



THE 3rd VIETNAM SCHOOL OF BIOLOGY (VSOB-3)

Bioinformatic Analysis For Bulk RNAseq Data

December 06th-08th, 2024, ICISE, Quy Nhon, Vietnam

RNA-seq: Upstream Analysis

Giảng viên:

TS. Trần Thị Thanh Tâm

TS. Đỗ Hoàng Đăng Khoa

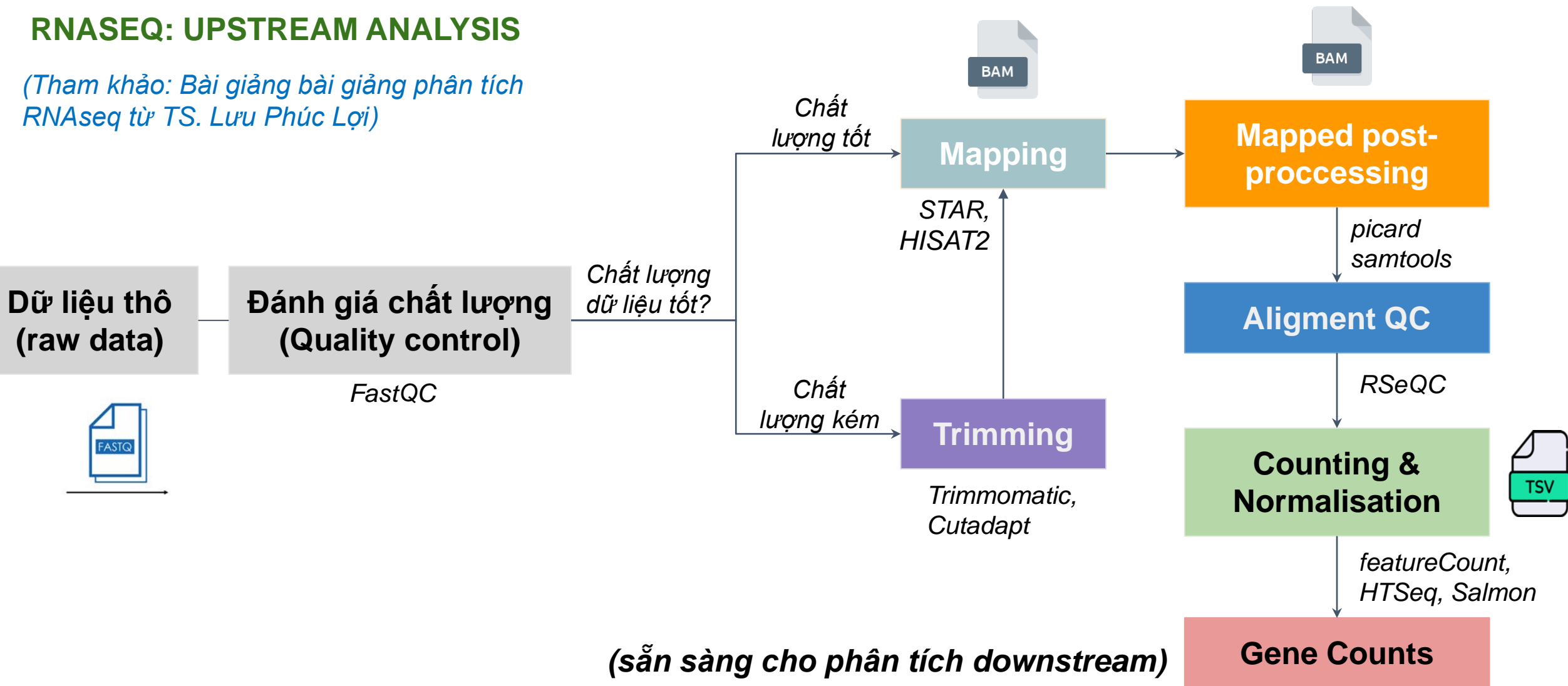
Nội dung

- Đánh giá chất lượng dữ liệu (*Quality control*)
- Tinh sạch dữ liệu (*Trimming*)
- So sánh dữ liệu giải trình tự với bộ gen tham chiếu (*Mapping*)
- Mapped post-processing
- Alignment QC

Quy trình phân tích dữ liệu

RNASEQ: UPSTREAM ANALYSIS

(Tham khảo: Bài giảng bài giảng phân tích RNAseq từ TS. Lưu Phúc Lợi)



1. Dữ liệu giải trình tự

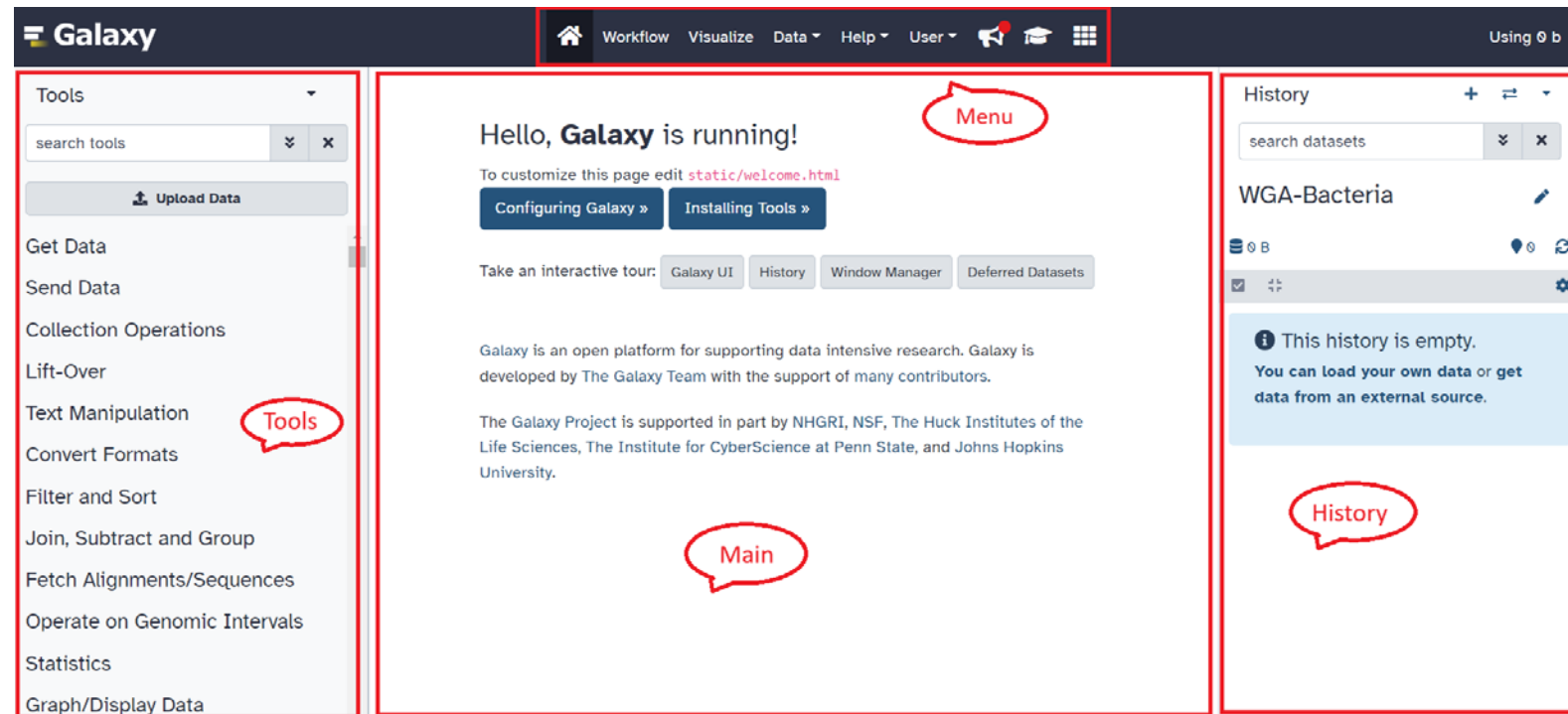
Phân tích dữ liệu với Galaxy

- Đăng nhập & tạo tài khoản tại mục “Log in or Register” trên Galaxy web server tại Úc

<https://usegalaxy.org.au/join-training/vsob3/>

Giao diện của Galaxy gồm 4 panels:

- Tools:** Liệt kê tất cả các công cụ, phần mềm
- Main:** hiển thị các phân tích và các kết quả phân tích.
- History:** lưu trữ lịch sử phân tích của bạn.
- Menu:** gồm các tabs khác nhau Workflow, Data, Help



Tạo lịch sử theo từng dự án phân tích

- Tại mục **History**: lưu trữ lịch sử phân tích của bạn.
- Tạo lịch sử mới -> đặt tên theo từng dự án, ví dụ: *RNAseq_upstream*

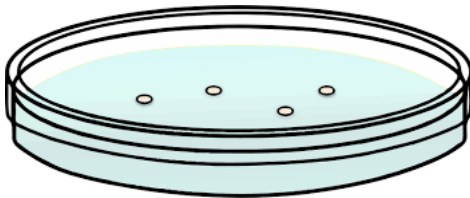
The screenshot shows the 'History' section of a software interface. It includes a search bar labeled 'search datasets', a list of history items, and a 'Save' button. Red arrows and text annotations guide the user through the steps to create a new history item.

