RNAseq in a nutshell

MSU Translational Bioinformatics Workshop

Ke Liu
(liuke2@msu.edu)

Outline

I. Introduction of RNAseq (Day 3)

II. From RNA to short reads (Day 3)

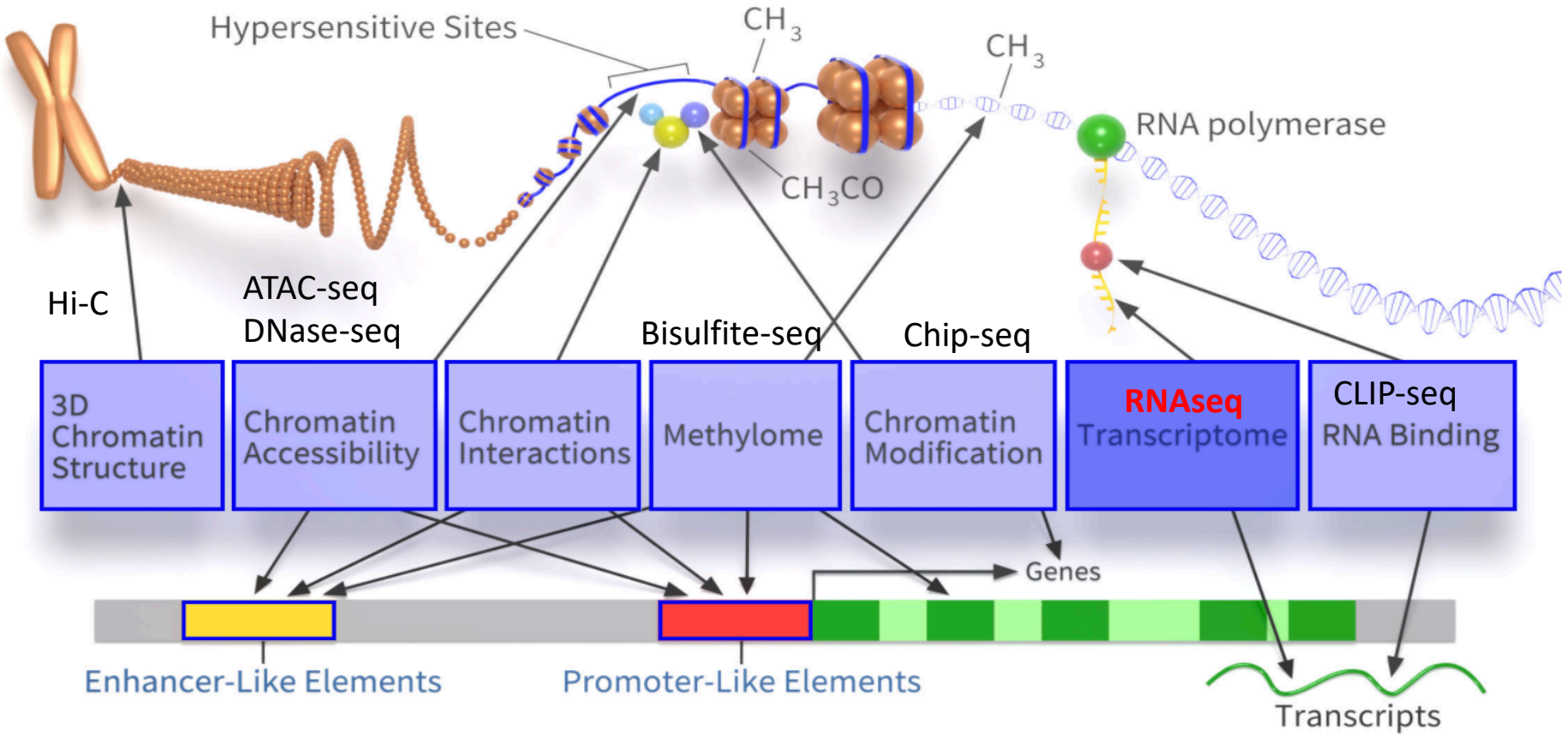
III. RNAseq data analysis (Day 3)

IV. Demo (Day 4)

V. Advanced topics (Day 4)

I. Introduction of RNAseq

Probing cells with sequencing technologies



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

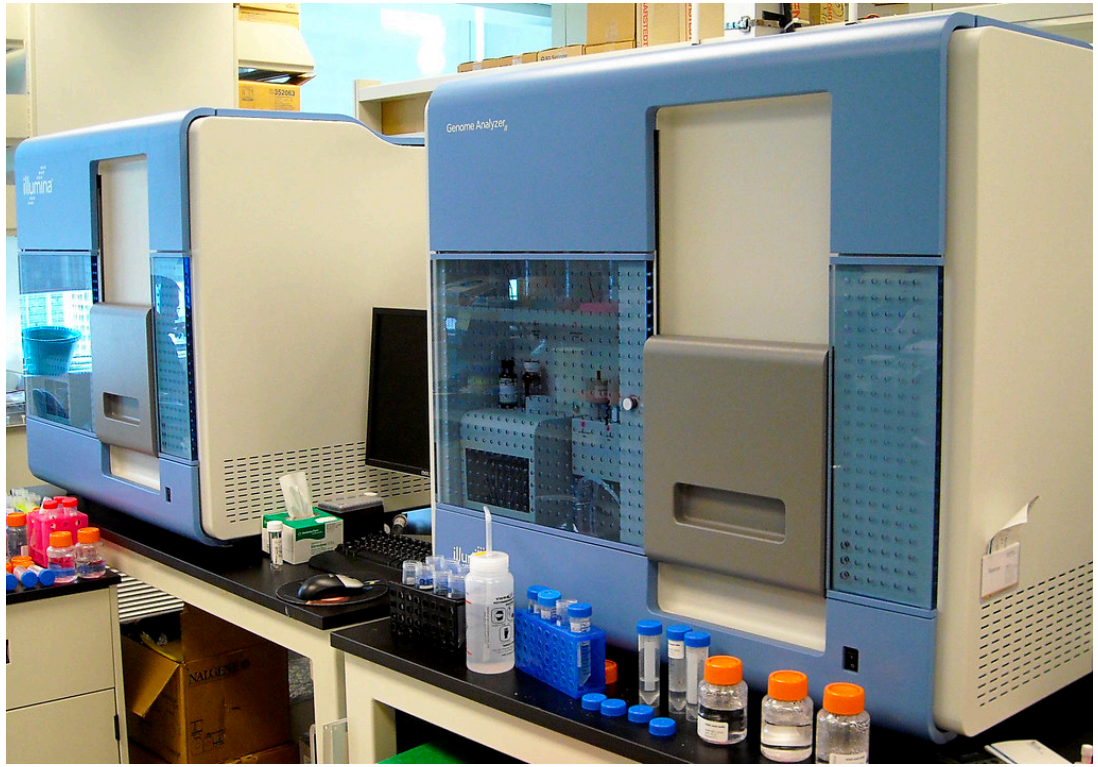
Types of RNA

mRNA
lncRNA (long non-coding RNA)

microRNA (small RNAseq)
rRNA
tRNA
snoRNA
piRNA

Long read vs short read

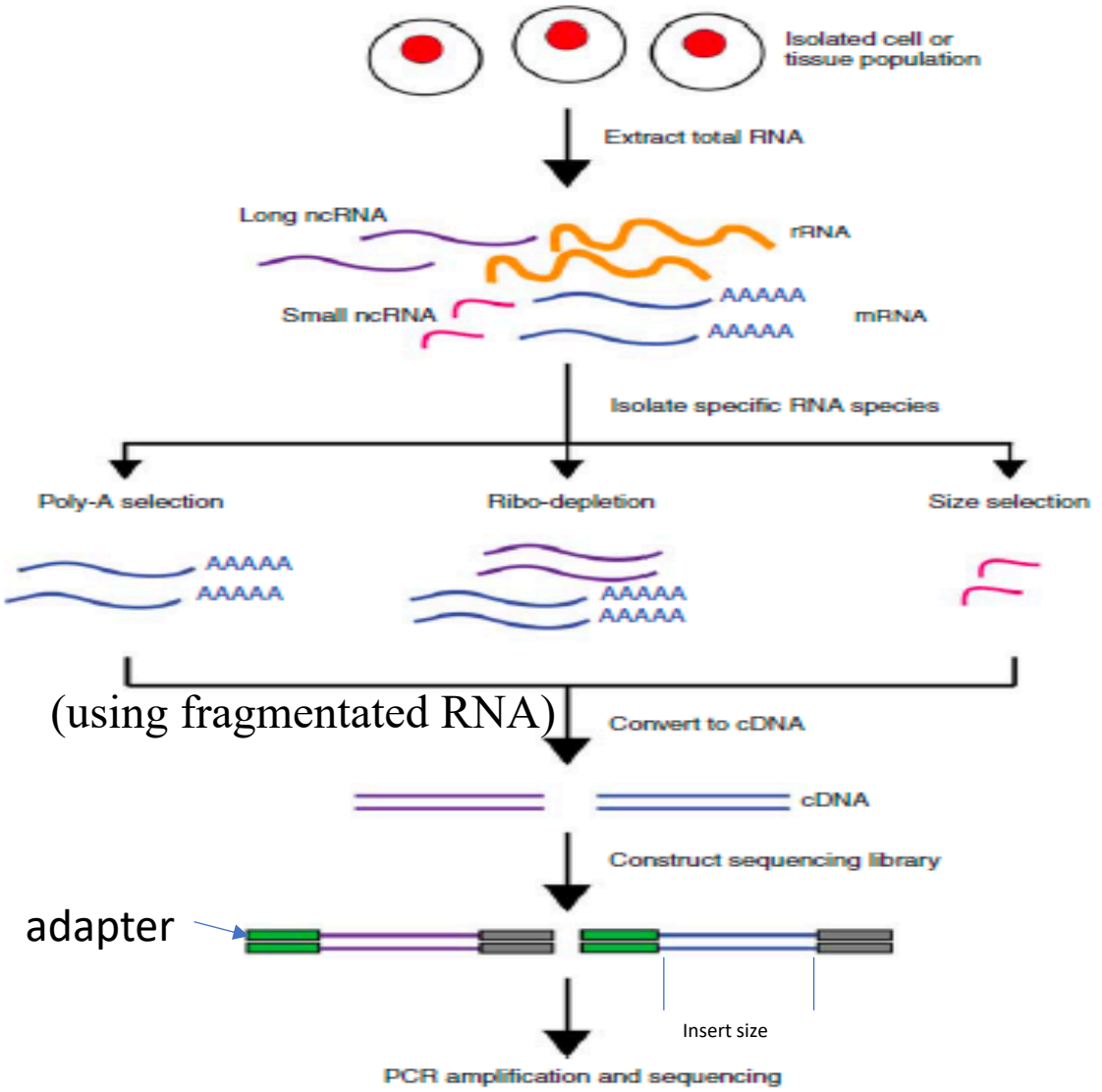
Oxford Nanopore
(long read)



