Club bioinformatique - Contrôle Qualité

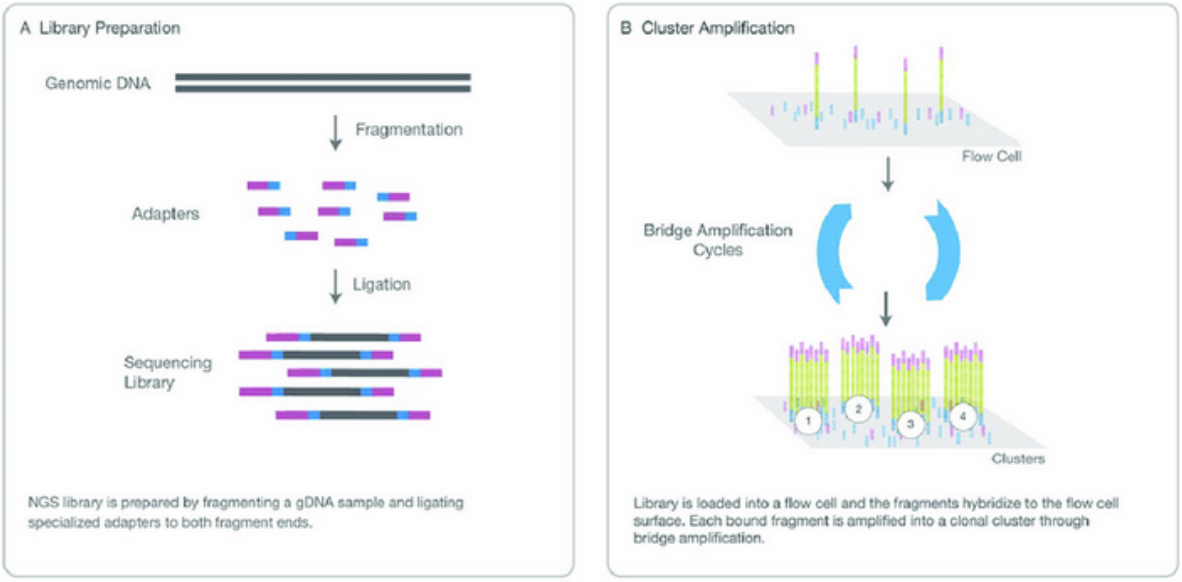
Bénédicte Condamine et Marie Petitjean

Résumé

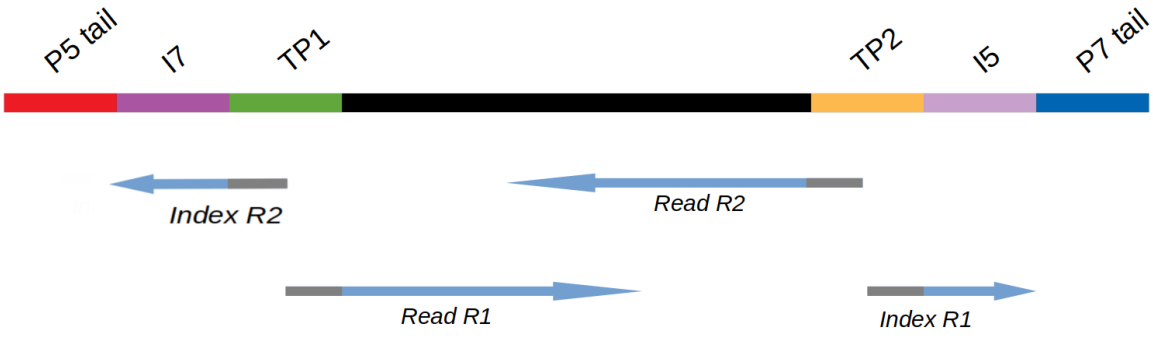
Le contrôle qualité de séquençage Illumina. Comment lancer soi-même son analyse ?

1 Le séquençage

1.1 Illumina



Pour séquencer un échantillon d'ADN, on prépare une library. Cette library est obtenue grâce à une fragmentation de l'ADN et l'ajout d'adaptateurs aux extrémités des fragments.

