Tin Sinh học
Bioinformatics

Data Wrangling and Processing for Genomics

Background

https://datacarpentry.org/wrangling-genomics

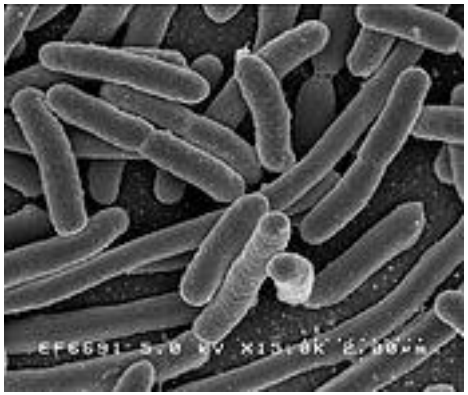
1. Đặt vấn đề
2. Assessing Read Quality
3. Trimming and Filtering

1. Đặt vấn đề

- Questions
  - What data are we using?
  - Why is this experiment important?
- Objectives
  - Why study E. coli?
  - Understand the data set.
  - What is hypermutability?

Đặt vấn đề

- Vi khuẩn Escherichia coli: doubling its population every 20 minutes
- Thí nghiệm trong điều kiện giới hạn glucose, bổ sung citrate.
- Phát hiện các đột biến between 31,000 and 31,500 generations
- Giải trình tự gen tại các thời điểm 5000, 15000, 50000.
- Mục tiêu: khám phá các đột biến thúc đẩy quá trình thích nghi với môi trường



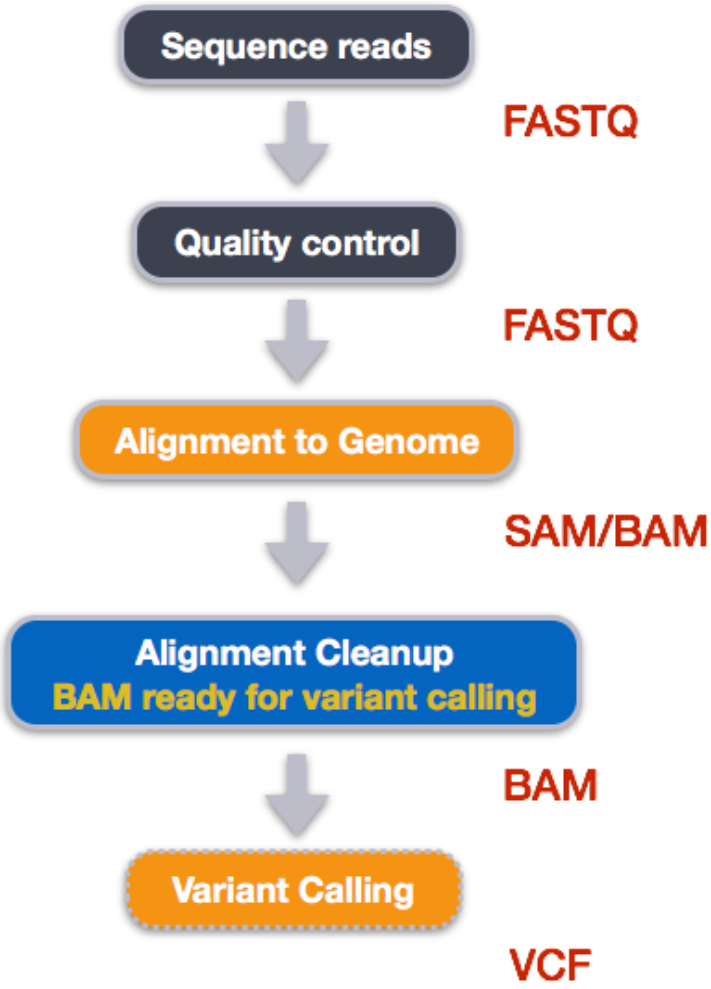
Tìm hiểu file metadata

- Ecoli_metadata_composite.csv
- Trả lời các câu hỏi:
  - How many different generations exist in the data?
  - How many rows and how many columns are in this data?
  - How many citrate+ mutants have been recorded in Ara-3?
  - How many hypermutable mutants have been recorded in Ara-3?

Solution

- 25 different generations
- 62 rows, 12 columns
- 10 citrate+ mutants
- 6 hypermutable mutants
- Key Points: It is important to record and understand your experiment's metadata.

Workflow phát hiện đột biến



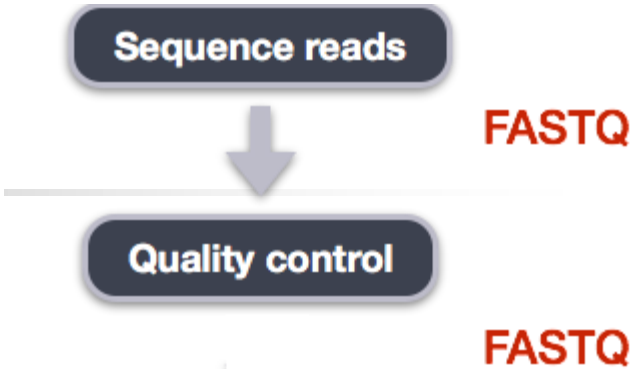
Chuẩn bị dữ liệu

```
mkdir -p ~/dc_workshop/data/untrimmed_fastq/
cd ~/dc_workshop/data/untrimmed_fastq
curl -O
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/004/SRR2589044/SRR2589044_1.fastq.gz
curl -O
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/004/SRR2589044/SRR2589044_2.fastq.gz
curl -O
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/003/SRR2584863/SRR2584863_1.fastq.gz
curl -O
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/003/SRR2584863/SRR2584863_2.fastq.gz
curl -O
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/006/SRR2584866/SRR2584866_1.fastq.gz
curl -O
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/006/SRR2584866/SRR2584866_2.fastq.gz

$ gunzip SRR2584863_1.fastq.gz
```

2. Assessing Read Quality

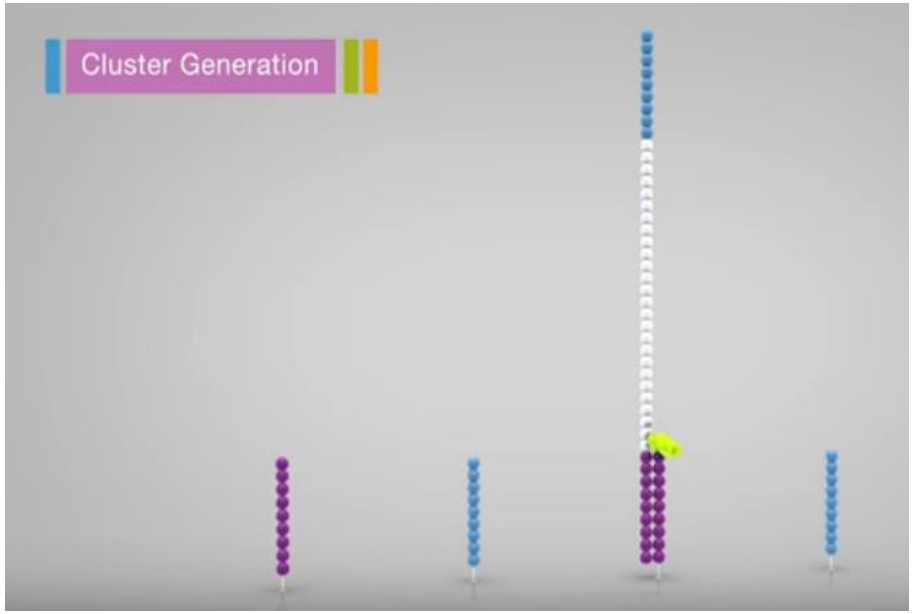
2. Assessing Read Quality

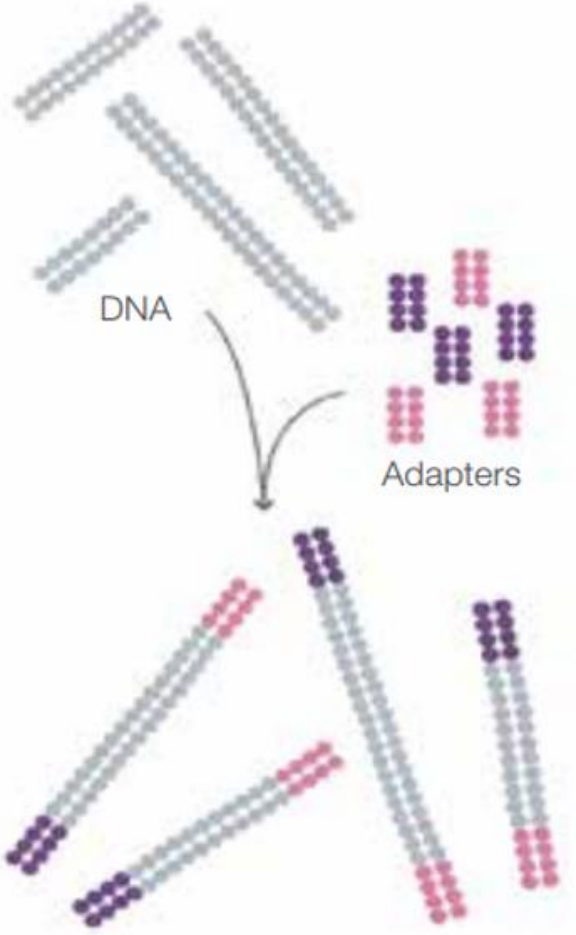


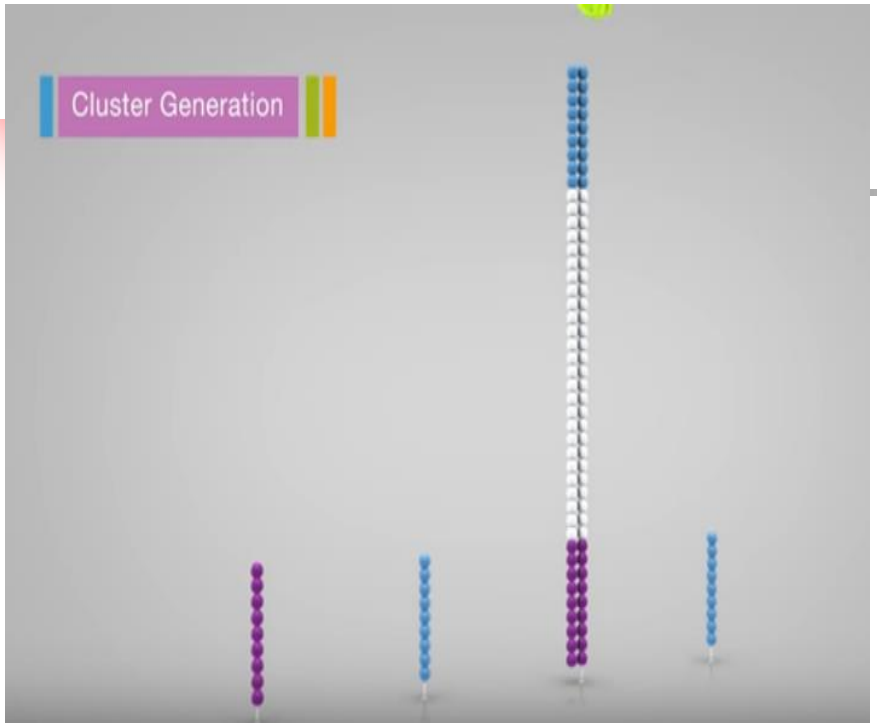
- Questions
  - How can I describe the quality of my data?
- Objectives
  - Explain how a FASTQ file encodes per-base quality scores.
  - Interpret a FastQC plot summarizing per-base quality across all reads.
  - Use for loops to automate operations on multiple files.