Illumina genome analyzer
(short read)

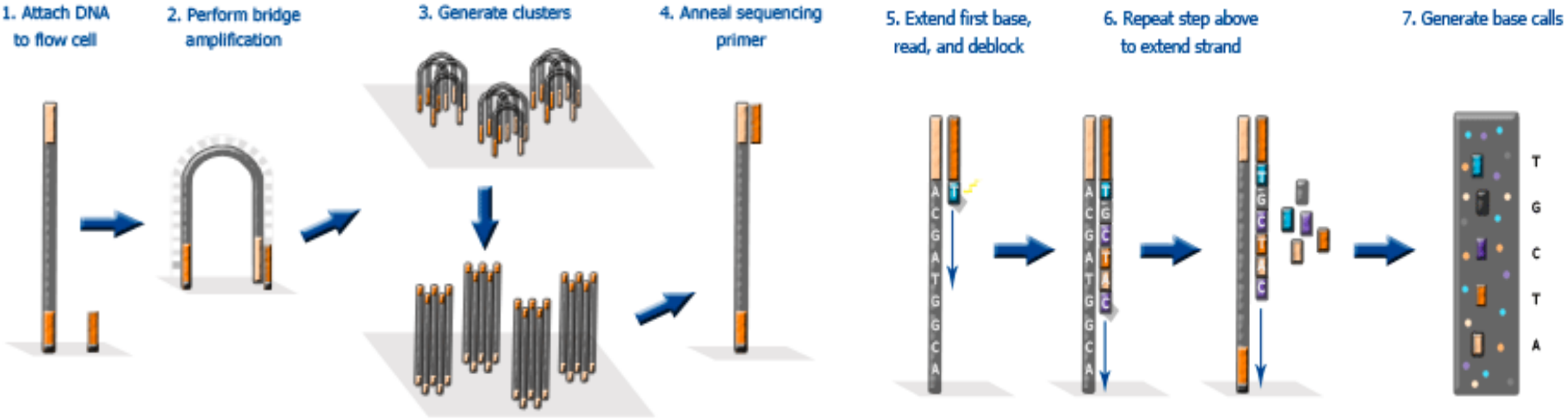
II. From RNA to short reads

RNAseq library construction



https://www.rna-seqblog.com/introduction-to-rna-sequencing-and-analysis/

Overview of illumina sequencing



Sequencing by Synthesis

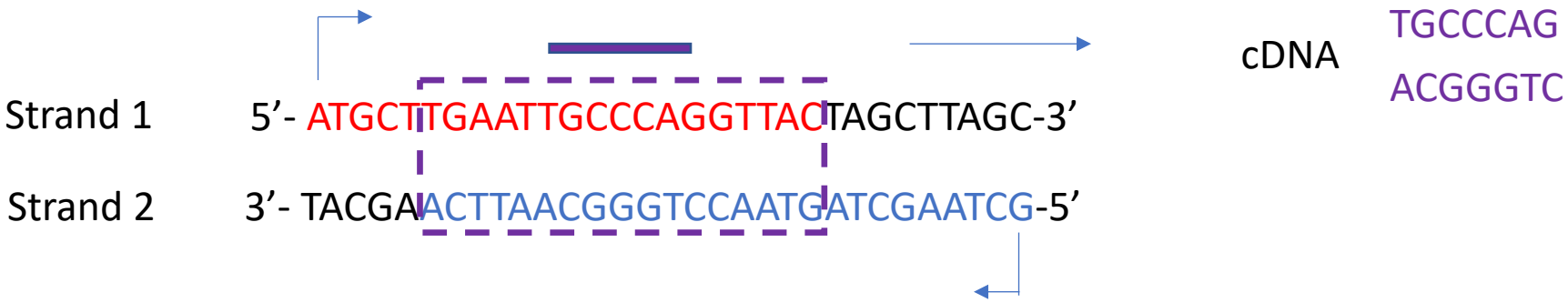
Picture is from https://www.eurofinsgenomics.co.in/en/eurofins-genomics/product-faqs/next-generation-sequencing/general-technical-questions/what-is-the-principal-of-the-illumina-sequencing-technology.aspx

https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf (Documentation from illumina)