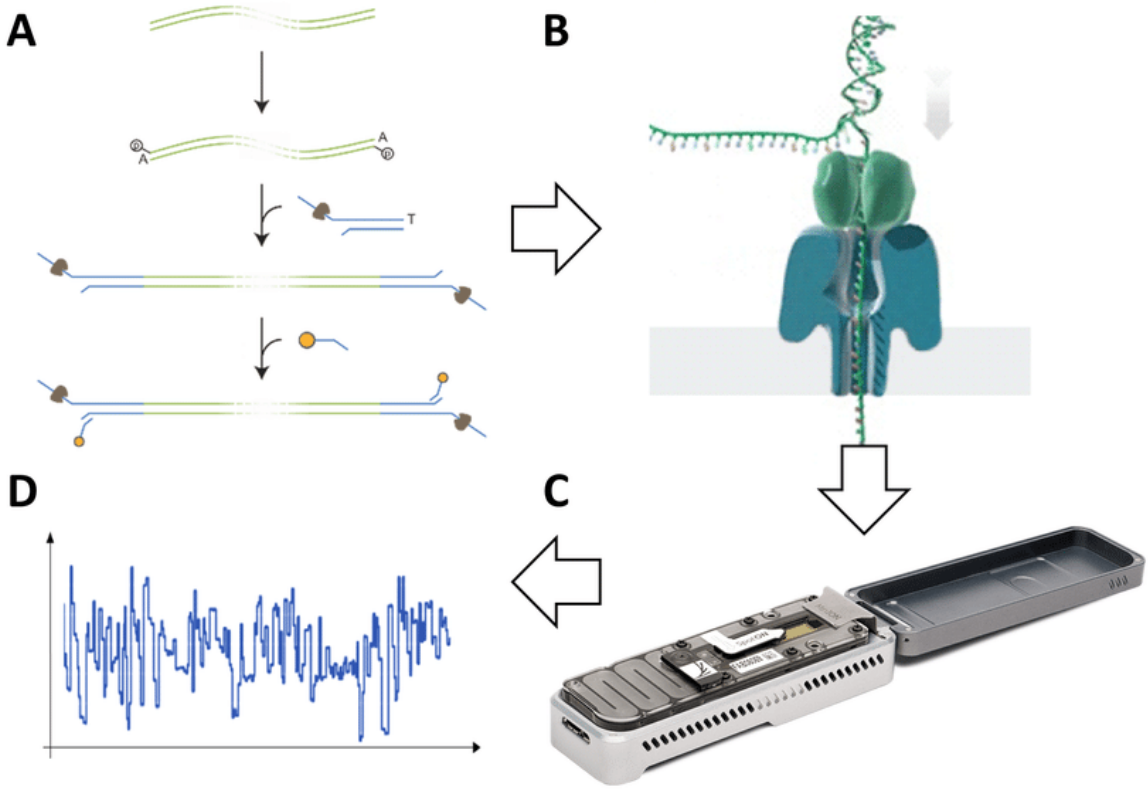


- P5 et P7 tail lient la flowcell et le fragment.
- I5 et I7 sont les Index (8 pb)
- TP1 et TP2 sont les séquences transposases.

Une fois les fragments fixés sur la flowcell par hybridation des P5 et P7 tail, les ponts vont être crées et les séquences complémentaires aux fragments sont synthétisés.

- Depuis NextSeq system Illumina, il y a une réduction du nombre d'image par cycle, deux filtres : A+C et T+C
- Absence de Signal = G

1.2 Nanopore



Pour le séquençage nanopore, on a également une library à préparer (A). Préparation des extrémités des fragments, ajout des adaptateurs et de la protéine "moteur" qui va permettre de se fixer sur le pore et de faire passer le brin d'ADN (B). Le tout est déposé dans le séquençeur (C) et on obtient la résultat sous la forme d'un signal électrique qui sera ensuite convertit en nucléotides (basecalling).

2 Format Fastq

Succession de 4 lignes :

- L'identifiant commençant par '@'
- La séquence nucléique
- Une ligne commençant par '+'
- Les scores de qualité

Exemple de Fastq Illumina :

```
@MN00979:83:000H3G7NF:1:11102:11735:1096 1:N:0:11
GGCTGGCGGGCTGGGGCAATGTCGCCTTTACTGGCGGCAGAGACATTCCTTACTTCCNGGTGCCGCTTTTNGGCNCTATCNTTNGCGCGATTNTAGGTGCAN
NTGCCNNCCGCANNCTNATTGGTCGCNATNTNCCTTGNNATAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF#FFFFFFFFFFFF#FFF#FFFFF#FF#FFFFFFFF#FFFFFFFF#
#AFFF##AFFFF##FF#FFFFFFFFF#FF#F#FFFFF##FFFF
```

```
@MN00979:83:000H3G7NF:1:11102:11735:1096 2:N:0:11
TCATATTACAGCGAAGCTTTTTGTTCTGAAGGAGTTGTGGTCTCCTTTTCTTCCACAACACAGATATCGCAAGGCAAATGGCGACCAATCAGTTTGCGGTAG
GCAAATGCACCTACAATCGCGCCAACGATAGGGCCAAAAAGCGGCACCA
+
FFFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

MN00940 Nom de l'instrument

83 Numéro du run

000H3G7NF ID de la flow-cell

1 Numéro de la lane

11102 Numéro de la tile

11735 Position X

1096 Position Y

1 Reads 1 ou 2

N Filtré (sinon Y)

0 Numéro de contrôle

11 Index

Exemple de Fastq Nanopore :

