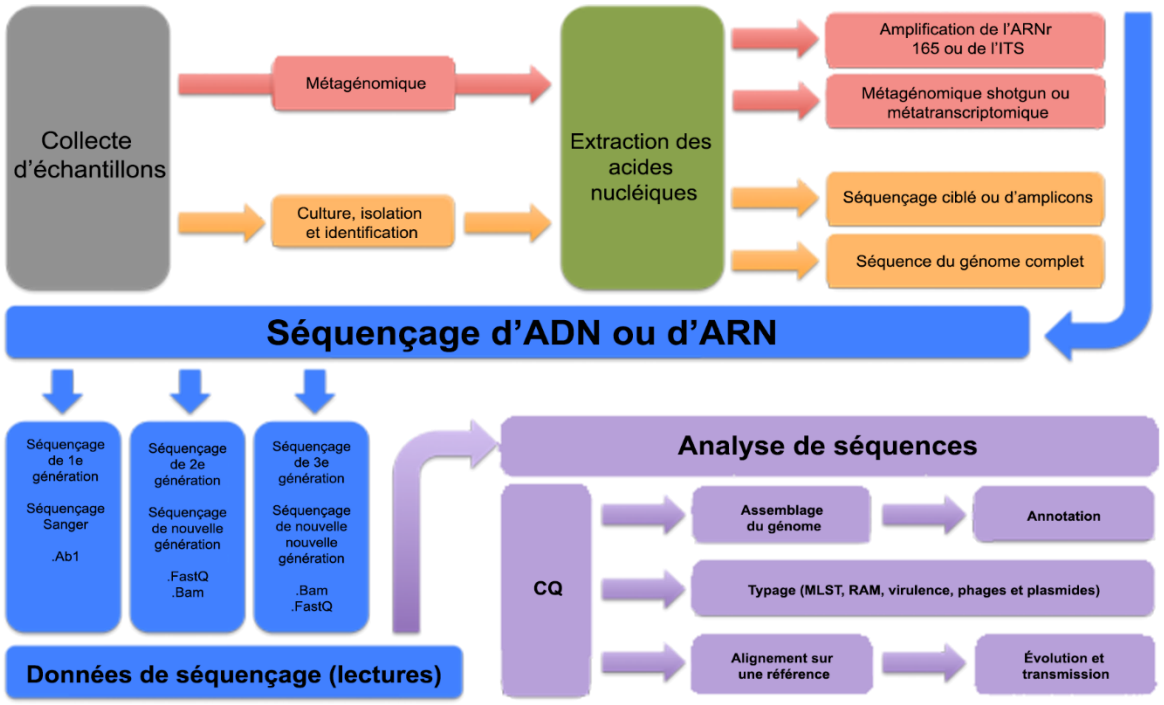
Le contenu original en anglais de cette présentation a été élaboré par l'Institut national de l'alimentation de l'Université technique du Danemark (DTU) dans le cadre du Fleming Fund Regional Grant SeqAfrica financé par Département de la Santé et de la Protection sociale du Royaume-Uni.

Introduction

- Les séquenceurs modernes à haut débit peuvent produire des millions de séquences en un seul run.

- Avant d'analyser une séquence pour en tirer des conclusions biologiques, vous devez toujours effectuer un contrôle qualité (CQ) simple pour vérifier l'exactitude des données brutes et éviter les problèmes ou les biais dans vos données, qui sont susceptibles d'avoir un impact sur l'utilisation de ces dernières.

- La plupart des séquenceurs créent un rapport de CQ lors du processus d'analyse, mais il se limite généralement aux problèmes générés par le séquenceur lui-même.