1. Create new history (Tạo lịch sử mới) - points to the '+' icon in the History header.
2. Edit (Sửa tên) - points to the edit icon (pencil) next to the 'RNAseq_upstream' item.
3. Điền tên theo dự án phân tích (Fill name according to the analysis project) - points to the text input field containing 'RNAseq_upstream'.
4. Lưu lại tên mới (Save the new name) - points to the 'Save' button.

Below the 'Save' button, there is a message box stating: "This history is empty. You can load your own data or get data from an external source."

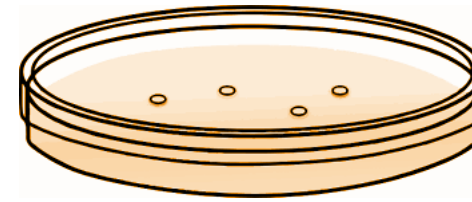
Dữ liệu thực hành

Comparative gene expression profiling analysis of RNA-seq data for *S. cerevisiae* cells treated with edelfosine for 60 minutes (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE227381>)



Control (X3)

```
WT_C_1_R1.fastq.gz  
WT_C_1_R2.fastq.gz  
WT_C_2_R1.fastq.gz  
WT_C_2_R2.fastq.gz  
WT_C_3_R1.fastq.gz  
WT_C_3_R2.fastq.gz
```



Edelfosine treatment (X3)

```
WT_E_1_R1.fastq.gz  
WT_E_1_R2.fastq.gz  
WT_E_2_R1.fastq.gz  
WT_E_2_R2.fastq.gz  
WT_E_3_R1.fastq.gz  
WT_E_3_R2.fastq.gz
```

- Mỗi mẫu gồm 2 files dữ liệu giải trình tự từ chiều xuôi và chiều ngược (R1 & R2)
- Dữ liệu được lưu dưới dạng **Fastq** file bao gồm cả điểm chất lượng (quality score) và trình tự nucleotide

Upload dữ liệu lên Galaxy server

1. **Download dữ liệu giải trình tự**, link download:

https://drive.google.com/drive/folders/14k-lzmrjOdmzaA2a6vyQZwnoL6i_ABEa?usp=sharing

Chọn 1 cặp dữ liệu từ 1 mẫu (v.d. WT_E_2_R1_fastq_gz.gz & WT_E_2_R2_fastq_gz.gz)

2. **Upload dữ liệu lên Galaxy server**: Click **Upload** -> **Choose local file**: chọn file upload -> **Start** -> **Close**

Upload from Disk or Web to **RNAseq_upstream**

Regular Composite Collection Rule-based

Please wait...2 out of 2 remaining...

File	Size	Type	Search	Reference	Progress	Actions
WT_E_2_R1.fastq.gz	481.6 MB	Auto-detect	Q unspecified (?)		69%	
WT_E_2_R2.fastq.gz	547.2 MB	Auto-detect	Q unspecified (?)		0%	

Type (set all): Auto-detect Reference (set all): unspecified (?)

2 Choose local file Choose remote files Paste/Fetch data Start 3 Reset Close

History

search datasets

RNAseq_upstream

1.08 GB 2

2: WT_E_2_R2.fastq.gz

1: WT_E_2_R1.fastq.gz

- Quan sát định dạng file dữ liệu
- Loại file: **Fastq**

Dữ liệu thô (Raw data)

- Dữ liệu giải trình tự Illumina dạng paired-end gồm dữ liệu mỗi xuôi và dữ liệu từ mỗi ngược
- Định dạng tệp là FASTQ (chứa trình tự DNA và điểm chất lượng (quality score)) cho mỗi nucleotide.

Mỗi đoạn đọc gồm 4 dòng

1. **Định danh:** Tên đoạn trình tự bắt đầu “@”
2. **Trình tự nucleotide:** chứa trình tự thu được sau giải trình tự
3. **“+”:** Phân cách trình tự và chất lượng
4. **Điểm chất lượng** (theo bảng mã ASCII)

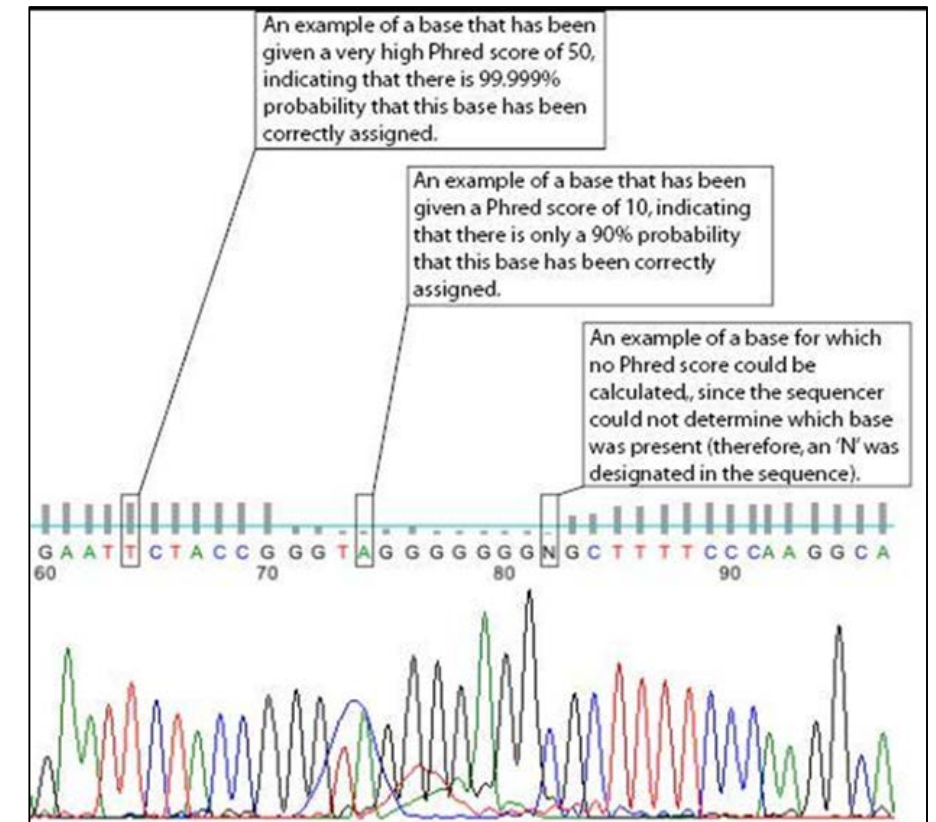
Read 1	Identifier	● @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
	Sequence	● TTGCCTGCCTATCATTTCAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
	'+' sign	● +
Read 2	Quality scores	● hhhhhhhhhghhghhhhhfhhhhfffffe'ee['X]b[d[ed'[Y[^Y
	Identifier	● @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
	Sequence	● GATTGTATGAAAGTATACAACTAAACTGCAGGTGGATCAGAGTAAGTC
	'+' sign	● +
	Quality scores	● hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd

Điểm chất lượng (Phred quality score)

Điểm chất lượng (Phred quality score) là một giá trị số nguyên đại diện cho xác suất ước tính của lỗi của nucleotide

$$Q = -10 \log_{10} P \quad \longrightarrow \quad P = 10^{-\frac{Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%



Ví dụ: Điểm chất lượng (Quality Score) = 30 tương đương với độ chính xác 99.9%

https://en.wikipedia.org/wiki/Phred_quality_score

Điểm chất lượng (Phred quality score)

Điểm chất lượng là giá trị số nguyên được chuyển đổi sang dạng 1 kí tự với **bảng mã ASCII**