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

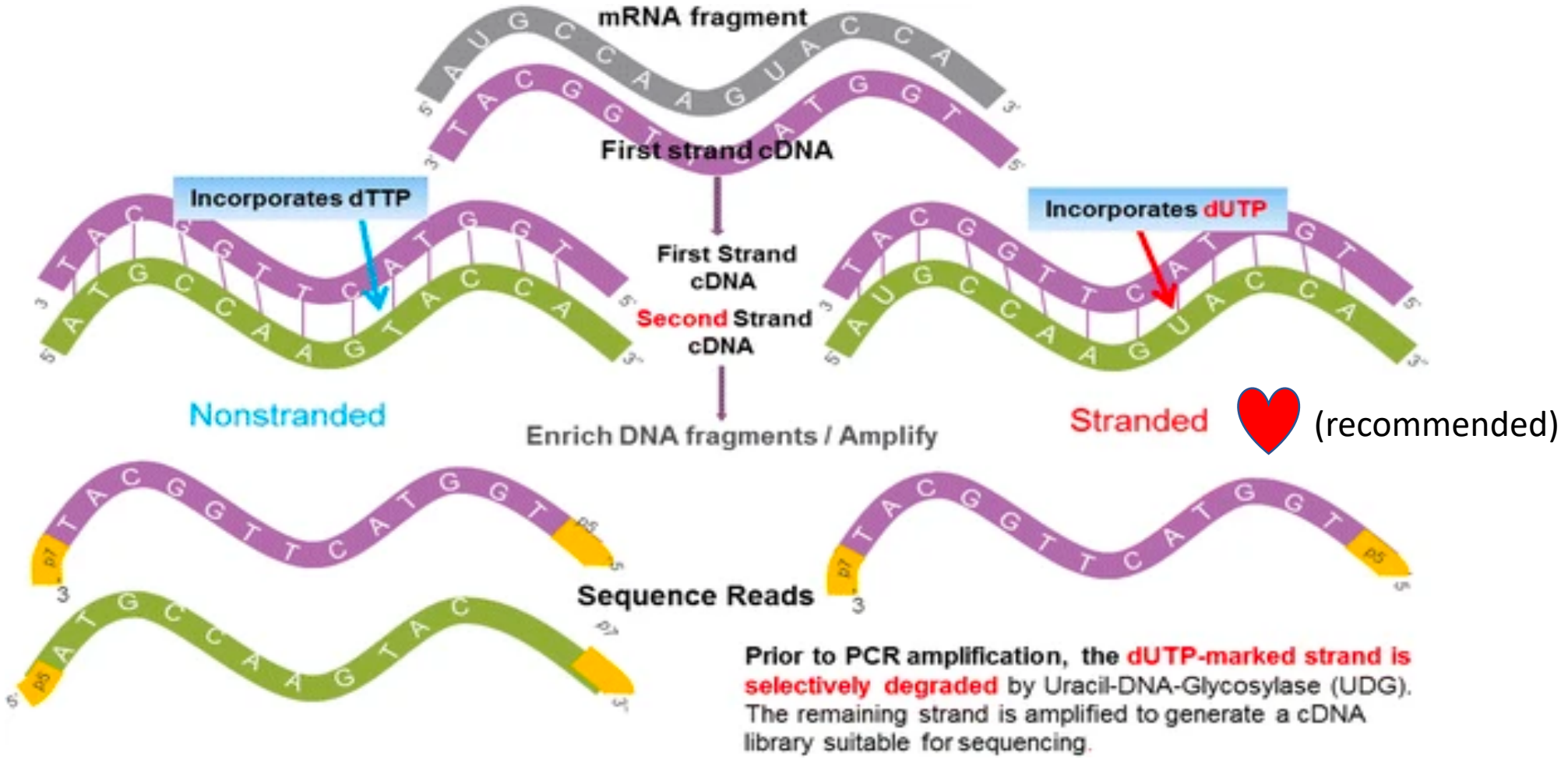
Strandness matters

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



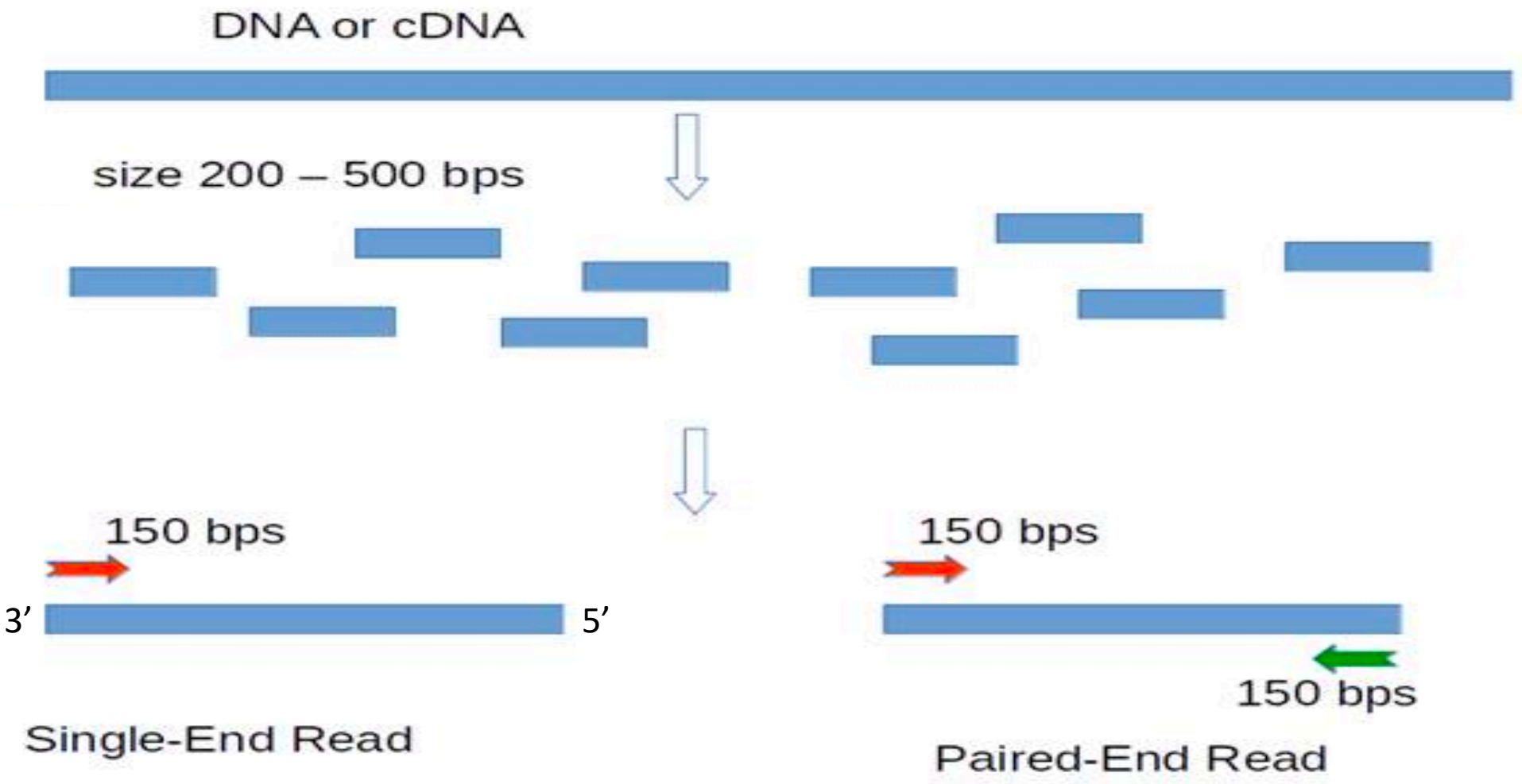
If an RNA fragment comes from the dashed region, it is hard to know its original gene without strand information.