```
@d28277cb-cb7f-443b-a8e5-d00e85fe939a runid=e602a585cb2df9736d5f48192291c9cd72afc999 read=44 ch=32 st
art_time=2021-06-10T15:22:14Z flow_cell_id=agd430 protocol_group_id=plasmid-clone210610 sample_id=no_
sample barcode=barcode01
TTGTACTTCGTTCAATTTATTTCAGATGATGGTGTTTAACAAGAAAGTTGTCAGTGTCTTTGTGGTTTTCGCATTTATCGTGAAACGCTTTCGCGTTTTTC
GTGCGCCGCTTCACCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGA
TCGCTGTTAAAGGACAATTACAAACAGGAATCCGAATGCAACCAGCGCGGAACACTGCACCAGCGCATCAACAATATTTTCACCTGAATCAGTATTCTTCT
AATACCTGGAATGCTGTTTTCAGGATCGCAGTGGTGATAACCATGCATCATCAGGGTACGGATAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAG
CCGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCTTACAATCGATA
GATTGTCGCACCTGATTACCCGACATTATCGCCCGAGCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCTTTCGAGCAAGACGT
TTCCCGTTTGAATATGGCTCATAACACCCTTGTATTACTGTTTTTATGTAAGCAGACAGTTTTTATTGTTCATGAT
+
'-+..666-467+/&*3+(,,+,%+%%&%&%(;<AA60)47;-.,-3=BD@7.?CIE>EJ=BF@BKMLGB>ACFGB5D<;D40*C@@>>E<07CBHLID/-
,+)(+++'167:>G6C-AD=?A7D;;B88:=/D>8;DGDC26B7/17;<('(4>@@BJ=GH269:>FKFBGHFGFJE;EHHFFGG87??<;9D9<FI5455
<DAD>7;-6/$.::A?BE=89ABA81B9<8<565=EE=;458)$)/**%*'()'')&&$%666C=E?EE,/C5-1/.0233;=>A98;:''*1/0=2<9;/
5*(*/-47EH@B;8/1681/-$48457%%$%1/*%&&&;==?:>62>C=?F<,;))779?@@?:<83,=>ADAAH>A?:=?=//1-@<?FC>=@><97(%$
$#&.760155<?@=?@@:05@BAB=BCDDFID:FDD@@1134<:=A@ADD:;=HI?AA((/0881413><::21,&*2/0),-++++.'+-,/46<667-*
3$$'0;988:<1212,,%%-+57*&>@>A<.(+*++0612350$$-4599>=>45**4,(5AC@>==<B@C@>(DC?75.73321((+,6,(0828:99@G
GIED@=@;3'%'(%%%'555$+)()*/1138=;<;>EA8:IFLFBCB85:;2:9;<4.::CD86.167@D=820'$
```

d28277cb-cb7f-443b-a8e5-d00e85fe939a Identifiant de la machine

read=44 Numéro du read

ch=32 Numéro du canal (channel)

start_time=2021-06-10T15 :22 :14Z Horaire de début du run

flow_cell_id=agd430 Identifiant de la flowcell

protocol_group_id=plasmid-clone210610 Nom du run (donné par l'utilisateur)

sample_id=no_sample Identifiant de l'échantillon