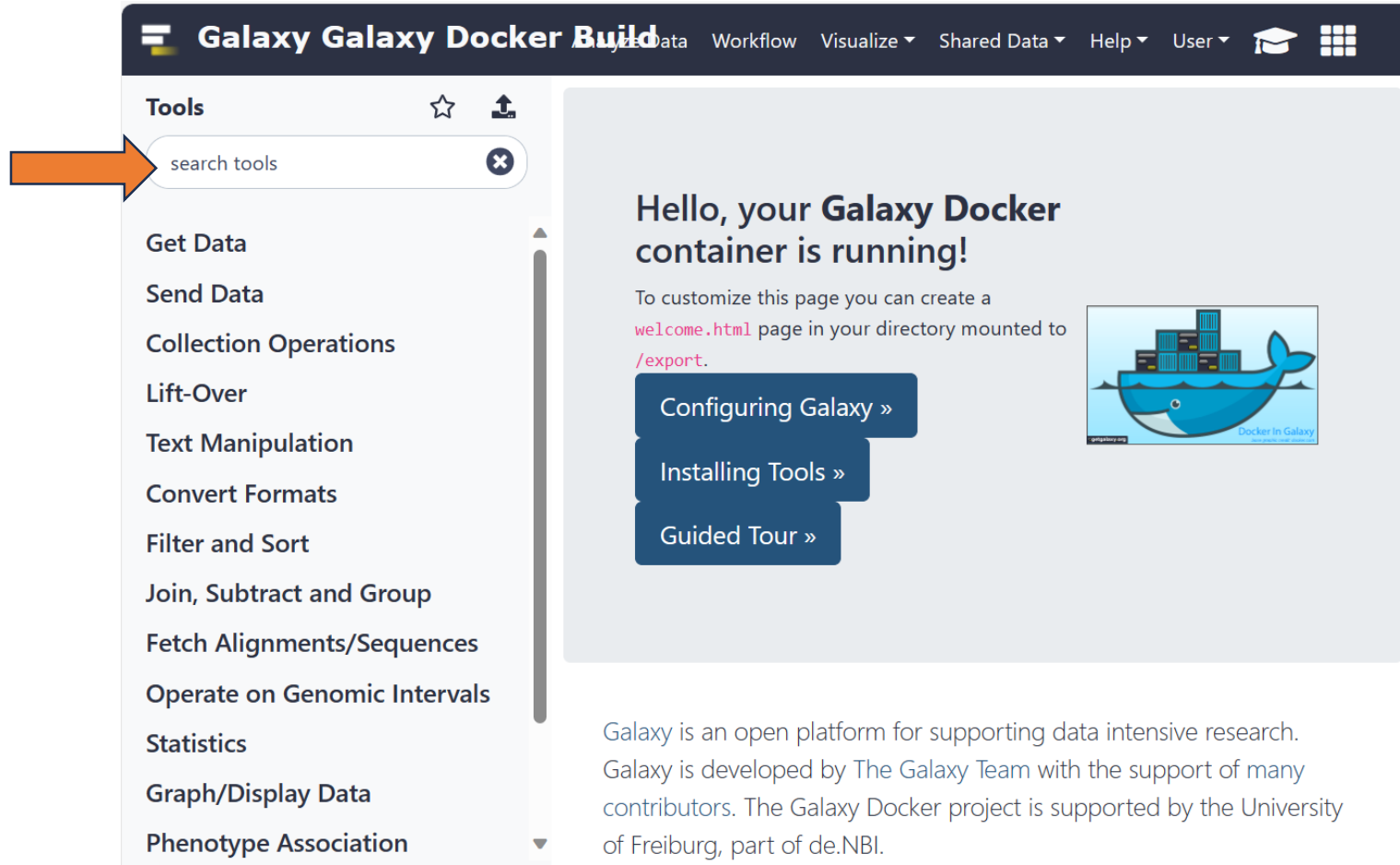
ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

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			

2. Đánh giá chất lượng dữ liệu giải trình tự (Quality control)

FastQC: Đánh giá chất lượng giải trình tự

Nhập tên công
cụ cần tìm kiếm



The screenshot shows the Galaxy Docker Build interface. On the left, a sidebar titled 'Tools' contains a search bar and a list of tool categories. An orange arrow points to the search bar. The main content area displays a welcome message and a list of links for configuration, installation, and a guided tour.

Tools

- Get Data
- Send Data
- Collection Operations
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association

Hello, your Galaxy Docker container is running!

To customize this page you can create a `welcome.html` page in your directory mounted to `/export`.

- [Configuring Galaxy »](#)
- [Installing Tools »](#)
- [Guided Tour »](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors. The Galaxy Docker project is supported by the University of Freiburg, part of de.NBI.

Thực hành đánh giá chất lượng giải trình tự với FastQC

- Lựa chọn công cụ: **FastQC**
- Trong mục **Raw read data from your current history** chọn **Single dataset** → Chọn file dữ liệu fastq -> **Run Tool**
- Thực hiện đối với cả file dữ liệu *forward read* và *reverse read*

The screenshot displays the Galaxy web interface for the FastQC tool. On the left, a sidebar lists tools, with 'fastqc' selected. The main panel shows the 'FastQC Read Quality reports' tool configuration. Under 'Tool Parameters', the 'Raw read data from your current history' section has a dropdown menu showing '1: WT_E_2_R1.fastq.gz', which is highlighted with an orange arrow. The 'Contaminant list' section is optional and currently shows 'Nothing selected'. A 'Run Tool' button is visible at the top right of the tool configuration area, also highlighted with an orange arrow.

➔ Xem kết quả tại:
“**Webpage**” file

Thống kê cơ bản (Basic Statistics)

Bảng tổng quan về dữ liệu:

- Tên tập dữ liệu
- Định dạng dữ liệu
- Máy giải trình tự
- **Số lượng đoạn trình tự**
- Số trình tự có chất lượng thấp
- **Độ dài các đoạn trình tự**
- Tỷ lệ % GC

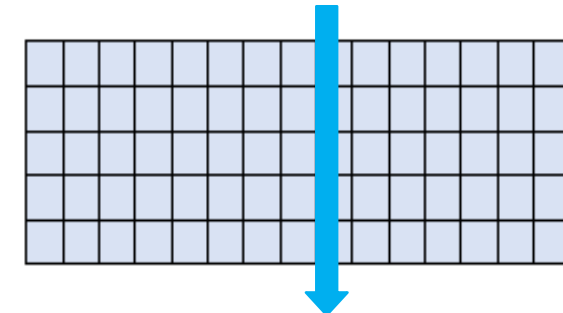


Basic Statistics

Measure	Value
Filename	WT_E_2_R1_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10391006
Total Bases	783.2 Mbp
Sequences flagged as poor quality	0
Sequence length	35-76
%GC	41

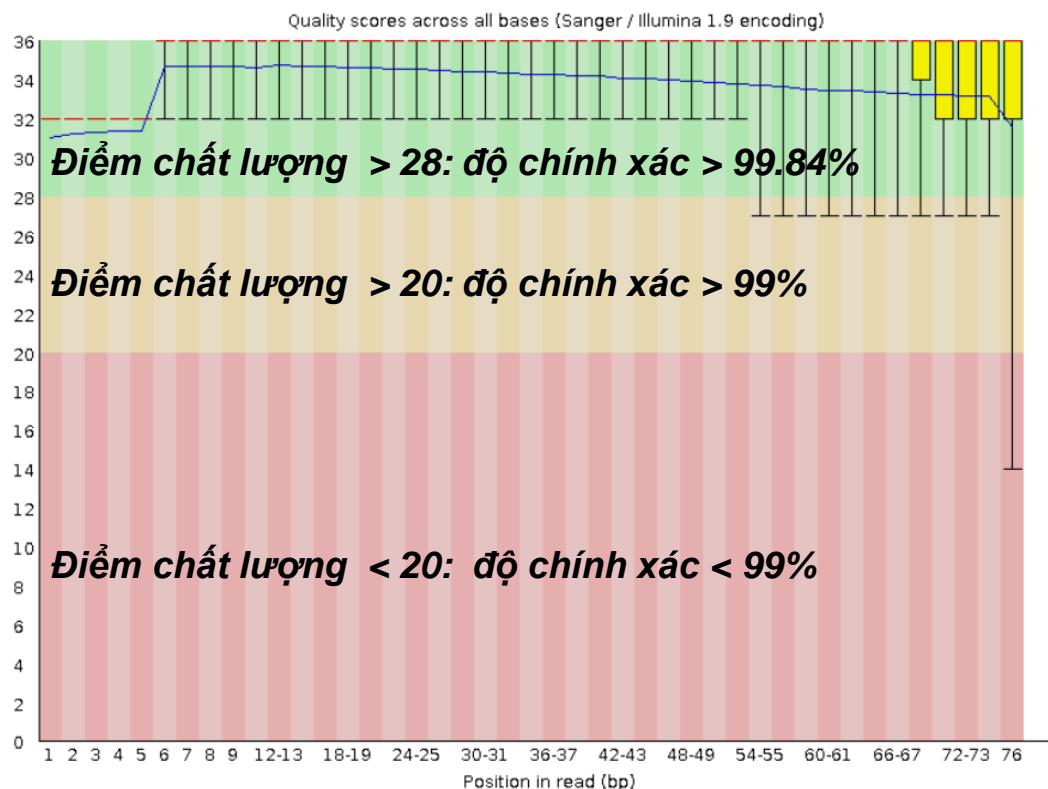
Kết quả từ FastQC (Webpage file)

