

# RNAseq Data Analysis

MSU Workshop

Rama Shankar

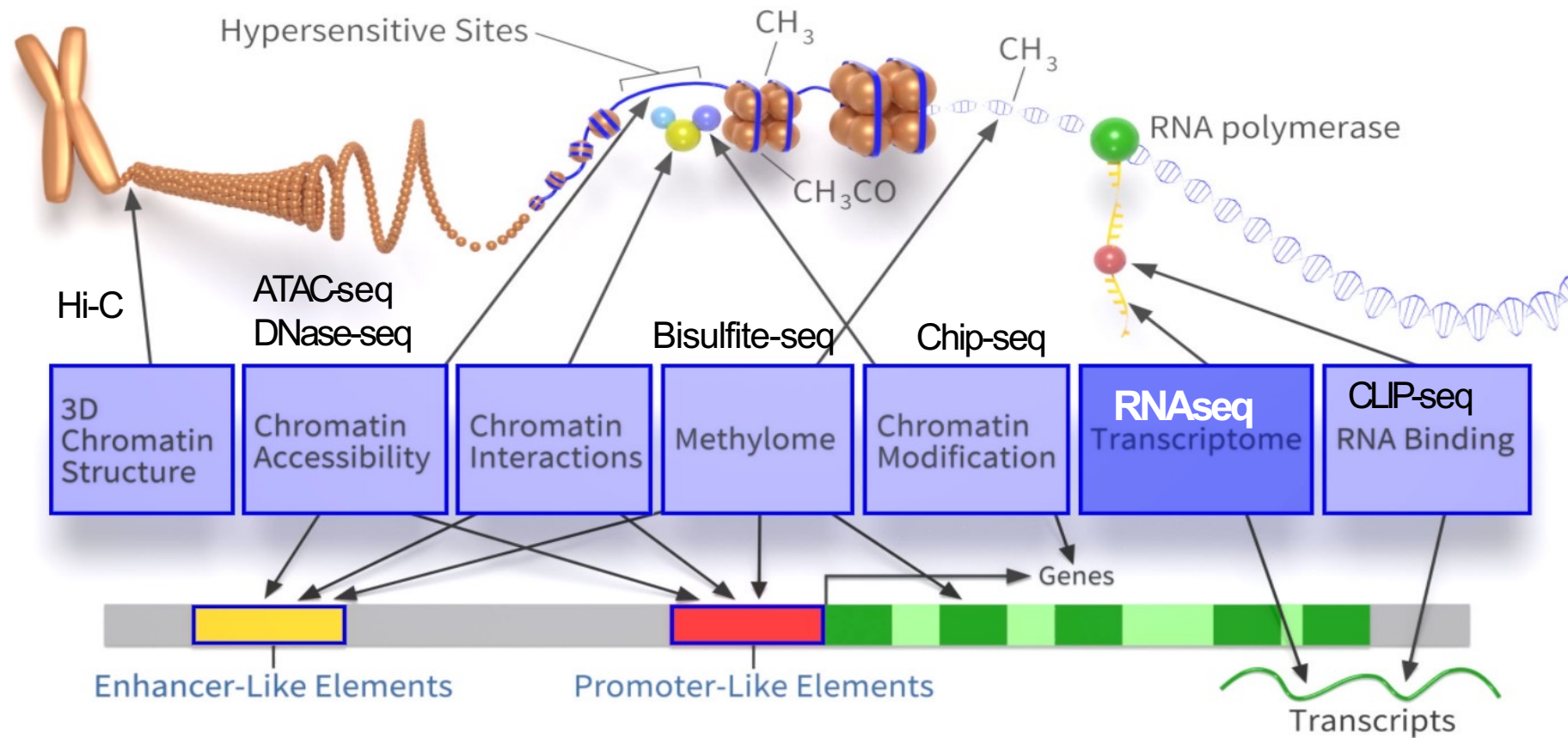
[ramashan@msu.edu](mailto:ramashan@msu.edu)

# Outline

- I. Introduction of RNAseq
- II. RNA isolation and library preparation
- III. RNAseq data analysis
- IV. Demo of RNAseq data analysis on MSU HPCC

# **I. Introduction of RNAseq**

# Probing cells with sequencing technologies



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

# Types of RNA

mRNA

lncRNA (long non-coding RNA)

microRNA (small RNAseq)

Circular RNA

rRNA

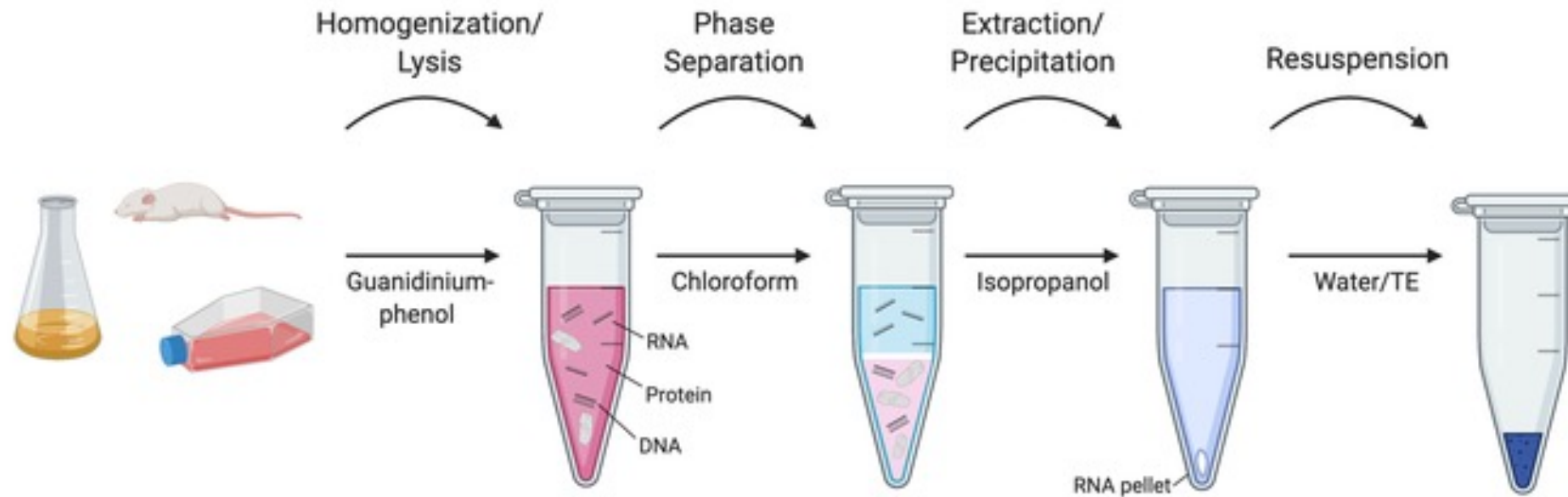
tRNA

snoRNA

piRNA

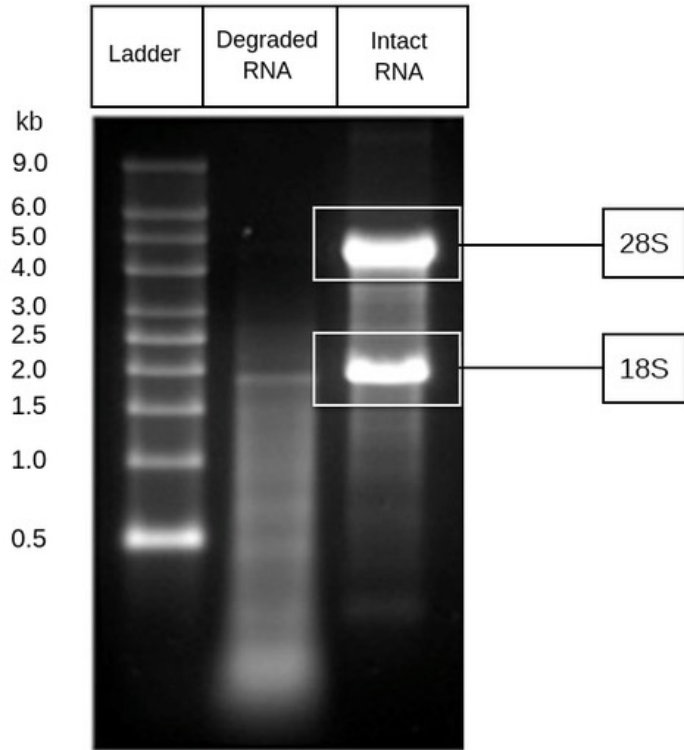
## **II. RNA isolation and library preparation**