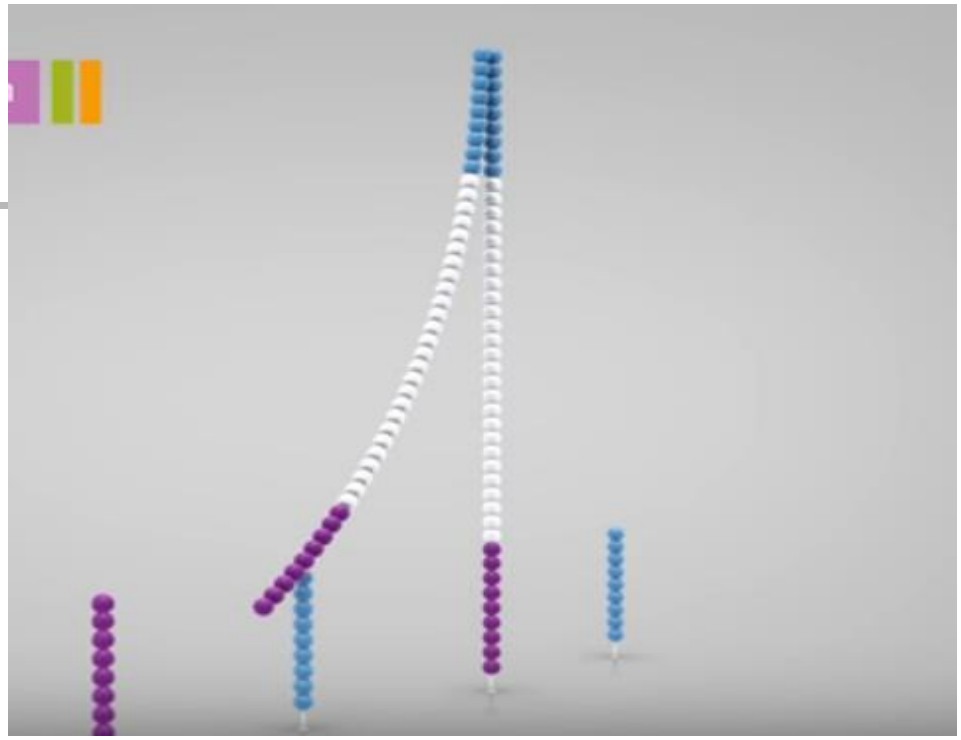
Nguyên lý giải trình tự Illumina

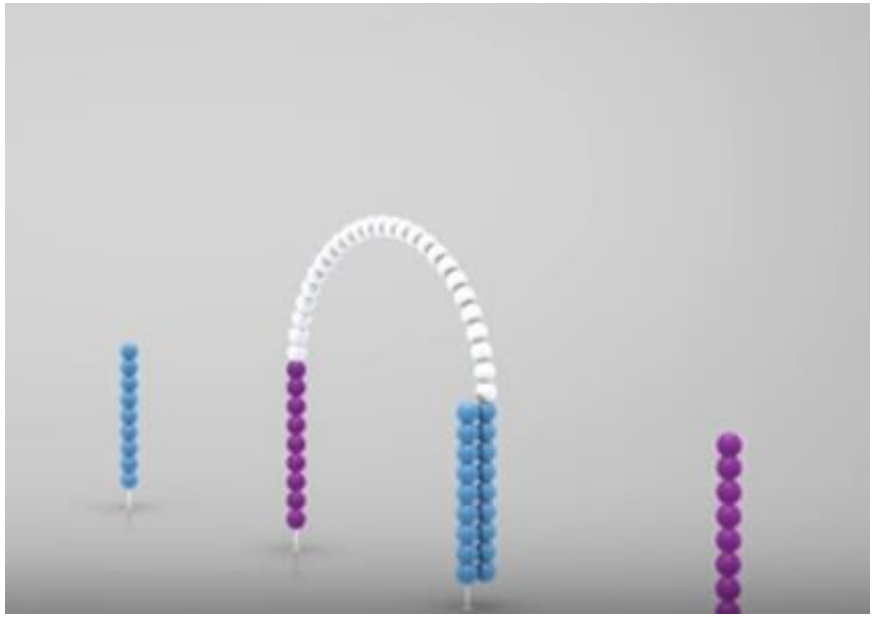
- Randomly fragment genomic DNA and ligate adapters to both ends of fragments
- Tạo các cụm clusters theo nguyên tắc "bắc cầu"

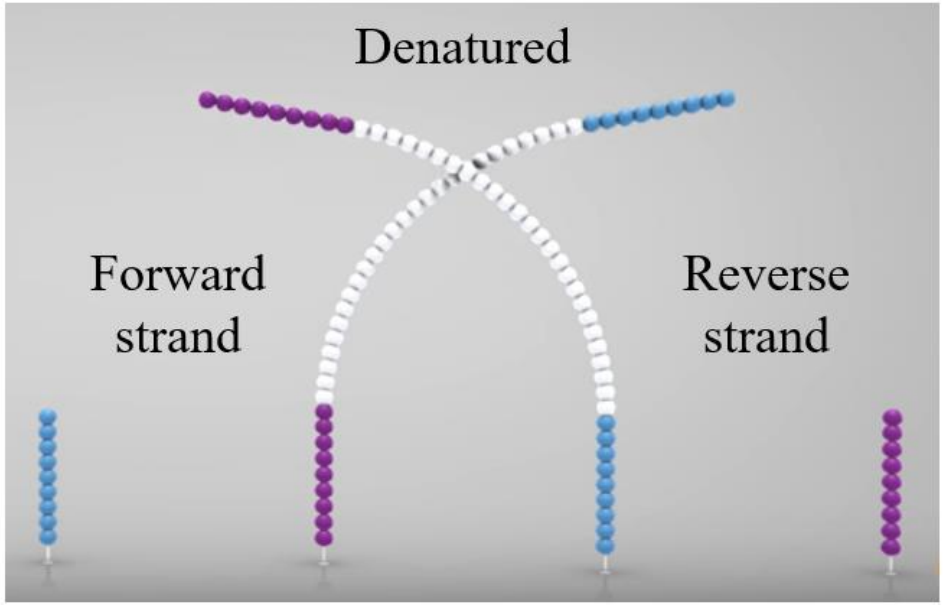




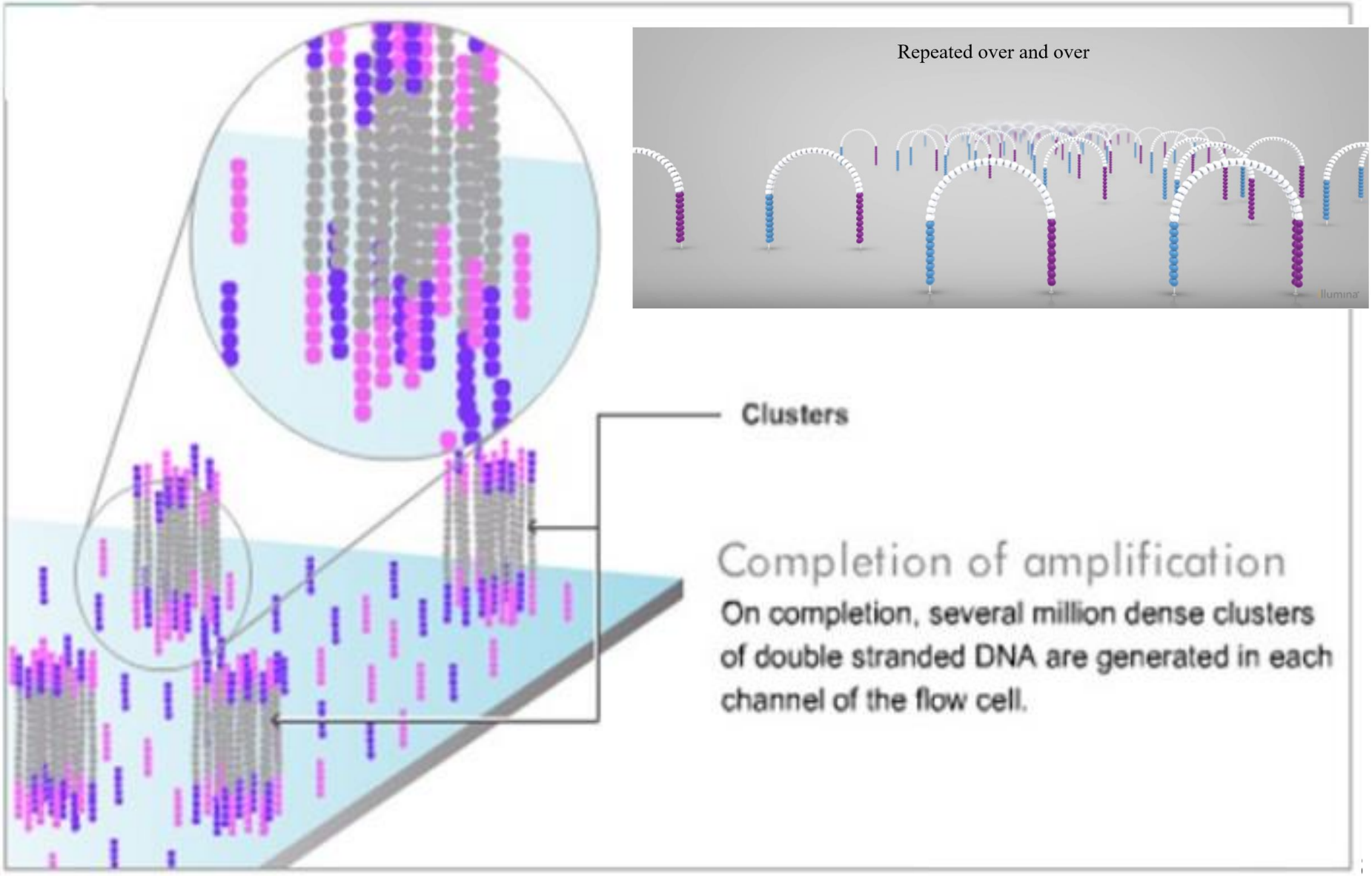

Cluster Generation





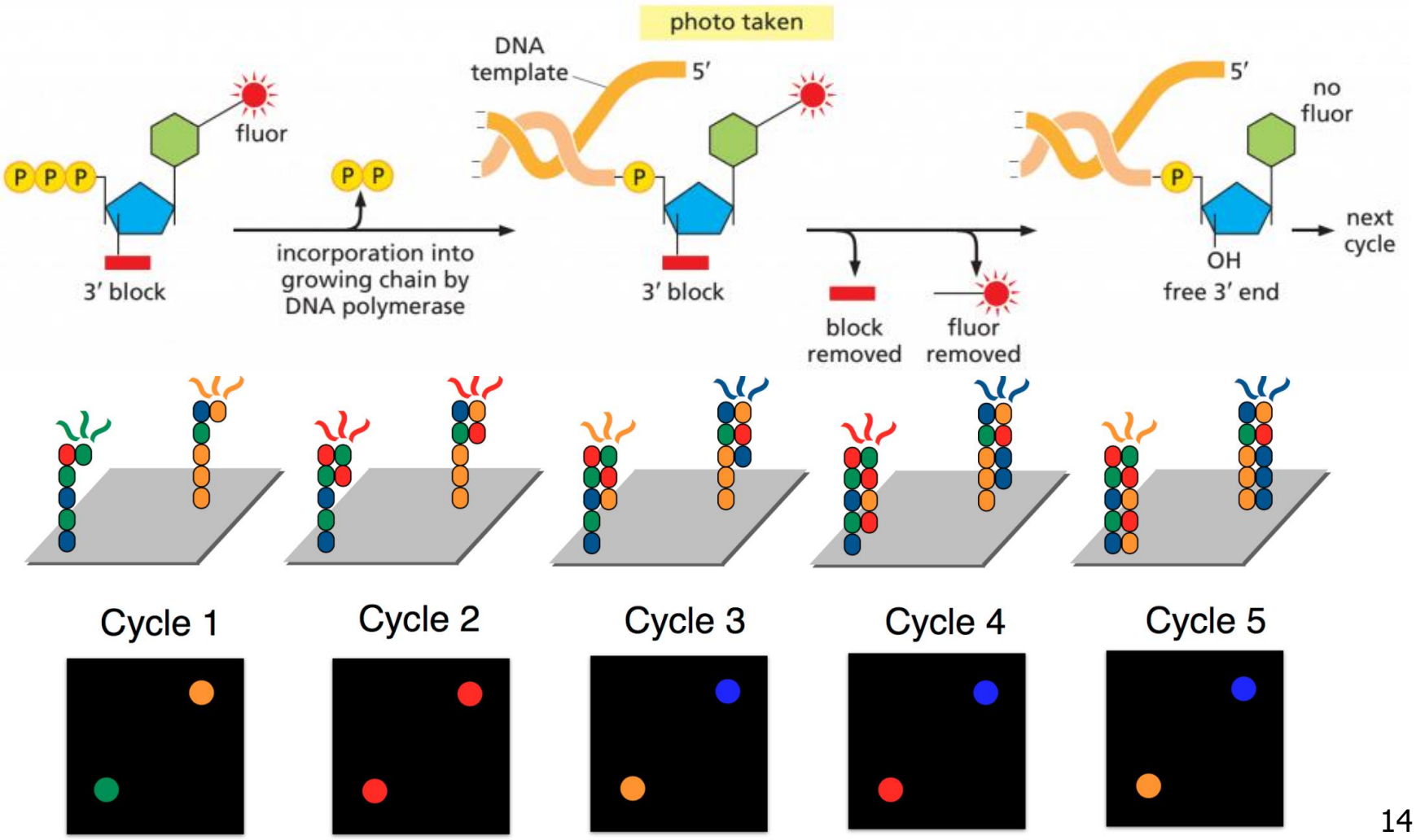

Denatured
Forward strand
Reverse strand

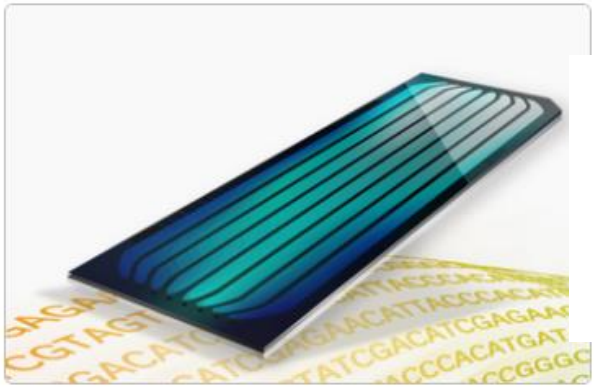
Repeated over and over
Illumina

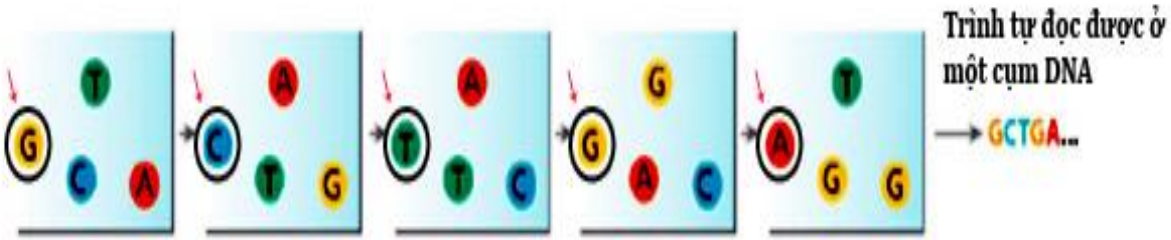

Clusters
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.