Format FASTQ

- Le format FASTQ est depuis longtemps devenu la norme pour stocker les données de séquençage à lectures courtes.
- Dans le cas d'un processus à **lecture unique**, un fichier FASTQ Read1 (R1) est créé pour chaque échantillon de chaque piste sur support solide ("flow cell").
- Pour un processus **en paire**, un fichier FASTQ R1 et un fichier Read2 (R2) sont créés pour chaque échantillon de chaque piste.
- Les fichiers FASTQ contiennent les séquences de nucléotides et les scores de qualité des bases de millions de lectures.
- Les fichiers FASTQ sont créés et compressés avec l'extension *.fastq.gz.
- Bien que la taille du fichier dépende du nombre de lectures, les fichiers FASTQ sont généralement volumineux (de l'ordre de plusieurs mégaoctets et gigaoctets).
- La plupart des outils qui prennent en charge les fichiers FASTQ peuvent les traiter au format compressé. Par conséquent, afin d'économiser de l'espace de stockage, il est recommandé de ne pas les décompresser.

TV17_S12_L001_R1_001.fastq.gz

- **TB17:** nom de l'échantillon
- **S12:** ordre de l'échantillon dans la feuille d'échantillon
- **L001:** numéro de piste
- **R1:** Read 1
- **001:** dernier segment
- **.fastq:** fichier FASTQ
- **.gz:** compressé

- Le fichier contient des millions d'enregistrements
  - Chaque enregistrement comprend quatre lignes et représente UNE séquence
    - Ligne 1 – le nom, commence par @
    - Ligne 2 – la séquence, **commence à la ligne**
    - Ligne 3 – autres éléments facultatifs, commence par +
    - Ligne 4 – les scores de qualité, **commence à la ligne**