# RNA isolation

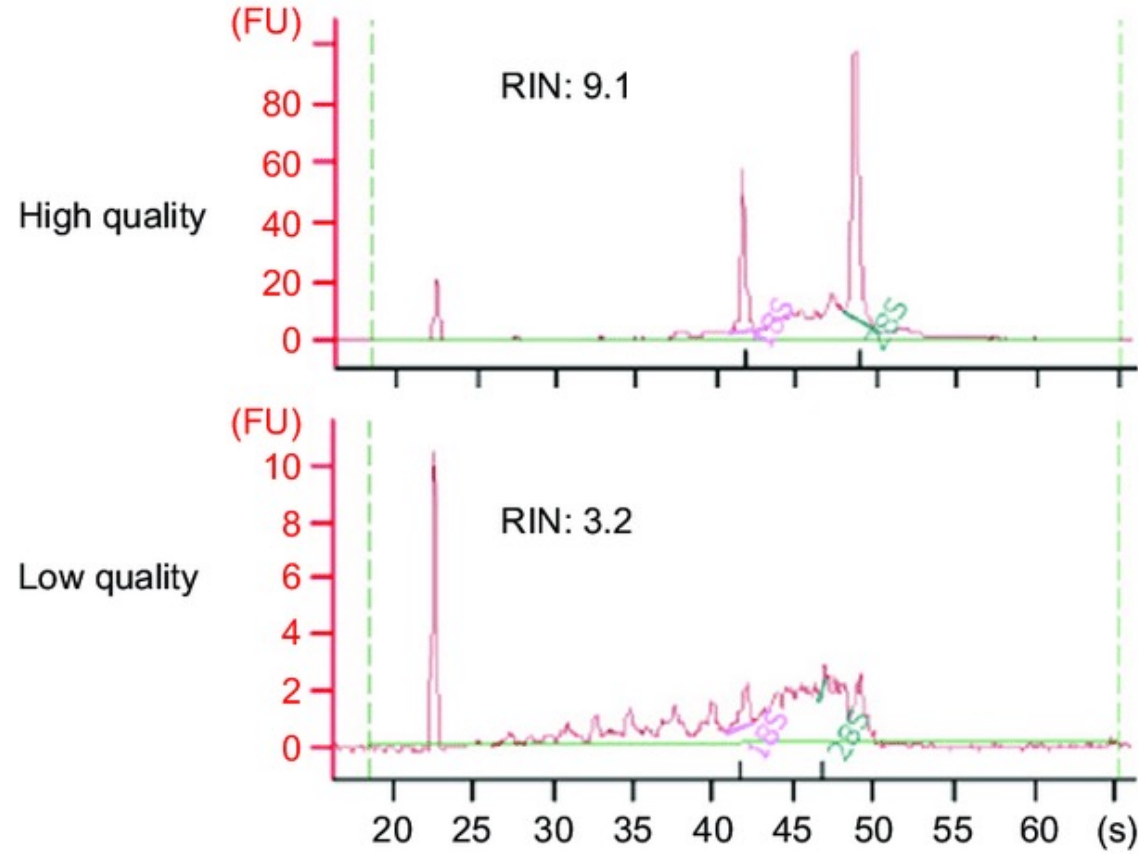


<https://www.addgene.org/protocols/kit-free-rna-extraction/>

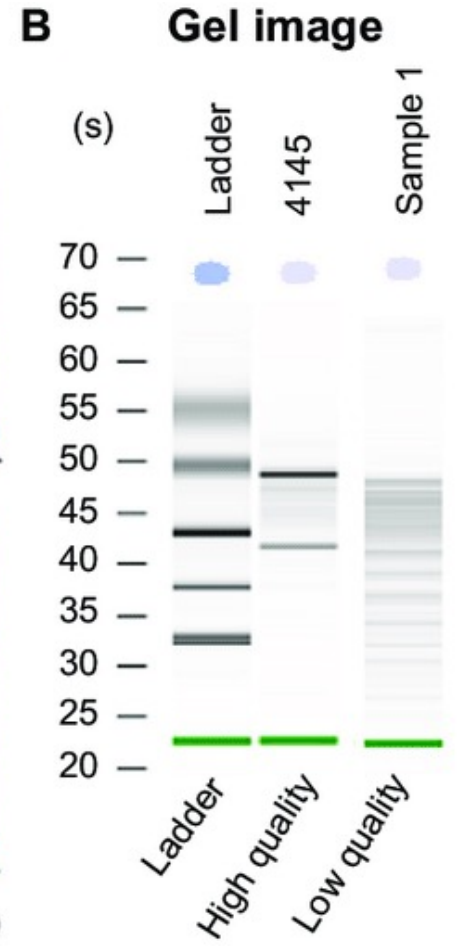
# Quality check of RNA



**A**

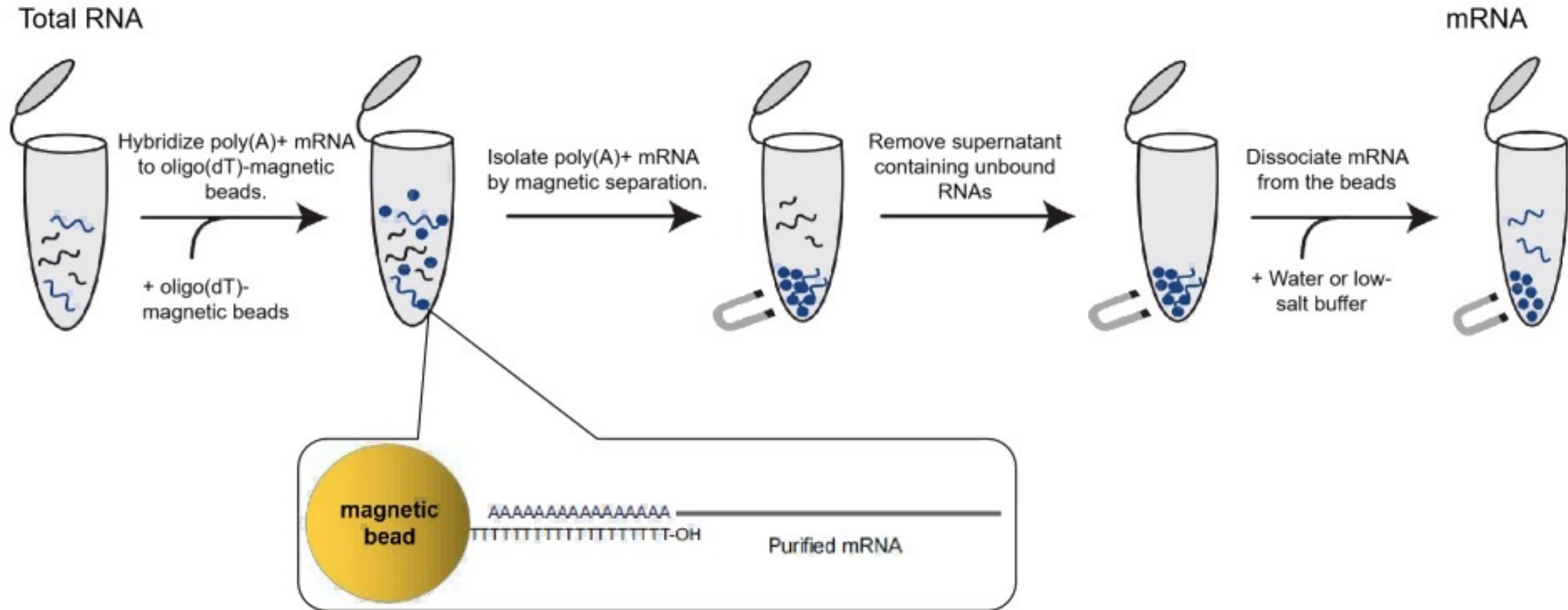


**B**

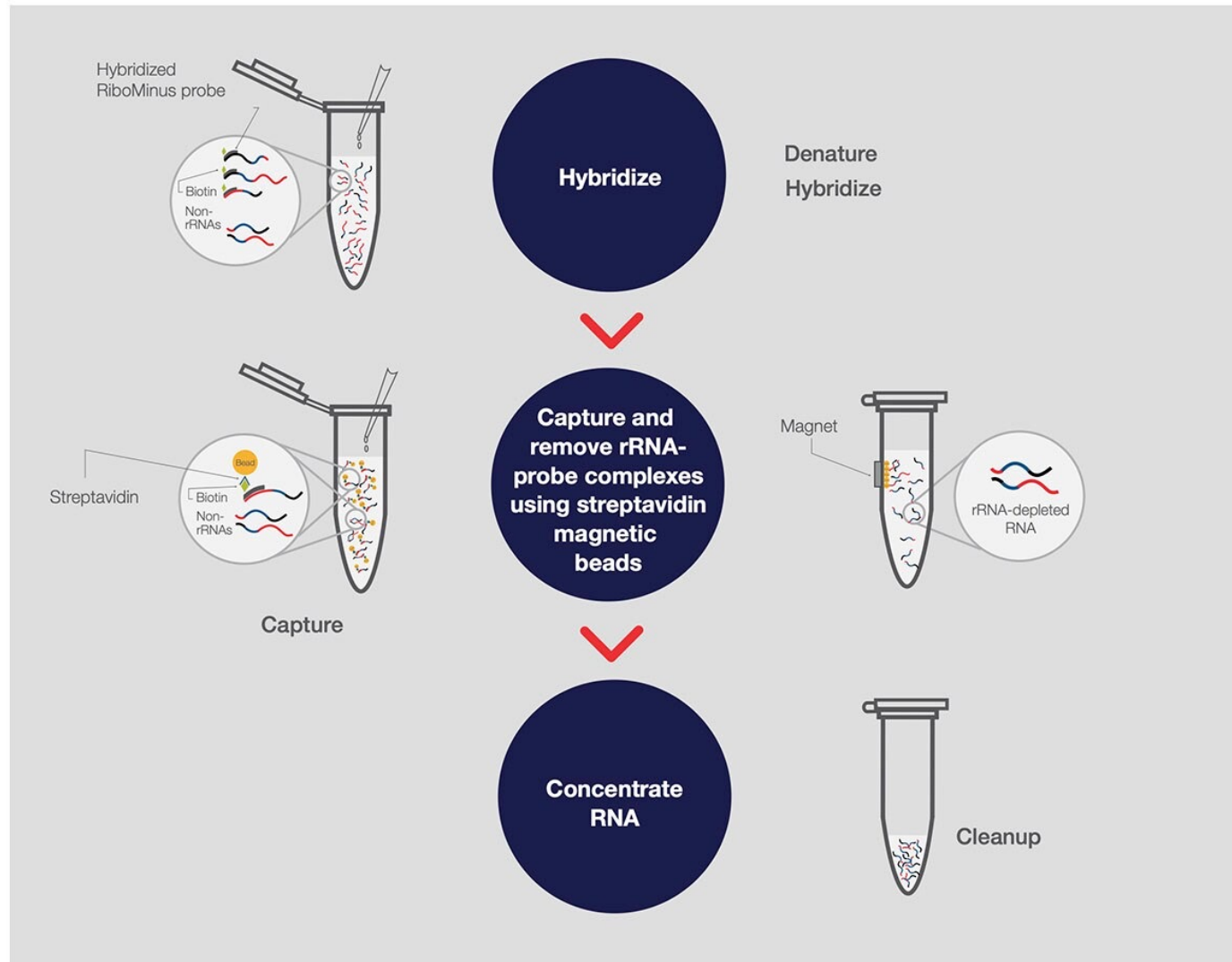




# mRNA enrichment using poly dT beads



# mRNA enrichment using ribosomal RNA depletion

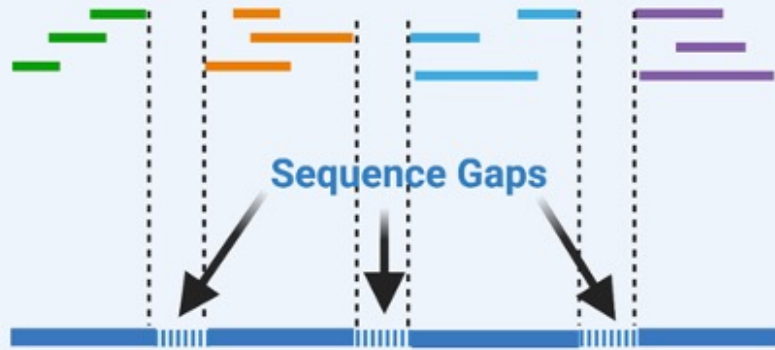


# Short reads vs long reads

## ① Short Reads



Reference Genome



Sequence Gaps

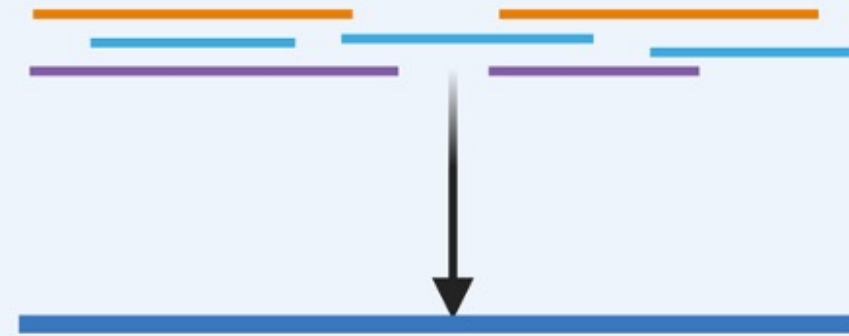
Disease Gene

Missing sequence data leads to gaps  
in genome coverage and limits variant detection

## ② Long Reads



Reference Genome



Disease Gene

Long reads map uniquely and  
span large variants providing  
comprehensive variant detection

# Long read vs short read platforms



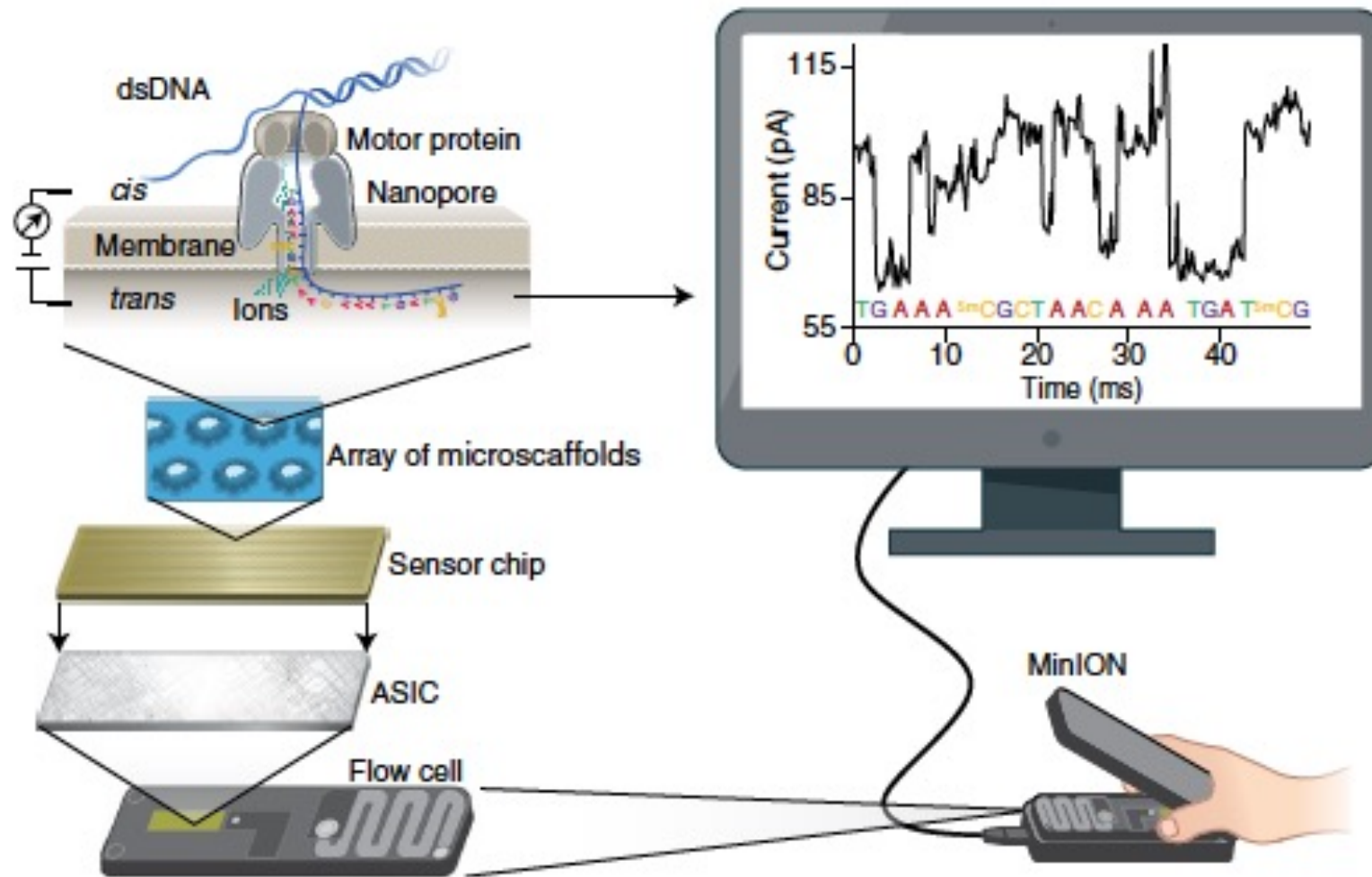
Oxford Nanopore  
(long read)



Illumina genome analyzer  
(short read)

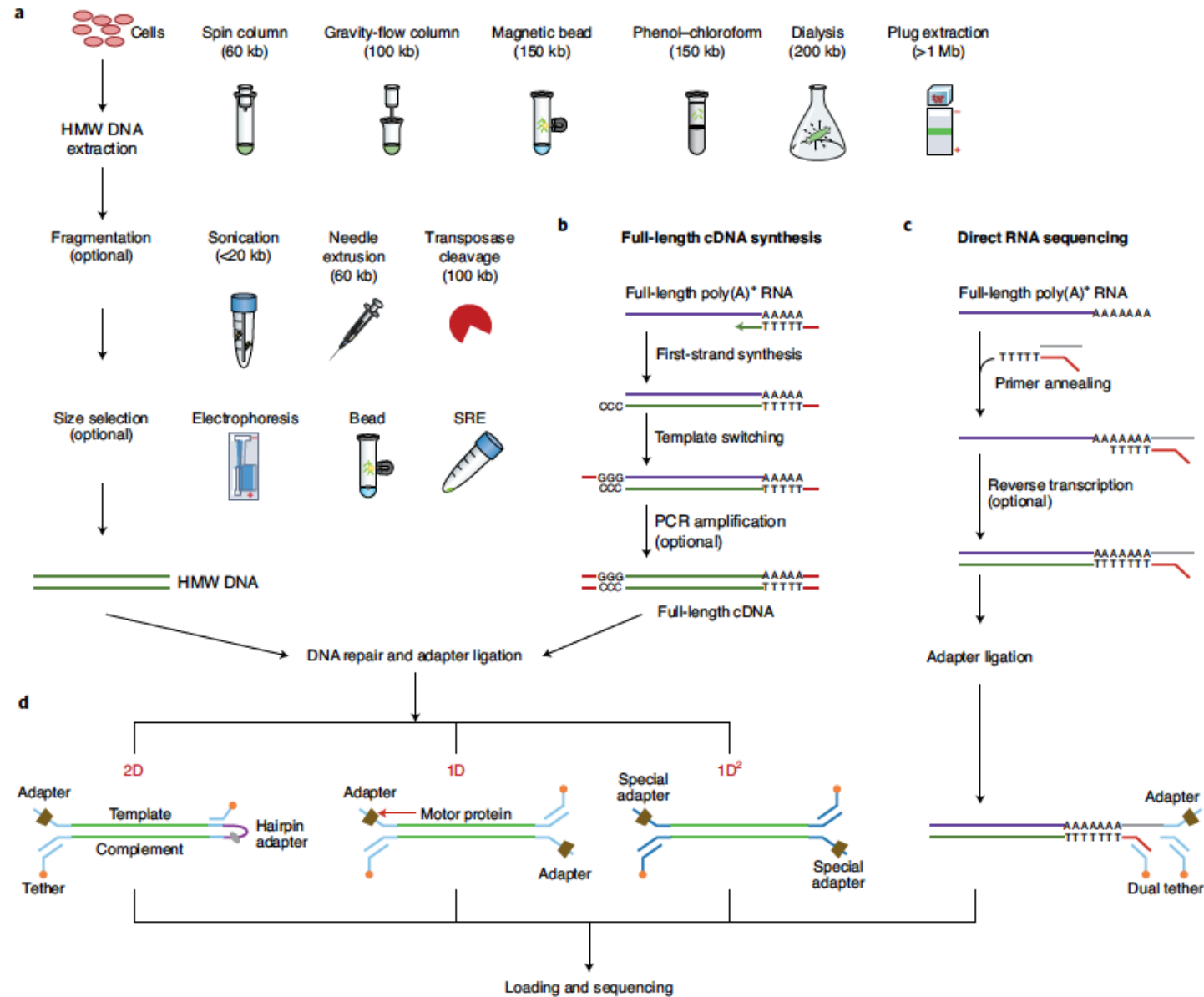
## **IIA. From RNA to long reads**

# Principle of Nanopore sequencing





# Library preparation and sequencing workflow for Nanopore

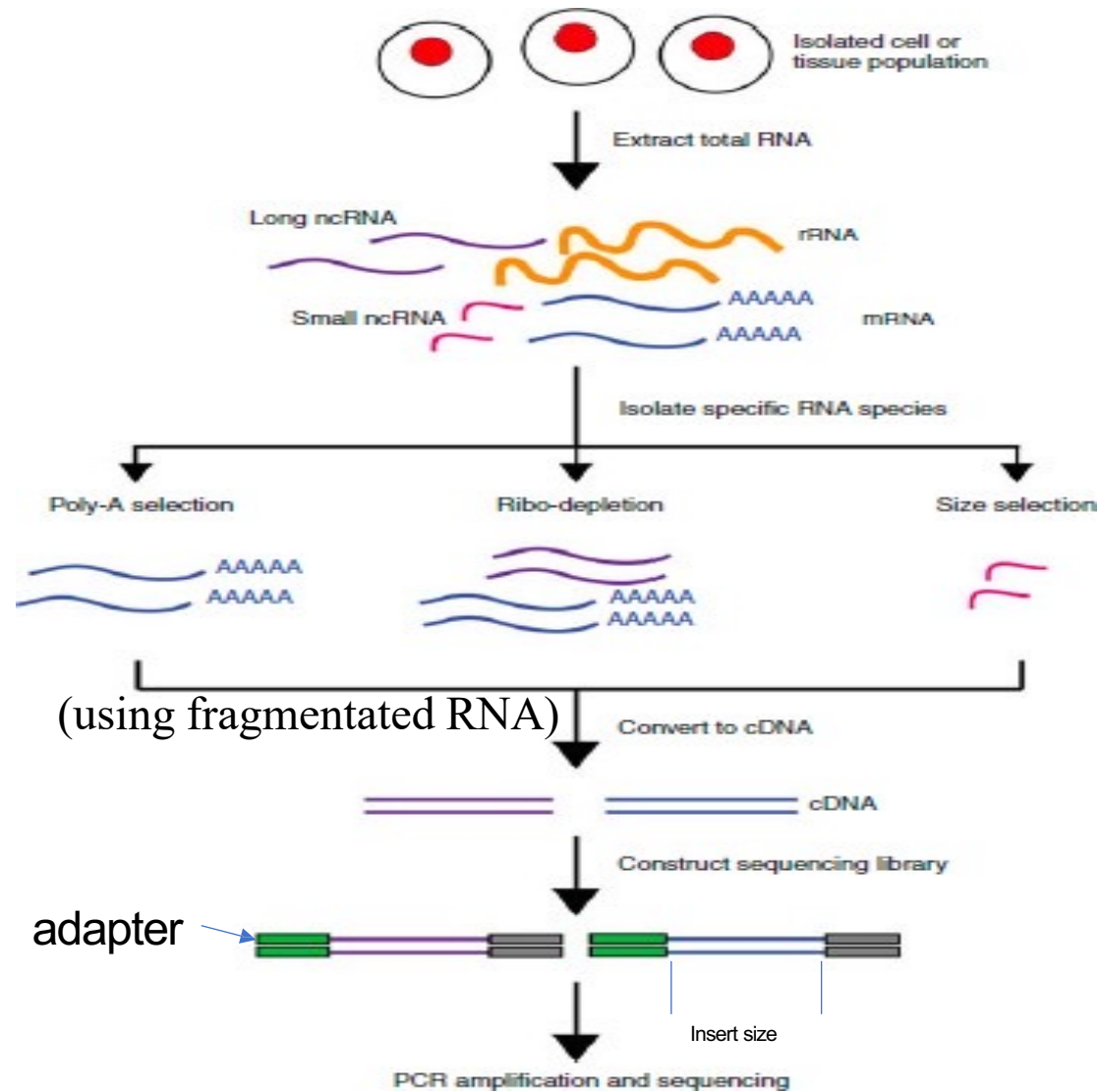


2D (template strand is sequenced), 1D (where each strand is ligated with an adapter and sequenced independently) and 1D<sup>2</sup> (special adapter ligated to sequence one followed by another one)