Chất lượng trình tự theo đơn vị base (Per base sequence quality)



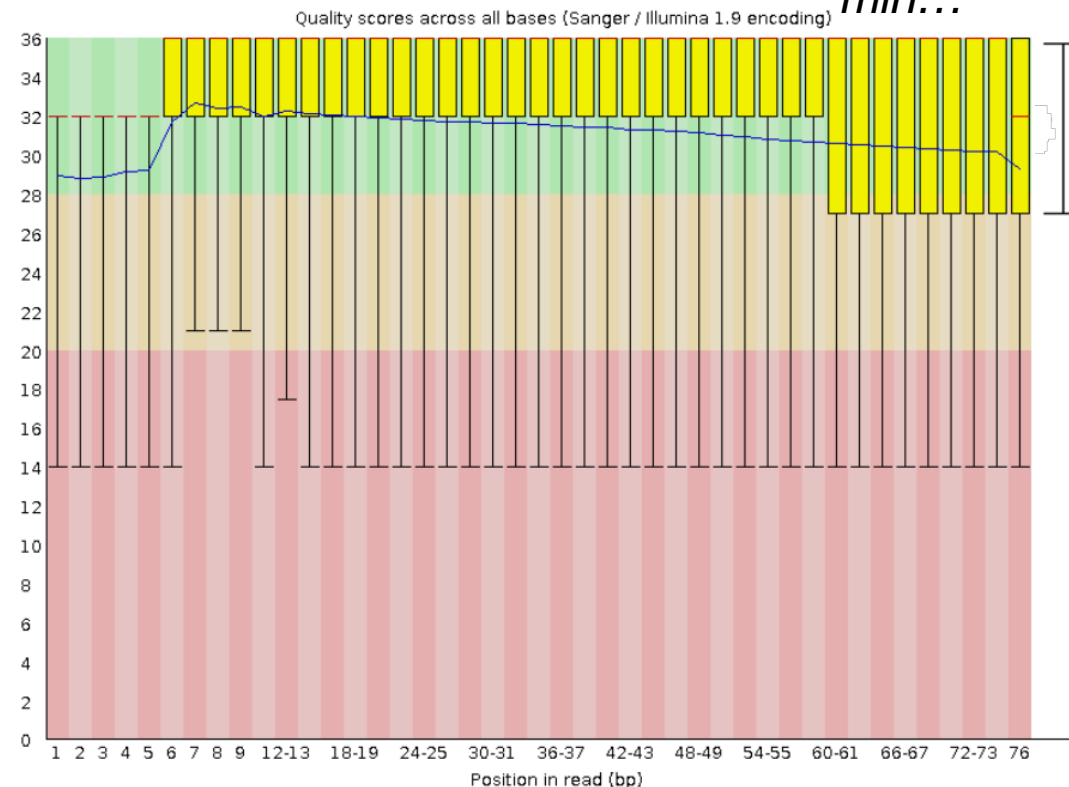
Mean, median, max,
min...

Per base sequence quality



R1: WT_E_2_R1_fastq_gz.gz

Per base sequence quality



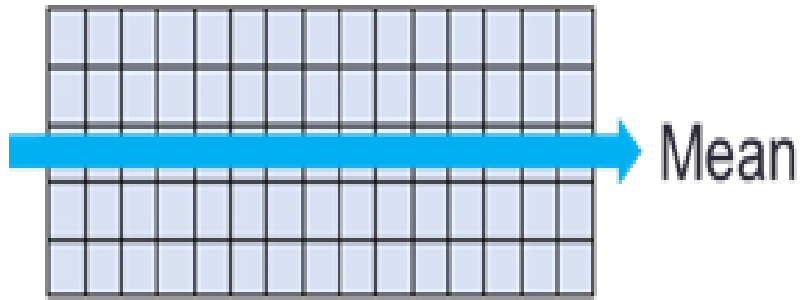
R2: WT_E_2_R2_fastq_gz.gz

IQR: tứ phân vị thứ nhất (Q1) -> tứ phân vị thứ 3 (Q3)

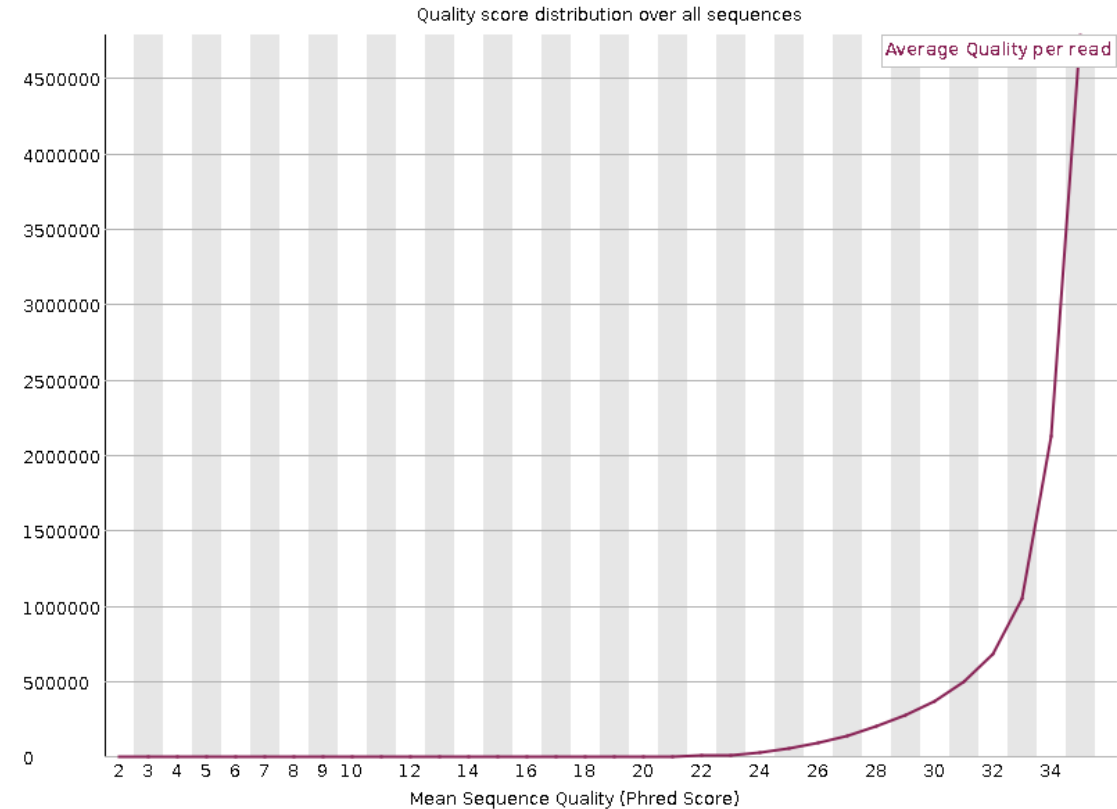
- Đường “**màu xanh**”: Trung bình điểm chất lượng
- Đường “**màu đỏ**”: Trung vị điểm chất lượng

Chất lượng trình tự trung bình (Per sequence quality scores)