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
NAAAGCACATACCAAGGCCACCACACACCACCTGTCCAAAAAGGCCTTCGATACGGGATAATCCTATTTATTACCTCAGAAGTTTTTTTCTTCGCAGGAT
+SRR4822549.1 UNC13-SN749_0135:5:1101:1389:1958 length=100
#1=B7BDFFFHFDGGEBHGIIGEGCHGIIIIGAG?DFFIGGIIIHG3BGHG@@EDHHEHFFBDFFFEEEDCEDCC;>@CCCD5:AC?B@DDDDDDD@B@C
@SRR4822549.2 UNC13-SN749_0135:5:1101:1498:1960 length=100
NTTCTTCAATTTCTTGCCTTCTTCCTTGGAGGCTGGAAGAATCATGGCAAGGTAGGGCCCATCAACCTCAAAAAAGATGCTGTTCTCTGAGCGGGTGACG
+SRR4822549.2 UNC13-SN749_0135:5:1101:1498:1960 length=100
#1=DFFFFHHHHHJJIJJJJJJJJJJJJJHIIJJJJIIJJJIJIIIJJJJJJJBGIJJJJJIJJJJJJHHHHHFFDDDDDDDDDCDEDDDDDDDDD<BBDB
```

Preparation of RNAseq experiments

1. Sequencing depth (library size): deeper is better! Usually > 5 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. Read length: longer is better.

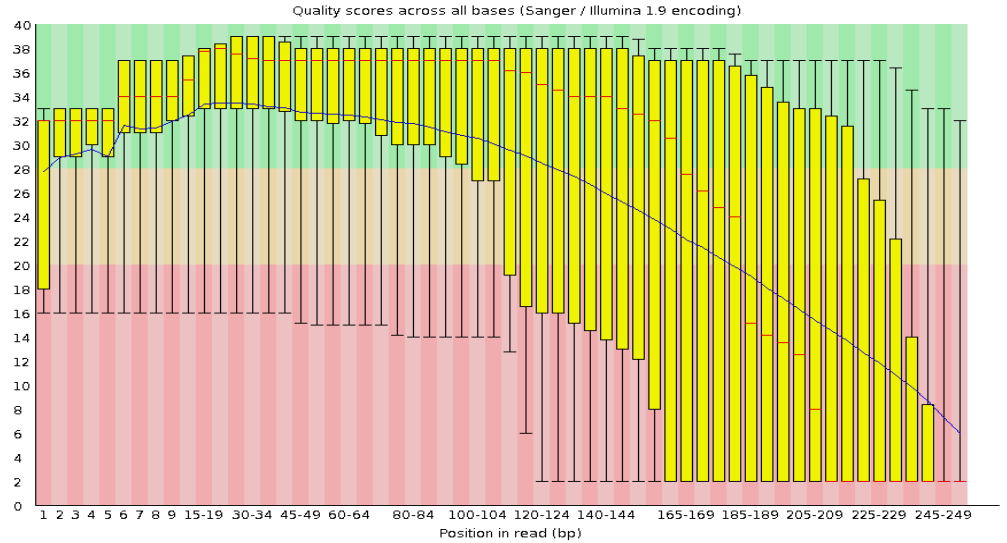
III. RNAseq data analysis

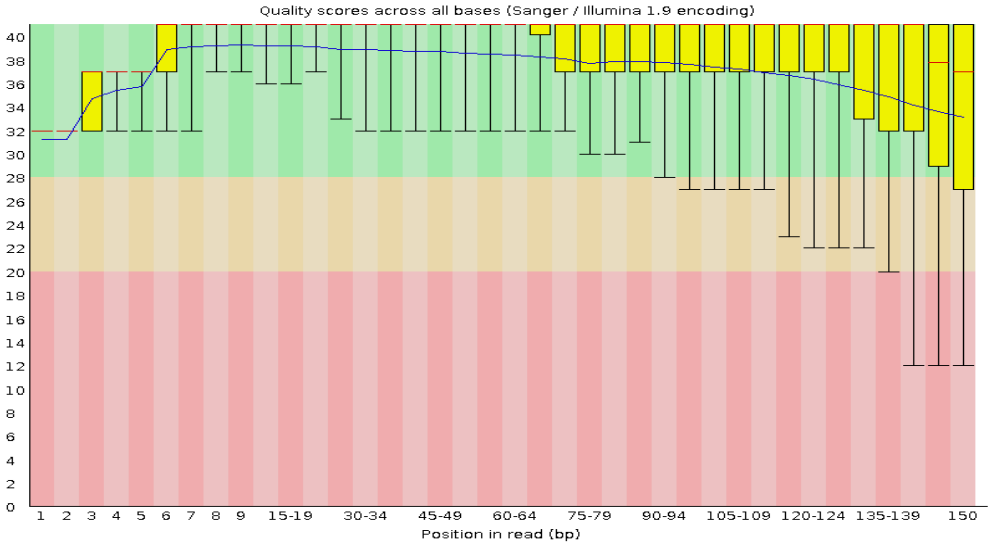
Core analysis

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

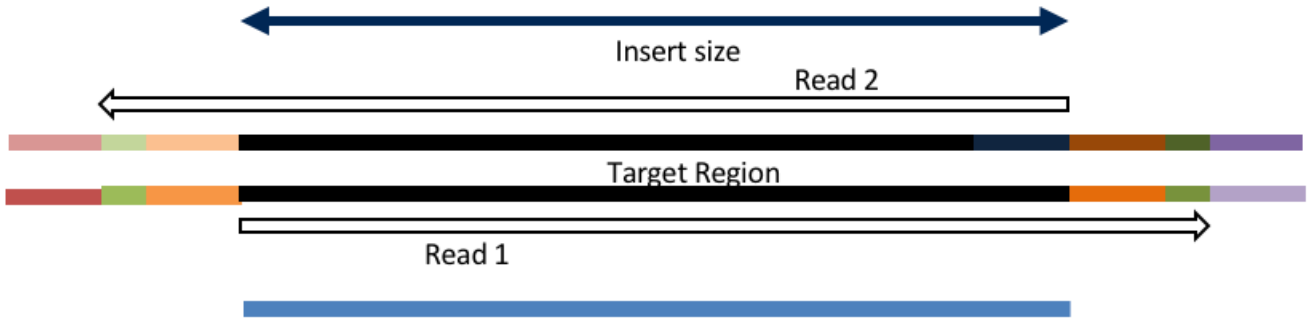
Data quality checking

RNAseq reads quality





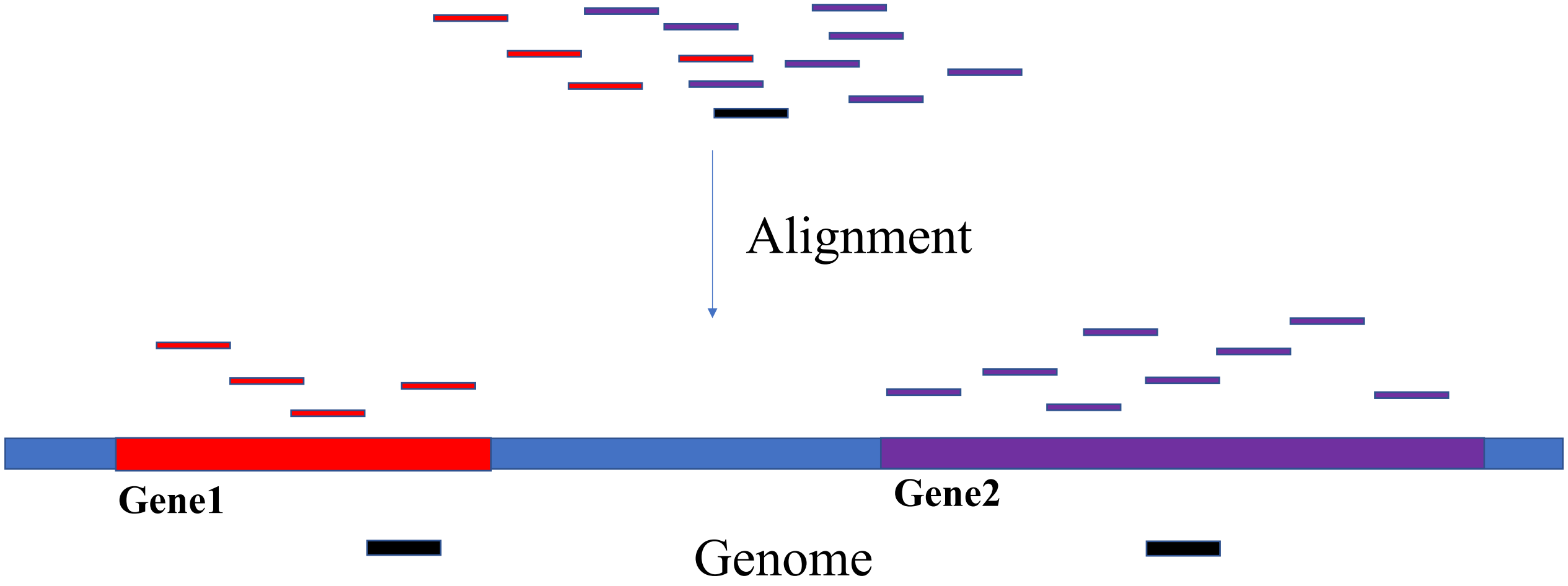
Adapter removal



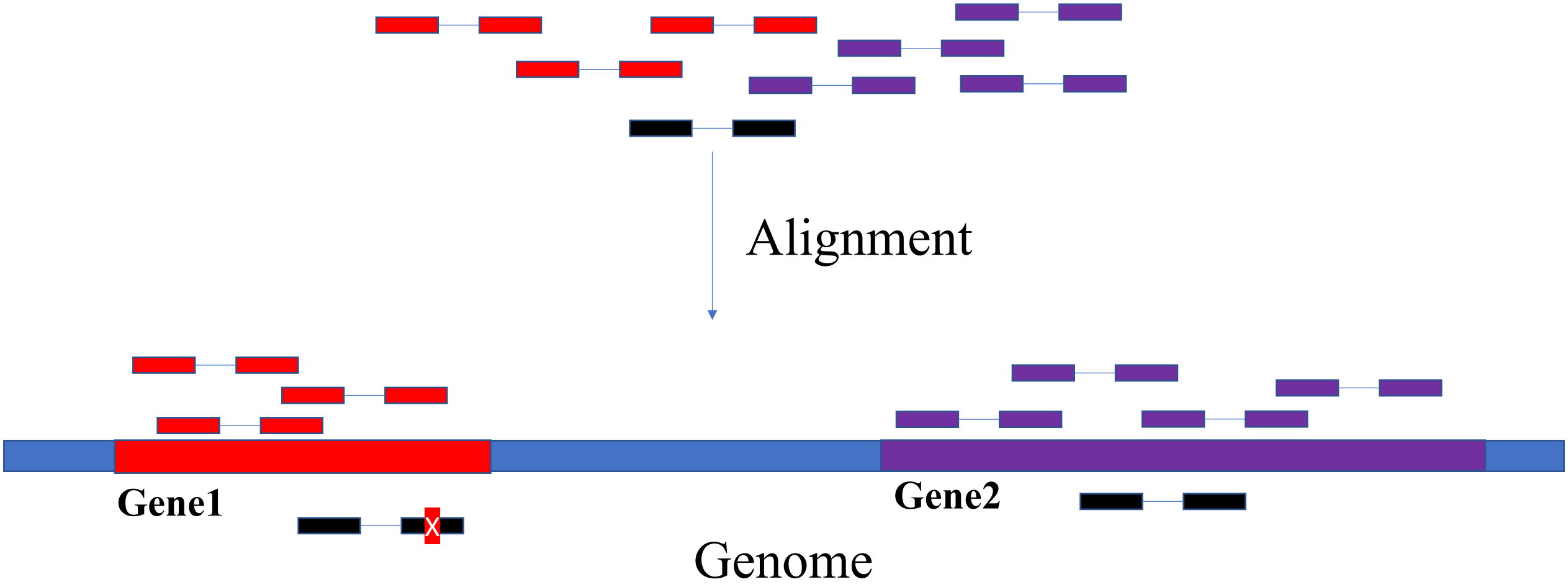
Insert size < read length

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)

Genome

Expression of **Gene1**: 3

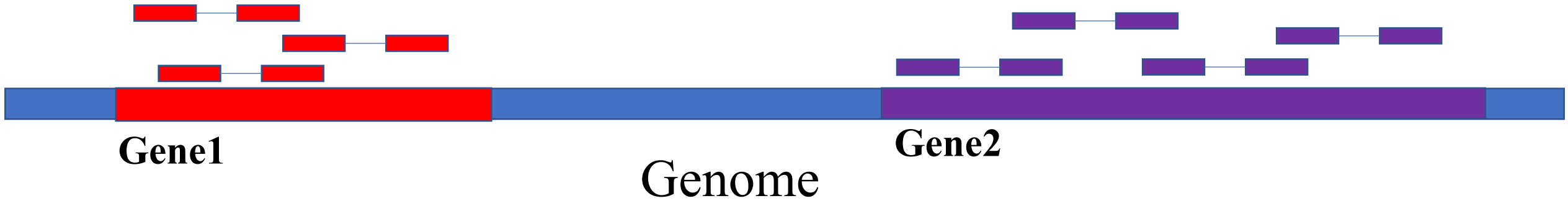
Expression of **Gene2**: 4

Problem of raw read count

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

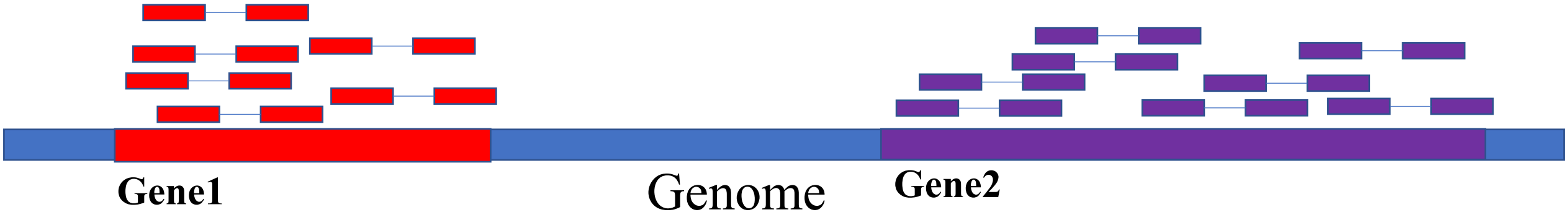
Raw-read-count is proportional to library.size * gene.length,
normalization is needed!

Rep1 (library size = 7)



Genome

Rep2 (library size = 14)



Genome

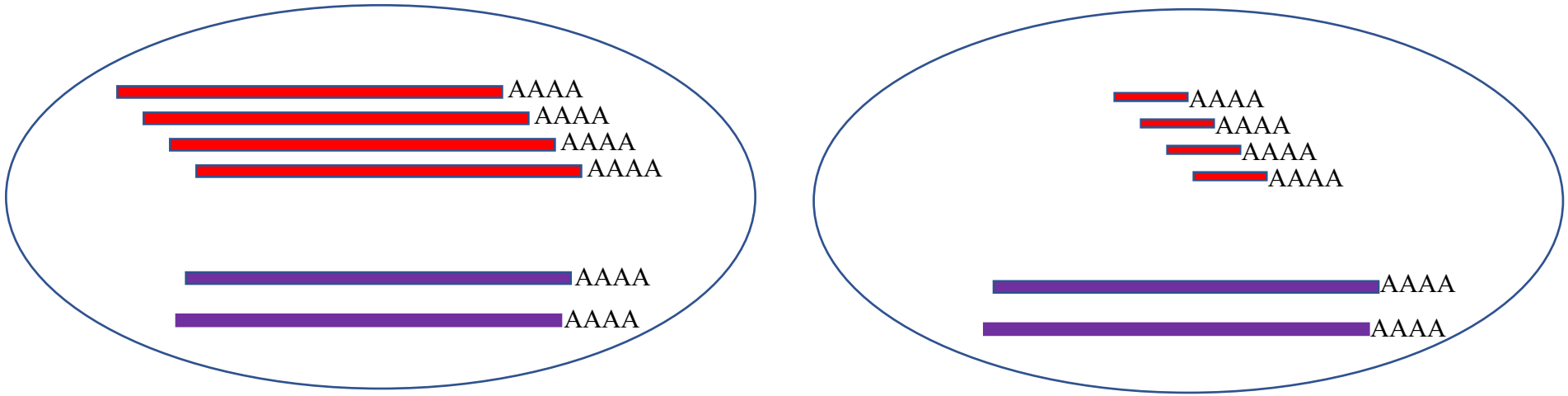
Quantification of gene expression (RPKM FPKM)

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

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

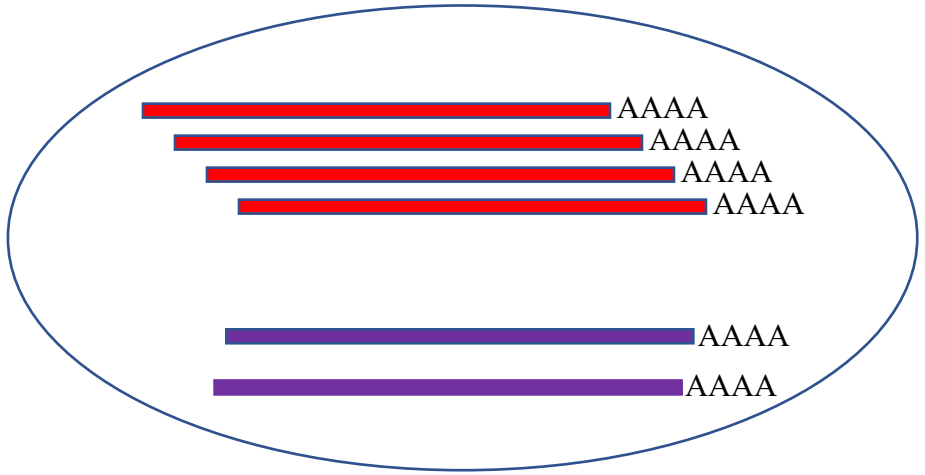
Problem of RPKM (FPKM)



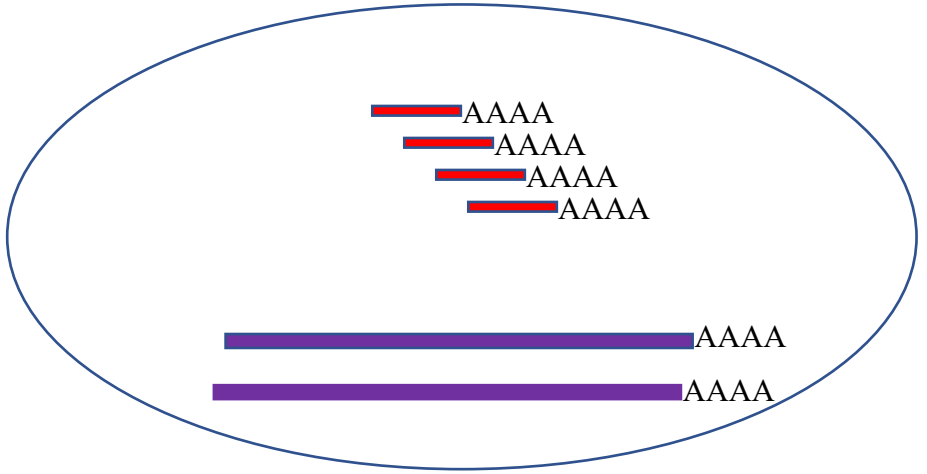
Gene1 $E_1 = (4 / 6)$

Gene2 $E_2 = (2 / 6)$

Isoform matters



Exp1



Exp2

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

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

RPKM1 = 10^9 * (P1/(P1+R1)) * 1 / Lp
RPKM2 = 10^9 * (P2/(P2+R2)) * 1 / Lp

Quantification of gene expression (TPM)

TPM: Transcripts Per Million.

Given a Gene Gi, 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

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