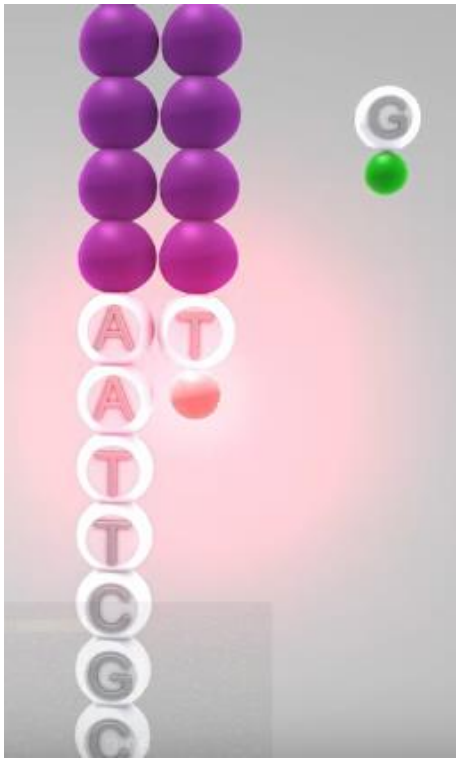
Nguyên lý giải trình tự Illumina

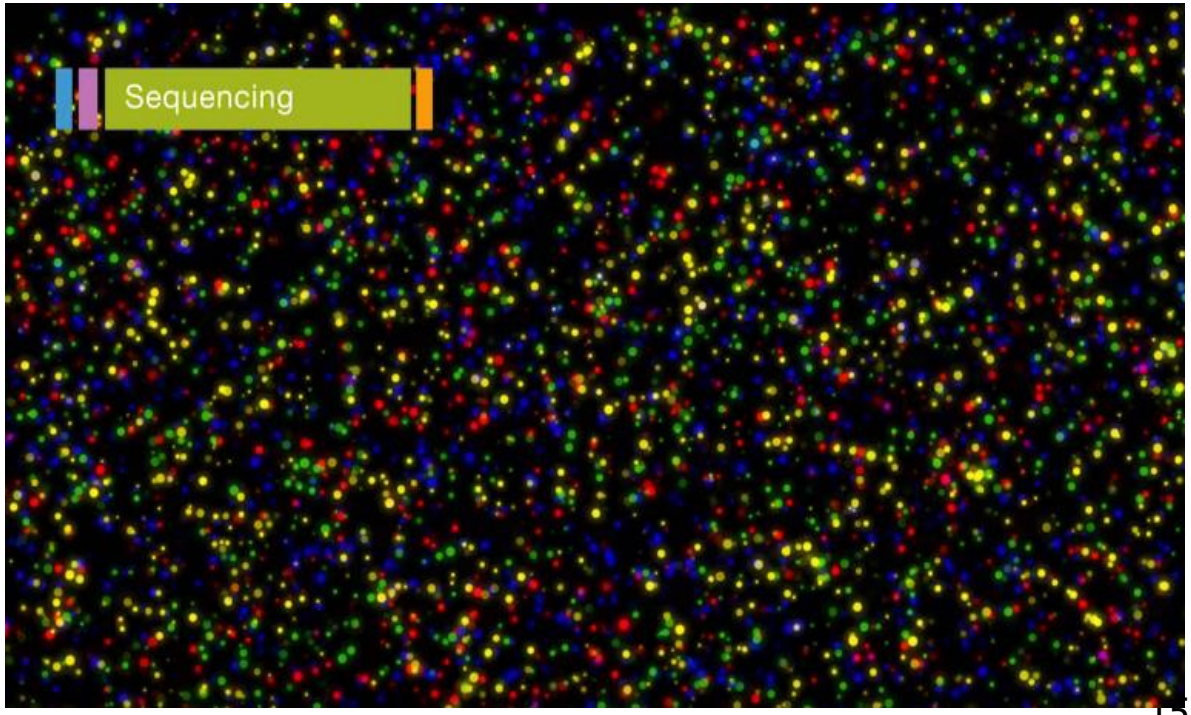
- Giải trình tự theo nguyên lý bổ sung





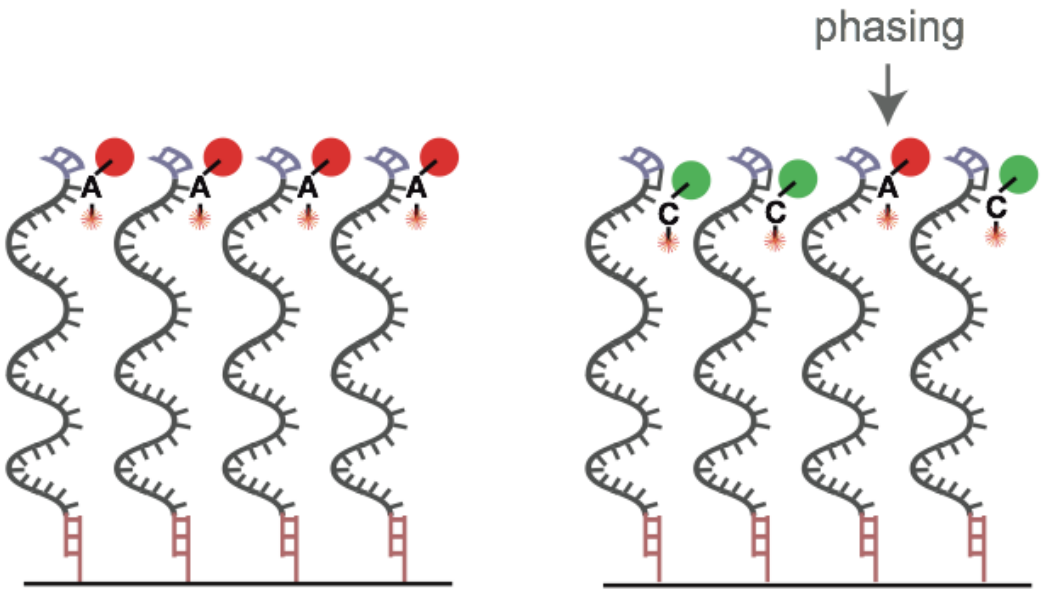

Trình tự đọc được ở một cụm DNA
GCTGA...

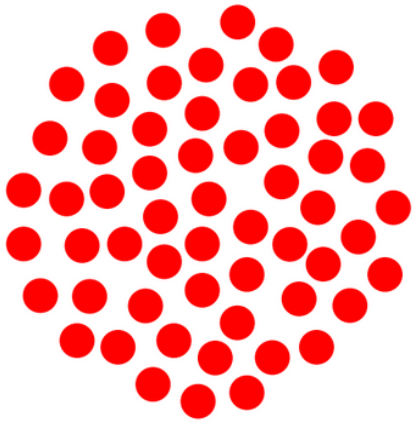



Sequencing

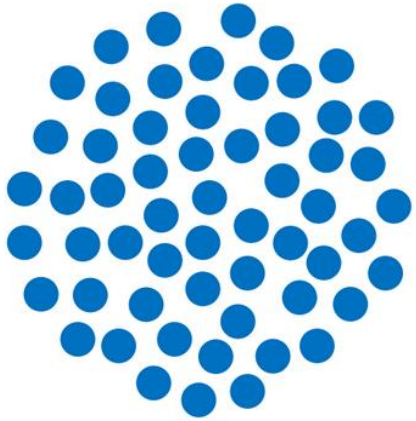
Phát sinh lỗi trong quá trình giải trình tự

- Why does the per base sequence quality decrease over the read in Illumina?
- Phasing means that the blocker of a nucleotide is not correctly removed after signal detection.



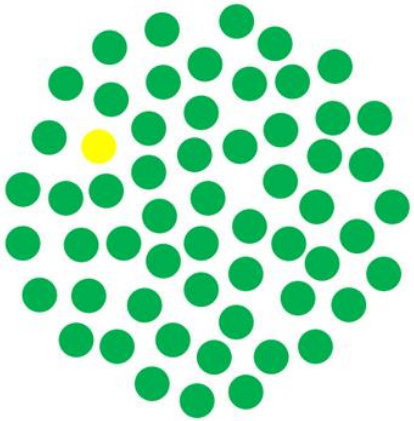


Cycle 1: C

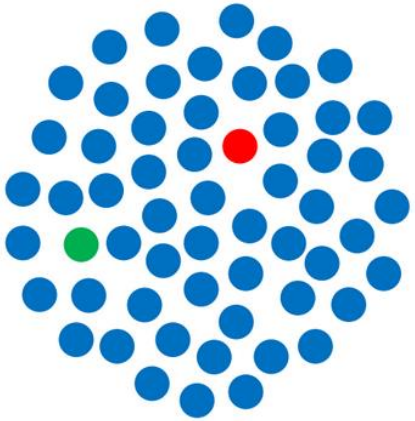


Cycle 2: CG

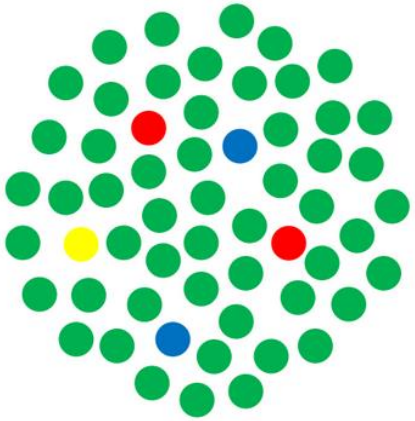
Cycle 3: CGA



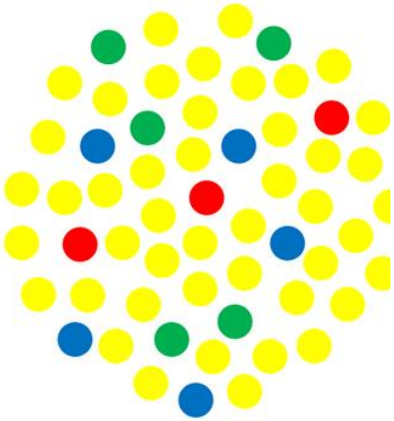
Cycle 4: CGAT



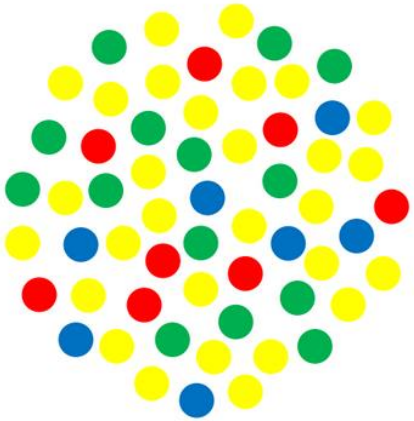
Cycle 20:
CGATGAC…G



Cycle 50:
CGATGAC.....G
.....T(probably)



Cycle 150:
CGATGAC.....G....
..T.....A(maybe?)



Cycle 200:
CGATGAC.....G....
..T.....A.....???