barcode=barcode01 Barcode

3 Le Phred score

Un score de qualité est attribué à chaque acide nucléique :

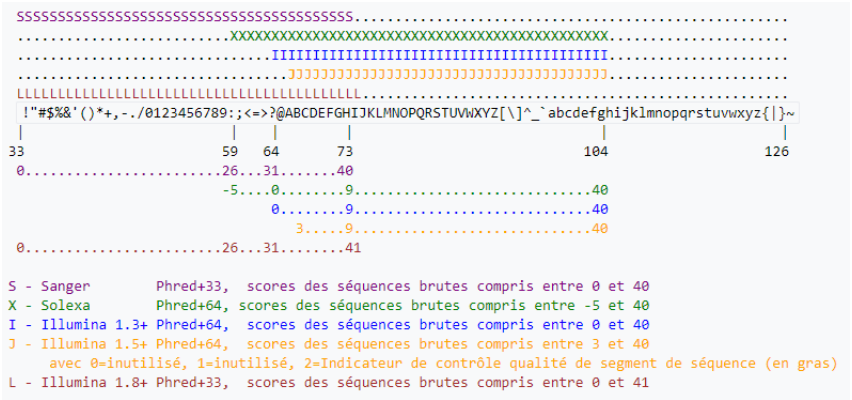
$$Q = -10\ log_{10}\ P$$

P étant la probabilité d'identifier une base.

Score de qualité phred	Probabilité d'une identification incorrecte	Précision de l'identification d'une 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 %

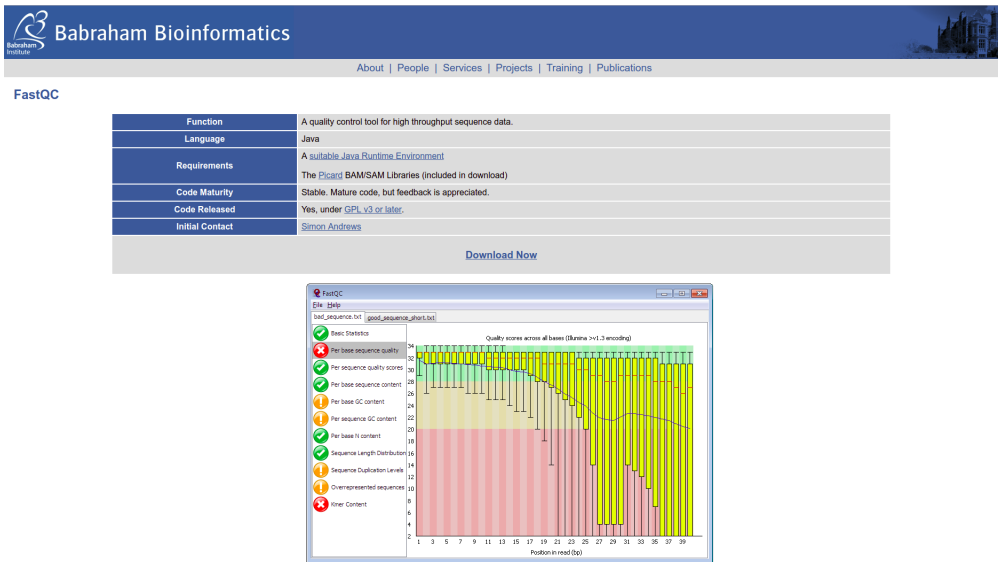
Le score de qualité de chaque base est encodé sous forme de caractère ASCII : Score de qualité Phred

Le score Phred va de 0 à 93 et est encodé en ASCII de 33 à 126



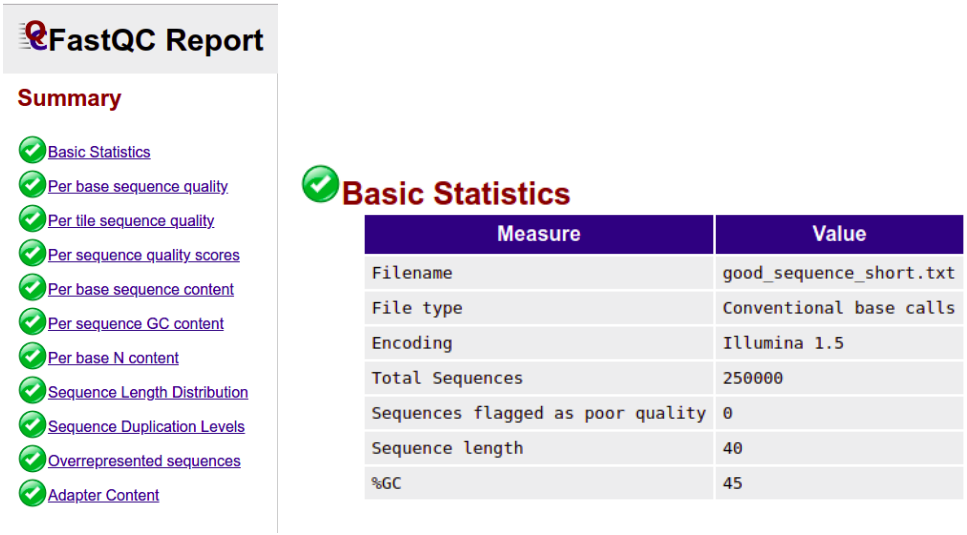
4 Contrôle de qualité : FastQC et MultiQC

4.1 Le logiciel FastQC

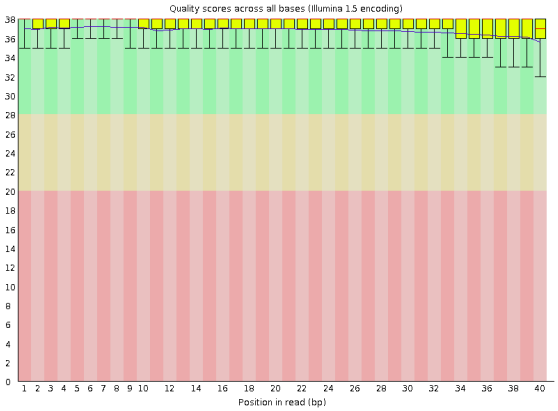


http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

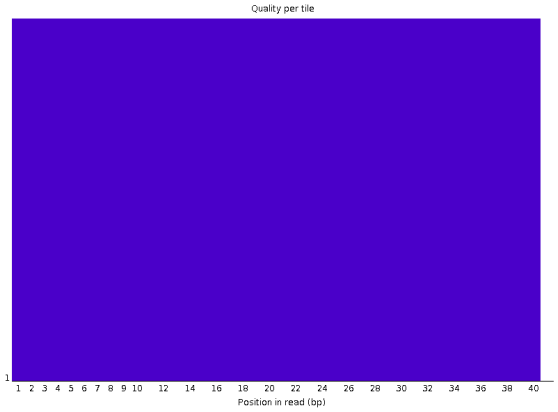
4.2 Les résultats attendus - Illumina



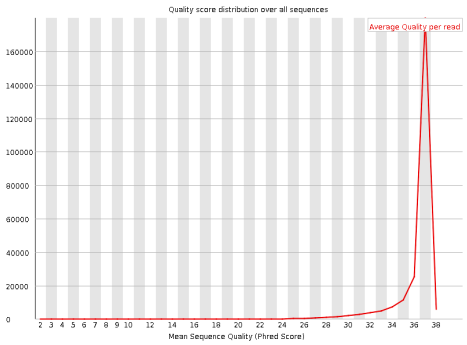
Per base sequence quality



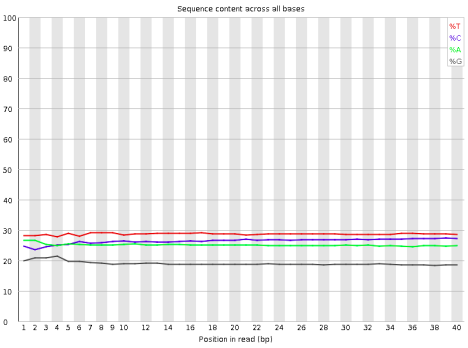
Per tile sequence quality



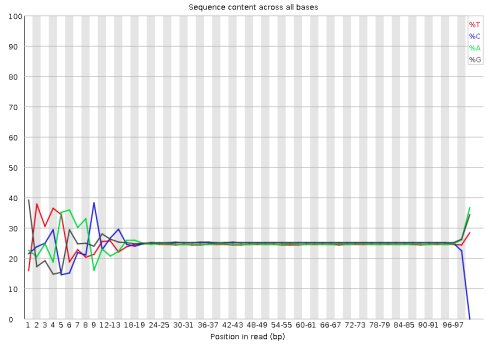
Per sequence quality scores



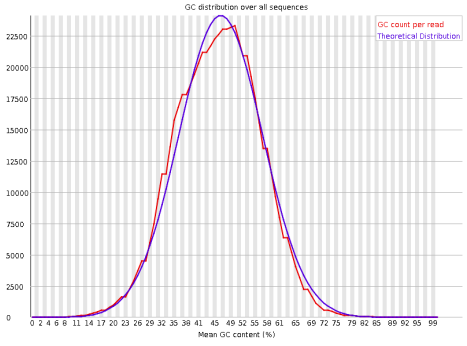
Per base sequence content



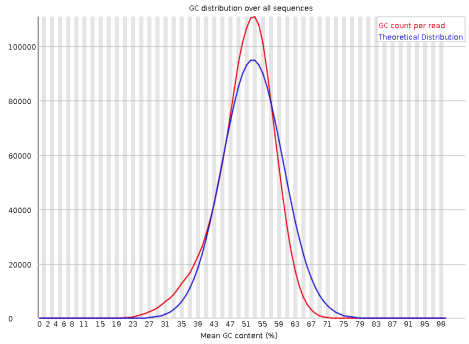
Per base sequence content



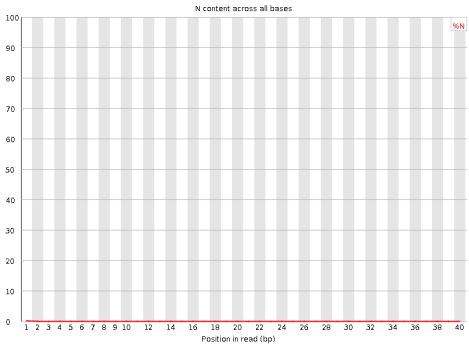
Per sequence GC content



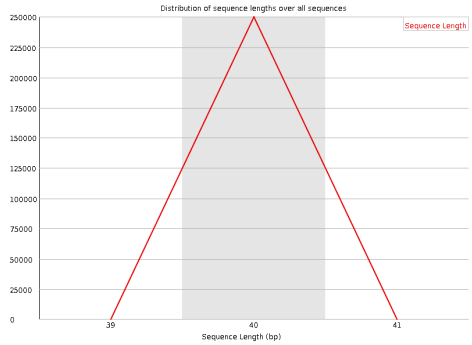
Per sequence GC content



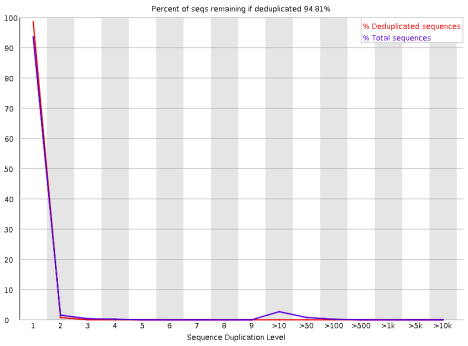
Per base N content



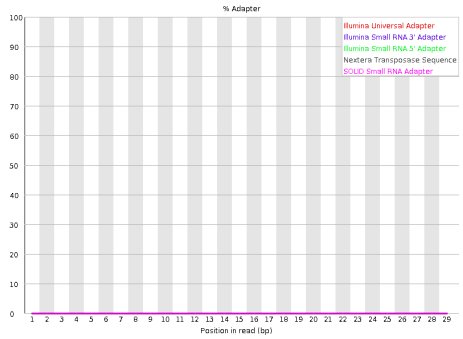
Sequence Length Distribution



Sequence Duplication Levels



Adapter Content



4.3 Les résultats attendus - Nanopore

FastQC Report

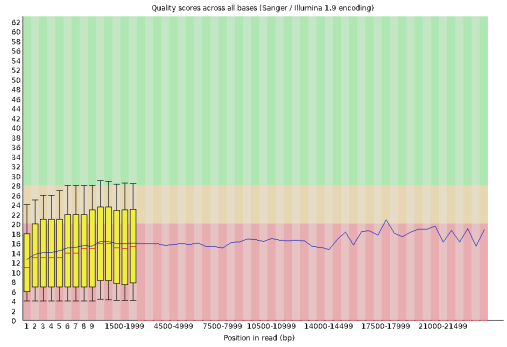
Summary

- Basic Statistics
- Per base 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

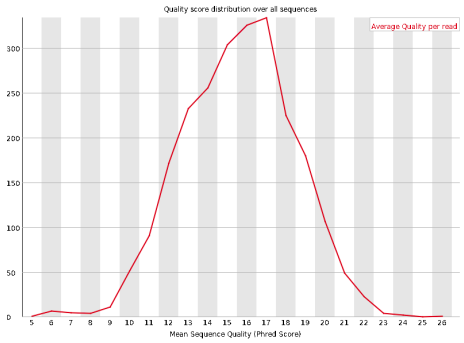
Basic Statistics

Measure	Value
Filename	barcode02.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2387