✓ Per sequence quality scores

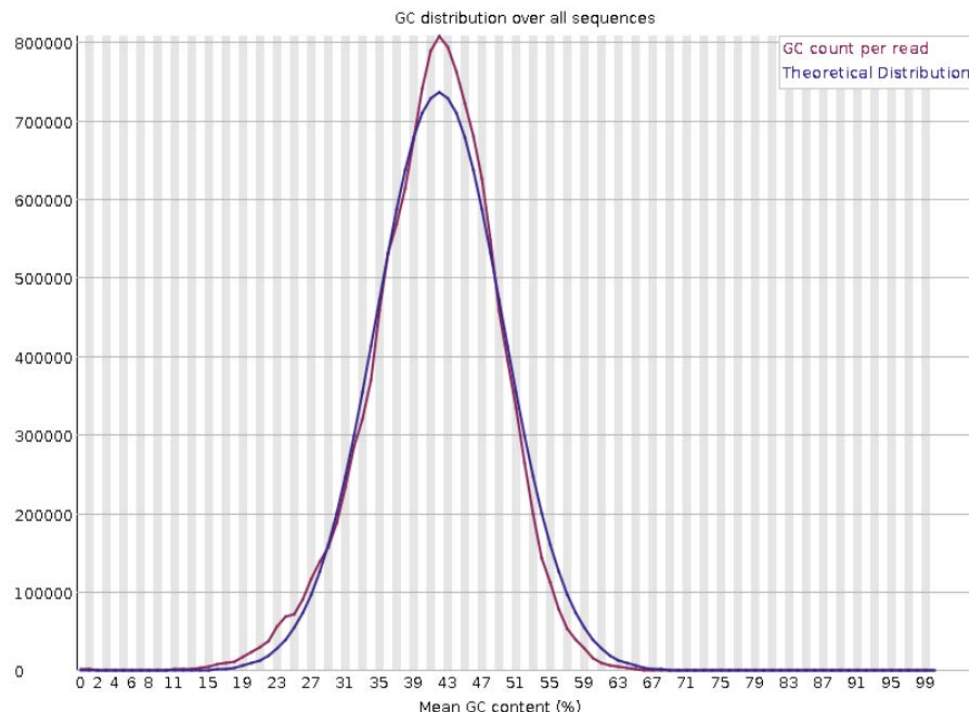


- Đánh giá điểm chất lượng (QC) trung bình của mỗi đoạn đọc (reads).
- Chất lượng thấp nhất -> Cao nhất, v.d. QC=17 -> 35
- Đa số trình tự có chất lượng trung bình rất cao, v.d. 35



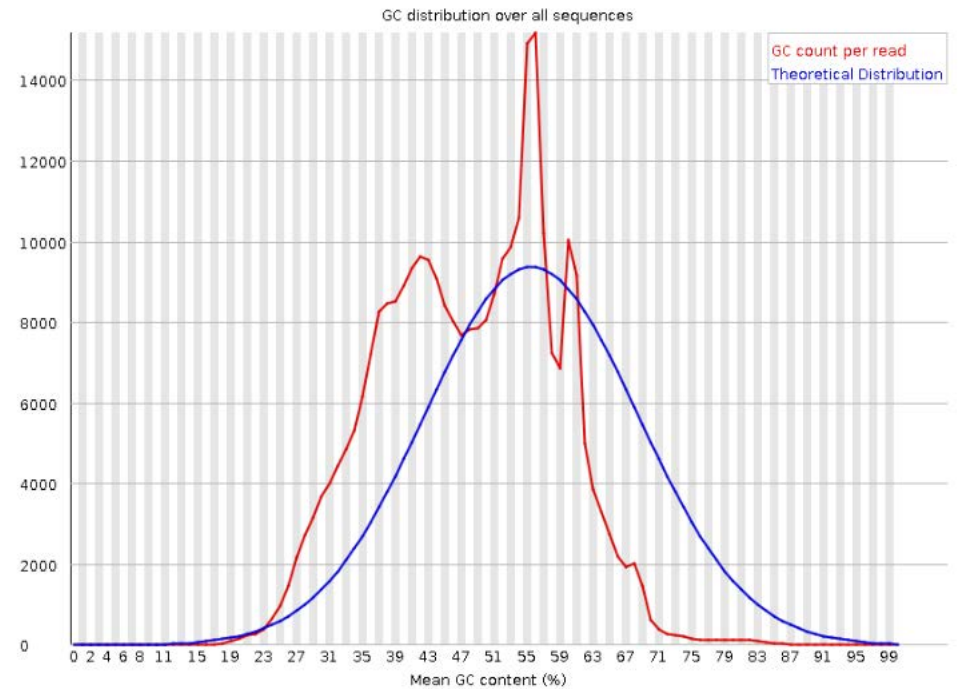
Tỷ lệ GC (Per sequence GC content)

✔ Per sequence GC content



Tỷ lệ GC phân bố theo phân phối chuẩn
(dạng quả chuông)

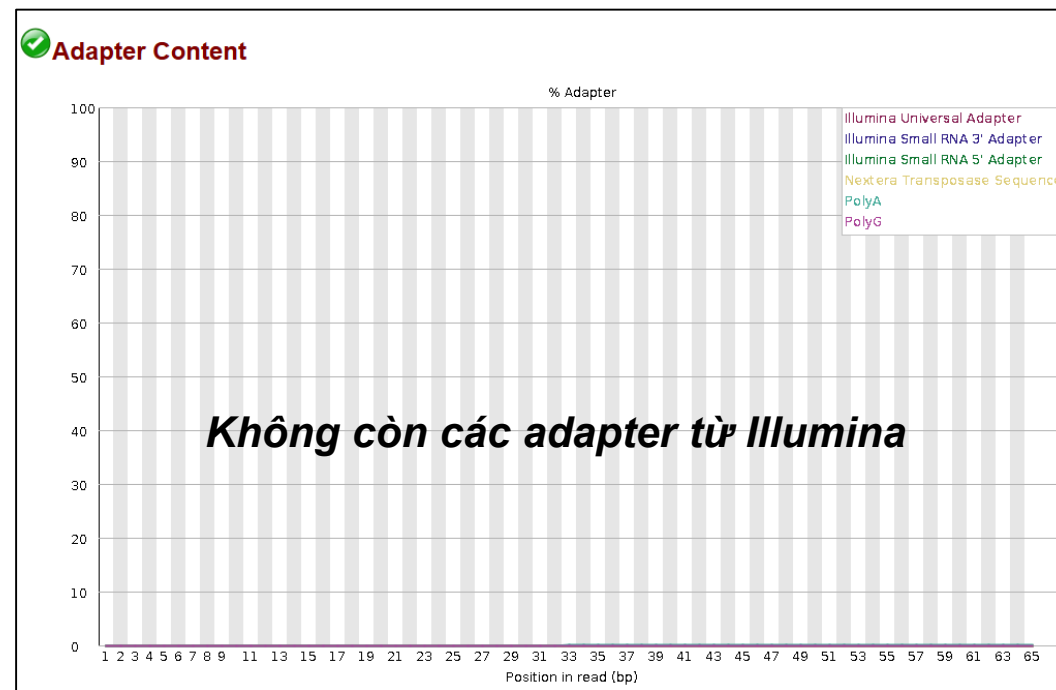
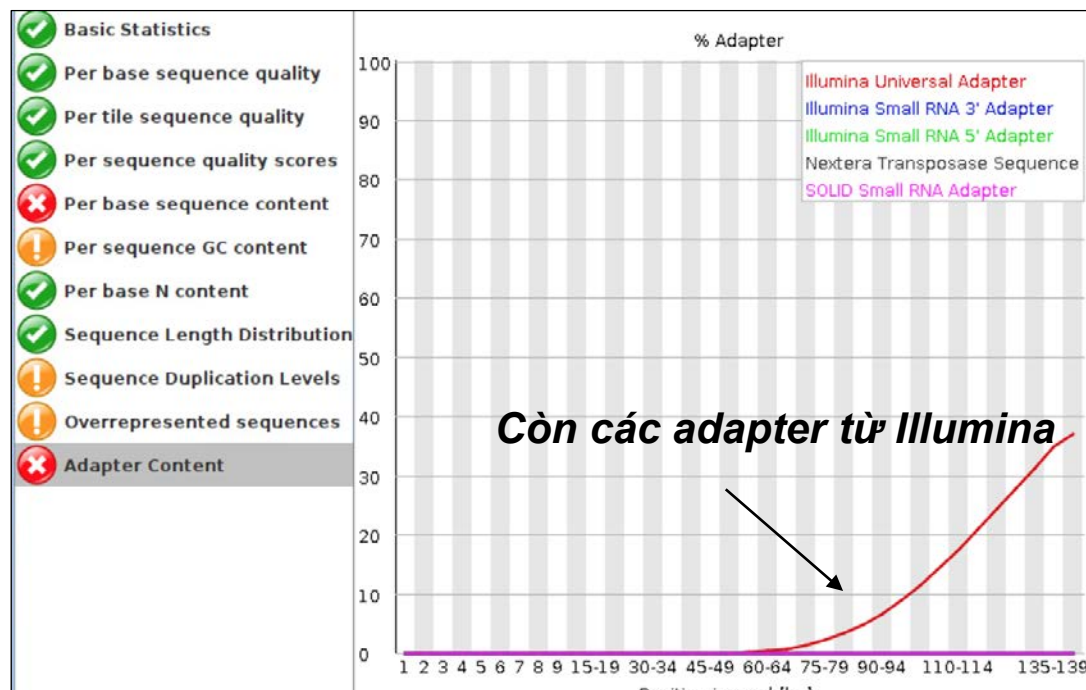
✘ Per sequence GC content



Tỷ lệ GC phân bố tạo thành các đỉnh nhọn
→ có thể bị nhiễm hoặc gen được biểu hiện quá mức.