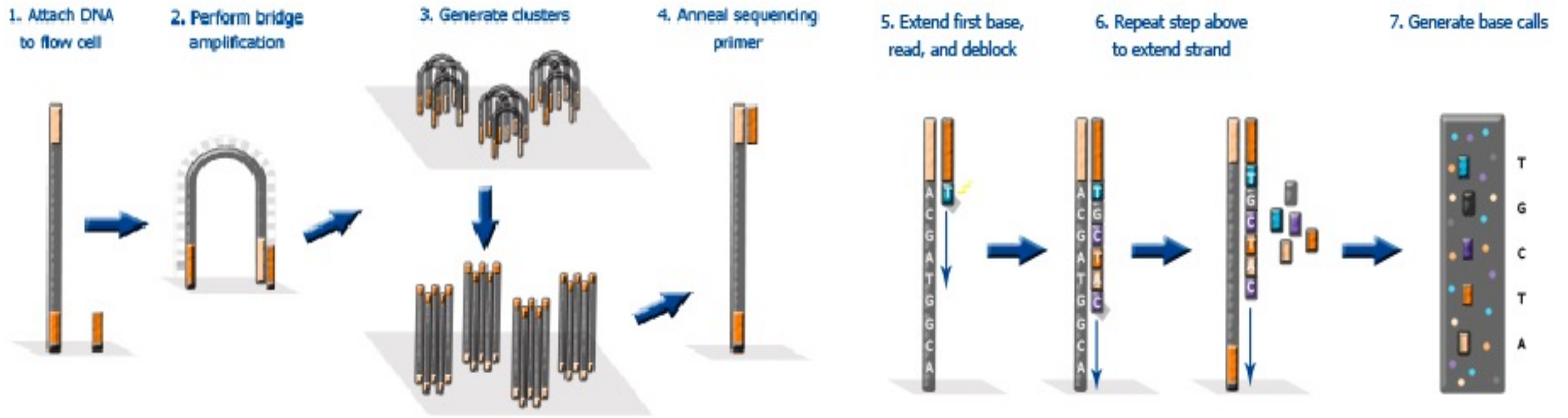
## **IIB. From RNA to short reads**



# RNAseq library construction for Illumina (short reads)



# Overview of Illumina sequencing



## Sequencing by Synthesis

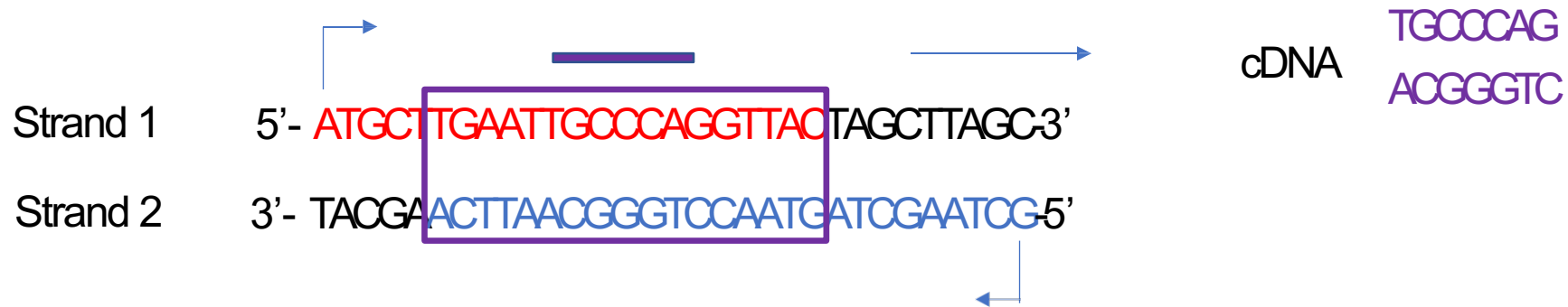
Picture is from <https://www.eurofinngenomics.co.in/en/eurofins-genomics/product-fags/next-generation-sequencing/general-technical-questions/what-is-the-principal-of-the-illumina-sequencing-technology.aspx>

[https://www.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf) (Documentation from illumina)

<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

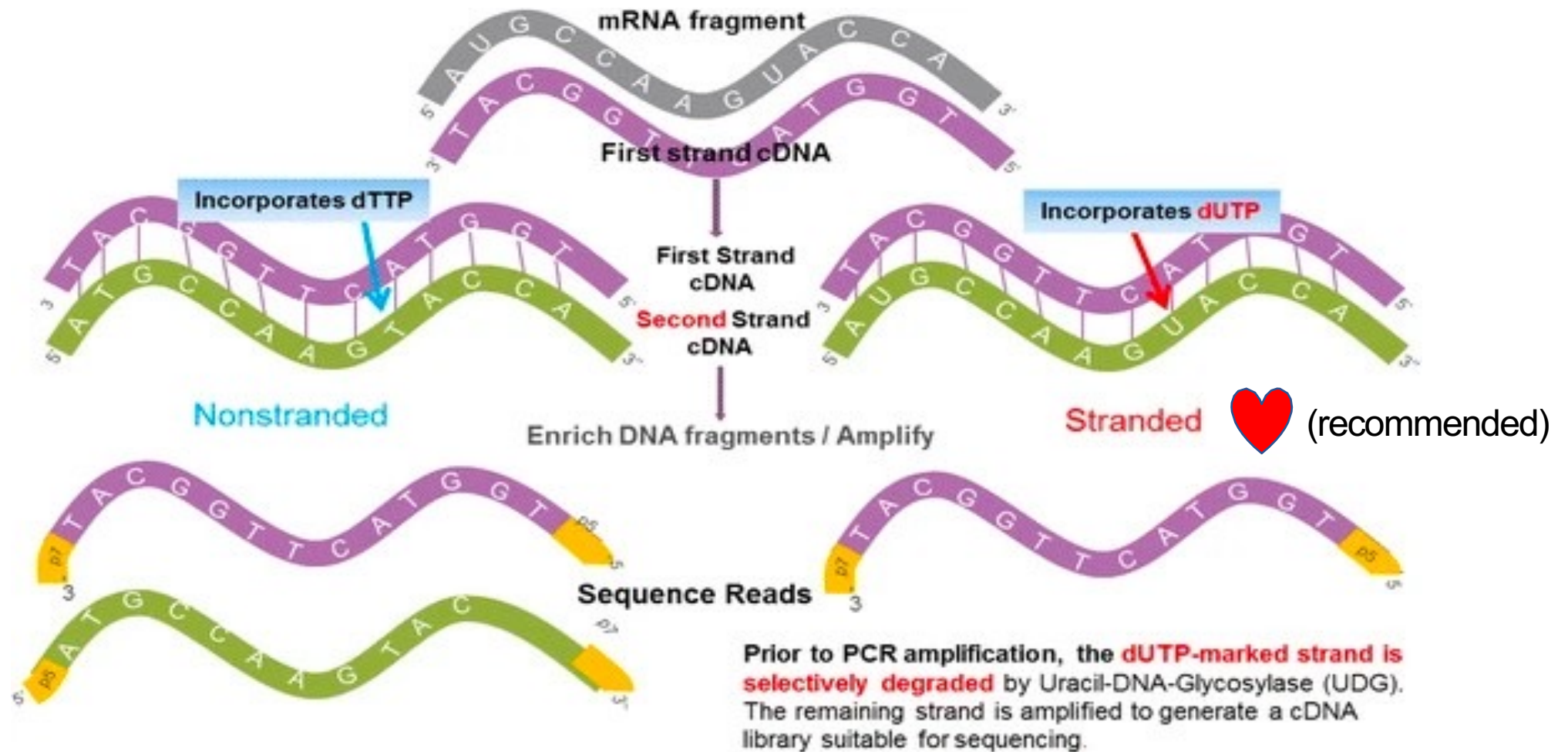
# Strandness matters

DNA is double-stranded, both strand could be transcribed.



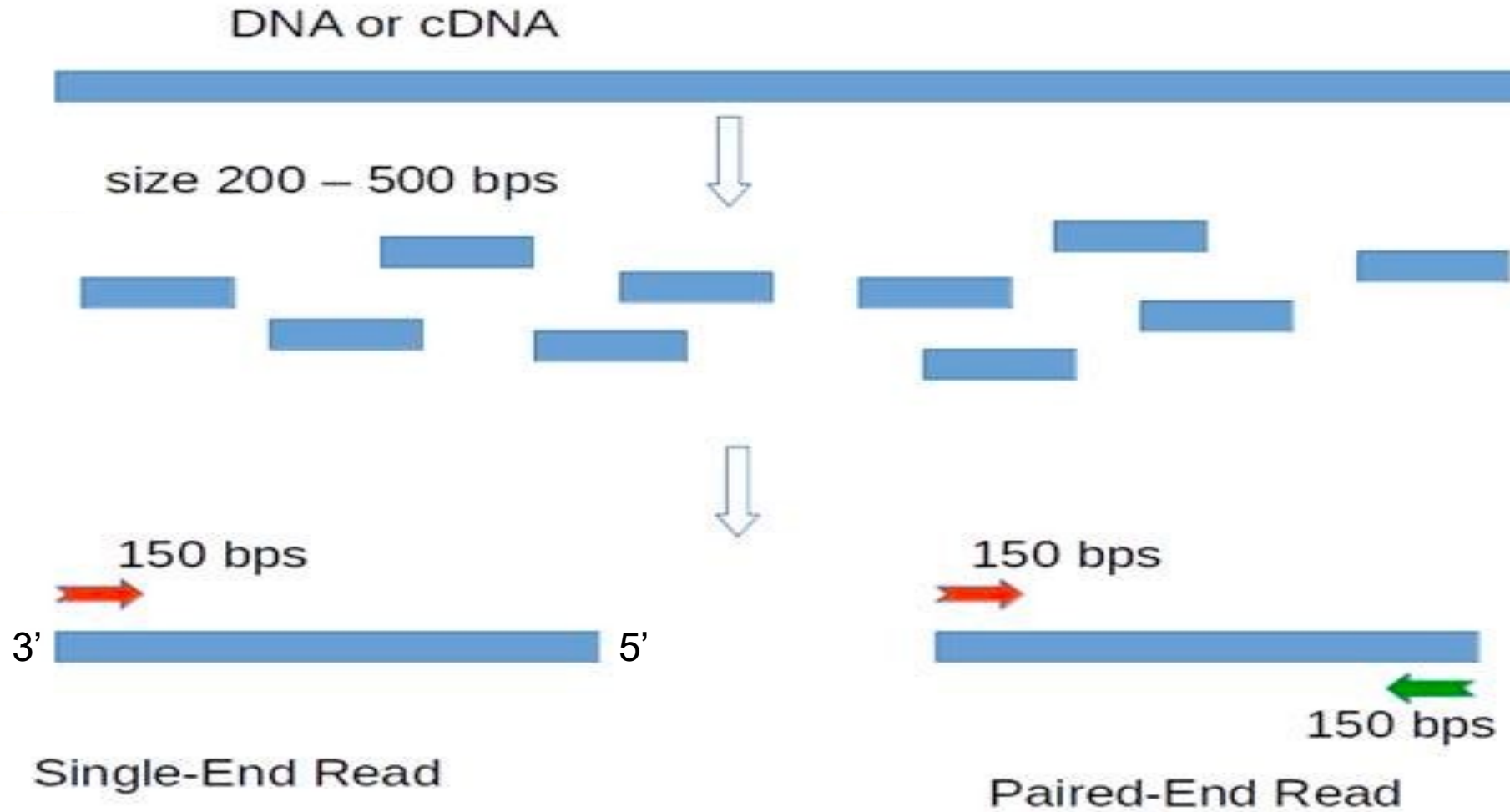
If an RNA fragment comes from the dashed region, it is hard to know its original gene without strand information.

# Nonstranded vs Stranded library



Ref: Zhao *et al.* Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. (BMC Genomics, 2015)

# Single-end vs Paired-end



# Output of RNAseq

1. According to sequencing mode, single-end or paired-end read.
2. Typical read length: 50bp, 75bp, 100bp.
3. Reads are usually stored in FASTQ format.

```
@SRR4822549.1 UNC13-SN749_0135:5:1101:1389:1958 length=100
NAAAGCACATACCAAGGCCACCACACACCACCTGTCCAAAAGGCCTTCGATACGGGATAATCCTATTTATTACCTCAGAAGTTTTTTCTTCGCAGGAT
+SRR4822549.1 UNC13-SN749_0135:5:1101:1389:1958 length=100
#1=B7BDFFFHFDGGEBHGIIGEGCHGIIIIIGAG?DFFIGGIIIHG3BGHG@@EDHHEHFFBDFFFEEEDCEDCC;>@CCCD5:AC?B@DDDDDDD@B@C
@SRR4822549.2 UNC13-SN749_0135:5:1101:1498:1960 length=100
NTTCTTCAATTTCTTGCCTTCTTCCTTGGAGGCTGGAAGAATCATGGCAAGGTAGGGCCCATCAACCTCAAAAAGATGCTGTTCTCTGAGCGGGTGACG
+SRR4822549.2 UNC13-SN749_0135:5:1101:1498:1960 length=100
#1=DFFFFHHHHHJJJJJJJJJJJJJJHIIJJJJIIJJJJJJJJJJJBGIJJJJJJJJJJJJHHHHHFFDDDDDDDDDCDEDDDDDDDDDD<BBDB
```

# Preparation of RNAseq experiments

1. Sequencing depth (library size): deeper is better! Usually  $> 20$  million reads should be OK (from illumina).
2. Stranded vs Non-stranded: Strand-specific is recommended.
3. SE vs PE: consider PE if you have enough budget. For gene expression analysis, SE also works well.
4. Technical replicates, the more the better (at least three) .
5. Longer read length is better.