Sequences flagged as poor quality	0
Sequence length	7-23504
%GC	46

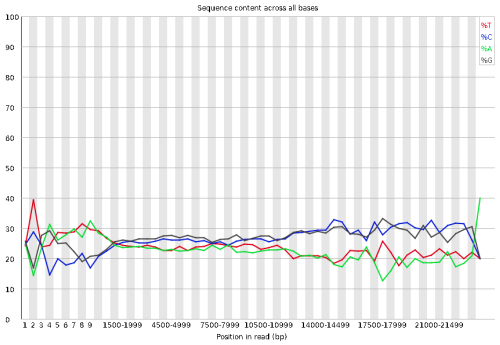
Per base sequence quality



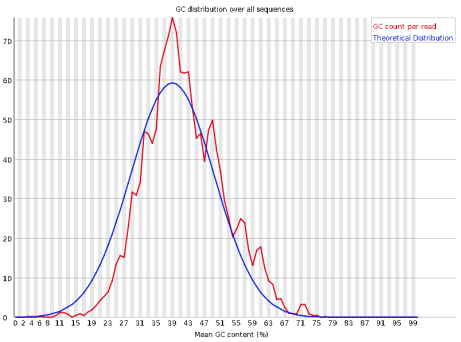
Per sequence quality scores

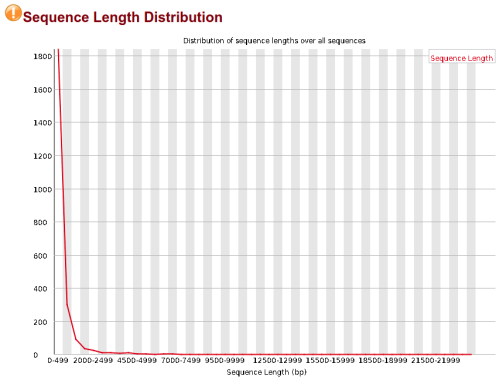


Per base sequence content



Per sequence GC content





4.4 Exemples de problèmes

4.4.1 Nombre de reads

Basic Statistics

Measure	Value
Filename	[redacted]_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1129012
Sequences flagged as poor quality	0
Sequence length	100
%GC	49

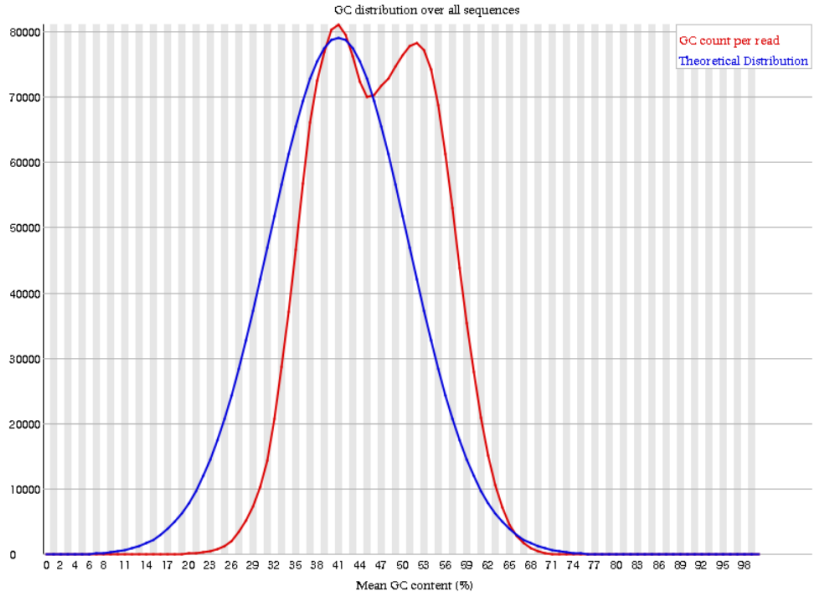
Basic Statistics

Measure	Value
Filename	[redacted]_R2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1129063
Sequences flagged as poor quality	0
Sequence length	100
%GC	49

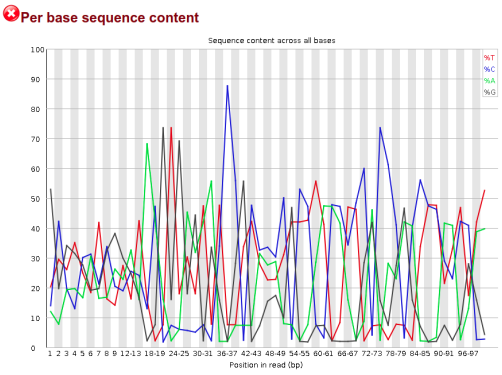
4.4.2 Hétérogénéité

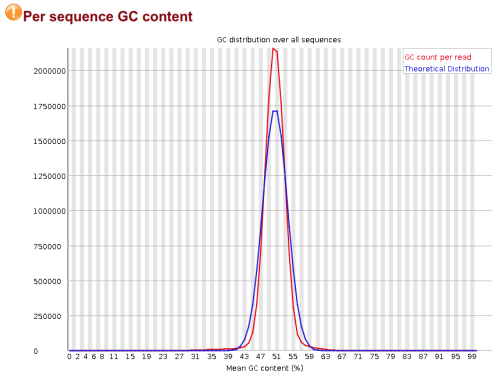
Sample name	Raw total reads	Raw read pairs	Adapter clipped total reads	Adapter clipped read pairs
Sample 1	895 834	447 917	895 770	447 885
Sample 2	1 078 044	539 022	1 078 002	539 001
Sample 3	1 142 624	571 312	1 142 546	571 273
Sample 4	1 224 938	612 469	1 224 866	612 433
Sample 5	1 690 118	845 059	1 690 022	845 011
Sample 6	1 790 346	895 173	1 790 224	895 112
Sample 7	2 081 916	1 040 958	2 081 794	1 040 897
Sample 8	2 077 218	1 038 609	2 077 158	1 038 579
Sample 9	2 437 230	1 218 615	2 437 064	1 218 532
Sample 10	2 550 894	1 275 447	2 550 656	1 275 328
Sample 11	2 672 188	1 336 094	2 672 050	1 336 025
Sample 12	3 029 394	1 514 697	3 029 224	1 514 612
Sample 13	2 840 550	1 420 275	2 840 372	1 420 186
Sample 14	2 810 482	1 405 241	2 810 460	1 405 230
Sample 15	3 027 302	1 513 651	3 027 110	1 513 555
Sample 16	2 926 418	1 463 209	2 926 270	1 463 135
Sample 17	2 993 100	1 496 550	2 992 924	1 496 462
Sample 18	2 942 202	1 471 101	2 942 010	1 471 005
Sample 19	3 101 274	1 550 637	3 101 112	1 550 556
Sample 20	3 120 158	1 560 079	3 119 958	1 559 979