Base = T
Score = F = 37

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 1:N:0:12
CCTAAATGGTGCCATGCTAGGAGGCCGTGCCCTTCTTGAAAAGTTGTATGTGAA
+
BBBFFFFFFFBFFFIIIIFI<FFIIIIIFIIIIFBFIIIIIIIIIFFFIIIIFIII
```

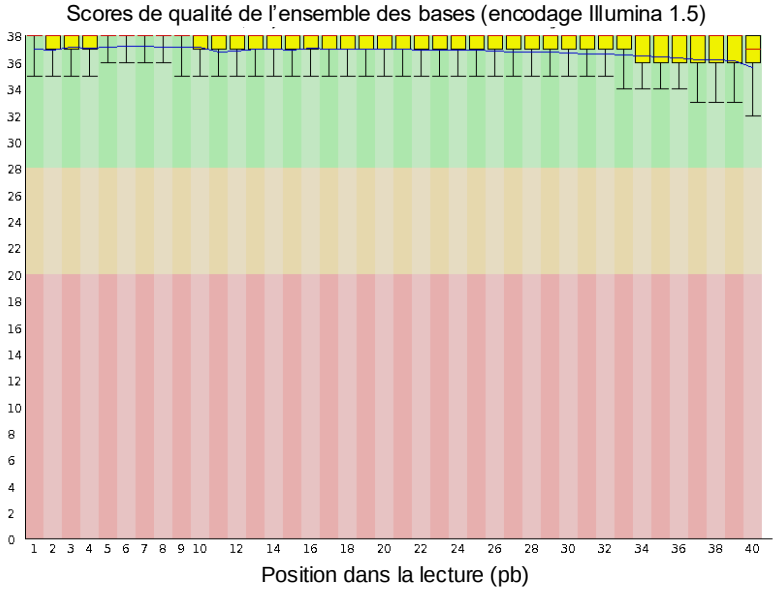
Symbole	Code ASCII	Score Q	Symbole	Code ASCII	Score Q	Code ASCII	Score Q	Codigo ASCII
!	33	0	/	47	14	^	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

Score de qualité Phred

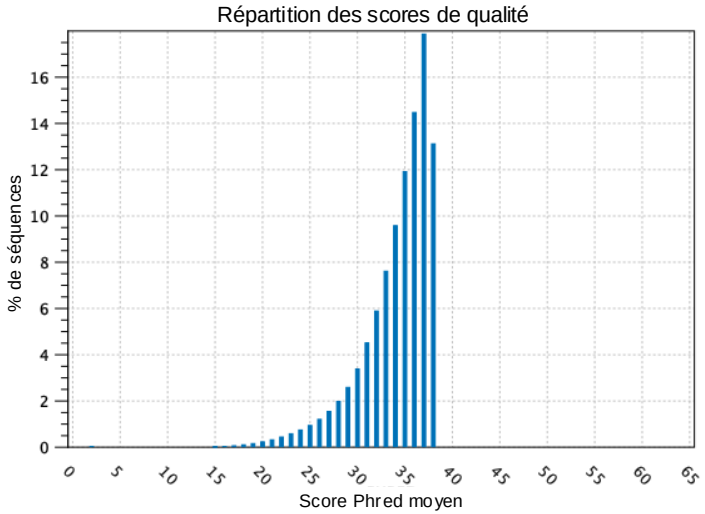
- Le score de qualité Phred (Q) est une mesure de la précision de l'identification des bases azotées dans le cadre du séquençage automatique de l'ADN.

- Si le score de qualité Phred d'une base est de 30 (Q30), la probabilité que cette base soit identifiée incorrectement est de 1 sur 1 000.

Score de qualité Phred	Probabilité d'identification incorrecte de la base	Précision d'identification des bases