**Q&A**

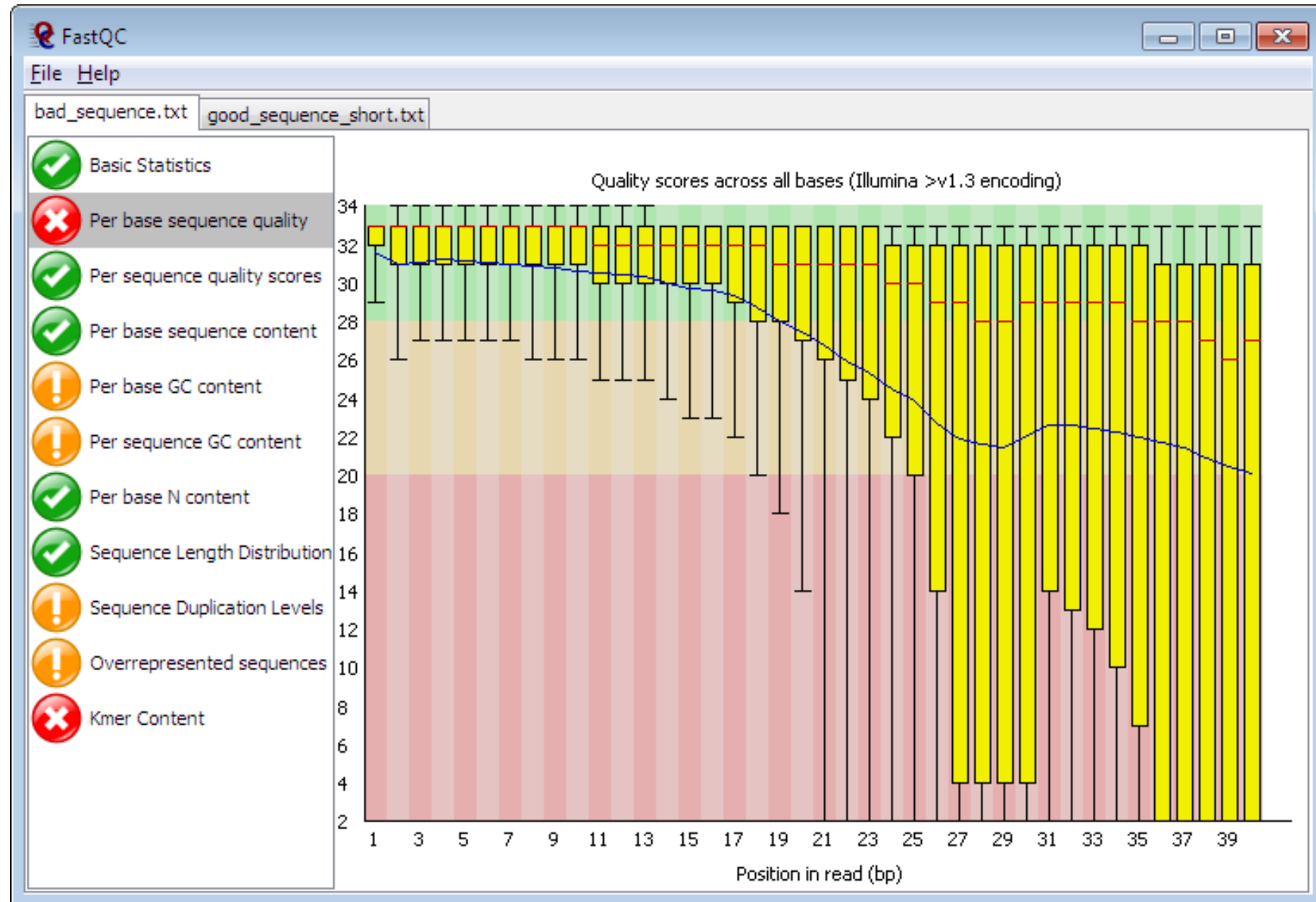


# **III. RNAseq data analysis**

# RNA-seq data analysis

1. Data quality checking.
2. Read mapping.
3. Quantification of gene expression.
4. Differential gene expression analysis.
5. Interpretation of DE analysis results.

# Quality check of reads using FastQC



# Phred quality score

```
+SEQ_ID
```

```
! ' ' * ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * *
```

A quality value  $Q$  is an integer representation of the probability  $p$  that the corresponding base call is incorrect.

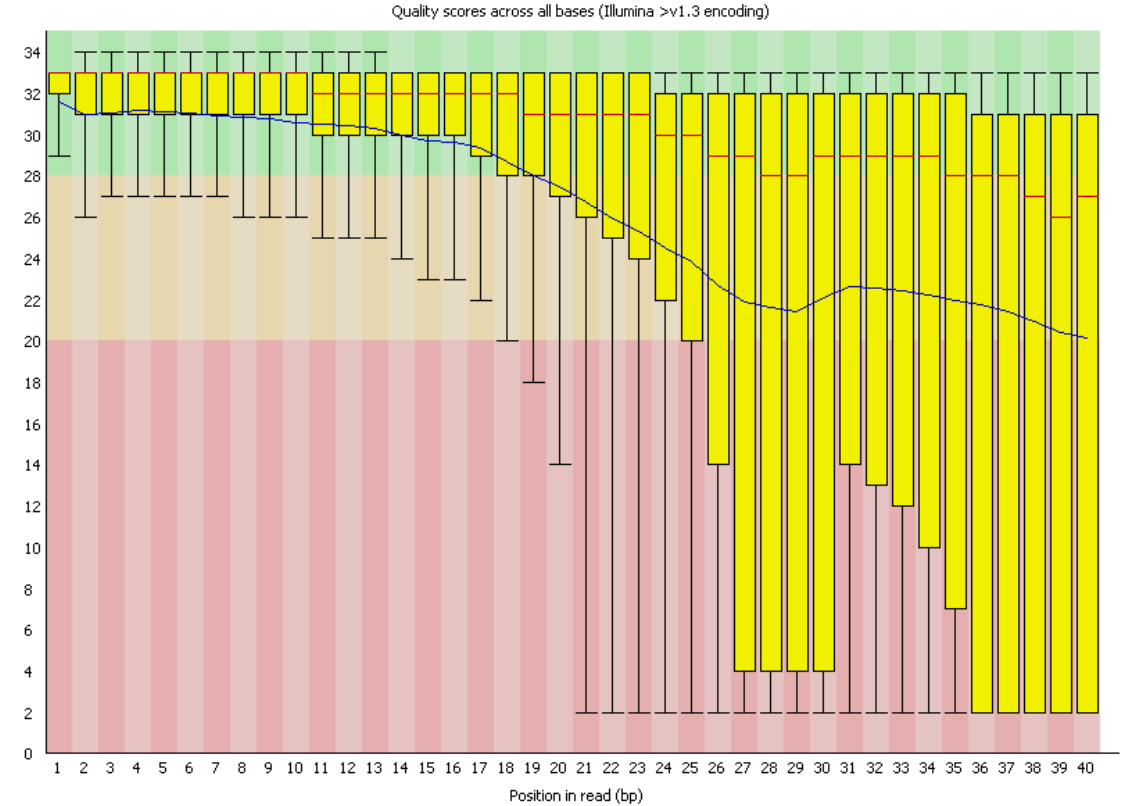
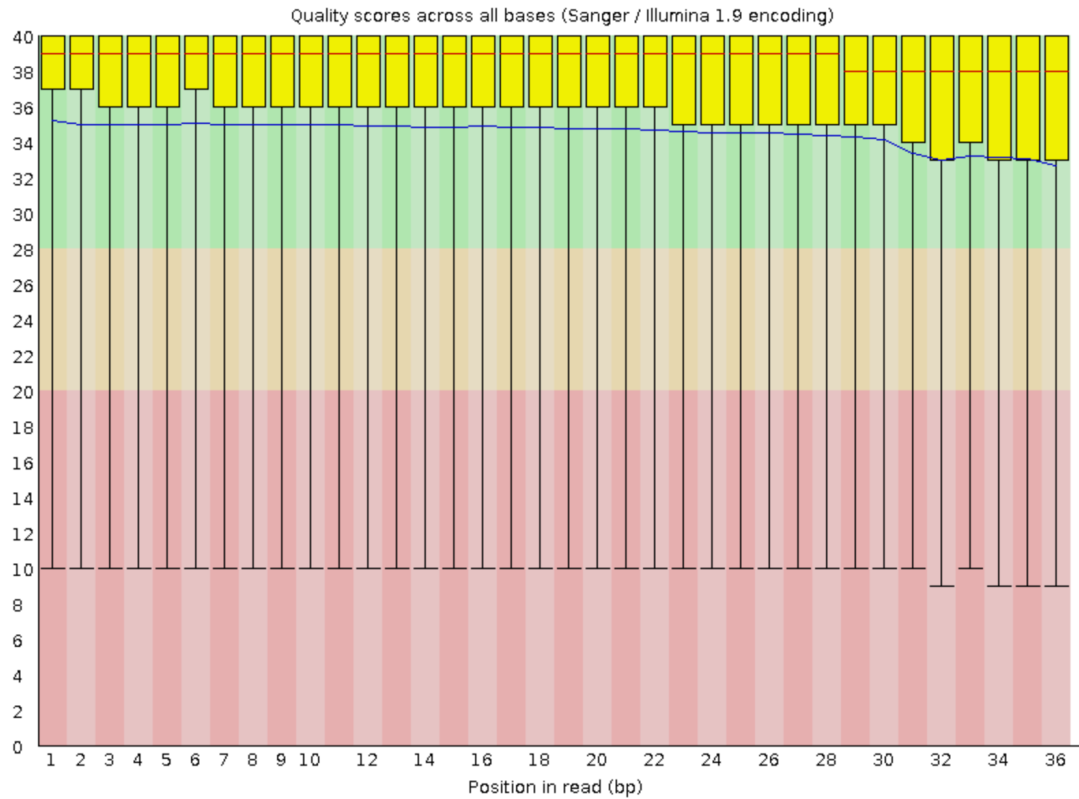
$$Q = -10 \log_{10} P \quad \longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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%

# Good and bad quality of sequencing reads

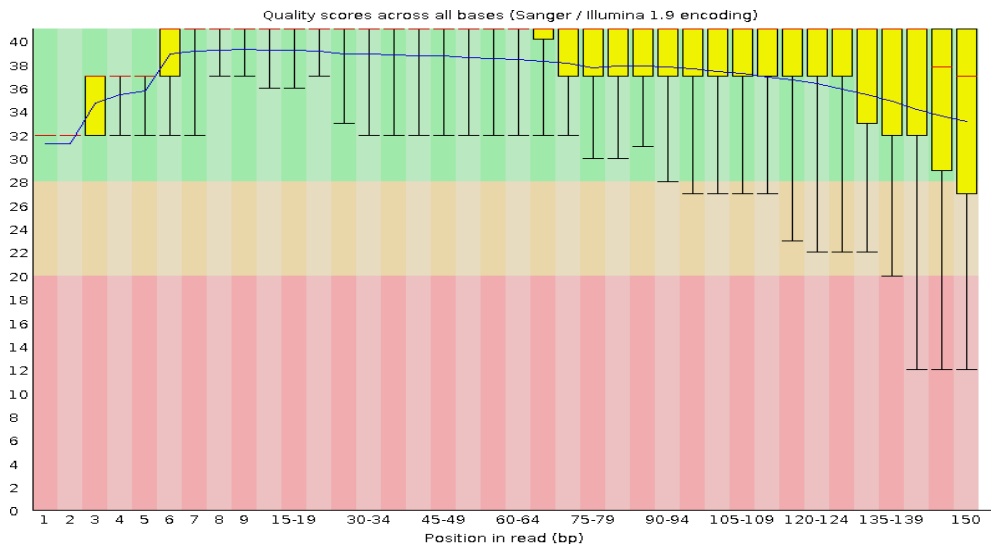
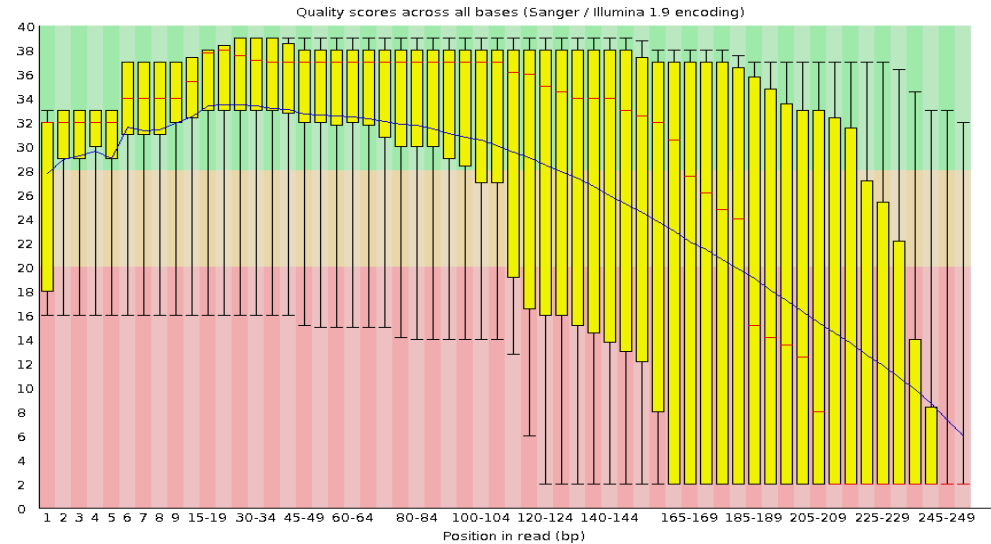