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

Relationship between RPKM (FPKM) and TPM

$$TPM_i = RPKM_i / Sum\ (RPKM)$$

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

Differential gene expression analysis

Performing **statistical** analysis on RNAseq data to discover quantitative changes in expression levels between experimental groups. For example, we use statistical testing to decide whether, for a given gene, **an observed difference in gene expression is significant, that is, whether it is greater than what would be expected just due to natural random variation.**

https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/differential-gene

Thinking on DE analysis

1. Which measurement should we use, raw-read-count, RPKM (FPKM) or TPM?

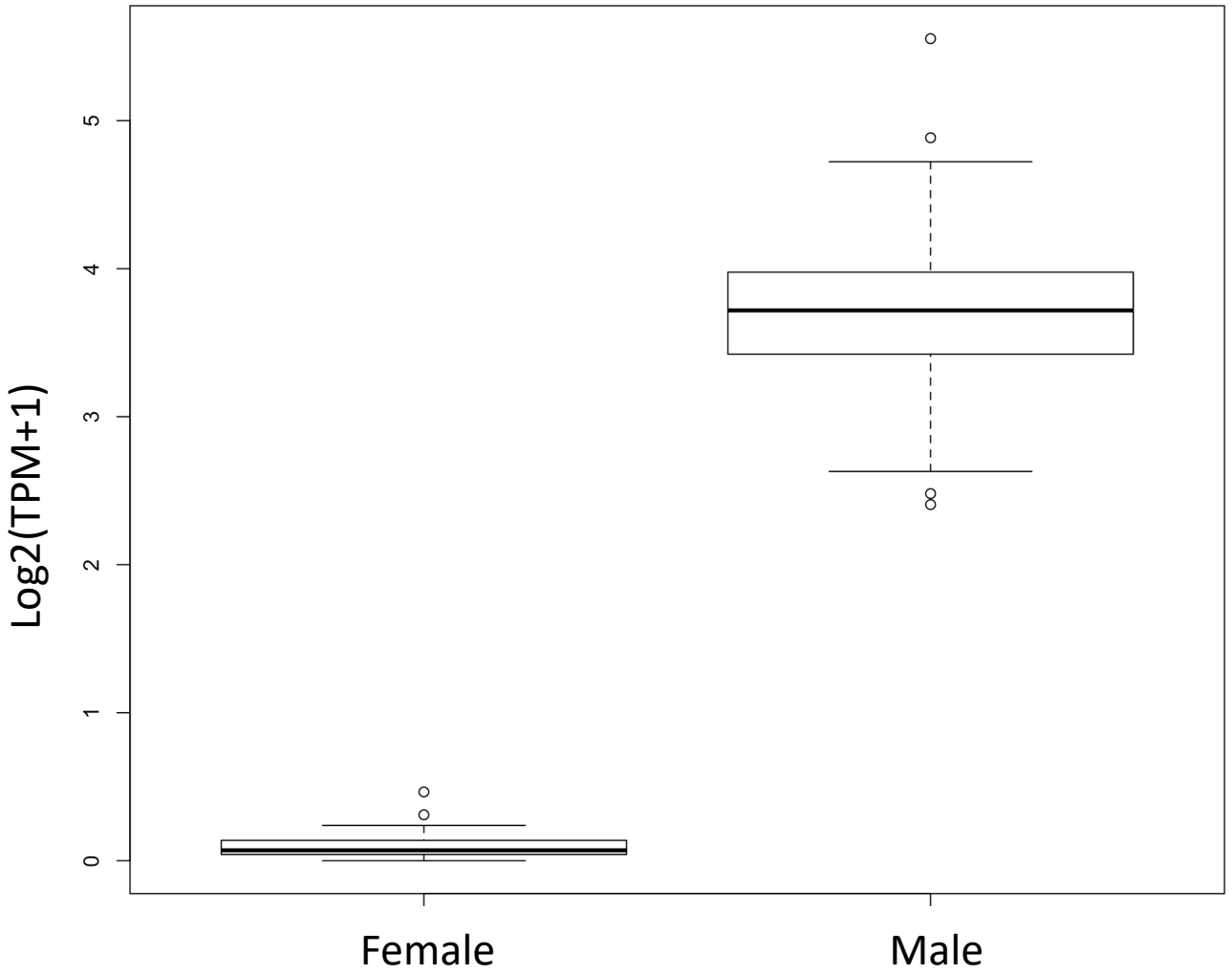
2. What statistical test should we use? (maybe t-test, or nonparametric test?)

3. In the field, people **prefer** to use **raw-read-count**. (WAIT! Raw read count is not comparable between different experiments, why do you use it?)

4. RPKM (FPKM), TPM can also be used in DE analysis!

Example

I want to know whether the expression of KDM5D is significantly higher in male.

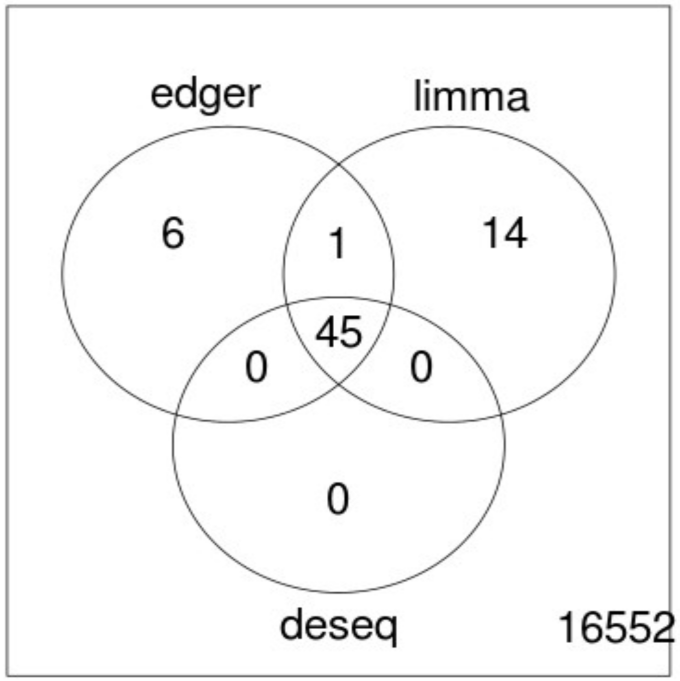


p-value < 2.2e-16
(t-test and wilcoxon rank-sum test)

DE analysis with R

1. Many R packages are available for DE analysis. For example, DESeq2, edgeR, and limma + voom.
   limma + voom is faster.
   DESeq2 may be the most popular one (works well in our study)

2. Output: a list of differentially expressed genes



Overview of DESeq2

Model and normalization

The read count K_{ij} for gene i in sample j is described with a GLM of the negative binomial family with a logarithmic link:

$$\begin{aligned} K_{ij} &\sim \mathrm{NB}\left(\mathrm{mean} = \mu_{ij}, \mathrm{dispersion} = \alpha_i\right) \\ \mu_{ij} &= s_{ij} q_{ij} \end{aligned} \tag{1}$$

$$\log q_{ij} = \sum_r x_{jr} \beta_{ir}. \tag{2}$$

For notational simplicity, the equations here use the natural logarithm as the link function, though the *DESeq2* software reports estimated model coefficients and their estimated standard errors on the log2 scale.

By default, the normalization constants s_{ij} are considered constant within a sample, $s_{ij} = s_j$, and are estimated with the median-of-ratios method previously described and used in *DESeq* [4] and *DEXSeq* [30]:

$$s_j = \underset{i:\ K_i^{\mathrm{R}} \neq 0}{\mathrm{median}} \frac{K_{ij}}{K_i^{\mathrm{R}}} \quad \text{with} \quad K_i^{\mathrm{R}} = \left(\prod_{j=1}^{m} K_{ij} \right)^{1/m}.$$

Michael *et al*. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 (Genome Biology, 2014).

Interpretation of DE analysis results

Now we have a list of differentially expressed genes, what is next? (time to plug in some domain knowledge!)

Enrichment analysis could help us tell whether DE genes tend to share some common biological characteristics. For example, do the DE genes tend to be on a specific signaling pathway (e.g, Wnt pathway, TGF-beta pathway)?

Hypergeometric distribution

Suppose there is a box, inside it there are K while balls and N – K black balls. Now if we **randomly** draw n balls without replacement, then what is the probability we get k **white** balls?

Use X to denote the number of white balls we get, then

$$f(k; N, K, n) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

How to utilize this distribution to perform enrichment analysis? Just consider the genes as balls!

Pathway enrichment analysis

White balls: all the genes that are on the pathway you are interested in (e.g, TGF-beta signaling)

Black balls: all the genes that are NOT on the pathway you are interested in (e.g, TGF-beta signaling)

$$\text{P-value} = 1 - \text{sum}\left(f(I; N, K, n) \right) \quad k <= I <= n$$

K : number of pathway genes
N : number of all the genes
n: number of DE genes
k: number of DE genes which are also on the pathway

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).

3. Using raw read count to perform differential gene expression analysis.

4. Enrichment analysis.

Q & A

Demo

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).

3. Using raw read count to perform differential gene expression analysis.

4. Enrichment analysis.

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

RNAseq reads
(FASTQ)

↓

Align to genome
(Tophat, STAR, HISAT2)

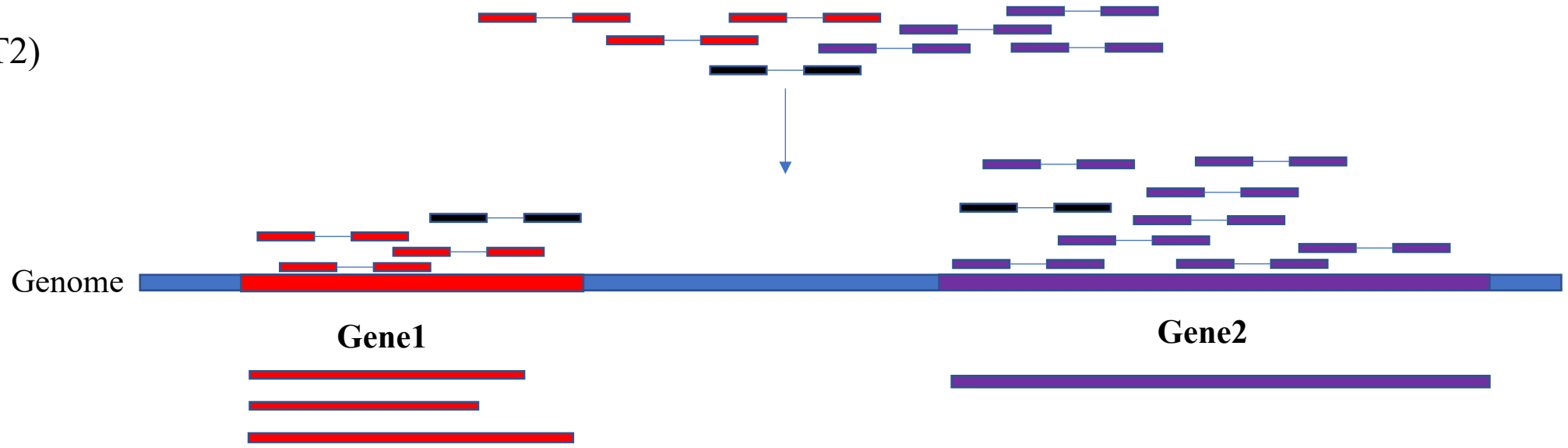
↓

BAM files

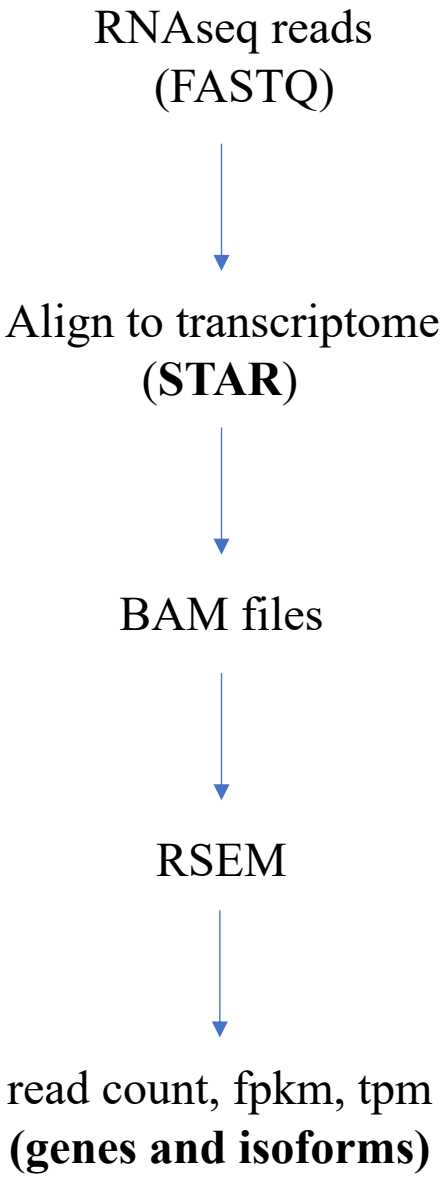
↓

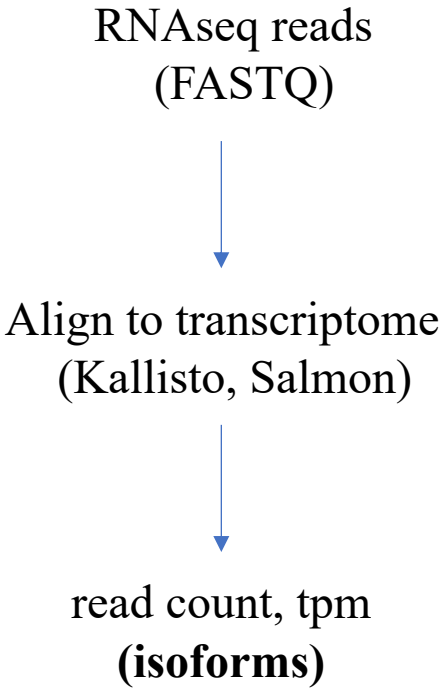
read count matrix
(featureCounts, HTSeq)

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.

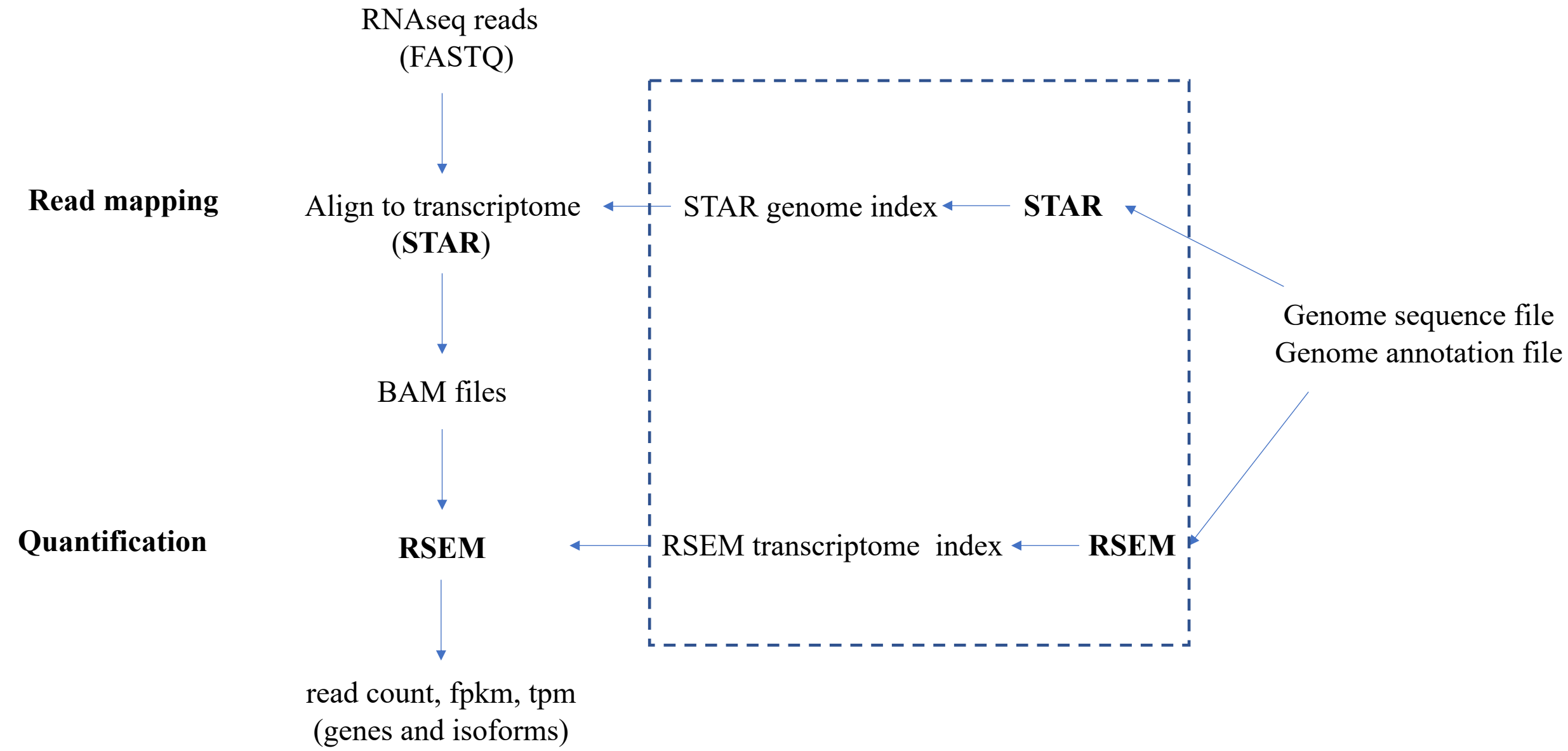


Quantification of gene expression





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**

Genome annotation file (from GENCODE):
ftp://ftp.ebi.ac.uk/pub/databases/gencode/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
```