Sample 21	2 958 580	1 479 290	2 958 536	1 479 268
Sample 22	3 214 464	1 607 232	3 214 312	1 607 156
...	...	...	...	...
Sample 52	4 116 902	2 058 451	4 116 620	2 058 310
Sample 53	4 210 322	2 105 161	4 210 080	2 105 040
Sample 54	4 137 522	2 068 761	4 137 282	2 068 641
Sample 55	4 153 332	2 076 666	4 152 964	2 076 482
Sample 56	4 248 366	2 124 183	4 248 116	2 124 058
Sample 57	4 286 048	2 143 024	4 285 784	2 142 892
Sample 58	4 329 860	2 164 930	4 329 602	2 164 801
Sample 59	4 298 018	2 149 009	4 297 780	2 148 890
Sample 60	4 263 254	2 131 627	4 263 140	2 131 570
Sample 61	4 319 716	2 159 858	4 319 554	2 159 777
Sample 62	4 479 606	2 239 803	4 479 020	2 239 510
Sample 63	4 337 394	2 168 697	4 337 332	2 168 666
Sample 64	4 751 710	2 375 855	4 751 508	2 375 754
Sample 65	4 744 802	2 372 401	4 744 536	2 372 268
Sample 66	4 740 820	2 370 410	4 740 522	2 370 261
Sample 67	4 871 668	2 435 834	4 871 314	2 435 657
Sample 68	5 056 696	2 528 348	5 056 330	2 528 165
Sample 69	5 021 208	2 510 604	5 020 832	2 510 416
Sample 70	5 130 750	2 565 375	5 130 188	2 565 094
Sample 71	5 187 338	2 593 669	5 187 060	2 593 530
Sample 72	5 221 446	2 610 723	5 221 250	2 610 625
Sample 73	5 349 372	2 674 686	5 349 018	2 674 509
Sample 74	5 331 956	2 665 978	5 331 658	2 665 829
Sample 75	5 206 590	2 603 295	5 206 226	2 603 113
Sample 76	5 113 990	2 556 995	5 113 730	2 556 865
Sample 77	5 374 852	2 687 426	5 374 730	2 687 365
Sample 78	5 629 768	2 814 884	5 629 188	2 814 594
Sample 79	5 876 912	2 938 456	5 876 536	2 938 268
Sample 80	6 582 280	3 291 140	6 581 894	3 290 947
Sample 81	6 896 034	3 448 017	6 895 648	3 447 824
Sample 82	7 031 750	3 515 875	7 031 150	3 515 575

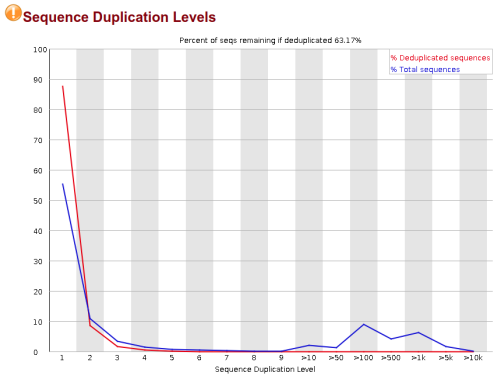
4.4.3 Taux de GC

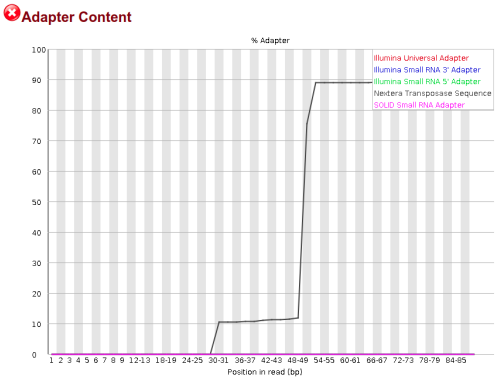


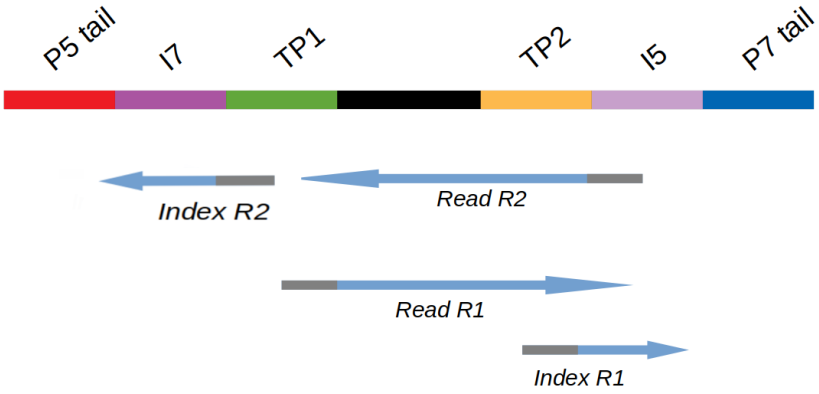
4.4.4 Adaptateurs











4.5 Lancement de fastQC en ligne de commande

fastqc : lancer fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