Định dạng file FASTQ

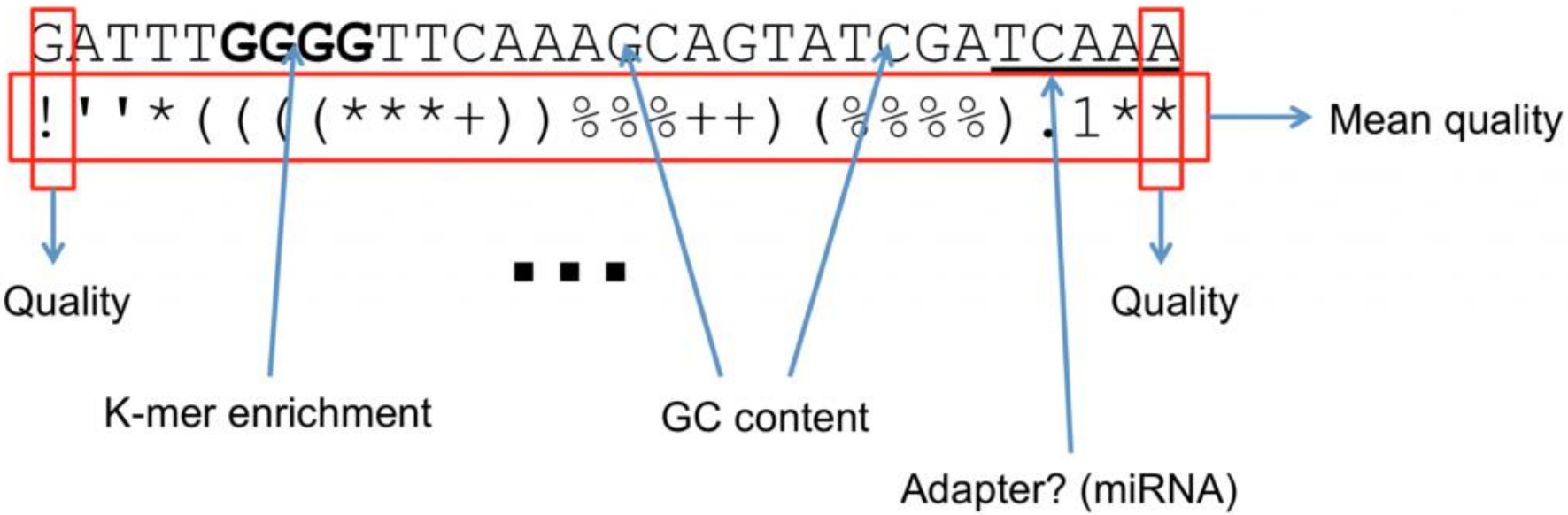
- $ head -n 4 SRR2584863_1.fastq

```
@SRR2584863.1 HWI-ST957:244:H73TDADXX:1:1101:4712:2181/1
TTCACATCCTGACCATTCAGTTGAGCAAAATAGTTCTTCAGTGCCTGTTTAACCGAGTCACG
CAGGGGTTTTTGGGTTACCTGATCCTGAGAGTTAACGGTAGAAACGGTCAGTACGTCAGAA
TTTACGCGTTGTTCGAACATAGTTCTG
+
CCCFFFFFGHHHHJIJJJJIJJJIIJJJJIIIJJGFIIIJEDDFEGGJIFHHJIJJDECCGGEGIIJFHFFF
ACD:BBBDDACCCCAA@@CA@C>C3>@5(8&>C:9?8+89<4(:83825C(:A#######
##################
```

Output $Q = -10 \log_{10} P$ **Phred +33 encoded**

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
                   |         |         |         |         |
Quality score:    01........11........21........31........41
```

Phân tích quality reads



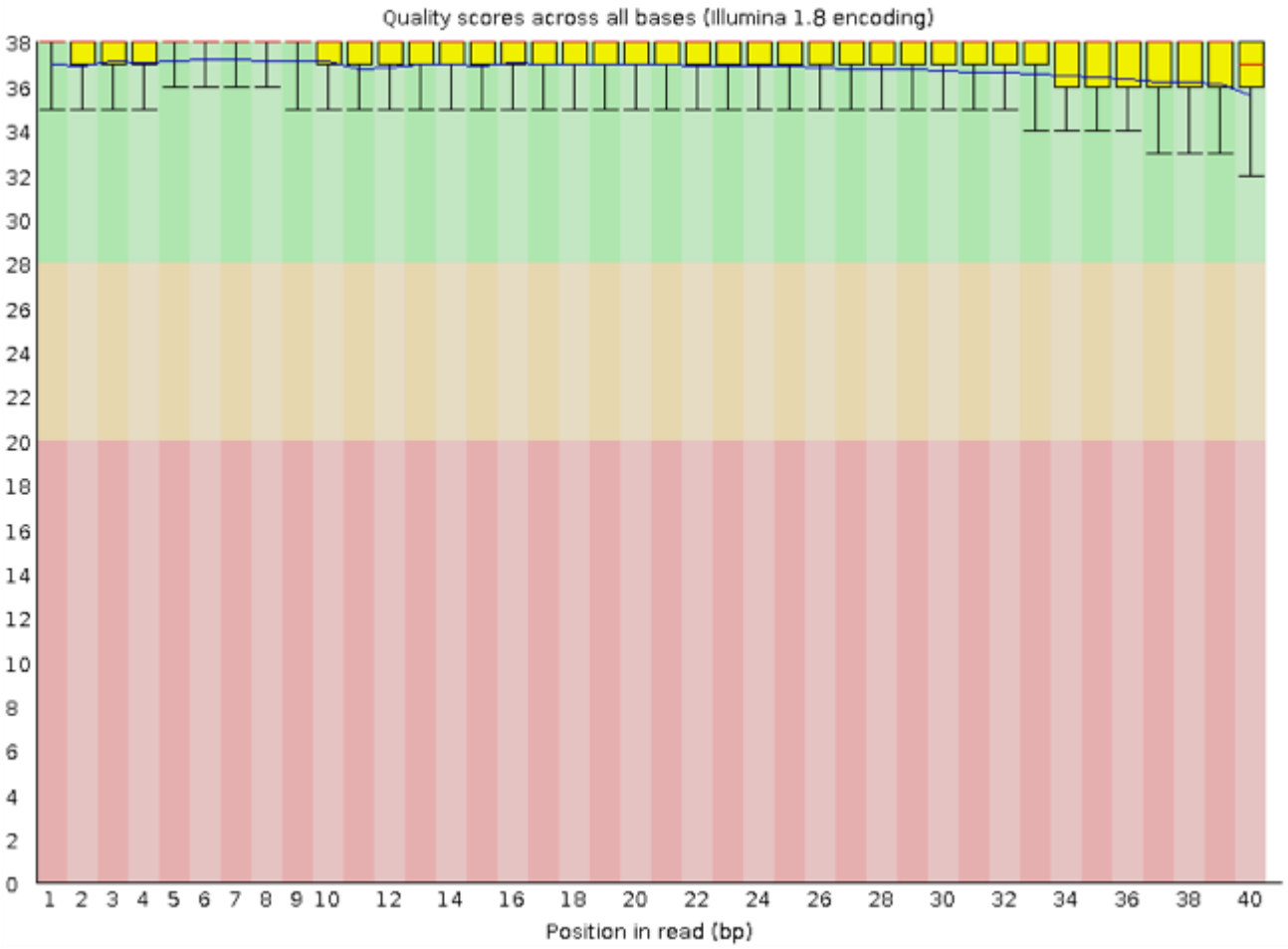
FastQC

$ fastqc -h

- FastQC - A high throughput sequence QC analysis tool
- FastQC reads a set of sequence files and produces from each one a quality control report consisting of a number of different modules, each one of which will help to identify a different potential type of problem in your data.
- sudo apt-get install fastqc

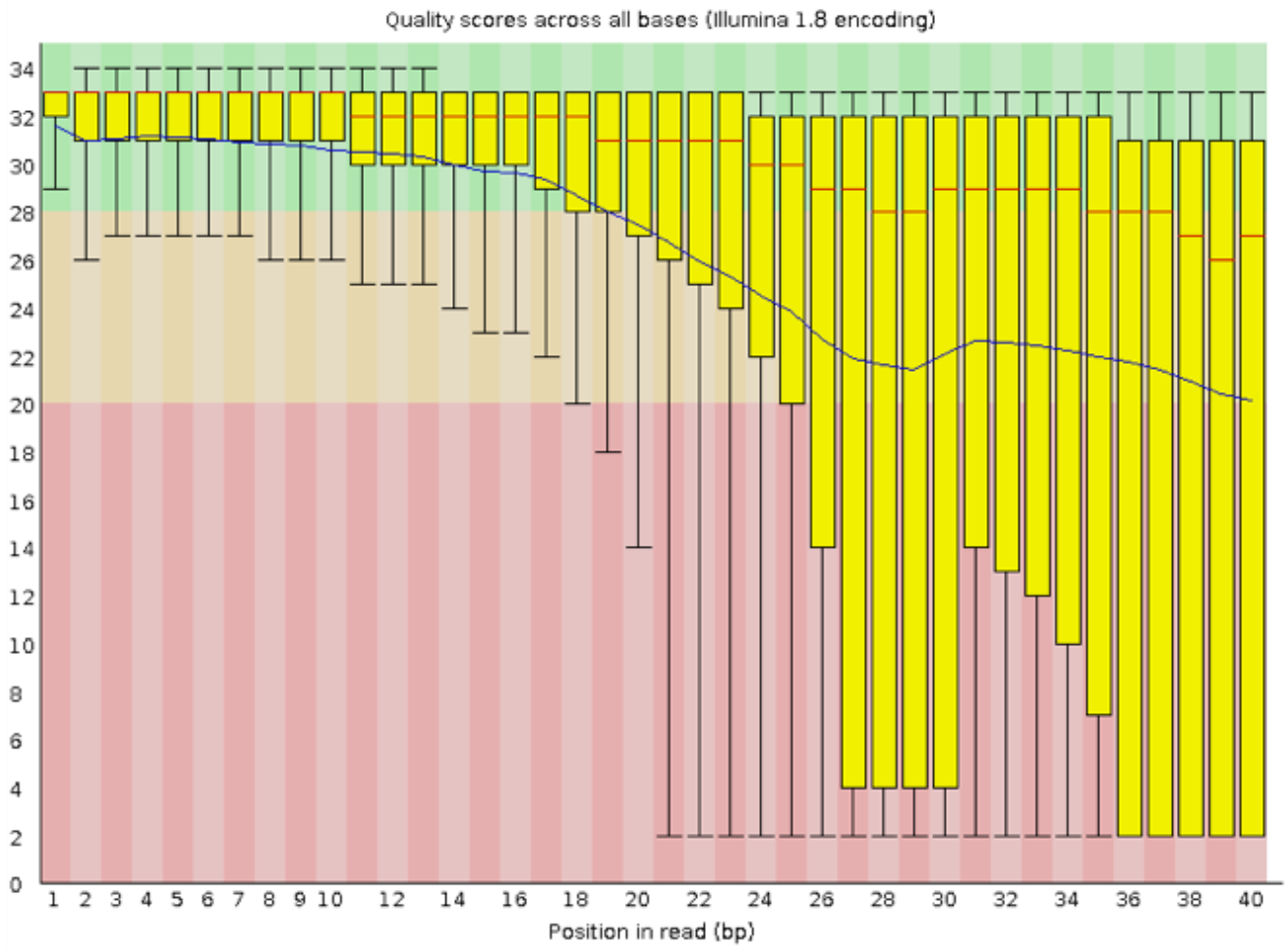
chất lượng theo từng vị trí của cặp bazo

Per base sequence quality



Good quality sequences / Bad quality sequences ?

Bad quality sequences



Basic Statistics

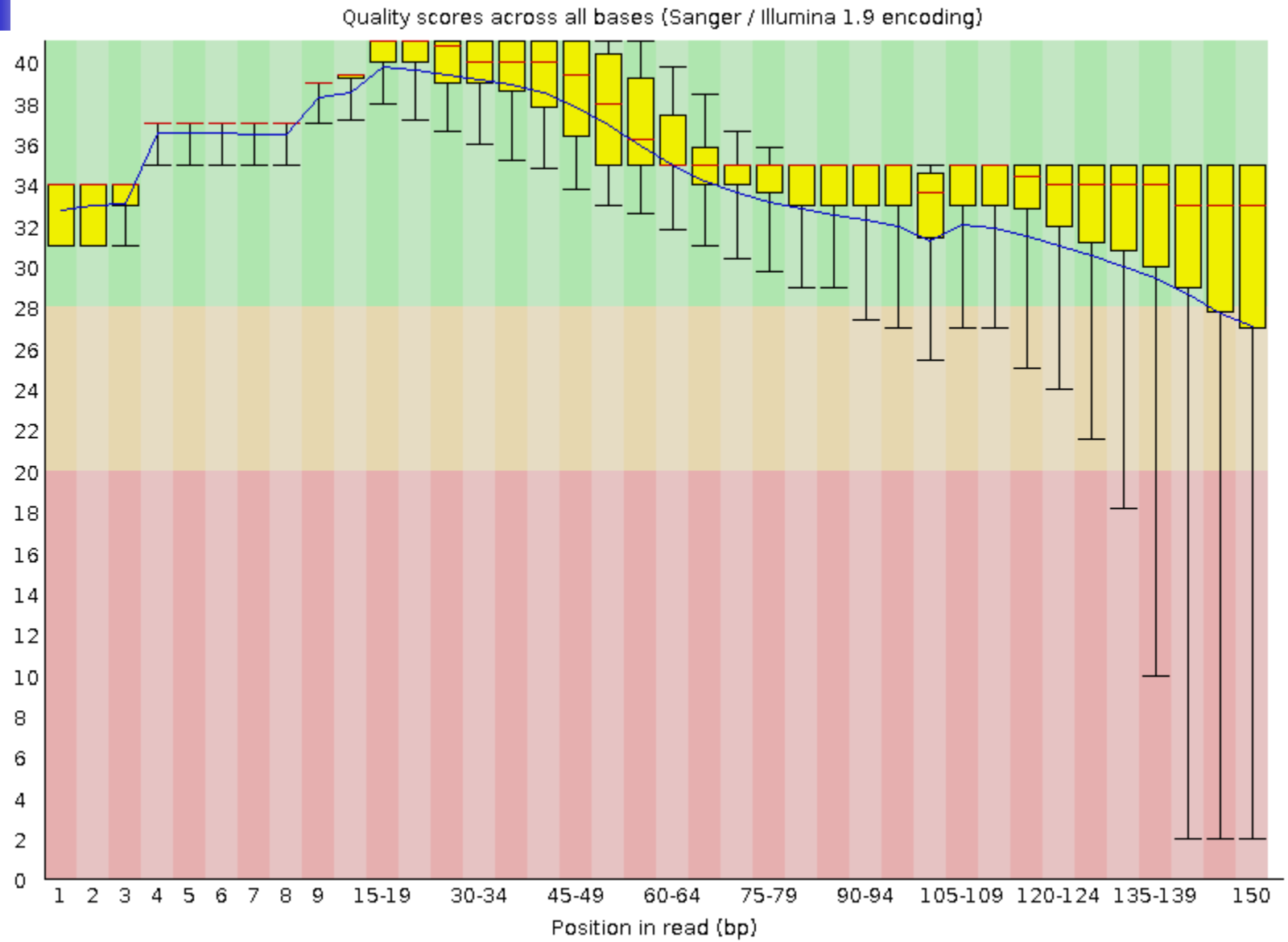
FastQC Report

Measure	Value
Filename	SRR2584863_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1553259
Sequences flagged as poor quality	0
Sequence length	150
%GC	50

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

SRR2584863_1.fastq.gz



Running FastQC

$ cd ~/dc_workshop/data/untrimmed_fastq/

- Exercise
- How big are the files?
- Hint: Look at the options for the ls command to see how to show file sizes.
- $ ls -l -h