DE analysis (male.vs.female, human liver tissue)

1. Install required packages

2. Prepare read count matrix and meta data

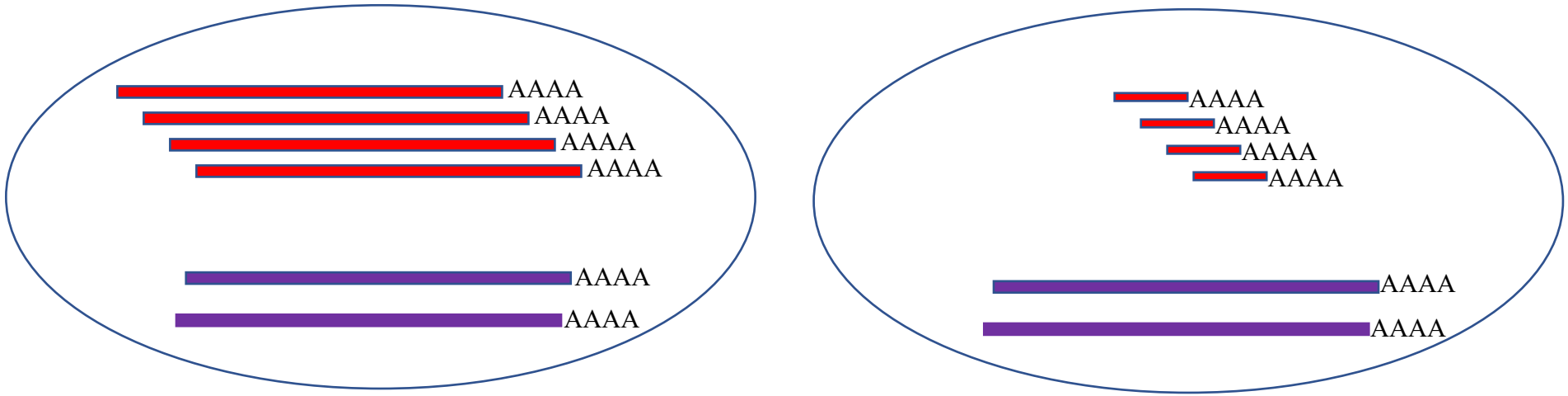
3. Run DE analysis

4. Extract DE genes

5. Using enrichR server to run enrichment analysis

Advanced topics

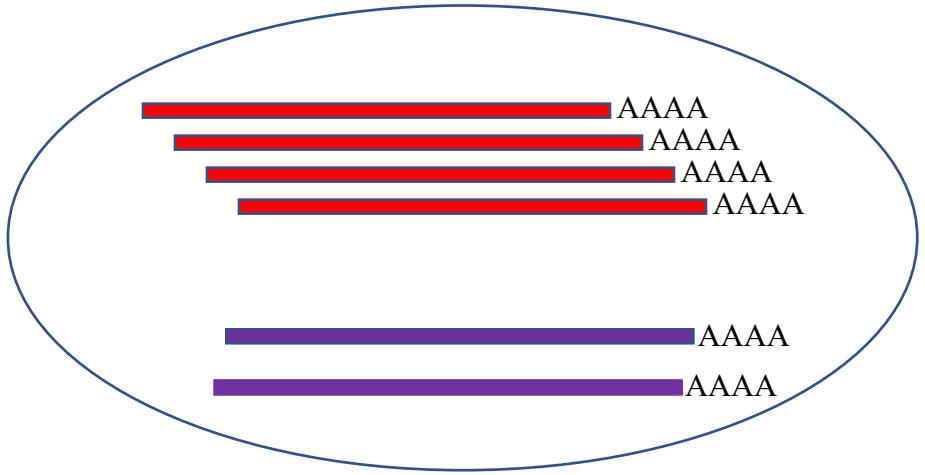
Problem of RPKM (FPKM)