10	1 sur 10	90 %
20	1 sur 100	99 %
30	1 sur 1 000	99,9 %
40	1 sur 10 000	99,99 %
50	1 sur 100 000	99,999 %
60	1 sur 1 000 000	99,9999 %



2.4 Répartition des scores de qualité



Nettoyage FASTQ

- Nettoyage qualité basé sur les scores de qualité
  - **≥ Q20**
- Suppression d'ambiguïté pour retirer, p. ex., des portions de N
  - ~~NNNNNNNNNNNN~~TACTACATAGATACAGATACTTA~~NNNNNN~~
- Suppression d'adaptateurs
  - CTGTCTCTTATACACATCTTACTACATAGATACAGATACTTACTGTCTCTTATACACATCT
- Suppression de bases pour retirer un nombre spécifique de bases à l'extrémité 3' ou 5'
  - ATACACATCTTACTACATAGATACAGATACTTACTGTCTCTTA
- Suppression de longueur pour retirer des lectures plus courtes ou plus longues que la limite définie
  - TACTACATAGA **≤ 15**
  - TACTACATAGA **≥ 1 000**

Trimm-omatic

PRINSEQ

FASTX toolkit

FASTX-Toolkit
FASTQ/A short-reads pre-processing tools

Babraham Institute
Babraham Bioinformatics

Trim Galore

FastQC

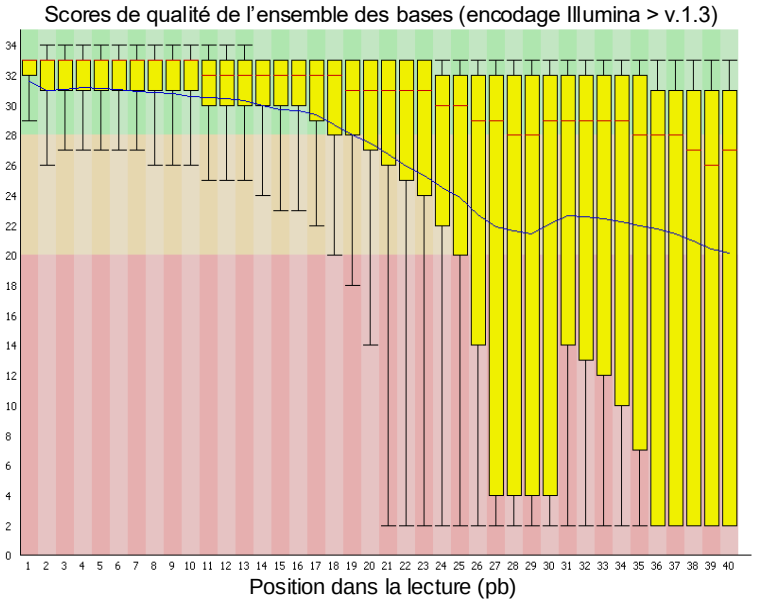
- FastQC produit un rapport de CQ identifiant les problèmes qui proviennent soit du séquenceur, soit du matériel de la banque de départ.
- FastQC peut être utilisé dans l'un des deux modes suivants:
  - Il peut être utilisé comme une application interactive autonome pour l'analyse immédiate d'un nombre limité de fichiers FASTQ
  - ou il peut être utilisé en mode non interactif en vue d'être intégré dans un processus d'analyse plus large pour le traitement systématique d'un grand nombre de fichiers.
- FastQC produit les résultats au format HTML.

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

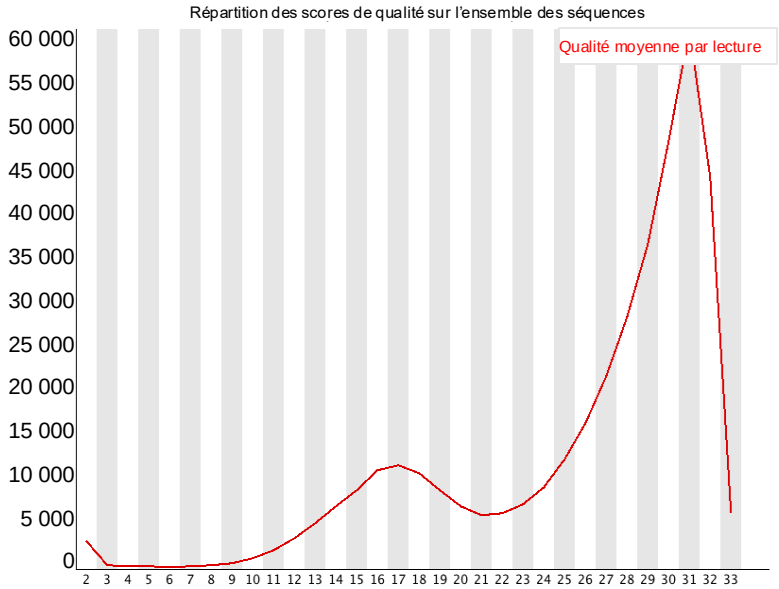
FastQC: Modules d'analyse

Mesure	Valeur
Nom du fichier	Good_sequence_short.txt
Type de fichier	Identification conventionnelle des bases