```
-rw-rw-r-- 1 dcuser dcuser 545M Jul  6 20:27 SRR2584863_1.fastq
-rw-rw-r-- 1 dcuser dcuser 183M Jul  6 20:29 SRR2584863_2.fastq.gz
-rw-rw-r-- 1 dcuser dcuser 309M Jul  6 20:34 SRR2584866_1.fastq.gz
-rw-rw-r-- 1 dcuser dcuser 296M Jul  6 20:37 SRR2584866_2.fastq.gz
-rw-rw-r-- 1 dcuser dcuser 124M Jul  6 20:22 SRR2589044_1.fastq.gz
-rw-rw-r-- 1 dcuser dcuser 128M Jul  6 20:24 SRR2589044_2.fastq.gz
```

Bài tập

- Đếm số reads trong dữ liệu đã giải trình tự của bài thực hành

```
-rw-rw-r-- 1 dcuser dcuser 545M Jul  6 20:27 SRR2584863_1.fastq
-rw-rw-r-- 1 dcuser dcuser 183M Jul  6 20:29 SRR2584863_2.fastq.gz
-rw-rw-r-- 1 dcuser dcuser 309M Jul  6 20:34 SRR2584866_1.fastq.gz
-rw-rw-r-- 1 dcuser dcuser 296M Jul  6 20:37 SRR2584866_2.fastq.gz
-rw-rw-r-- 1 dcuser dcuser 124M Jul  6 20:22 SRR2589044_1.fastq.gz
-rw-rw-r-- 1 dcuser dcuser 128M Jul  6 20:24 SRR2589044_2.fastq.gz
```

Running FastQC

- $ fastqc *.fastq*