$ fastqc -o dossier_resultats sample1_R1.fq.gz

4.6 Lancement de multiQC en ligne de commande

multiqc : lancer multiqc (https://multiqc.info/)
$ multiqc -ip dossier_resultats/*

(?) Créer un script listant les fastq, lançant les fastQC puis le multiQC
(?) Trouver le bon docker et lancer le script dedans

5 Nettoyage des reads : Trim galore

Le nettoyage de read permet de retirer des bases selon des seuils fixés par l'utilisateur.

Trimgalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) est un script wrapper c'est à dire qu'il regroupe plusieurs logiciels (cutadapt et fastqc) pour effectuer un contrôle qualité des données issues de séquençage haut débit. Pour toute analyse, un contrôle qualité est obligatoire pour s'assurer d'une bonne qualité des reads.

Trimgalore permet :

- de retirer les bases de mauvaises qualités (phred score)
- de retirer les séquences adaptatrices
- de retirer les reads avec une taille inférieure à un seuil (fixée par l'utilisateur ou par défaut fixé à 20 bases)

Trimming des bases → les bases de faibles qualités sont supprimées depuis l'extrémité 3' du read avant la suppression de séquence adaptatrice.

Trimming adaptateur → cutadapt cherche une séquence adaptatrice depuis l'extrémité 3' du read et supprime cette séquence ainsi que les bases suivantes. Cutadapt peut détecter

différents adaptateurs :

- Illumina : AGATCGGAAGAGC
- Small RNA : TGGAATTCTCGG
- Nextera : CTGTCTCTTATA
- sequence entrée par l'utilisateur

Si aucune sequence adaptatrice n'a été donnée, cutadapt recherchera dans le premier million de reads une des trois séquences qu'il connait et trimmera les séquences en fonction de celle reconnue.

Dans le cas de reads paired-end, le trimming est exécuté en fonction des options données sur les reads R1 puis les reads R2, et enfin s'il y a correspondance entre les deux reads, la paire de reads est conservée sinon celle-ci est suprimée sauf dans le cas d'option particulière (- -retain_unpaired).