Adapter content

- Dữ liệu có bao gồm trình tự adapter không?
- Nếu có --> Cần tinh sạch



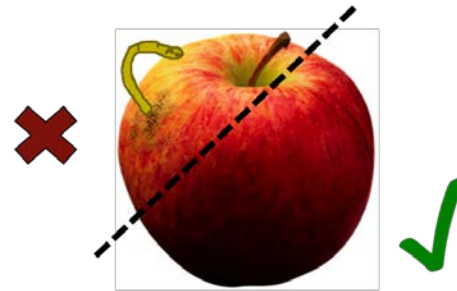
Câu hỏi:

1. Bạn đánh giá chất lượng dữ liệu giải trình tự của thí nghiệm này như thế nào? (Tốt hay Kém? Chỉ số đánh giá)
2. Chất lượng dữ liệu có giống nhau giữa dữ liệu từ môi xuôi và môi ngược?

3. Tinh sạch dữ liệu (Trimming)

Tinh sạch dữ liệu (*option*)

Loại bỏ các nucleotide với điểm chất lượng thấp,
ví dụ: < 20 tương ứng với độ chính xác $< 99\%$



Tiêu chí

- Độ dài các đoạn trình tự
- Chất lượng trình tự
- Tỷ lệ N (trình tự không biết rõ) trên các đoạn trình tự

Công cụ

- **Trimmomatics**: Phần mềm được phát triển dành riêng cho xử lý dữ liệu của Illumina.



```
@SRR638895.6046 6046 length=76
-----CAAACGAAACGTCAAGT(
+SRR638895.6046 6046 length=76
#EEEEEEEEEEEEEEEEEEEE
```

2

36

2

GTCCATTGCGCCCT

###



Thực hành tinh sạch dữ liệu

Công cụ: *Trimmomatic*

Input: paired-end (two separate input files)

Trimmomatic Operation:

- 1: SLIDNGWINDOW

Numer of base to average across: 4

Average quality required: 20

→ loại bỏ slide (4bp) có chất lượng trung bình < 20

- 2: MINLEN

Minimum length of read to be kept:

Ví dụ: 35 → loại bỏ tất cả trình tự < 35 bp

Output trim log messages: Yes

Lưu ý: Khi phát hiện dữ liệu có trình tự adapter ở bước FastQC có thể sử dụng **Perform initial ILLUMINACLIP step? Yes**

Trimmomatic
flexible read trimming tool for Illumina NGS data
(Galaxy Version 0.36.6)

Tool Parameters

Single-end or paired-end reads?
Paired-end (two separate input files)

Input FASTQ file (R1/first of pair) *
1: WT_E_2_R1.fastq.gz
accepted formats ▼

Input FASTQ file (R2/second of pair) *
2: WT_E_2_R2.fastq.gz
accepted formats ▼

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform
Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across *
4

Average quality required *
20

2: Trimmomatic Operation

Select Trimmomatic operation to perform
Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept *
35

+ Insert Trimmomatic Operation

Output trimlog file?
☐ No
(-trimlog)

Output trimmomatic log messages?
☒ Yes

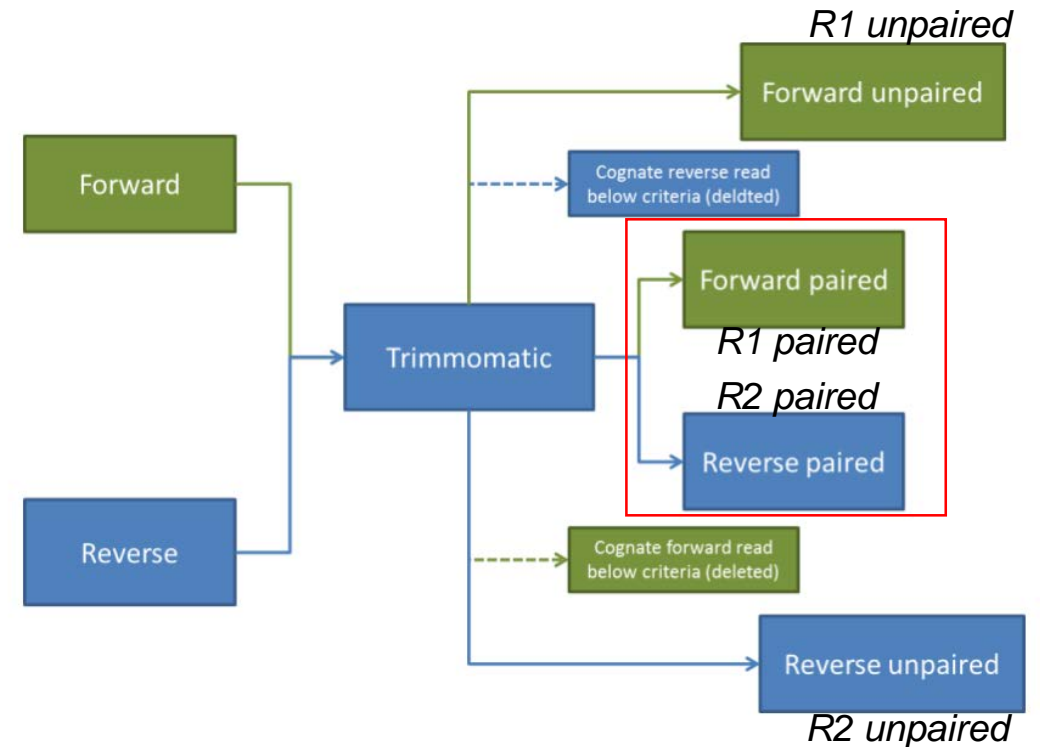
Kết quả chạy Trimmomatic từ dữ liệu paired-end data

Gồm 5 files:

- ***R1 paired, R2 paired***: Là hai tập dữ liệu fastq chất lượng tốt, paired.
- ***R1 unpaired, R2 unpaired***: Là hai tập fastq loại ra vì không tìm thấy paired-end.
- ***Log file***: thông tin đã chạy: Bao nhiêu read được giữ lại?

```
Picked up _JAVA_OPTIONS: -Xmx7G -Xms1G
TrimmomaticPE: Started with arguments:
  -threads 2 fastq_r1.fastqsanger.gz fastq_r2.fastqsanger.gz
fastq_out_r1_paired.fastqsanger.gz fastq_out_r1_unpaired.fastqsanger.gz
fastq_out_r2_paired.fastqsanger.gz fastq_out_r2_unpaired.fastqsanger.gz
SLIDINGWINDOW:4:20 MINLEN:35
Quality encoding detected as phred33
Input Read Pairs: 10391006 Both Surviving: 7505663 (72.23%) Forward Only
Surviving: 2229691 (21.46%) Reverse Only Surviving: 163802 (1.58%)
Dropped: 491850 (4.73%)
TrimmomaticPE: Completed successfully
```

Ví dụ: **WT_E_2**



Câu hỏi :