```
$ mkdir -p
~/dc_workshop/results/fastqc_untrimmed_reads
$ mv *.zip
~/dc_workshop/results/fastqc_untrimmed_reads/
$ mv *.html
~/dc_workshop/results/fastqc_untrimmed_reads/
```

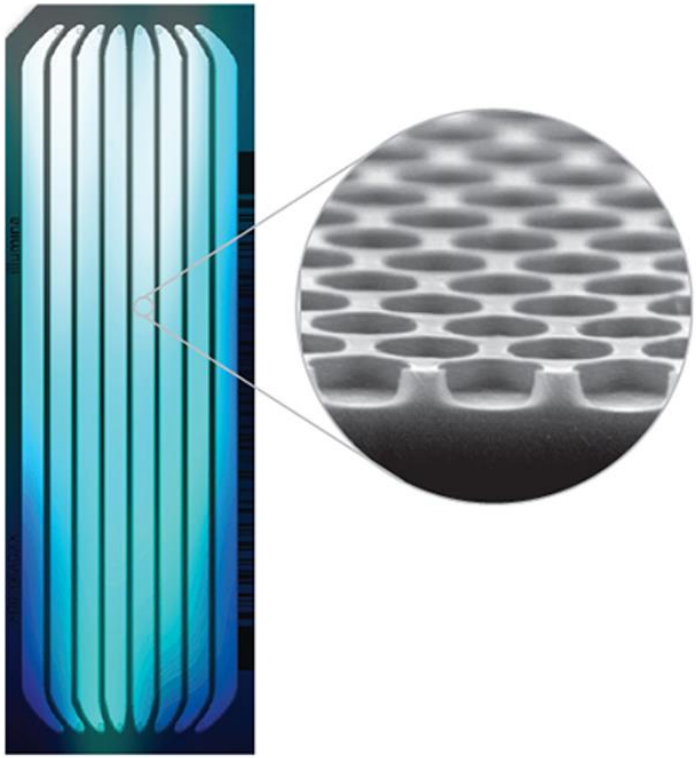
```
SRR2584863_1.fastq
SRR2584866_1_fastqc.html
SRR2589044_1_fastqc.html
SRR2584863_1_fastqc.html
SRR2584866_1_fastqc.zip
SRR2589044_1_fastqc.zip
SRR2584863_1_fastqc.zip
SRR2584866_1.fastq.gz
SRR2589044_1.fastq.gz
SRR2584863_2_fastqc.html
SRR2584866_2_fastqc.html
SRR2589044_2_fastqc.html
SRR2584863_2_fastqc.zip
SRR2584866_2_fastqc.zip
SRR2589044_2_fastqc.zip
SRR2584863_2.fastq.gz
SRR2584866_2.fastq.gz
SRR2589044_2.fastq.gz
```

- Exercise:
  - Discuss your results with a neighbor.
  - Which sample(s) looks the best in terms of per base sequence quality?
  - Which sample(s) look the worst?

chất lượng theo flowcell tile (ô)

Per tile sequence quality

- The machines that perform sequencing are divided into tiles
- This plot displays patterns in base quality along these tiles.
- Consistently low scores are often found around the edges, but hot spots can also occur in the middle if an air bubble was introduced at some point during the run.



- The first line, identifying the sequence, contains the following elements.

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>:<UMI>

<read>:<is filtered>:<control number>:<index>

- **@SIM:1:FCX**:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:0:ATCCGA
- @SRR2584863.1 HWI-ST957:244:H73TDADXX:1:1101:4712:2181/1

Element	Requirements	Description
@	@	Each sequence identifier line starts with @.
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID.
<run number>	Numerical	Run number on instrument.
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	

@SIM:1:FCX:**1:15:6329:1045**:GATTACT+GTCTTAAC 1:N:0:ATCCGA

@SRR2584863.1 HWI-ST957:244:H73TDADXX:**1:1101:4712:2181**/1

Element	**Requirements**	**Description**
`<lane>`	Numerical	Lane number.
`<tile>`	Numerical	Tile number.
`<x_pos>`	Numerical	X coordinate of cluster.
`<y_pos>`	Numerical	Y coordinate of cluster.
<UMI>	Restricted characters: A/T/G/C/N	Optional, appears when UMI is specified in sample sheet. UMI sequences for Read 1 and Read 2, seperated by a plus [+].

tuỳ chọn , xuất hiện khi UMI được chỉ định trong mẫu . chuỗi UMI cho read 1 và read 2 , chia cắt bởi dấu +

@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC **1:N**:0:ATCCGA

@SRR2584863.1 HWI-ST957:244:H73TDADXX:1:1101:4712:2181/1

Element	**Requirements**	**Description**
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end.
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise.

Số read

Y nếu read được lọc r, N nếu chưa

@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:**0:ATCCGA**

@SRR2584863.1 HWI-ST957:244:H73TDADXX:1:1101:4712:2181/1

Element	Requirements	Description
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. 0 nếu không có control bit được bật ,nếu không thì nó sẽ là số chẵn On HiSeq X and NextSeq systems, control specification is not performed and this number is always 0.