Lors de l'execution de trim galore sur les reads de l'échantillon "Echantillon_R1.fq.gz" et "Echantillon_R2.fq.gz", de nombreux fichiers sont générés :

- "Echantillon_R1_val_1.fq.gz" et "Echantillon_R2_val_2.fq.gz" qui sont les fichiers fastq des reads filtrés et trimmés.
- "Echantillon_R1.fq.gz_trimming_report.txt" et "Echantillon_R2.fq.gz_trimming_report.txt" qui sont des rapports sur le trimming.
- Dans le cas de l'uilisation de l'option permettant de récupérer les reads sans correspondances R1/R2 (- -retain_unpaired), les fichiers supplémentaires "Echantillon_R1_unpaired_1.fq.gz" et "Echantillon_R2_unpaired_2.fq.gz" sont crées.

trimgalore
$ trim_galore [options] <filename(s)>
-q <nb> : retire les bases ayant une qualité (phred score) inférieure au nombre entré.
- -nextera : trimming en fonction de l'adaptateur nextera.
- -illumina : trimming en fonction de l'adaptateur illumina.
- -small_rna : trimming en fonction de l'adaptateur Small-RNA.
-a <sequence> : trimming en fonction d'une séquence spécifiée.
- -paired : trimming sur des fichiers paired-end.
- -retain_unpaired : conserve les reads sans correspondance entre le R1 et R2.
-o <dossier> : dossier dans lequel les fichiers seront écrits.
- -length <nbr> : supprime les reads dont la taille est inférieure.
$ trim_galore -q 30 - -nextera - -paired - -retain_unpaired -o outdir - -length 50 read1.fq.gz read2.fq.gz

(?) Lancer le trimming sur l'échantillon donné en retirant les bases ayant un phred score inférieur à 30, une taille de reads inférieur à 100 bases, les éventuels séquences adaptatrices, en conservant les reads unpaired lors du trimming. Écrire les fichiers résultats dans un dossier de votre espace personnel.

(?) Renommer les fichiers avec "val" en "Sample_R1.fastq.gz" et "Sample_R2.fastq.gz"

(?) Lancer un fastqc sur l'échantillon sample qui a été trimmé.