## Per base sequence quality

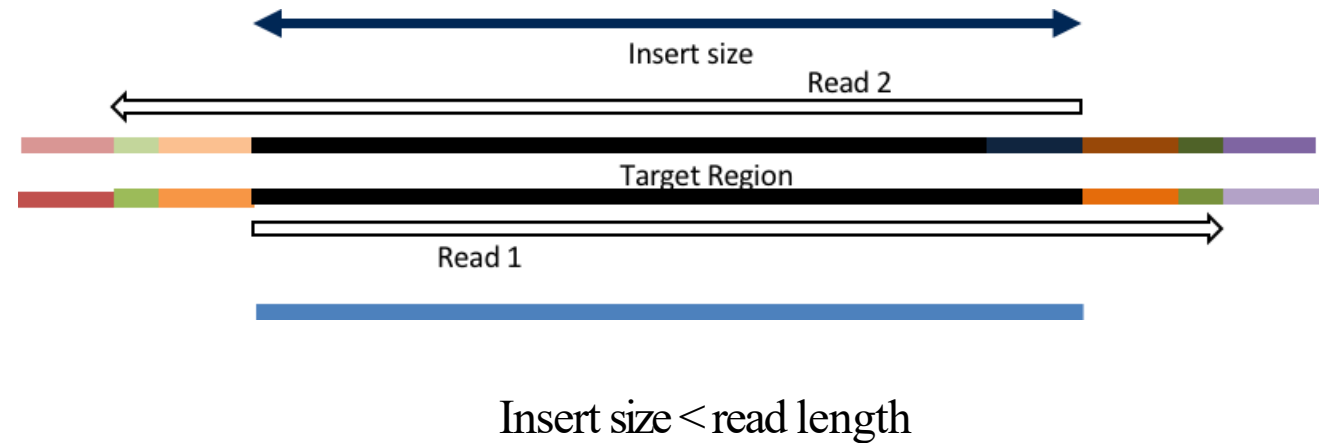


# Data quality checking and removing the contaminations

## RNAseq reads quality



## Adapter removal



## Trimming the reads

**For paired end:** Bbduk, Skewer, HTStream, and FASTP

**For single end:** Cutadapt, HTStream, and BBduk

# Software for data quality checking

1. We use fastqc to run data quality checking.

2. Available at

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

```
fastqc xx.fastq
```

# Tools used for long reads mapping

- Minimap2
- LAST
- NGMLR
- GraphMap

<https://www.nature.com/articles/s41587-021-01108-x>

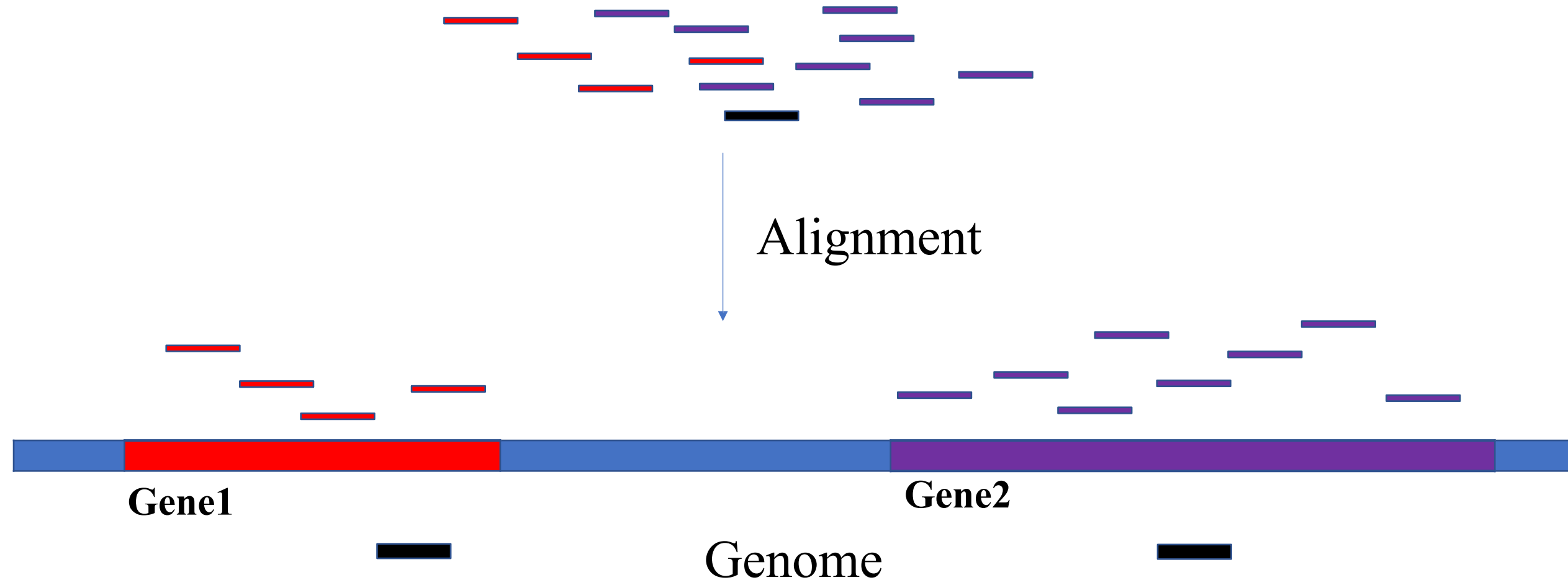


# Tools used for short reads mapping

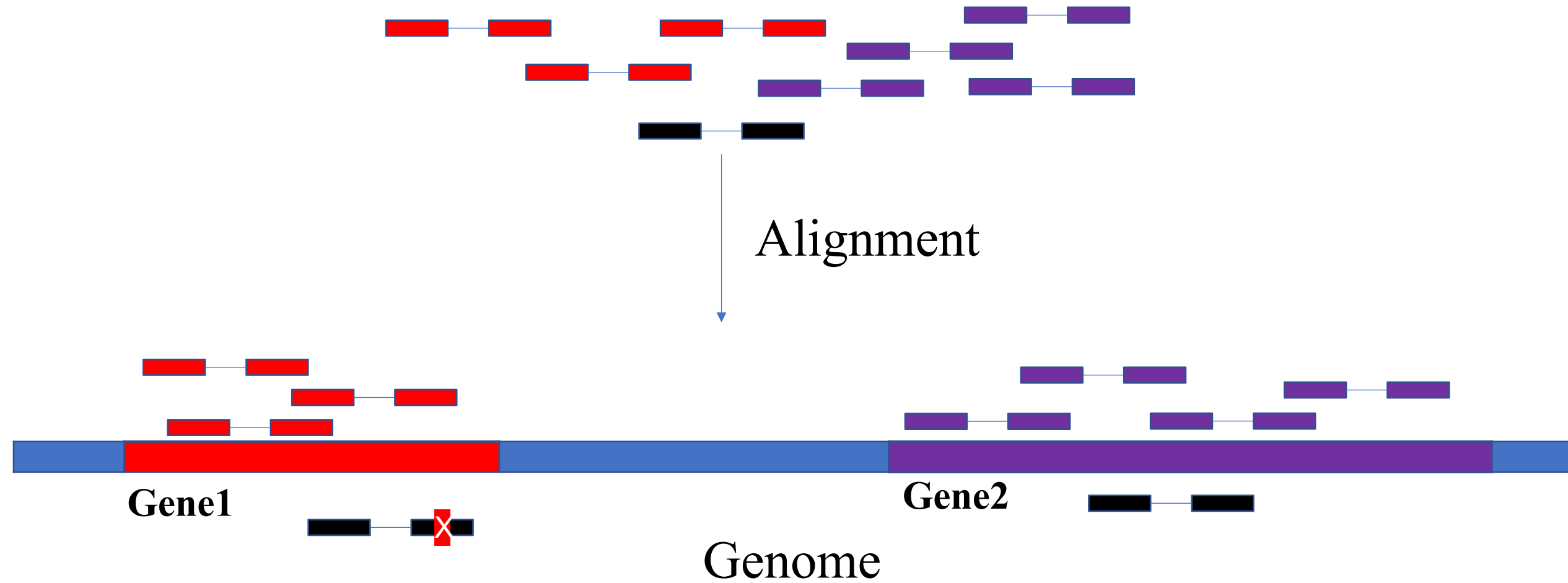
- STAR (Spliced Transcripts Alignment to a Reference)
- Kallisto
- BWA (Burrows-Wheeler Aligner)

# Read mapping (single-end)

The process to align short RNAseq reads with the genome (transcriptome) sequence.



# Read mapping (paired-end)



# Quantification of gene expression (raw read count)



Expression of **Gene1**: 3

Expression of **Gene2**: 4

# Problem of raw read count

1. NOT comparable between different experiments (library size matters).
2. NOT comparable between different genes (gene length matters).

Raw-read-count is proportional to  $\text{library.size} * \text{gene.length}$ ,  
**normalization** is needed!

Rep1 (library size = 7)



Rep2 (library size = 14)



# Quantification of gene expression (RPKM and FPKM)

RPKM: Reads Per Kilobase Per million mapped reads (single-end).

FPKM: Fragments Per Kilobase Per million mapped reads (paired-end).

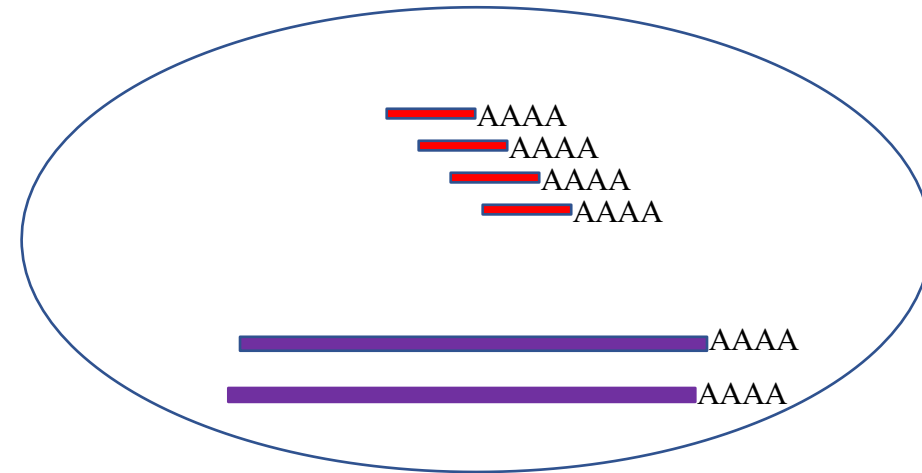
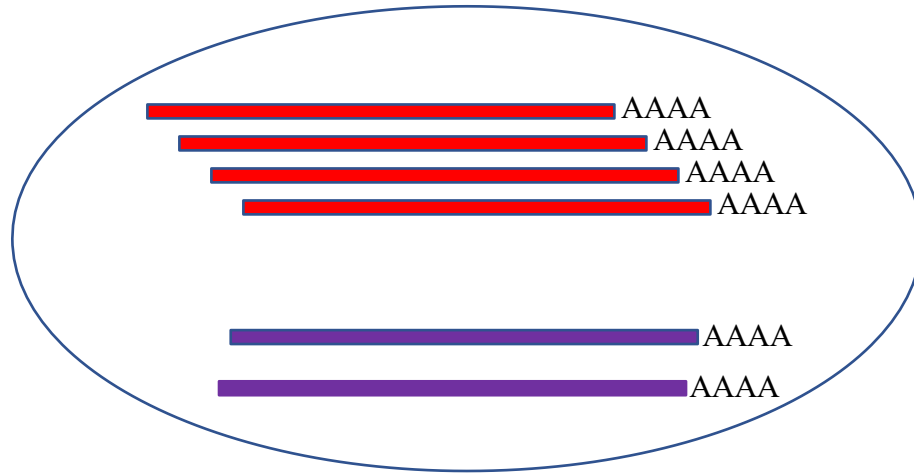
C = Number of reads (fragments) mapped to a gene

N = Total number of mapped reads (fragments) in the experiment

L = Exon length in base-pairs for a gene

$$\text{RPKM} = (10^9 * C) / (N * L)$$

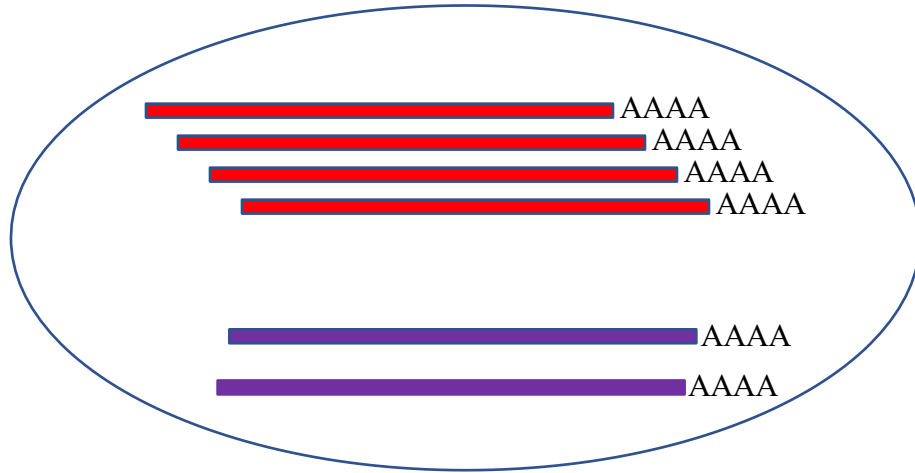
# Problem of RPKM (FPKM)



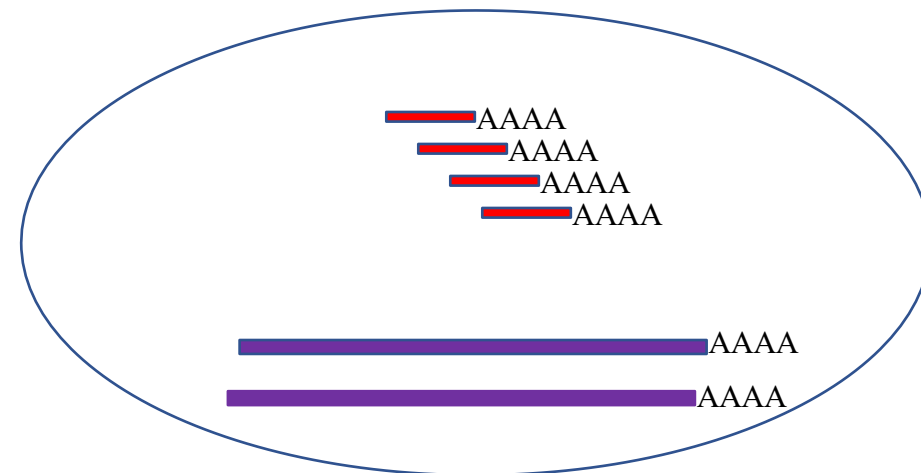
Gene1   $E_1 = (4 / 6)$

Gene2   $E_2 = (2 / 6)$

# Isoform matters



Exp1



Exp2

Suppose we want to compute RPKM for the **purple** gene

$$\text{RPKM} = (10^9 * C) / (N * L) = 10^9 * (C / N) * (1 / L)$$

C = Number of reads mapped to a gene

N = Total number of mapped reads in the experiment

L = Exon length in base-pairs for a gene

C / N: proportion of reads coming from a gene

rl: read length RNAseq experiment

Lp: isoform length (purple)

Lr: isoform length (red, exp1)

Lrs: isoform length (red, exp2)

$$P1 = Lp * 2 / rl$$

$$R1 = Lr * 4 / rl$$

$$P2 = Lp * 2 / rl$$

$$R2 = Lrs * 4 / rl$$

$$\text{RPKM1} = 10^9 * (P1 / (P1 + R1)) * 1 / Lp$$

$$\text{RPKM2} = 10^9 * (P2 / (P2 + R2)) * 1 / Lp$$



# Quantification of gene expression (TPM)

TPM: Transcripts Per Million.

Given a Gene  $G_i$ , compute  $T_i = C_i / L_i$

$C_i$ : Number of reads mapped to the gene

$L_i$ : Exon length in base-pairs for the gene

$$TPM_i = 10^6 * T_i / (T_1 + T_2 + \dots T_n)$$

Wagner *et al.* **Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.**

# TPM vs RPKM

Original data:

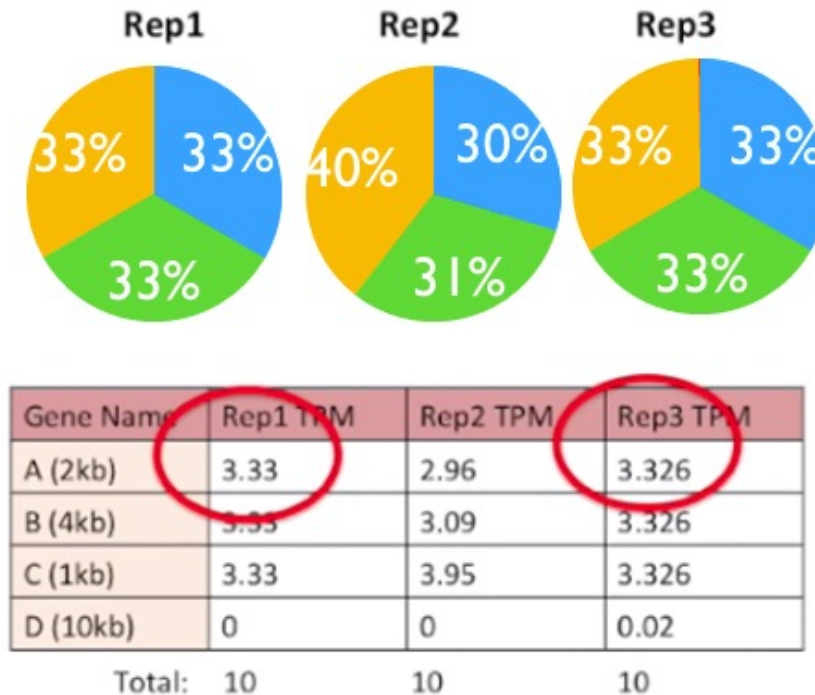
Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Consider 3 pies, each the same size (10).

A 3.33 sized slice is the same in each pie, and is always larger than 3.32.

TPM makes it clear that in Rep1, more of its total reads mapped to gene A than in Rep3.

TPM

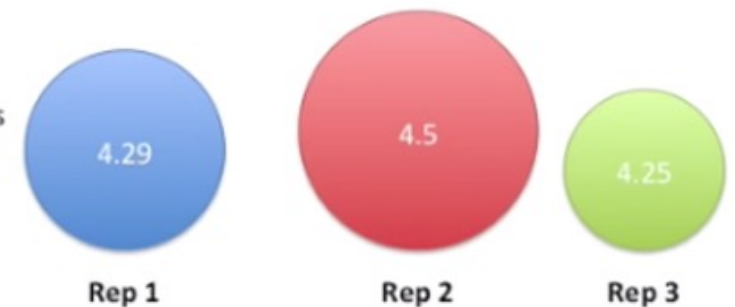


RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

With RPKM, it is harder to compare the proportion of total reads because each replicate has different total (each pie has a different size)

A 1.43 size slice represents a different proportion of reads in in different pies.



# Relationship between RPKM (FPKM) and TPM

$$\text{TPM}_i = \text{RPKM}_i / \text{Sum (RPKM)}$$

Ref: <https://rnajournal.cshlp.org/content/early/2020/04/13/rna.074922.120.full.pdf>

# Core analysis

1. Data quality checking.
2. Read mapping.
3. Quantification of gene expression.
4. Differential gene expression analysis (R).
5. Interpretation of DE analysis results.