<index>	Restricted characters: A/T/G/C/N	Index of the read.

- An example of a valid entry is as follows; note the space preceding the read number element:

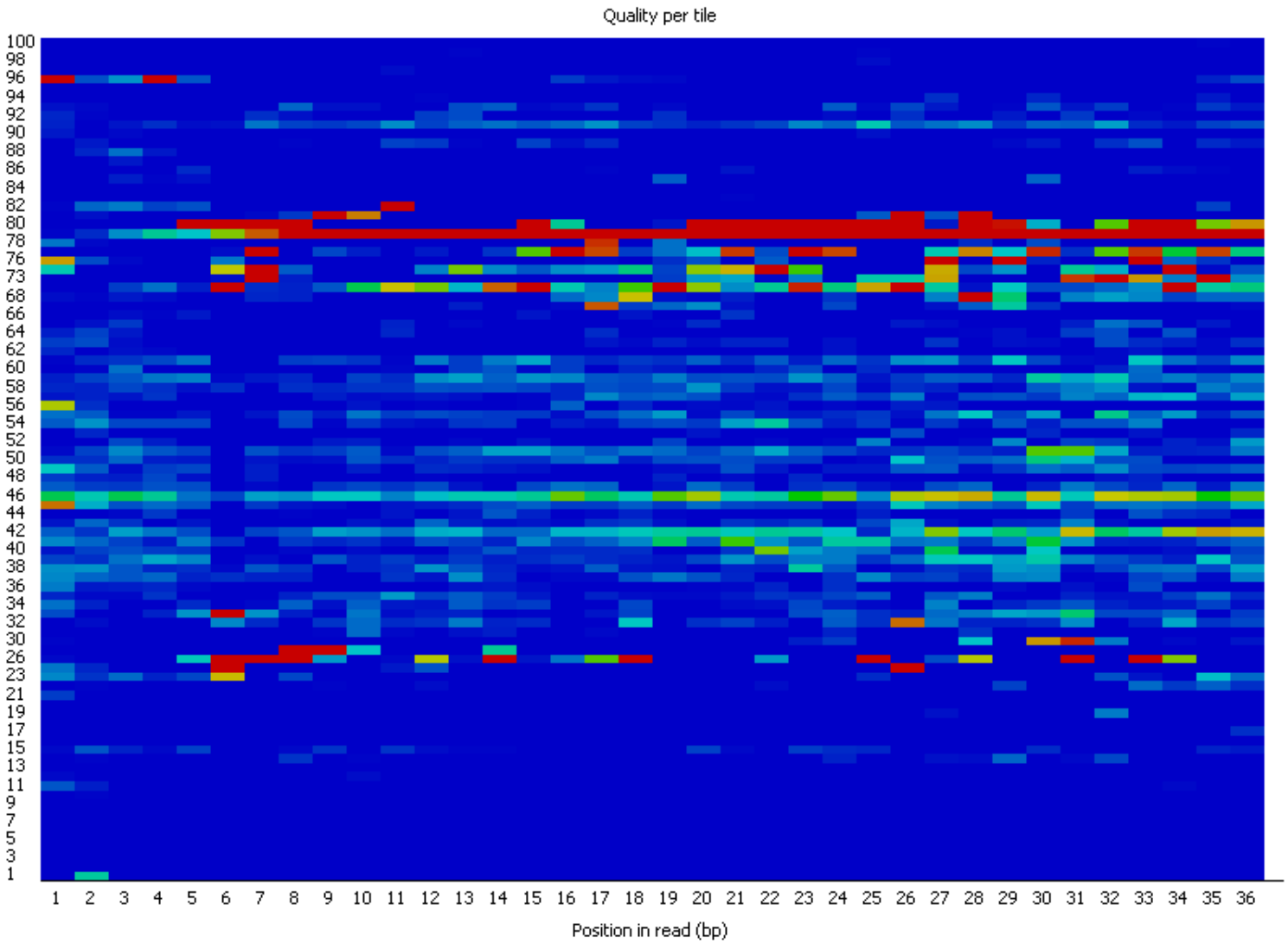
@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:0:ATCCGA

TCGCACTCAACGCCCTGCATATGACAAGACAGAATC

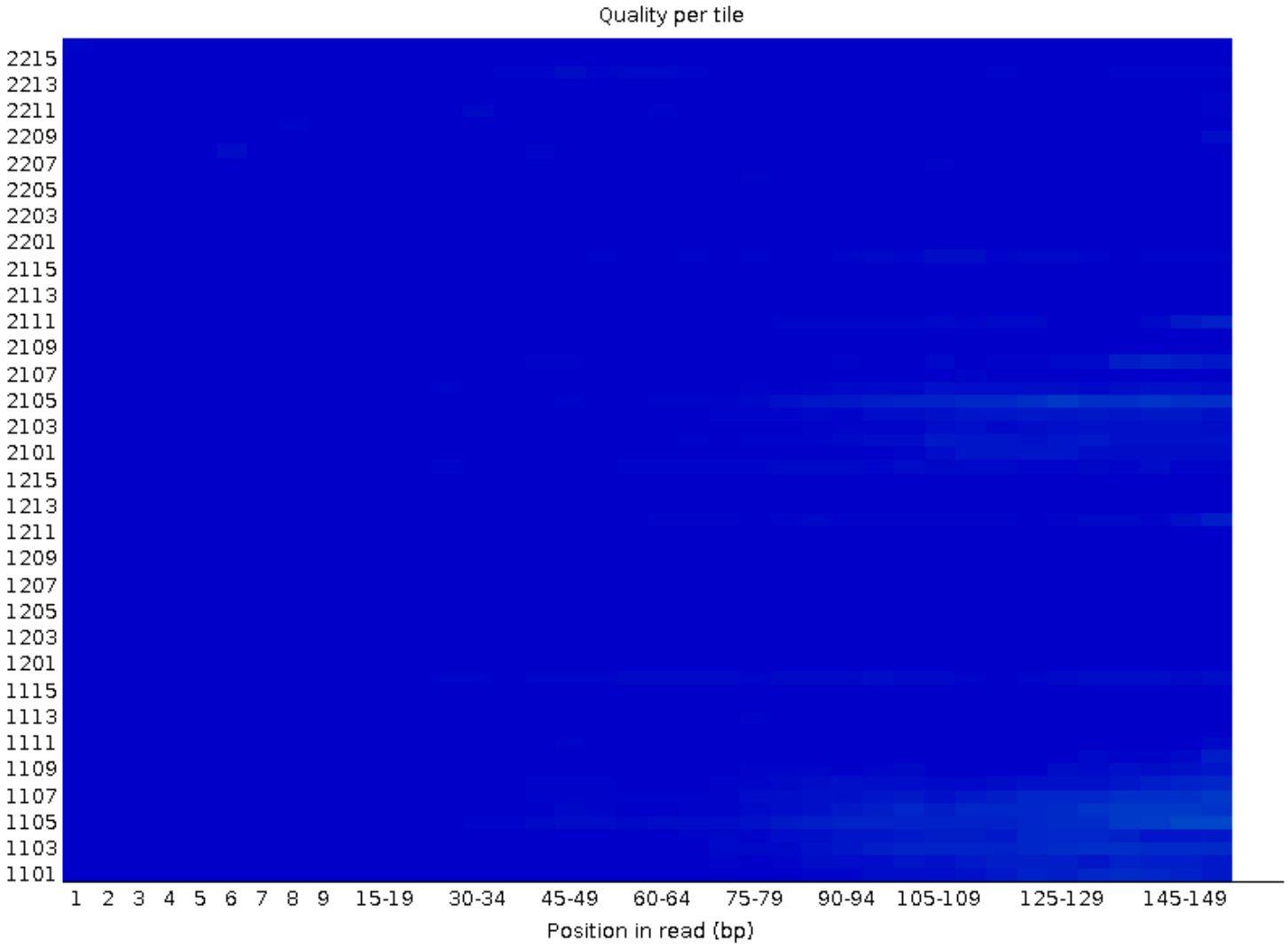
+

<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=

Per tile sequence quality



Per tile sequence quality SRR2584863_1.fastq.gz



Per tile sequence quality

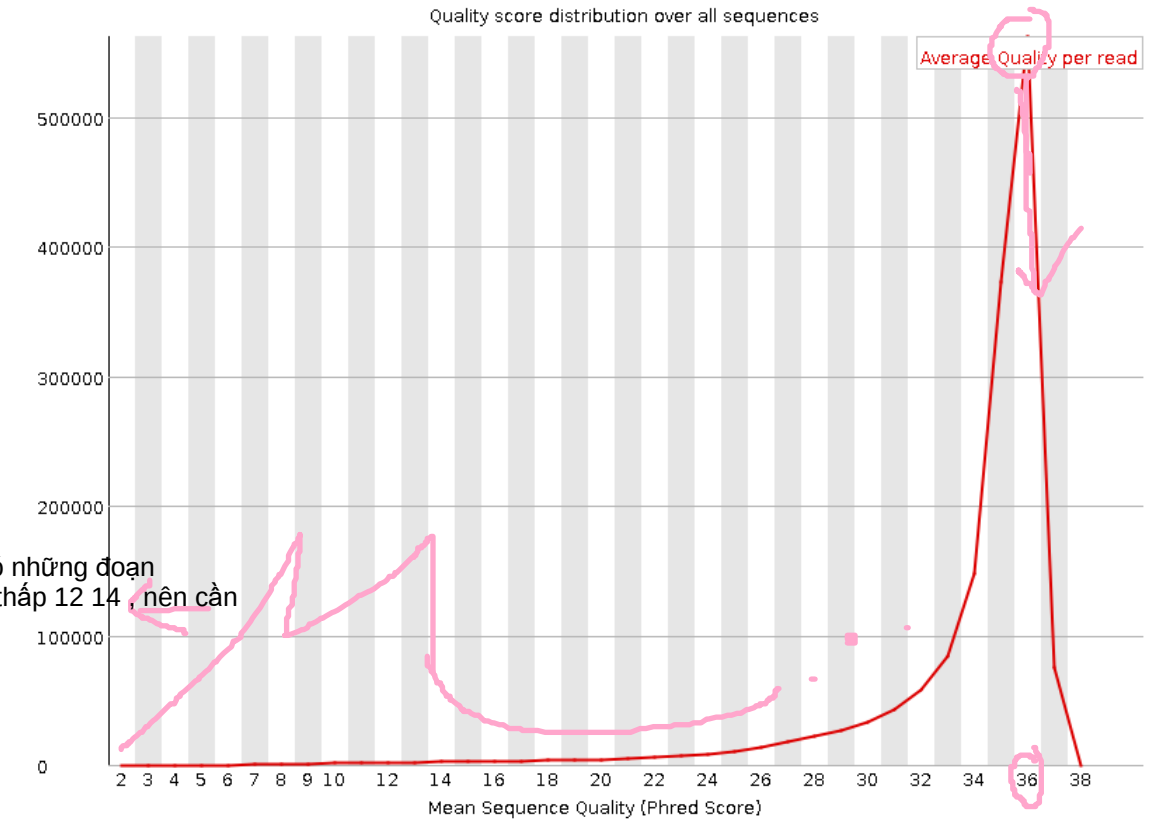
- **Warning**
  - This module will issue a warning if any tile shows a mean Phred score more than 2 less than the mean for that base across all tiles.
- **Failure**
  - This module will issue a warning if any tile shows a mean Phred score more than 5 less than the mean for that base across all tiles.

Chất lượng theo đoạn trình tự

Per sequence quality scores

Đánh giá điểm chất lượng trung bình của mỗi đoạn trình tự trong từng read

- A density plot of quality for all reads at all positions.
- This plot shows what quality scores are most common.



giả sử có những đoạn có điểm thấp 12 14 , nên cần loại bỏ

-> biểu đồ này cho thấy điểm chất lượng nào là phổ biến nhấ

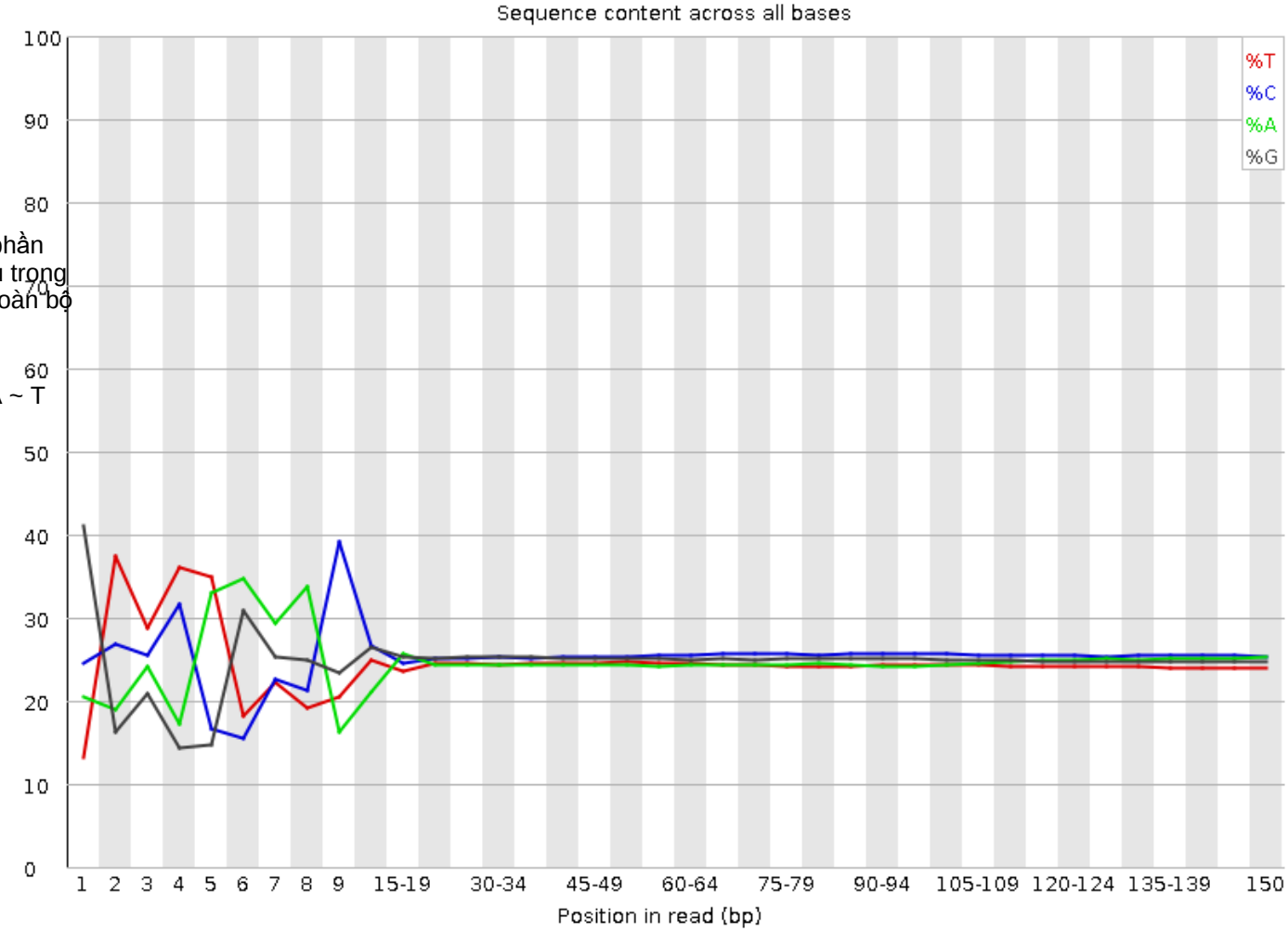
chủ yếu các đoạn trình tự có điểm chất lượng là 36

tỉ lệ A T G C trong dữ liệu

Per base sequence content

biểu đồ hiển thi phần trăm số lượng nu trong 1 position trong toàn bộ reads

nếu đẹp thì tỉ lệ của A ~ T và G ~C

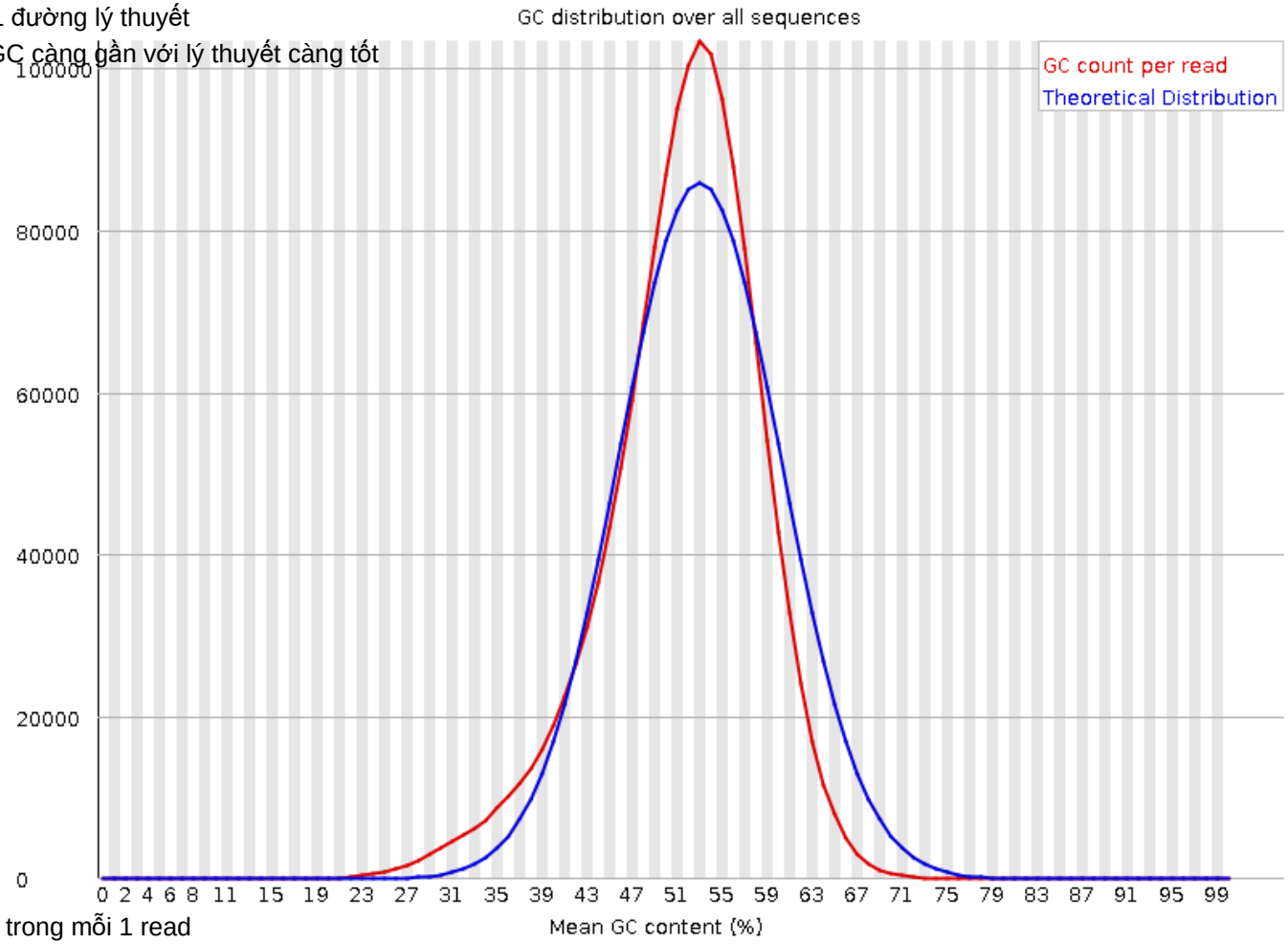


tỉ lệ GC trong dữ liệu

Per sequence GC content

thống kê tỉ lệ %GC ở mỗi đoạn trình tự trong 1 read
được so sánh với 1 đường lý thuyết
nếu dữ liệu có tỉ lệ GC càng gần với lý thuyết càng tốt

- A density plot of average GC content in each of the reads



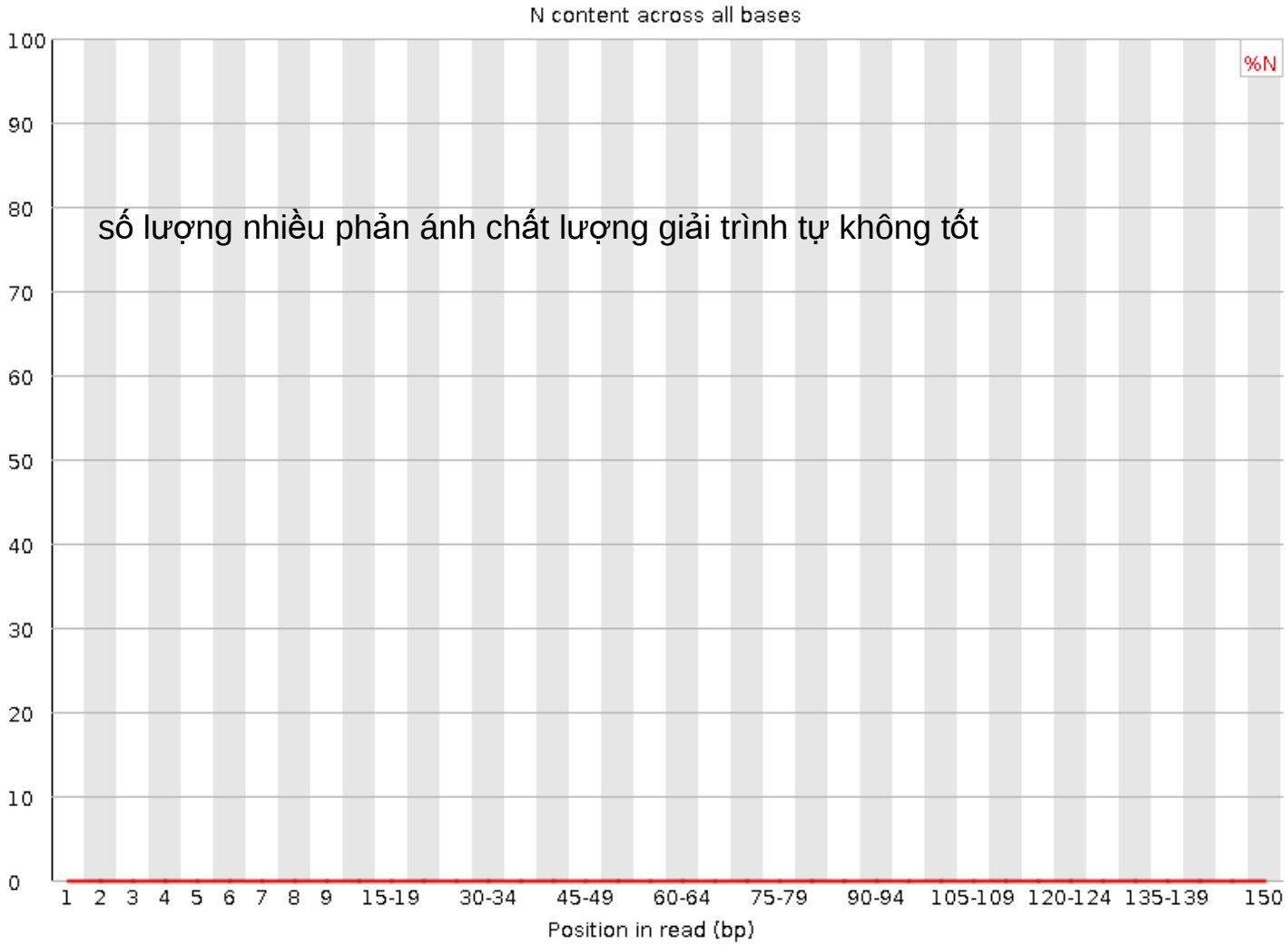
biểu đồ mật độ của GC trung bình trong mỗi 1 read

đường màu xanh : phân phối lý thuyết
Đường màu đỏ : phân phối trong toàn bộ reads

Số lượng cặp bazo không xác định (trình tự N) trong dữ liệu

Per base N content

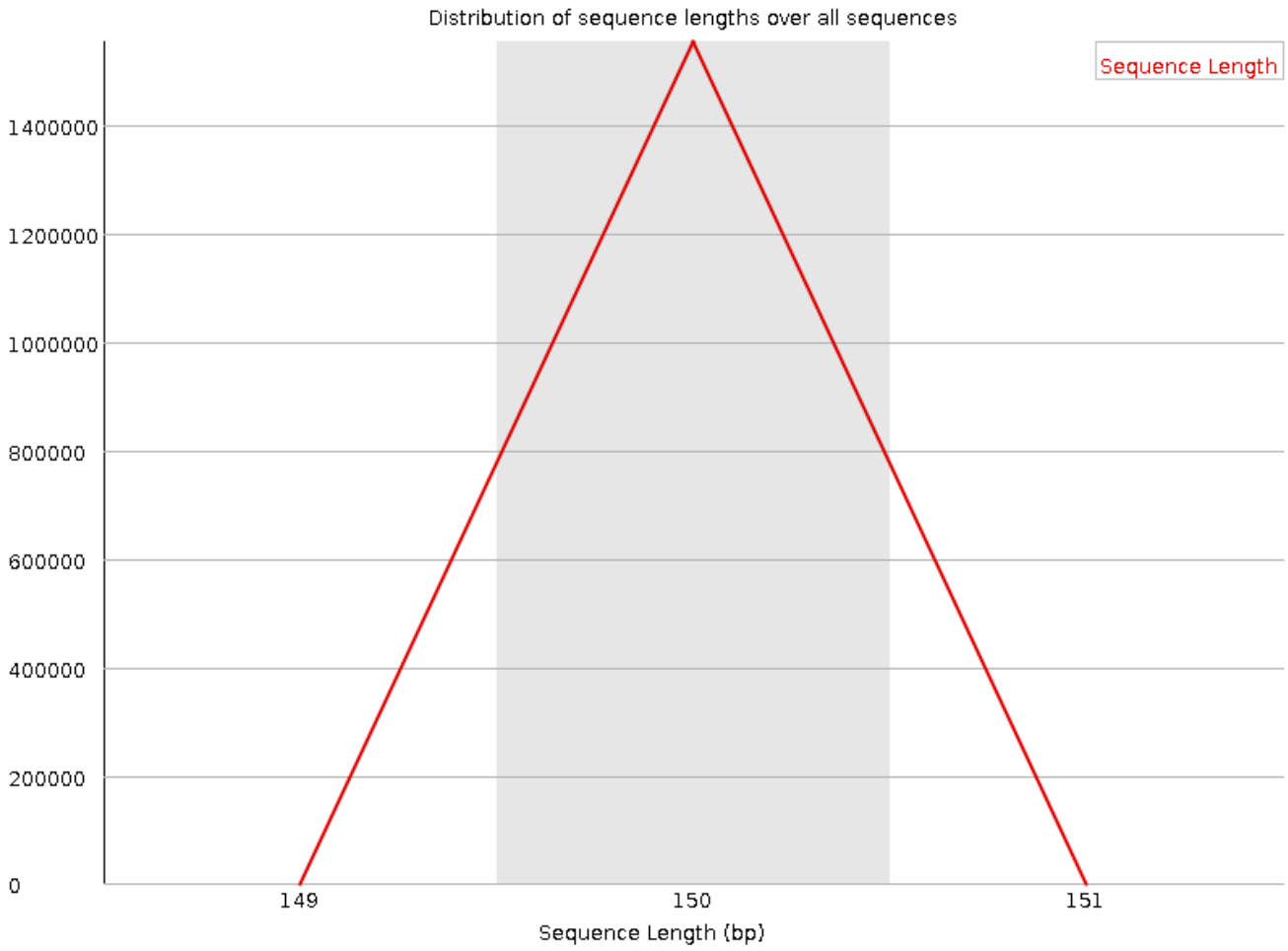
tính toán số lượng trình tự không xác định (N) trong dữ liệu



số lượng nhiều phản ánh chất lượng giải trình tự không tốt

Phân bố độ dài đoạn trình tự

Sequence Length Distribution



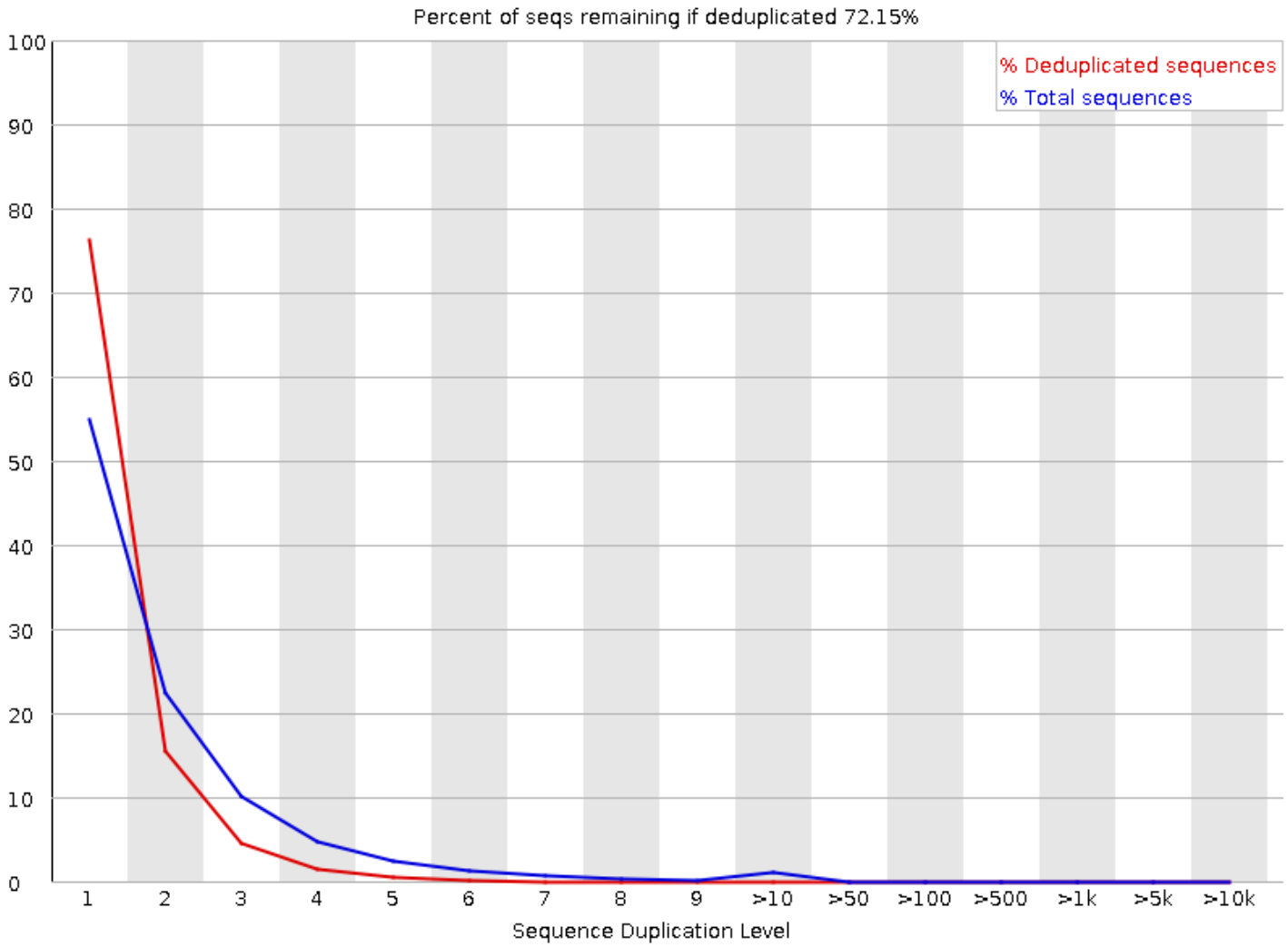
thống kê độ dài các đoạn trình tự trong dữ liệu

theo sơ đồ, hầu hết các đoạn trình tự có độ dài 150 base pair

nếu có một số đoạn trình tự ngắn cần loại bỏ

Mức độ lặp lại

Sequence Duplication Levels



đánh giá mức độ lặp lại của đoạn trình tự

Overrepresented sequences

- A list of sequences that occur more frequently than would be expected by chance.

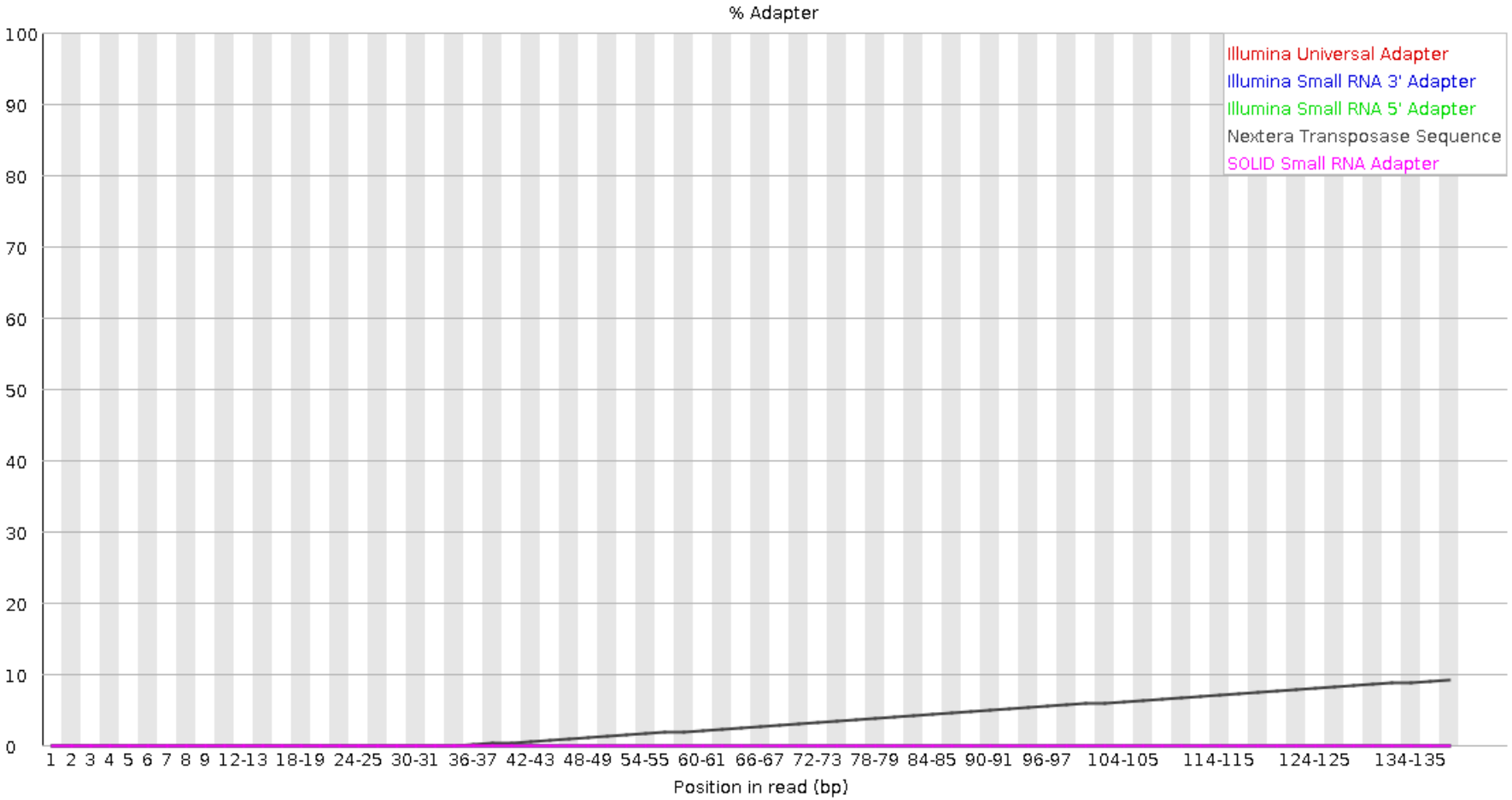
- No overrepresented sequences

Các adapter còn xót lại trong dữ liệu

Adapter Content

- a graph indicating where adapater sequences occur in the reads.

biểu đồ hiển thị chỗ nào adapter xuất hiện trong 1 read



Other files

$ cd ~/dc_workshop/results/fastqc_untrimmed_reads/