Encodage	Illumina 1.5
Total des séquences	250 000
Séquences identifiées comme de mauvaise qualité	0
Longueur des séquences	40
% GC	45

Statistiques de base



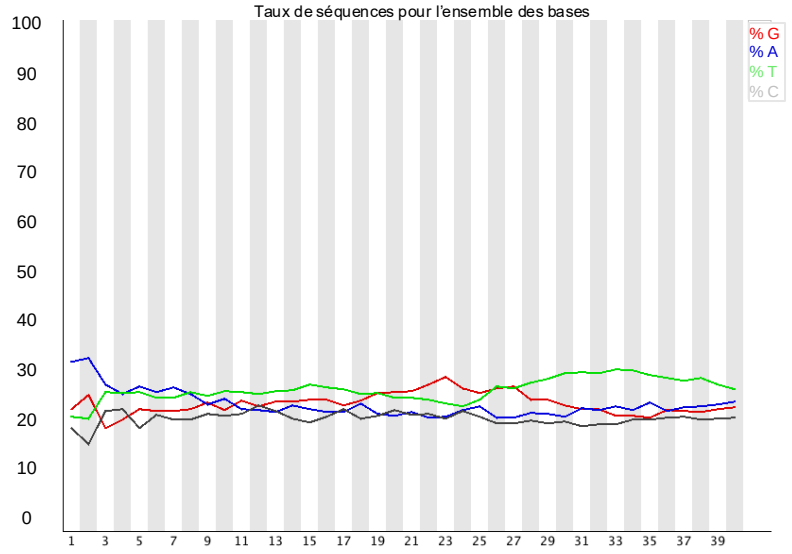
Qualité des séquences par base



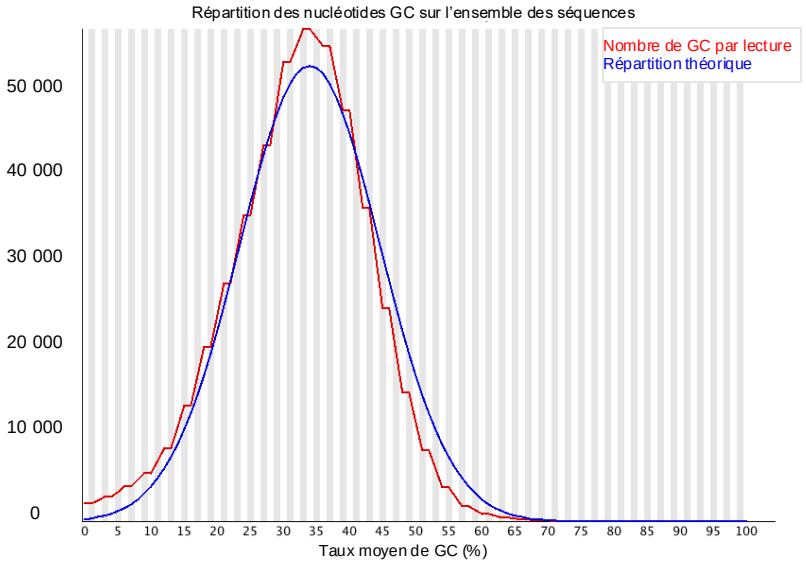
Scores de qualité par séquence

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

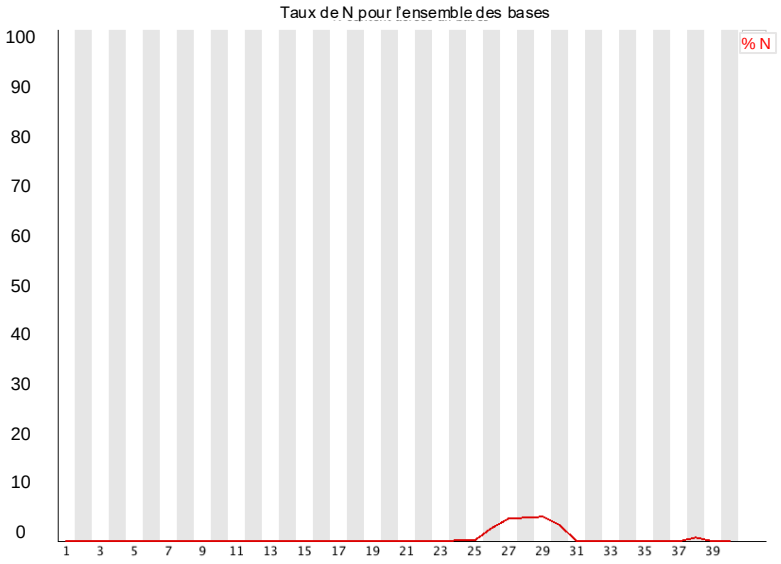
FastQC: Modules d'analyse



Taux de séquences par base



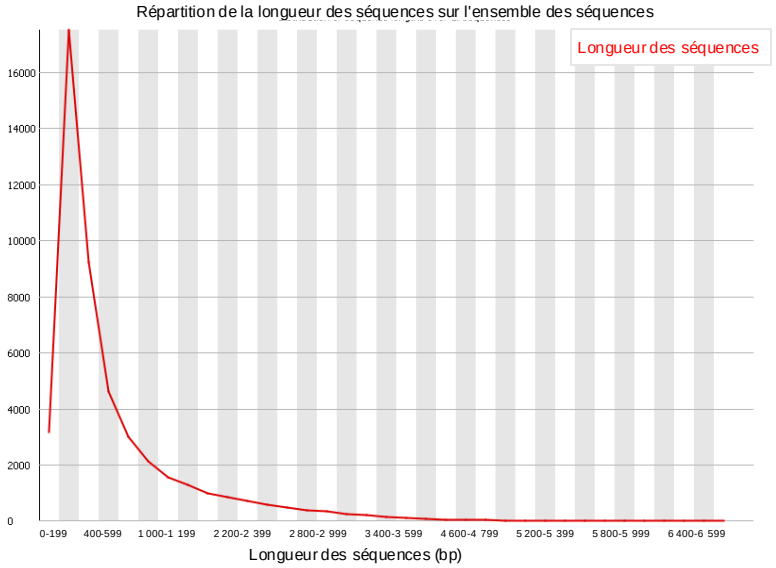
Taux de GC par séquence



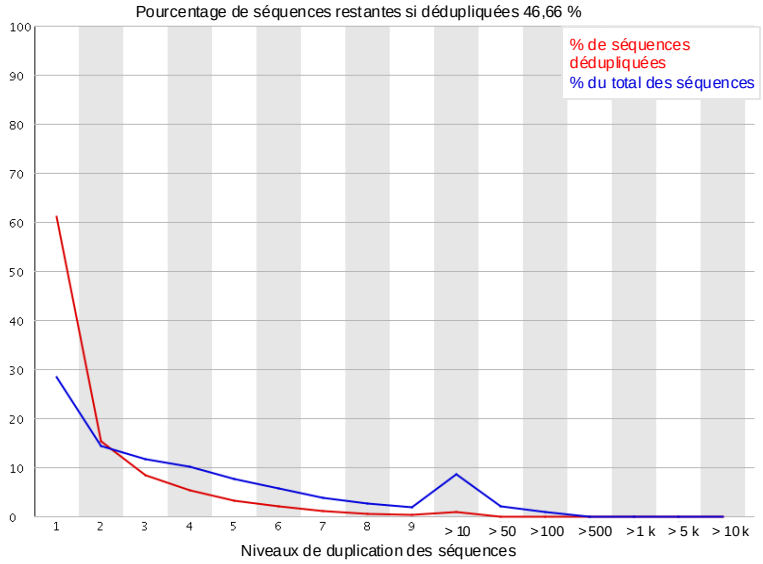
Taux de N par base

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

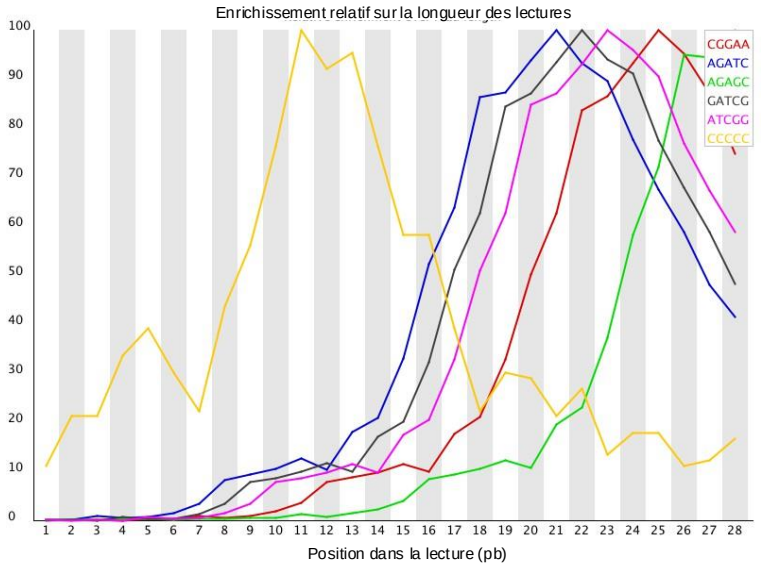
FastQC: Modules d'analyse



Répartition de la longueur des séquences



Séquences dupliquées



Taux de *k*-mères

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

Merci!