# Data quality checking

1. We use fastqc to run data quality checking.

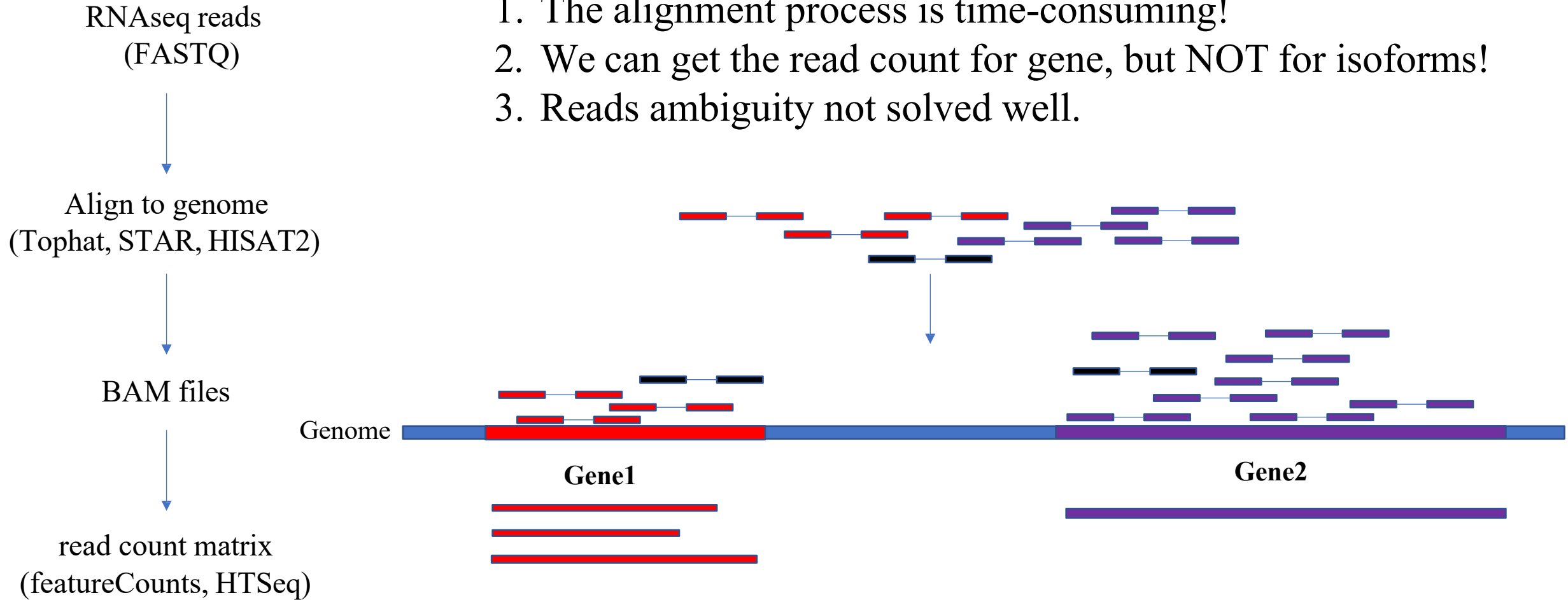
2. Available at

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

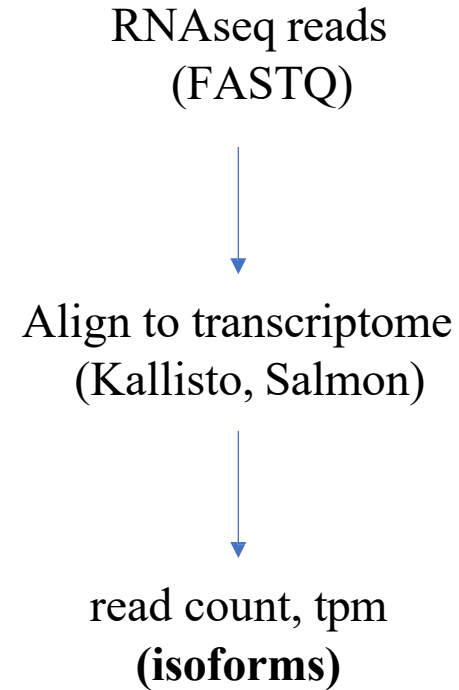
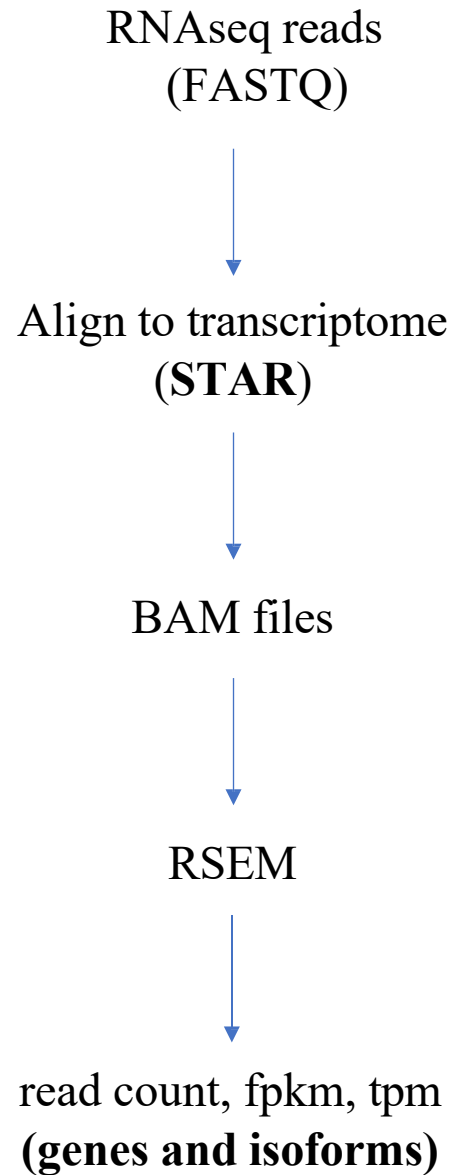
```
fastqc xx.fastq
```

# Quantification of gene expression

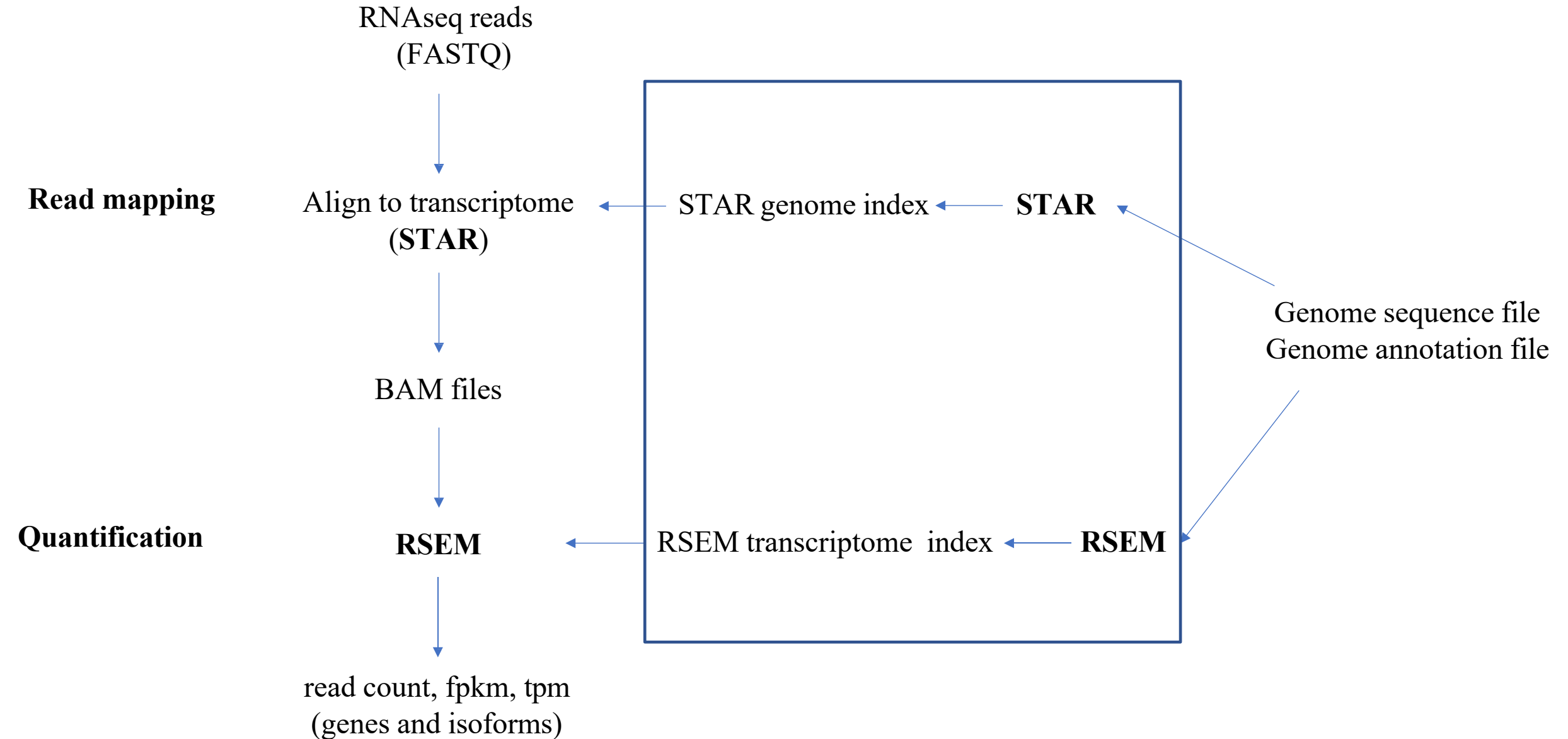
1. The alignment process is time-consuming!
2. We can get the read count for gene, but NOT for isoforms!
3. Reads ambiguity not solved well.



# Quantification of gene expression



# Quantification of gene expression





# Preparation of index files

Genome sequence (from NCBI):

[ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old\\_genbank/Eukaryotes/vertebrates\\_mammals/Homo\\_sapiens/GRCh38/seqs\\_for\\_alignment\\_pipelines/GCA\\_000001405.15\\_GRCh38\\_no\\_alt\\_analysis\\_set.fna.gz](ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_genbank/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh38/seqs_for_alignment_pipelines/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz)

Genome annotation file (from GENCODE):

[ftp://ftp.ebi.ac.uk/pub/databases/genocode/Gencode\\_human/release\\_23/gencode.v23.annotation.gtf.gz](ftp://ftp.ebi.ac.uk/pub/databases/genocode/Gencode_human/release_23/gencode.v23.annotation.gtf.gz)

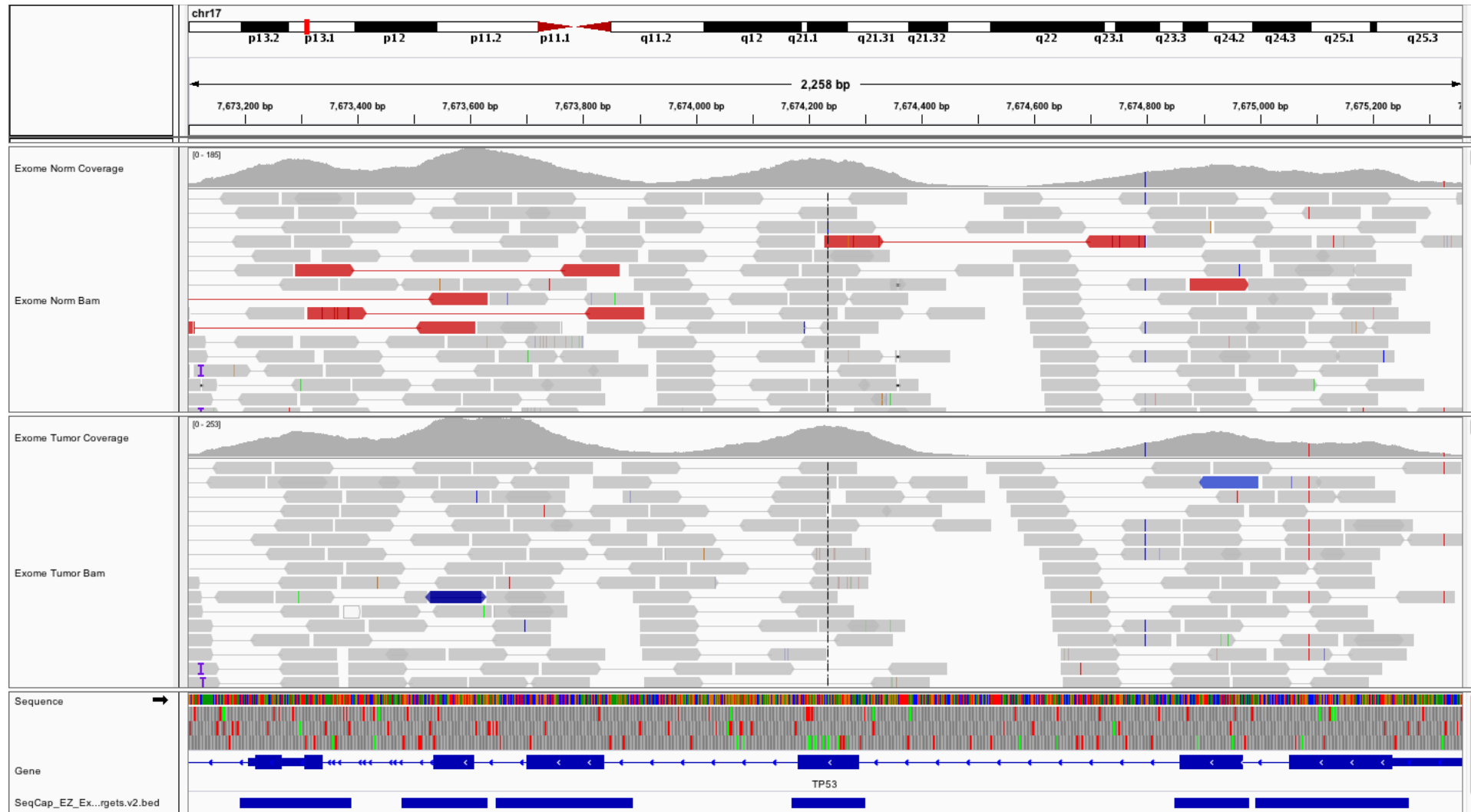
```
STAR --runThreadN 32 \
      --runMode genomeGenerate \
      --genomeDir STAR.index \
      --genomeFastaFiles GCA_000001405.15_GRCh38_no_alt_analysis_set.fna \
      --sjdbGTFfile gencode.v23.annotation.gtf
```

```
rsem-prepare-reference -p 4
                        --gtf gencode.v23.annotation.gtf
                        GCA_000001405.15_GRCh38_no_alt_analysis_set.fna
                        RSEM.index/hg38.RSEM.index
```

# Code to run the mapping using STAR aligner

```
STAR --genomeDir /ensembl38_STAR_index/  
--runThreadN 6 \  
--readFilesIn Mov10_oe_R1.subset.fq Mov10_oe_R2.subset.fq \  
--outFileNamePrefix ../results/STAR/Mov10_oe_1_ \  
--outSAMtype BAM SortedByCoordinate \  
--outSAMunmapped Within \  
--outSAMattributes Standard
```

# STAR aligner output



# Quantify the read count, TPM, and FPKM

```
Rsem-calculate-expression \  
--no-bam-output \  
--quiet \  
--no-qualities \  
-p 8 \  
--seed-length 25 \  
--bam \  
--paired-end tmp/sample/Aligned.toTranscriptome.out.bam \  
$RSEM_INDEX \  
RSEM.output/sample.txt
```