```
SRR2584863_1_fastqc.html  SRR2584866_1_fastqc.html  SRR2589044_1_fastqc.html
SRR2584863_1_fastqc.zip   SRR2584866_1_fastqc.zip   SRR2589044_1_fastqc.zip
SRR2584863_2_fastqc.html  SRR2584866_2_fastqc.html  SRR2589044_2_fastqc.html
SRR2584863_2_fastqc.zip   SRR2584866_2_fastqc.zip   SRR2589044_2_fastqc.zip
```

```
$ unzip *.zip

$ for filename in *.zip
> do
> unzip $filename
> done
```

```
$ ls -F SRR2584863_1_fastqc/

fastqc_data.txt
fastqc.fo
fastqc_report.html
Icons/
Images/
summary.txt
```

summary.txt

- $ less SRR2584863_1_fastqc/summary.txt

```
PASS    Basic Statistics        SRR2584863_1.fastq
PASS    Per base sequence quality       SRR2584863_1.fastq
PASS    Per tile sequence quality       SRR2584863_1.fastq
PASS    Per sequence quality scores     SRR2584863_1.fastq
WARN    Per base sequence content       SRR2584863_1.fastq
WARN    Per sequence GC content SRR2584863_1.fastq
PASS    Per base N content      SRR2584863_1.fastq
PASS    Sequence Length Distribution    SRR2584863_1.fastq
PASS    Sequence Duplication Levels     SRR2584863_1.fastq
PASS    Overrepresented sequences       SRR2584863_1.fastq
WARN    Adapter Content SRR2584863_1.fastq
```

Documenting our work

```
$ cat */summary.txt >
~/dc_workshop/docs/fastqc_summaries.txt
```

- Which samples failed at least one of FastQC's quality tests?
- What test(s) did those samples fail?

```
$ cd ~/dc_workshop/docs
$ grep FAIL fastqc_summaries.txt
```

```
FAIL    Per base sequence quality       SRR2584863_2.fastq.gz
FAIL    Per tile sequence quality       SRR2584863_2.fastq.gz
FAIL    Per base sequence content       SRR2584863_2.fastq.gz
FAIL    Per base sequence quality       SRR2584866_1.fastq.gz
FAIL    Per base sequence content       SRR2584866_1.fastq.gz
FAIL    Adapter Content SRR2584866_1.fastq.gz
FAIL    Adapter Content SRR2584866_2.fastq.gz
FAIL    Adapter Content SRR2589044_1.fastq.gz
FAIL    Per base sequence quality       SRR2589044_2.fastq.gz
FAIL    Per tile sequence quality       SRR2589044_2.fastq.gz
FAIL    Per base sequence content       SRR2589044_2.fastq.gz
FAIL    Adapter Content SRR2589044_2.fastq.gz
```

3. Trimming and Filtering

Trimming and Filtering

- Questions
  - How can I get rid of sequence data that does not meet my quality standards?
- Objectives
  - Clean FASTQ reads using Trimmomatic.
  - Select and set multiple options for command-line bioinformatic tools.
  - Write for loops with two variables.

Trimmomatic Example

$ trimmomatic PE -threads 4
SRR_1056_1.fastq

SRR_1056_2.fastq

SRR_1056_1.trimmed.fastq
SRR_1056_1un.trimmed.fastq

SRR_1056_2.trimmed.fastq

SRR_1056_2un.trimmed.fastq

ILLUMINACLIP:SRR_adapters.fa thực thi cắt adapter từ input file sử dujng sequence trong file SRR_adapters.fa
SLIDINGWINDOW:4:20 sử dụng cửa sổ trượt độ dài = 4 , sẽ xoá các ba zơ nếu điểm phred < 20

- While using FastQC we saw that Nextera adapters were present in our samples.

```
$ cd ~/dc_workshop/data/untrimmed_fastq
$ cp ~/.miniconda3/pkgs/trimmomatic-0.38-0/share/trimmomatic-0.38-0/adapters/NexteraPE-PE.fa .
```

https://github.com/timflutre/trimmomatic/blob/master/adapters/NexteraPE-PE.fa

NexteraPE-PE.fa

>PrefixNX/1

AGATGTGTATAAGAGACAG

>PrefixNX/2

AGATGTGTATAAGAGACAG

>Trans1

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

>Trans1_rc

CTGTCTCTTATACACATCTGACGCTGCCGACGA

>Trans2

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

>Trans2_rc

CTGTCTCTTATACACATCTCCGAGCCCACGAGAC

Thực hiện với trimmomatic

$ trimmomatic PE SRR2589044_1.fastq.gz

SRR2589044_2.fastq.gz

SRR2589044_1.trim.fastq.gz
SRR2589044_1un.trim.fastq.gz

SRR2589044_2.trim.fastq.gz
SRR2589044_2un.trim.fastq.gz

SLIDINGWINDOW:4:20

MINLEN:25 loại bỏ các read không có ít nhất 25 bazo sau khi được trim

ILLUMINACLIP:NexteraPE-PE.fa:2:40:15

Using PrefixPair: 'AGATGTGTATAAGAGACAG' and 'AGATGTGTATAAGAGACAG'

Using Long Clipping Sequence: 'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG'

Using Long Clipping Sequence: 'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG'

Using Long Clipping Sequence: 'CTGTCTCTTATACACATCTCCGAGCCCACGAGAC'

Using Long Clipping Sequence: 'CTGTCTCTTATACACATCTGACGCTGCCGACGA'

ILLUMINACLIP: Using 1 prefix pairs, 4 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences

Quality encoding detected as phred33

Input Read Pairs: 1107090

Both Surviving: 885220 (79.96%)

Forward Only Surviving: 216472 (19.55%)

Reverse Only Surviving: 2850 (0.26%)

Dropped: 2548 (0.23%)

TrimmomaticPE: Completed successfully

Exercise

- Use the output from your Trimmomatic command to answer the following questions.

1) What percent of reads did we discard from our sample?

2) What percent of reads did we keep both pairs?

- 1) 0.23% 2) 79.96%

```
$ ls SRR2589044*
SRR2589044_1.fastq.gz
SRR2589044_1un.trim.fastq.gz
SRR2589044_2.trim.fastq.gz
SRR2589044_1.trim.fastq.gz
SRR2589044_2.fastq.gz
SRR2589044_2un.trim.fastq.gz

$ ls SRR2589044* -l -h
```

```
$ for infile in *_1.fastq.gz
> do
>   base=$(basename ${infile} _1.fastq.gz)
>   trimmomatic PE ${infile} ${base}_2.fastq.gz \
>                ${base}_1.trim.fastq.gz
${base}_1un.trim.fastq.gz \
>                ${base}_2.trim.fastq.gz
${base}_2un.trim.fastq.gz \
>                SLIDINGWINDOW:4:20 MINLEN:25
ILLUMINACLIP:NexteraPE-PE.fa:2:40:15
> done
```

```
$ ls
NexteraPE-PE.fa
SRR2584866_1.fastq.gz
SRR2589044_1.trim.fastq.gz
SRR2584863_1.fastq.gz
SRR2584866_1.trim.fastq.gz
SRR2589044_1un.trim.fastq.gz
SRR2584863_1.trim.fastq.gz
SRR2584866_1un.trim.fastq.gz
SRR2589044_2.fastq.gz
```

SRR2584863_1un.trim.fastq.gz
SRR2584866_2.fastq.gz
SRR2589044_2.trim.fastq.gz

SRR2584863_2.fastq.gz
SRR2584866_2.trim.fastq.gz
SRR2589044_2un.trim.fastq.gz

SRR2584863_2.trim.fastq.gz
SRR2584866_2un.trim.fastq.gz

SRR2584863_2un.trim.fastq.gz
SRR2589044_1.fastq.gz

Exercise

- We trimmed our FASTQ files with Nextera adapters, but there are other adapters that are commonly used.
- What other adapter files came with Trimmomatic?

```
$ ls ~/miniconda3/pkgs/trimmomatic-0.38-0/share/trimmomatic-0.38-0/adapters/
NexteraPE-PE.fa
TruSeq2-SE.fa
TruSeq3-PE.fa
TruSeq2-PE.fa
TruSeq3-PE-2.fa
TruSeq3-SE.fa
```

```
$ cd ~/dc_workshop/data/untrimmed_fastq
$ mkdir ../trimmed_fastq
$ mv *.trim* ../trimmed_fastq
$ cd ../trimmed_fastq
$ ls
```

SRR2584863_1.trim.fastq.gz
SRR2584866_1.trim.fastq.gz
SRR2589044_1.trim.fastq.gz

SRR2584863_1un.trim.fastq.gz
SRR2584866_1un.trim.fastq.gz
SRR2589044_1un.trim.fastq.gz

Bonus exercise (advanced)

- Now that our samples have gone through quality control, they should perform better on the quality tests run by FastQC.
- Go ahead and re-run FastQC on your trimmed FASTQ files and visualize the HTML files to see whether your per base sequence quality is higher after trimming.

```
$ fastqc
~/dc_workshop/data/trimmed_fastq/*.fastq*
```

Bài tập

- Xác định số lượng trình tự đã bị loại bỏ
- Chạy lại FASTQ với các trình tự này và nhận xét kết quả