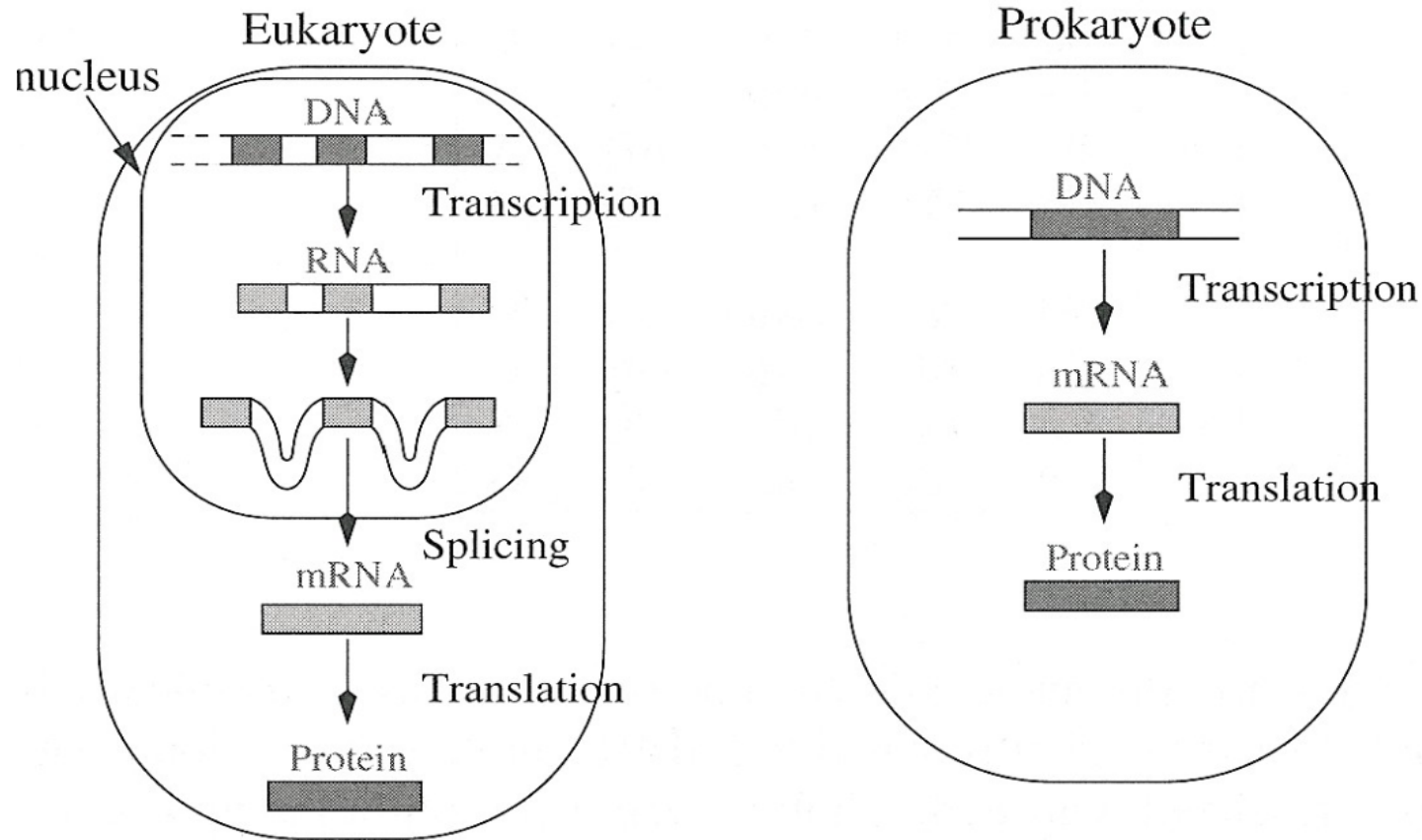
So sánh dữ liệu giải trình tự của thí nghiệm trước và sau khi được tinh sạch

- Chiều dài đoạn giải trình tự
- Điểm chất lượng
- Phần trăm số trình tự (reads) còn lại sau khi tinh sạch

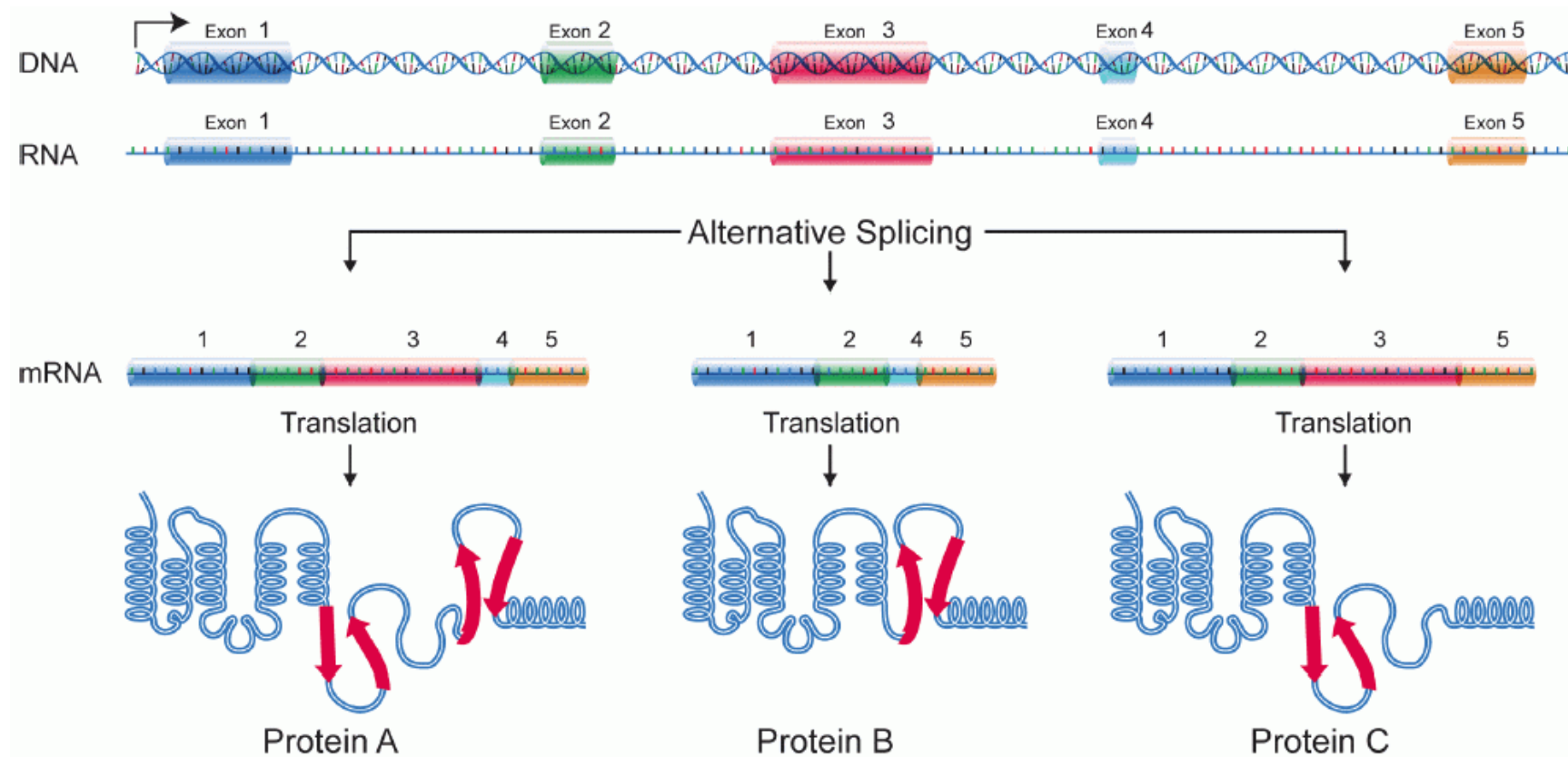
Gợi ý: Chạy FastQC cho mẫu sau khi tinh sạch và so sánh kết quả kết quả FastQC trước và sau khi tinh sạch

4. So sánh dữ liệu giải trình tự với bộ gen tham chiếu (Mapping)

Sự khác biệt phiên mã ở nhân sơ và nhân thực



Cắt nối RNA (Alternative splicing)



1 gene

=> Nhiều bản phiên mã (các cách ghép nối khác nhau)

=> Nhiều proteins (với chức năng khác nhau)

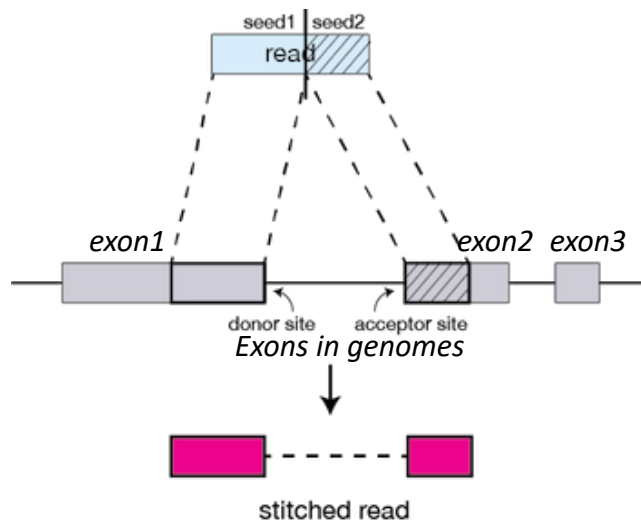
Chiến lược mapping

a. Genome mapping

Reads

Gapped mapper (STAR, HISAT2)