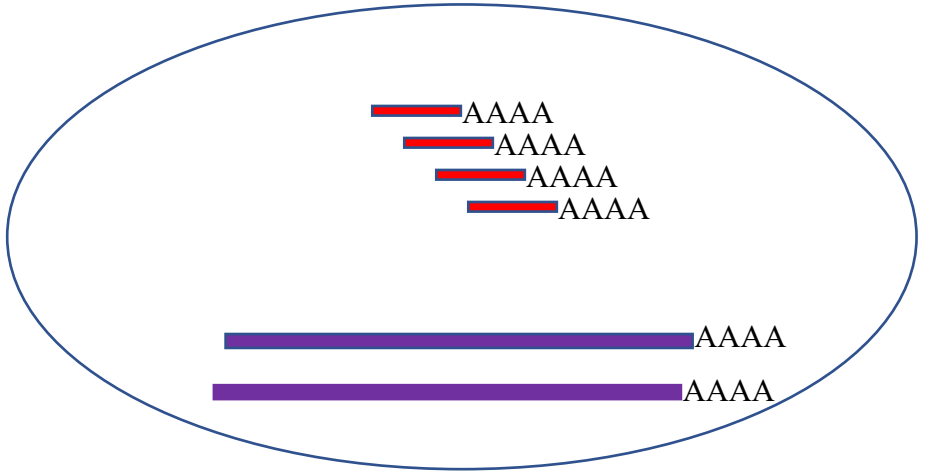
Gene1 $E_1 = (4 / 6)$

Gene2 $E_2 = (2 / 6)$

Isoform matters



Exp1



Exp2

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

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

RPKM1 = 10^9 * (P1/(P1+R1)) * 1 / Lp
RPKM2 = 10^9 * (P2/(P2+R2)) * 1 / Lp

GSEA analysis

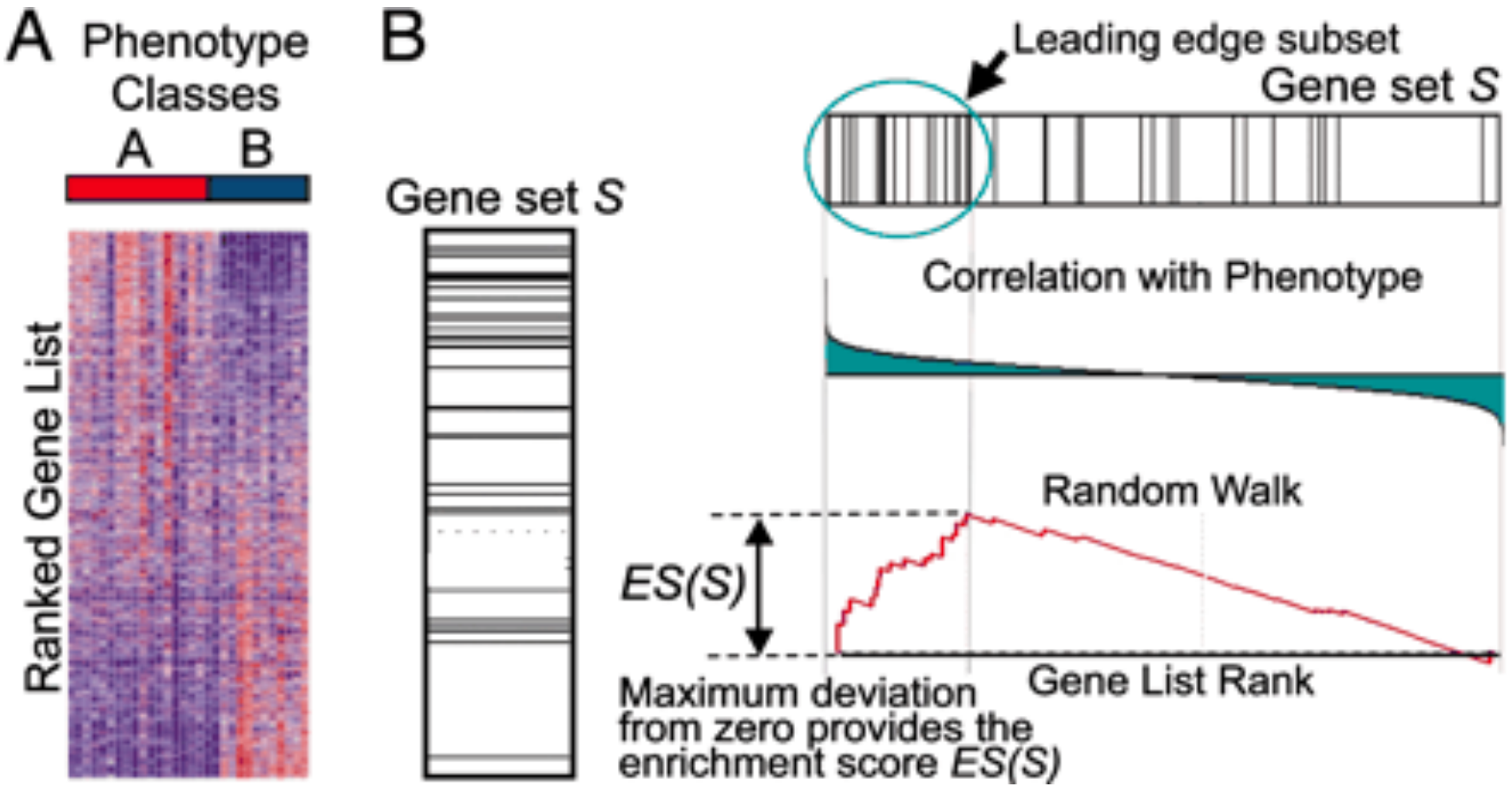
GSEA: gene set **enrichment** analysis.

We already have hypergeometric-distribution (HD) based enrichment analysis method, why GSEA?

1. HD-based method does NOT fully utilize the information derived from DE analysis.

2. The activity of biological processes may be perturbed at different level.

GSEA analysis



Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. (PNAS, 2005)

Batch effects

In RNAseq experiments, batch effect occurs when non-biological factors in an experiment cause changes in the data produced by the experiment (from Wiki).

RESEARCH ARTICLE

Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin, Yiing Lin, Joseph R. Nery, Mark A. Urich, Alessandra Breschi, Carrie A. Davis, Alexander Dobin, Christopher Zaleski, Michael A. Beer, William C. Chapman, Thomas R. Gingeras, Joseph R. Ecker, and Michael P. Snyder

PNAS December 2, 2014 111 (48) 17224-17229; first published November 20, 2014
https://doi.org/10.1073/pnas.1413624111

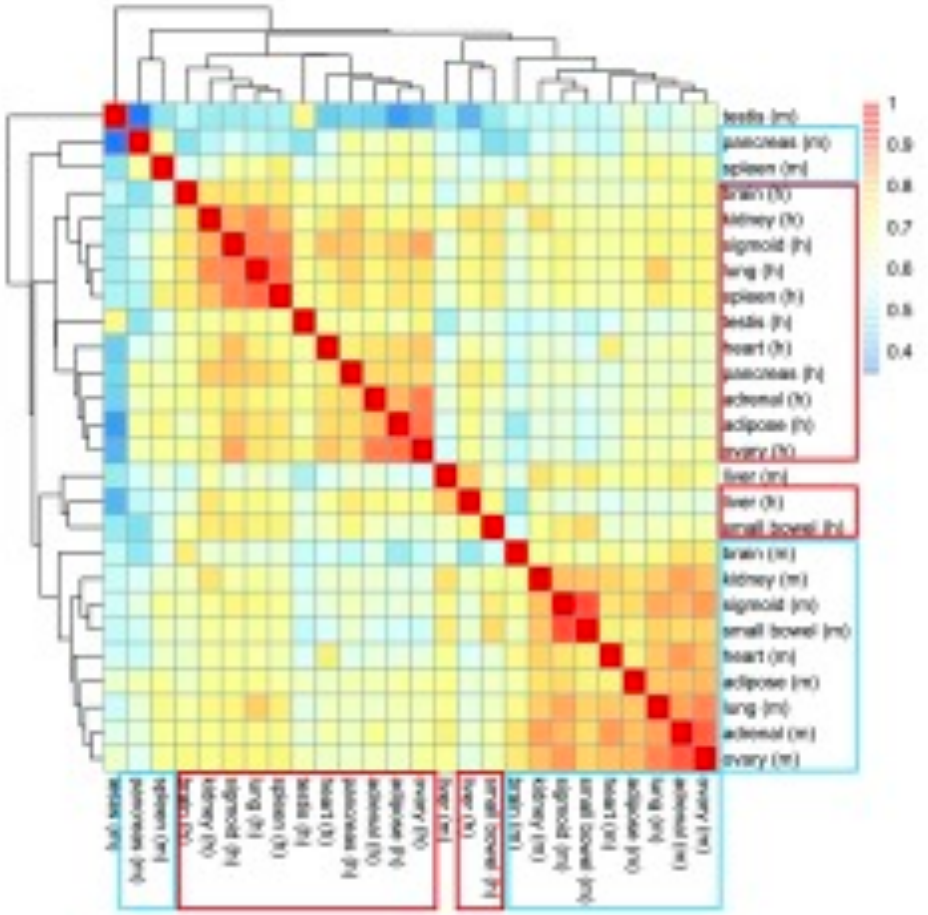
Contributed by Joseph R. Ecker, July 23, 2014 (sent for review May 23, 2014)

Version 1. F1000Res. 2015; 4: 121.
Published online 2015 May 19. doi: 10.12688/f1000research.6536.1
PMCID: PMC4516019
PMID: 26236466

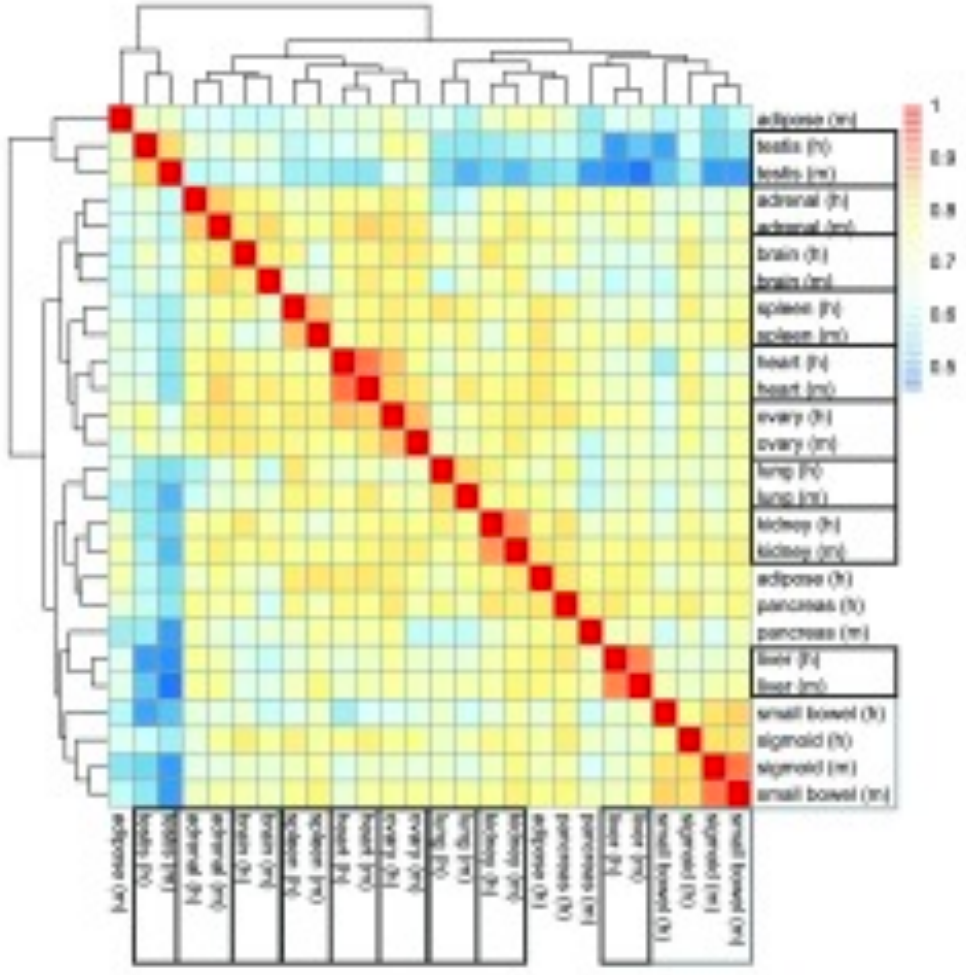
A reanalysis of mouse ENCODE comparative gene expression data

Yoav Gilad[a,1] and Orna Mizrahi-Man[1]

Batch effects



Before correction



After COMABT correction

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/).

5. DESeq2 tutorial:
http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

6. https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html (clusterProfiler)

Q & A