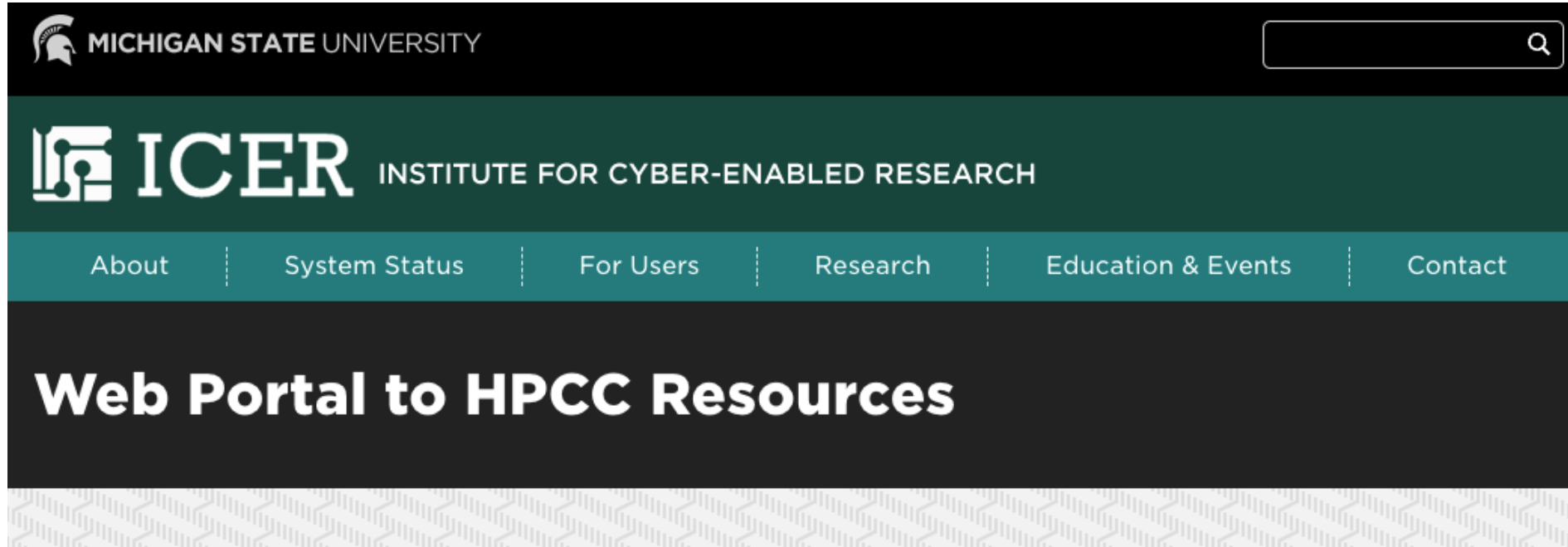
# RSEM-gene-count output

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
ENSG00000000003.14	ENST00000373020.8,ENST00000494424.1,ENST00000496771.5,ENST0000061215.1	2232.08	1949.92	507	21.37	16.94
ENSG00000000005.5	ENST00000373031.4,ENST00000485971.1	940.5	659.64	0	0	0
ENSG000000000419.12	ENST00000371582.8,ENST00000371584.8,ENST00000371588.9,ENST0000041308.1	1080.86	798.7	1030	106	84.01
ENSG000000000457.13	ENST00000367770.5,ENST00000367771.10,ENST00000367772.8,ENST000004236.1	3552.99	3270.82	322.57	8.11	6.42
ENSG000000000460.16	ENST00000286031.10,ENST00000359326.8,ENST00000413811.3,ENST000004597.1	2094.04	1811.95	266.43	12.09	9.58
ENSG000000000938.12	ENST00000374003.7,ENST00000374004.5,ENST00000374005.7,ENST0000039917.1	1735.29	1453.21	0	0	0
ENSG000000000971.15	ENST00000359637.2,ENST00000367429.8,ENST00000466229.5,ENST0000047091.1	2516.93	2234.77	491	18.06	14.31
ENSG000000001036.13	ENST00000002165.10,ENST00000367585.1,ENST00000451668.1	2310.54	2028.39	1246.83	50.53	40.04
ENSG000000001084.10	ENST00000229416.10,ENST00000504353.1,ENST00000504525.1,ENST000005051.1	2397.67	2116.53	759	29.48	23.36
ENSG000000001167.14	ENST00000341376.10,ENST00000353205.5	2873.33	2591.16	601	19.07	15.11
ENSG000000001460.17	ENST00000003583.12,ENST00000337248.8,ENST00000374409.5,ENST000004351.1	2419.78	2137.65	150	5.77	4.57
ENSG000000001461.16	ENST00000003912.7,ENST00000339255.2,ENST00000358028.8,ENST0000037439.1	3891.98	3609.82	907	20.65	16.37
ENSG000000001497.16	ENST00000374804.9,ENST00000374807.9,ENST00000374811.7,ENST0000046909.1	2360.35	2078.42	1536	60.75	48.14
ENSG000000001561.6	ENST00000321037.4	4651	4368.83	200	3.76	2.98
ENSG000000001617.11	ENST00000002829.7,ENST00000413852.5,ENST00000414301.5,ENST0000042083.1	3300.24	3018.08	258	7.03	5.57
ENSG000000001626.14	ENST00000003084.10,ENST00000426809.5,ENST00000429014.1,ENST000004468.1	5207.91	4925.92	2667	44.5	35.27
ENSG000000001629.9	ENST00000265742.7,ENST00000413588.1,ENST00000422095.1,ENST0000043988.1	4820.71	4538.85	1338	24.23	19.2
ENSG000000001630.15	ENST00000003100.12,ENST00000422867.1,ENST00000435873.1,ENST000004507.1	2898.46	2616.3	1178	37.01	29.33
ENSG000000001631.14	ENST00000340022.6,ENST00000394503.6,ENST00000394505.6,ENST0000039450.1	2597.49	2315.46	523.65	18.59	14.73
ENSG000000002016.16	ENST00000228345.9,ENST00000358495.7,ENST00000397230.6,ENST0000043009.1	1486.5	1207.08	75.14	5.12	4.06
ENSG000000002079.12	ENST00000413734.1,ENST00000425880.1,ENST00000429079.1,ENST0000043978.1	1368.26	1090.4	21	1.58	1.25
ENSG000000002330.13	ENST00000309032.7,ENST00000394531.3,ENST00000394532.7,ENST0000049214.1	892.39	610.71	604.95	81.42	64.53
ENSG000000002549.12	ENST00000226299.8,ENST00000503467.1,ENST00000504614.5,ENST0000050796.1	1908.17	1626	1027	51.92	41.14
ENSG000000002586.17	ENST00000381177.5,ENST00000381180.7,ENST00000381184.5,ENST0000038118.1	1194.92	912.76	549	49.44	39.18
ENSG000000002587.9	ENST00000002596.5,ENST00000510712.1,ENST00000514690.5	6776.5	6495.05	1076	13.62	10.79
ENSG000000002726.19	ENST00000360937.8,ENST00000416793.6,ENST00000460213.1,ENST0000046729.1	2572.25	2290.08	1888	67.77	53.7

**Q&A**

# **III. Demo of RNAseq data analysis on MSU HPCC**

# Starting OnDemand on HPCC



Dear HPCC users,

It is now possible for users to access HPCC resources via internet browsers. Many new users will attempt to log into HPCC via the wiki page: [docs.icer.msu.edu/](https://docs.icer.msu.edu/). However, HPCC users are not able to login to the HPCC via the wiki pages. Instead, to access HPCC via a web browser, use the following portal:

**HPCC OnDemand, login portal:** <https://ondemand.hpcc.msu.edu/> 

This is our new web access to HPCC resources. To see its features and how to use it, please visit "[Open OnDemand](#)" [wiki page](#).

<https://ondemand.hpcc.msu.edu/>



# Login on OnDemand HPCC



## Consent to Attribute Release ✓

HPCC OnDemand requests access to the following information. If you do not approve this request, do not proceed.

- Your CILogon user identifier
- Your name
- Your email address
- Your username and affiliation from your identity provider

## Selected Identity Provider

Michigan State University ▼ ⓘ

☐ Remember this selection ⓘ

Log On

Log On

By selecting "Log On", you agree to the [privacy policy](#).

# Request to start Interactive Desktop

[Home](#) / [My Interactive Sessions](#) / Interactive Desktop

## Interactive Apps

[Desktops](#)

 **Interactive Desktop**

 Interactive Desktop (Legacy)

[GUIs](#)

 **MATLAB**

 MATLAB (Legacy)

 **ParaView**

 ParaView (Legacy)

 **RStudio (Legacy)**

 **Stata**

 Stata (Legacy)

[Servers](#)

 **Code Server**

## Interactive Desktop

This app will launch an interactive desktop on one or more compute nodes. You will have full access to the resources these nodes provide. This is analogous to an interactive batch job.

Number of hours

Jobs shorter than four hours will schedule much faster

Number of cores per task

Amount of memory

E.g. 100GB or 500MB. 750MB per core if left blank.








☐ I would like to receive an email when the session starts

☐ Advanced Options

**Launch**

\* The Interactive Desktop session data for this session can be accessed under the [data root directory](#).

# Lunch Interactive Desktop on HPCC

Interactive Apps
Desktops
 Interactive Desktop
 Interactive Desktop (Legacy)
GUIs
 MATLAB
 MATLAB (Legacy)
 ParaView
 ParaView (Legacy)
 RStudio (Legacy)

## Interactive Desktop (40316109)

1 node | 10 cores | Running

Host: >\_skl-135

Cancel

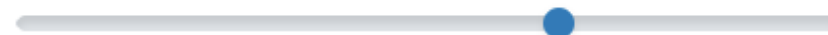
Created at: 2024-08-04 12:20:24 EDT

Time Remaining: 9 hours and 6 minutes

Session ID: 57dea186-d0f8-4ab1-ba51-703854ea9efe

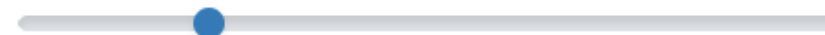
Problems with this session? [Submit support ticket](#)

Compression



0 (low) to 9 (high)

Image Quality

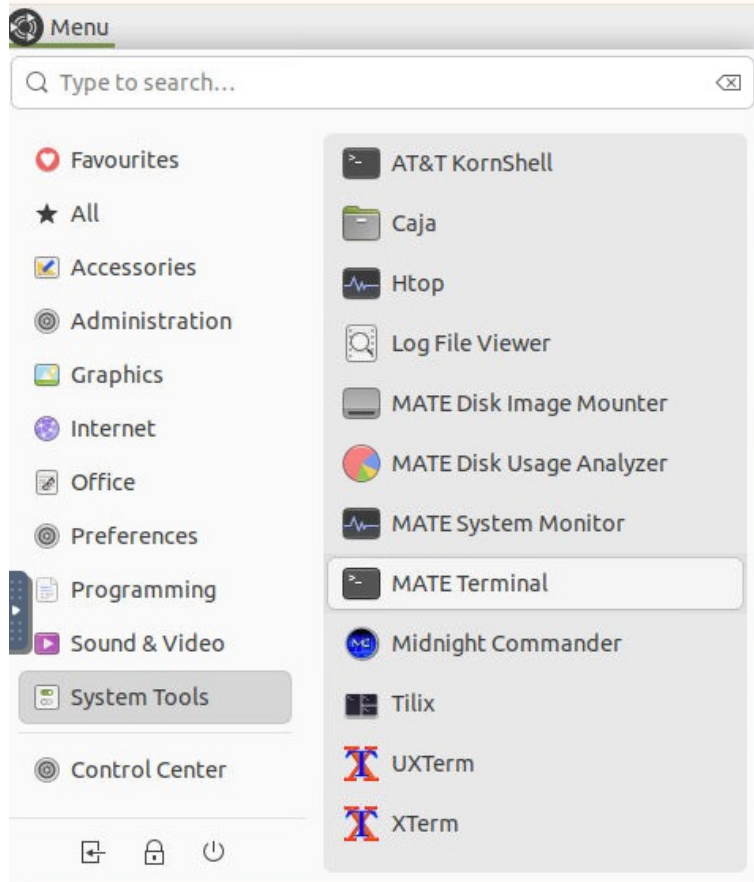


0 (low) to 9 (high)

Launch Interactive Desktop

View Only (Share-able Link)

# Access terminal on Interactive Desktop



# Available versions of STAR package on HPCC

```
[ramashan@css-075 ~]$ module spider STAR
```

```
-----  
STAR:
```

```
-----  
Description:
```

```
STAR aligns RNA-seq reads to a reference genome using uncompressed suffix arrays.
```

```
Versions:
```

```
STAR/2.6.0c
```

```
STAR/2.6.1c
```

```
STAR/2.7.2b
```

```
STAR/2.7.3a
```

```
STAR/2.7.9a
```

```
STAR/2.7.10b
```

```
Other possible modules matches:
```

```
stars
```

```
-----  
To find other possible module matches execute:
```

```
$ module -r spider '.*STAR.*'
```

```
-----  
For detailed information about a specific "STAR" package (including how to load the modules) use the module's full name.
```

```
Note that names that have a trailing (E) are extensions provided by other modules.
```

```
For example:
```

```
$ module spider STAR/2.7.10b  
-----
```

# Loading of STAR package on HPCC

```
[ramashan@css-075 ~]$ module spider STAR/2.7.10b
```

```
-----  
STAR: STAR/2.7.10b  
-----
```

Description:

STAR aligns RNA-seq reads to a reference genome using uncompressed suffix arrays.

You will need to load all module(s) on any one of the lines below before the "STAR/2.7.10b" module is available to load.

GCC/11.3.0

Help:

Description

=====

STAR aligns RNA-seq reads to a reference genome using uncompressed suffix arrays.

More information

=====

- Homepage: <https://github.com/alexdobin/STAR>

```
[ramashan@css-075 ~]$ module purge
```

```
[ramashan@css-075 ~]$ module load GCC/11.3.0 STAR/2.7.10b
```

# Loading of RSEM package on HPCC

```
[ramashan@css-075 ~]$ module spider RSEM
```

```
-----  
RSEM:
```

```
-----  
Description:
```

```
RNA-Seq by Expectation-Maximization
```

```
Versions:
```

```
RSEM/1.3.0
```

```
RSEM/1.3.1
```

```
RSEM/1.3.3
```

```
[ramashan@css-075 ~]$ module spider RSEM/1.3.3
```

```
-----  
RSEM: RSEM/1.3.3
```

```
-----  
Description:
```

```
RNA-Seq by Expectation-Maximization
```

You will need to load all module(s) on any one of the lines below before the "RSEM/1.3.3" module is available to load.

```
GCC/11.2.0  OpenMPI/4.1.1
```

```
GCC/11.3.0  OpenMPI/4.1.4
```

```
GCC/8.3.0   OpenMPI/3.1.4
```

```
[ramashan@css-075 ~]$ module load GCC/11.3.0  OpenMPI/4.1.4 RSEM/1.3.3
```

# Loading R on HPCC

```
[ramashan@css-075 ~]$ module spider R
```

Versions:

```
R/3.3.1
R/3.4.3-X11-20160819
R/3.4.3-X11-20171023
R/3.4.3xF
R/3.4.3xS
R/3.4.4-X11-20180131
R/3.5.0-X11-20180131
R/3.5.1-X11-20180131
R/3.5.1-X11-20180604
R/3.6.0-X11-20180604
R/3.6.2-X11-20180604
R/3.6.2
R/3.6.3
R/4.0.0-X11-20180604
R/4.0.0
R/4.0.2.bak
R/4.0.2.test
R/4.0.2-X11-20180604
R/4.0.2
R/4.0.3
R/4.1.0
R/4.1.2
R/4.2.1
R/4.2.2
R/4.3.1
```

Other possible modules matches:

```
ADMIXTURE  AMDuProf  APR  APR-util  Abaqus_parallel  AdapterRemoval  Advisor  ...
```

```
[ramashan@css-075 ~]$ module spider R/4.2.1
```

```
-----
R: R/4.2.1
-----
```

Description:

R is a free software environment for statistical computing and graphics.