Mapping to genome



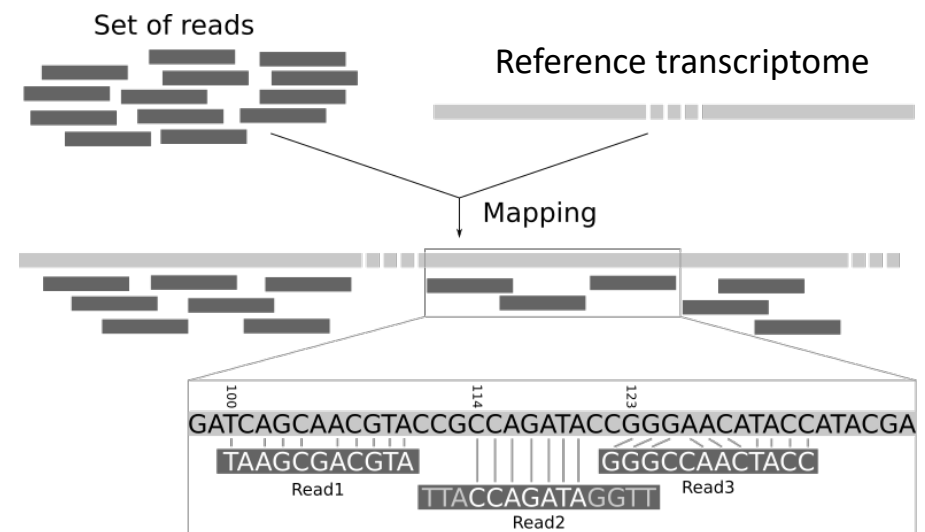
Mapping
with STAR

b. Transcriptome mapping

Reads

Ungapped mapper (Bowtie2)

Mapping to transcriptome



4.1. Thực hành mapping với HISAT2

- **Source for the reference genome:** Use a built-in genome
“*Select a reference genome*”: chọn bộ gen tham chiếu, vd. sacCer3
- **Is this a single or paired library:** Pair-end
Lựa chọn dữ liệu đã tinh sạch hoặc dữ liệu thô (nếu không trimming): R1 paired và R2 paired
- **Summary Options:**
“*Output alignment summary in a more machine-friendly style.*”: Yes
“*Print alignment summary to a file.*”: Yes

Summary Options

Output alignment summary in a more machine-friendly style.
→ ☒ Yes
Select this option for compatibility with MultiQC (--new-summary)

Print alignment summary to a file.
→ ☒ Yes
Output alignment summary to a file in addition to stderr. (--summary-file)

HISAT2 A fast and sensitive alignment program
(Galaxy Version 2.2.1+galaxy1)

Tool Parameters

Source for the reference genome
→ Use a built-in genome
Built-in references were created using default options
Select a reference genome *
→ Yeast (Saccharomyces cerevisiae): sacCer3
If your genome of interest is not listed, contact the Galaxy team

Is this a single or paired library
Paired-end

FASTA/Q file #1 *
→ →
accepted formats ▼
Must be of datatype "fastqsanger" or "fasta"

FASTA/Q file #2 *
→ →
accepted formats ▼

Kết quả chạy HISAT2

Aligned reads (BAM)

Kết quả so sánh

Mapping summary

Tổng kết % reads mapped với bộ gen tham chiếu

Header section									
@HD VN:1.5 SO:coordinate									
@SQ SN:ref LN:45									
r001	99	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG *
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAGGATA *
r003	0	ref	9	30	5S6M	*	0	0	GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC *
r003	2064	ref	29	17	6H5M	*	0	0	TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001	147	ref	37	30	9M	=	7	-39	CAGCGGCAT * NM:i:1

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

HISAT2 summary stats:

- Total pairs: 7505663
 - Aligned concordantly or discordantly 0 time: 167883 (2.24%)
 - Aligned concordantly 1 time: 7095046 (94.53%)
 - Aligned concordantly >1 times: 235845 (3.14%)
 - Aligned discordantly 1 time: 6889 (0.09%)
- Total unpaired reads: 335766
 - Aligned 0 time: 270019 (80.42%)
 - Aligned 1 time: 61793 (18.40%)
 - Aligned >1 times: 3954 (1.18%)
- Overall alignment rate: 98.20%

4.2. Thực hành mapping với STAR Aligner

a. Download GFT file cho bộ gen tham chiếu

- Tìm bộ gen của loài quan tâm tại **Ensembl**, **NCBI genome**, hoặc **UCSC Genome Browser**
- Download **GTF** (Gene transfer format) file: lưu trữ thông tin về cấu trúc gen
- Upload GTF file lên Galaxy server

*Ví dụ: Tìm GFT file cho *Saccharomyces cerevisiae* sacCer3*

- ✓ Link download: <https://hgdownload.soe.ucsc.edu/goldenPath/sacCer3/bigZips/genes/>
- ✓ File: *sacCer3.ensGene.gtf.gz*

b. Mapping với RNA STAR (Galaxy)

Single-end or paired-end reads: paired-end (as individual datasets)

Lựa chọn dữ liệu đã tinh sạch hoặc dữ liệu thô (nếu không trimming):
R1 paired và R2 paired

Custom or built-in reference genome: Use a built-in index

- “*Reference genome with or without an annotation*”: use genome reference without builtin gene-model but provide a gtf
- “*Select reference genome*”: Yeast (*Saccharomyces cerevisiae*)
- “*Gene model (gff3,gtf) file for splice junctions*”: chọn file đã upload *sacCer3.ensGene.gtf.gz*
- “*Length of the genomic sequence around annotated junctions*”: 75
(This parameter should be length of reads – 1, see FASTQC result)
- “*Per gene/transcript output*”: Per gene read counts (GeneCounts)

Compute coverage: Yes in bedgraph format

RNA STAR
Gapped-read mapper for RNA-seq data
(Galaxy Version 2.7.11a+galaxy1)

Tool Parameters

Single-end or paired-end reads

Paired-end (as individual datasets)

RNA-Seq FASTQ/FASTA file, forward reads *

7: Trimmomatic on WT_E_2_R1.fastq.gz (R1 paired)

RNA-Seq FASTQ/FASTA file, reverse reads *

8: Trimmomatic on WT_E_2_R2.fastq.gz (R2 paired)

Custom or built-in reference genome

Use a built-in index

Built-ins were indexed using default options

Reference genome with or without an annotation

use genome reference without builtin gene-model but provide a gtf

Select the '... with builtin gene-model' option to select from the list of available splice junction information. Select the '... without builtin gene-model' option to select from available indexes without annotated splice junctions, and, optionally, provide your own annotations.

Select reference genome *

Yeast (*Saccharomyces cerevisiae*): sacCer3

If your genome of interest is not listed, contact the Galaxy team (--genomeDir)

Gene model (gff3,gtf) file for splice junctions *

17: sacCer3.ensGene.gtf.gz

Câu hỏi :

So sánh kết quả chạy mapping từ **HISAT2** và **RNA STAR** :

- Tỷ lệ phần trăm các đoạn đọc được mapped 1 lần với bộ gen tham chiếu
- Tỷ lệ phần trăm các đoạn đọc được mapped > 1 lần với bộ gen tham chiếu
- Tỷ lệ phần trăm các đoạn đọc không map với bộ gen tham chiếu

HISAT2 log file

```
HISAT2 summary stats:
  Total pairs: 7505663
    Aligned concordantly or discordantly 0 time: 167883 (2.24%)
    Aligned concordantly 1 time: 7095046 (94.53%)
    Aligned concordantly >1 times: 235845 (3.14%)
    Aligned discordantly 1 time: 6889 (0.09%)
  Total unpaired reads: 335766
    Aligned 0 time: 270019 (80.42%)
    Aligned 1 time: 61793 (18.40%)
    Aligned >1 times: 3954 (1.18%)
  Overall alignment rate: 98.20%
```

RNA STAR log file

```
Started job on | Nov 18 11:02:19
Started mapping on | Nov 18 11:03:20
Finished on | Nov 18 11:04:36
Mapping speed, Million of reads per hour | 355.53

Number of input reads | 7505663
Average input read length | 144
UNIQUE READS:
Uniquely mapped reads number | 7138333
Uniquely mapped reads % | 95.11%
Average mapped length | 143.99
Number of splices: Total | 96538
Number of splices: Annotated (sjdb) | 87363
Number of splices: GT/AG | 94391
Number of splices: GC/AG | 487
Number of splices: AT/AC | 9
Number of splices: Non-canonical | 1651
Mismatch rate per base, % | 0.17%
Deletion rate per base | 0.01%
Deletion average length | 1.34
Insertion rate per base | 0.00%
Insertion average length | 1.10
MULTI-MAPPING READS:
Number of reads mapped to multiple loci | 210168
% of reads mapped to multiple loci | 2.80%
Number of reads mapped to too many loci | 45076
% of reads mapped to too many loci | 0.60%
UNMAPPED READS:
Number of reads unmapped: too many mismatches | 0
% of reads unmapped: too many mismatches | 0.00%
Number of reads unmapped: too short | 110911
% of reads unmapped: too short | 1.48%
Number of reads unmapped: other | 1175
% of reads unmapped: other | 0.02%
CHIMERIC READS:
Number of chimeric reads | 0
% of chimeric reads | 0.00%
```



We are happy to help you!

Dữ liệu chạy thử (Share History/Galaxy):

<https://usegalaxy.org.au/u/tam-tran/h/rnasequpstreamtest>