You will need to load all module(s) on any one of the lines below before the "R/4.2.1" module is available to load.

```
GCC/11.3.0  OpenMPI/4.1.4
```

```
[ramashan@css-075 ~]$ module load GCC/11.3.0  OpenMPI/4.1.4 R/4.2.1
```



# Loading the packages with the new OS on HPCC

```
ramashan@skl-135:~$ module purge  
ramashan@skl-135:~$ module load R  
ramashan@skl-135:~$ module load STAR  
ramashan@skl-135:~$ module load RSEM
```

# Test the loaded module

```
ramashan@dev-amd20-v100:~/data/RNASeq_test$ STAR
Usage: STAR [options]... --genomeDir /path/to/genome/index/ \
--readFilesIn R1.fq R2.fq
```

Spliced Transcripts Alignment to a Reference (c) Alexander Dobin, 2009–2022

STAR version=2.7.11b

STAR compilation time,server,dir=2024-07-18T13:18:20-04:00 dev-intel14:/tmp/  
panchyni/easybuild/easybuild/build/STAR/2.7.11b/GCC-13.2.0/STAR-2.7.11b/source

For more details see:

<<https://github.com/alexdobin/STAR>>

<<https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf>>

To list all parameters, run STAR --help

# Create reference index file for STAR and RSEM

```
[ramashan@skl-162 RNASeq_test]$ STAR --runThreadN 32 --runMode genomeGenerate --genomeDir Genome/ --genomeFastaFiles Genome/fastq/genome.fa --sjdbGTFfile Genome/genes/genes.gtf
STAR --runThreadN 32 --runMode genomeGenerate --genomeDir Genome/ --genomeFastaFiles Genome/fastq/genome.fa --sjdbGTFfile Genome/genes/genes.gtf
STAR version: 2.7.10b compiled: 2023-09-22T17:53:50-0400 dev-intel18:/tmp/panchyni/EASYBUILD/STAR/2.7.10b/GCC-11.3.0/STAR-2.7.10b/source
Aug 01 15:34:04 ..... started STAR run
Aug 01 15:34:04 ... starting to generate Genome files
Aug 01 15:34:48 ..... processing annotations GTF
Aug 01 15:35:15 ... starting to sort Suffix Array. This may take a long time...
Aug 01 15:35:30 ... sorting Suffix Array chunks and saving them to disk...
Aug 01 16:24:53 ... loading chunks from disk, packing SA...
Aug 01 16:25:56 ... finished generating suffix array
Aug 01 16:25:56 ... generating Suffix Array index
Aug 01 16:28:56 ... completed Suffix Array index
Aug 01 16:28:57 ..... inserting junctions into the genome indices
Aug 01 16:32:13 ... writing Genome to disk ...
Aug 01 16:32:14 ... writing Suffix Array to disk ...
Aug 01 16:32:23 ... writing SAindex to disk
Aug 01 16:32:24 ..... finished successfully
```

```
[ramashan@css-075 ~]$ rsem-prepare-reference -p 4 --gtf Genome/genes/genes.gtf
Genome/fastq/Genome.fa hg38
```

# Running bulk.rna.seq.pipeline.R to process the samples

```
ramashan@skl-135:/mnt/research/IPSTP_2024/RNAseq_analysis$ Rscript chenlab.bulk.rna.seq.pipeline.R sample.name.txt
Loading required package: data.table
Loading required package: plyr
Loading required package: dplyr

Attaching package: 'dplyr'

The following objects are masked from 'package:plyr':

  arrange, count, desc, failwith, id, mutate, rename, summarise,
  summarize

The following objects are masked from 'package:data.table':

  between, first, last

The following objects are masked from 'package:stats':

  filter, lag

The following objects are masked from 'package:base':

  intersect, setdiff, setequal, union

Loading required package: foreach
Loading required package: parallel
Loading required package: doParallel
Loading required package: iterators
[1] "sample HT29_Ctr1 finished!"
[1] "sample HT29_Ctr2 finished!"
[1] "sample HT29_Ctr3 finished!"
[1] "sample HT29_CPN_OE1 finished!"
[1] "sample HT29_CPN_OE2 finished!"
[1] "sample HT29_CPN_OE3 finished!"
```



# Raw read count in processed samples

	HT29_CPN_OE1	HT29_CPN_OE2	HT29_CPN_OE3	HT29_Ctr1	HT29_Ctr2	HT29_Ctr3
ENSG000000000003	9.192293	9.269127	9.330917	8.988685	8.965784	9.157347
ENSG000000000005	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ENSG000000000419	10.051209	10.317413	10.323055	10.009829	10.125413	10.317413
ENSG000000000457	8.463851	8.578826	8.552439	8.337934	8.354602	8.283551
ENSG000000000460	8.128871	8.236636	7.071750	8.063018	8.122414	8.310158
ENSG000000000938	0.000000	1.000000	0.000000	0.000000	1.000000	0.000000
ENSG000000000971	9.330917	9.560333	9.467606	8.942515	9.019591	8.960002
ENSG00000001036	10.278275	10.410271	10.383564	10.285206	10.229816	10.310567
ENSG00000001084	9.479780	9.622052	9.632995	9.569856	9.567956	9.471391
ENSG00000001167	9.105909	9.385862	9.442943	9.233620	9.182394	9.038919
ENSG00000001460	7.011227	7.383704	6.930737	7.238405	7.139551	7.228819
ENSG00000001461	10.080818	10.166163	9.981567	9.826548	9.823367	9.829723
ENSG00000001497	10.634811	10.760720	10.676839	10.585901	10.544964	10.502832
ENSG00000001561	7.451211	7.807355	7.918863	7.651052	8.055282	7.584963
ENSG00000001617	7.189825	7.383704	7.118941	8.016808	7.826548	8.027906
ENSG00000001626	11.611025	11.811776	11.706496	11.381543	11.442943	11.123475
ENSG00000001629	10.590587	10.716819	10.568906	10.386940	10.454299	10.090112
ENSG00000001630	10.592457	10.827343	10.743151	10.203348	10.197217	10.286558

# FPKM values in processed samples

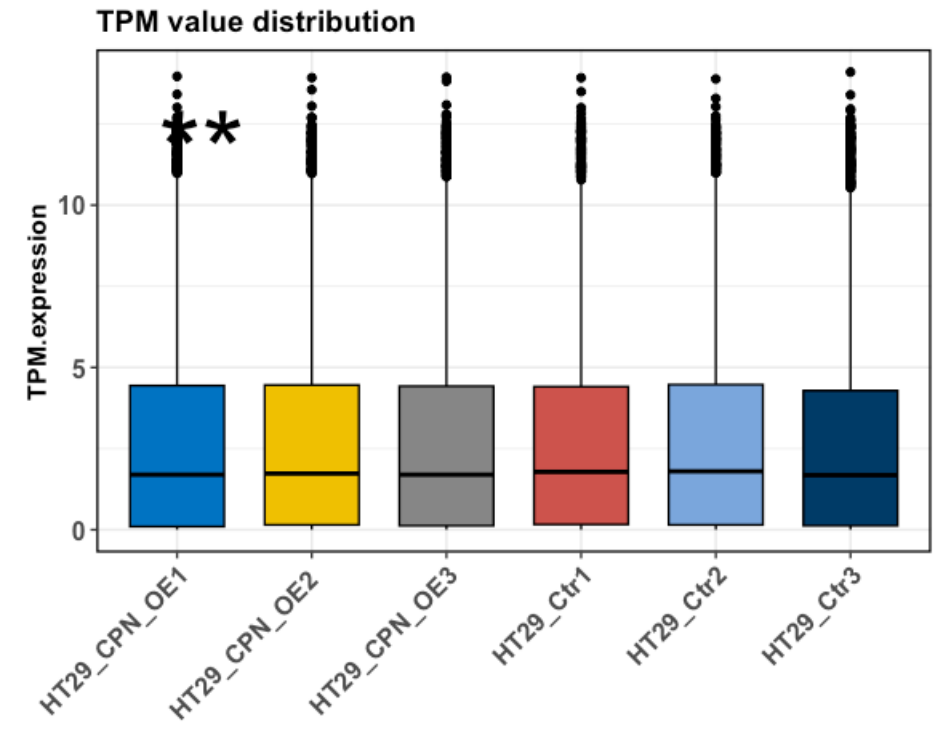
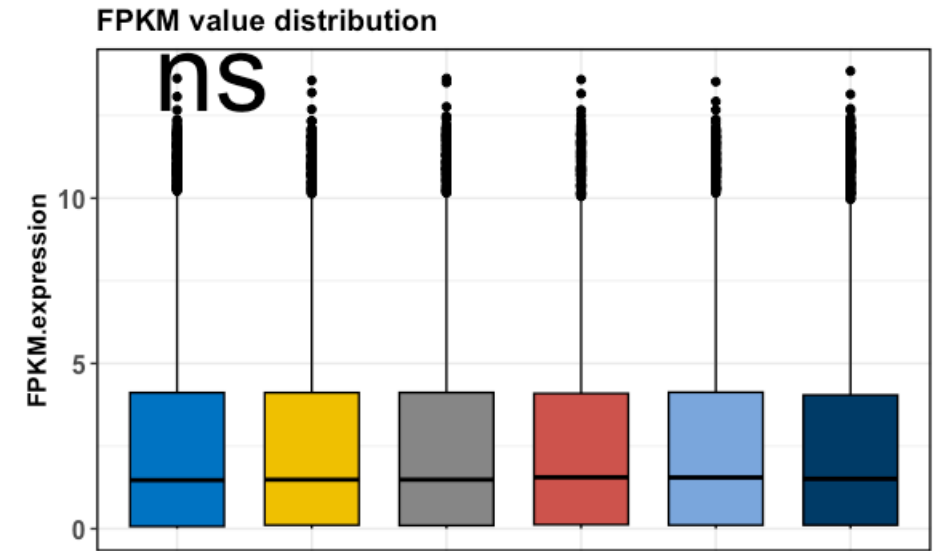
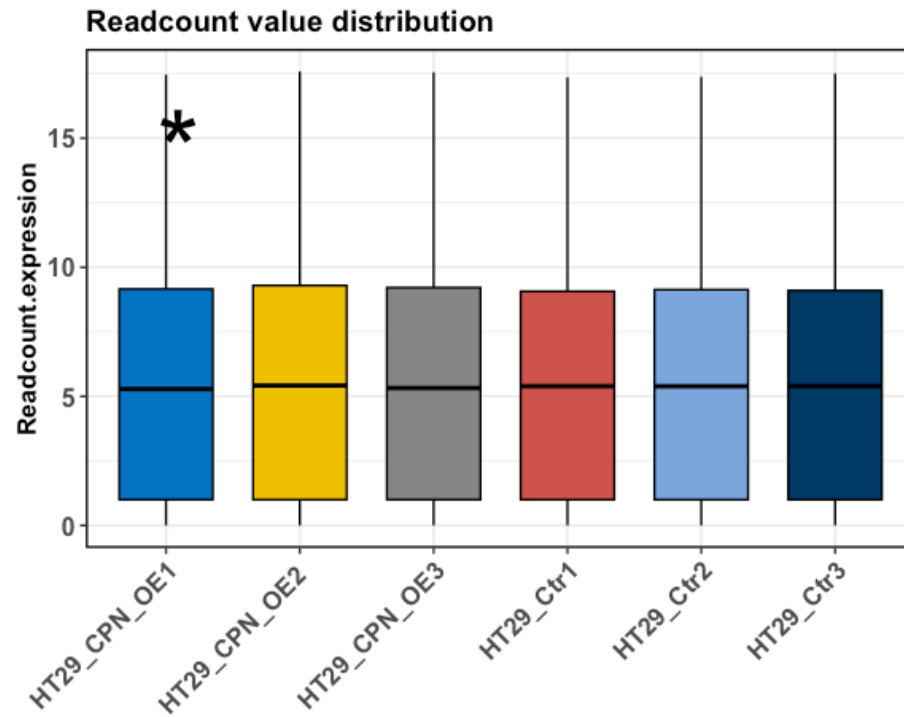
	HT29_CPN_OE1	HT29_CPN_OE2	HT29_CPN_OE3	HT29_Ctr1	HT29_Ctr2	HT29_Ctr3
ENSG000000000003	4.38474059	4.29278175	4.40190347	4.16510799	4.14812063	4.33199178
ENSG000000000005	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
ENSG000000000419	6.36457243	6.50937915	6.59155975	6.40956065	6.49441561	6.64745843
ENSG000000000457	2.86987141	3.00719550	2.92219785	2.89141919	2.88557436	2.75060650
ENSG000000000460	3.28540222	3.38818954	3.11935618	3.40326772	3.35614381	3.51222689
ENSG000000000938	0.00000000	0.04264434	0.00000000	0.00000000	0.05658353	0.00000000
ENSG000000000971	4.22650853	4.24031433	4.27575205	3.93640238	3.91838623	4.02236781
ENSG00000001036	5.28429203	5.28798935	5.33378150	5.35895883	5.26190686	5.33449677
ENSG00000001084	4.64558639	4.78816366	5.03033608	4.60644223	4.85199884	4.79233481
ENSG00000001167	3.82171022	3.89239103	4.12928302	4.00988459	3.96347412	3.64616266
ENSG00000001460	2.33342373	2.76765480	2.15380534	2.47767733	2.36176836	2.63691458
ENSG00000001461	4.39574833	4.34411833	4.27127626	4.11852585	4.19298317	4.00898878
ENSG00000001497	5.56955185	5.61470984	5.56437817	5.61882595	5.49952702	5.48284828
ENSG00000001561	1.79077204	1.95977016	2.09761080	1.99276843	2.28688115	1.90303827
ENSG00000001617	1.94860085	1.99276843	2.30742853	2.71589337	2.58255600	2.69599381
ENSG00000001626	5.25247621	5.32624970	5.40803247	5.18070484	5.13996057	4.75381844
ENSG00000001629	4.78920758	4.36946648	4.85399565	4.33628339	4.59514557	3.97361128
ENSG00000001630	5.25738784	5.31542132	5.35473424	4.92267359	4.93545975	4.90400232
ENSG00000001631	4.45549162	4.25851892	4.26077843	3.97544677	4.23496109	4.02768488

\* Data was paired end

# TPM values in processed samples

	HT29_CPN_OE1	HT29_CPN_OE2	HT29_CPN_OE3	HT29_Ctr1	HT29_Ctr2	HT29_Ctr3
ENSG00000000003	4.71314590	4.63981098	4.70984202	4.48349335	4.49313492	4.57470705
ENSG00000000005	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
ENSG00000000419	6.70431868	6.86888427	6.90977310	6.74146699	6.85436974	6.89953819
ENSG00000000457	3.17152711	3.33055840	3.20633065	3.18745105	3.20476675	2.97085365
ENSG00000000460	3.59693514	3.71918344	3.40599236	3.71039319	3.68818036	3.74631277
ENSG00000000938	0.00000000	0.05658353	0.00000000	0.00000000	0.07038933	0.00000000
ENSG00000000971	4.55397481	4.58616425	4.58255600	4.25247621	4.26077843	4.26228281
ENSG00000001036	5.61970645	5.64299031	5.64789009	5.68734069	5.61676937	5.58315800
ENSG00000001084	4.97682185	5.13955135	5.34269696	4.92979100	5.20437551	5.03826058
ENSG00000001167	4.14323013	4.23342794	4.43362717	4.32696871	4.30597052	3.88166462
ENSG00000001460	2.61353165	3.08406426	2.40871186	2.75915583	2.66220550	2.85199884
ENSG00000001461	4.72410452	4.69097574	4.57773093	4.43629512	4.53853816	4.24868663
ENSG00000001497	5.90665013	5.97108366	5.87946072	5.94836723	5.85574059	5.73199779
ENSG00000001561	2.04614178	2.23878686	2.35049725	2.25096157	2.58255600	2.09423607
ENSG00000001617	2.21101219	2.27202319	2.56803210	3.00539999	2.89141919	2.91264986
ENSG00000001626	5.58796499	5.68116812	5.72246602	5.50779464	5.49441561	4.99954909
ENSG00000001629	5.12142991	4.71644224	5.16551002	4.65706830	4.94532678	4.21256934
ENSG00000001630	5.59305492	5.67044381	5.66902677	5.24830712	5.28835856	5.15015346
ENSG00000001631	4.78502736	4.60466442	4.56681515	4.29204549	4.58135125	4.26753580

# Visualization of raw read count, FPKM, and TPM





# Take-home message

1. Know your RNAseq library clearly (SE or PE, Stranded or Non-stranded).
2. Three gene expression quantification measurements (raw-read-count, RPKM (FPKM), TPM).

# Resources

# Resources

1. Wang *et al.* RNA-Seq: a revolutionary tool for transcriptomics. (*Nature Reviews Genetics*).
2. Ana *et al.* A survey of best practices for RNA-seq data analysis (*Genome Biology*).
3. Lior Pachter's blog (<https://liorpachter.wordpress.com/>).
4. RNASeq blog (<https://www.rna-seqblog.com/>).

**Thank you**

**Q & A**