June, 2022

RNAseq analysis without coding

BIOVIT, CIGENE
Marie Saitou

Goal of today's class

- Learn how to analyze RNA-seq sequence data on

We will learn

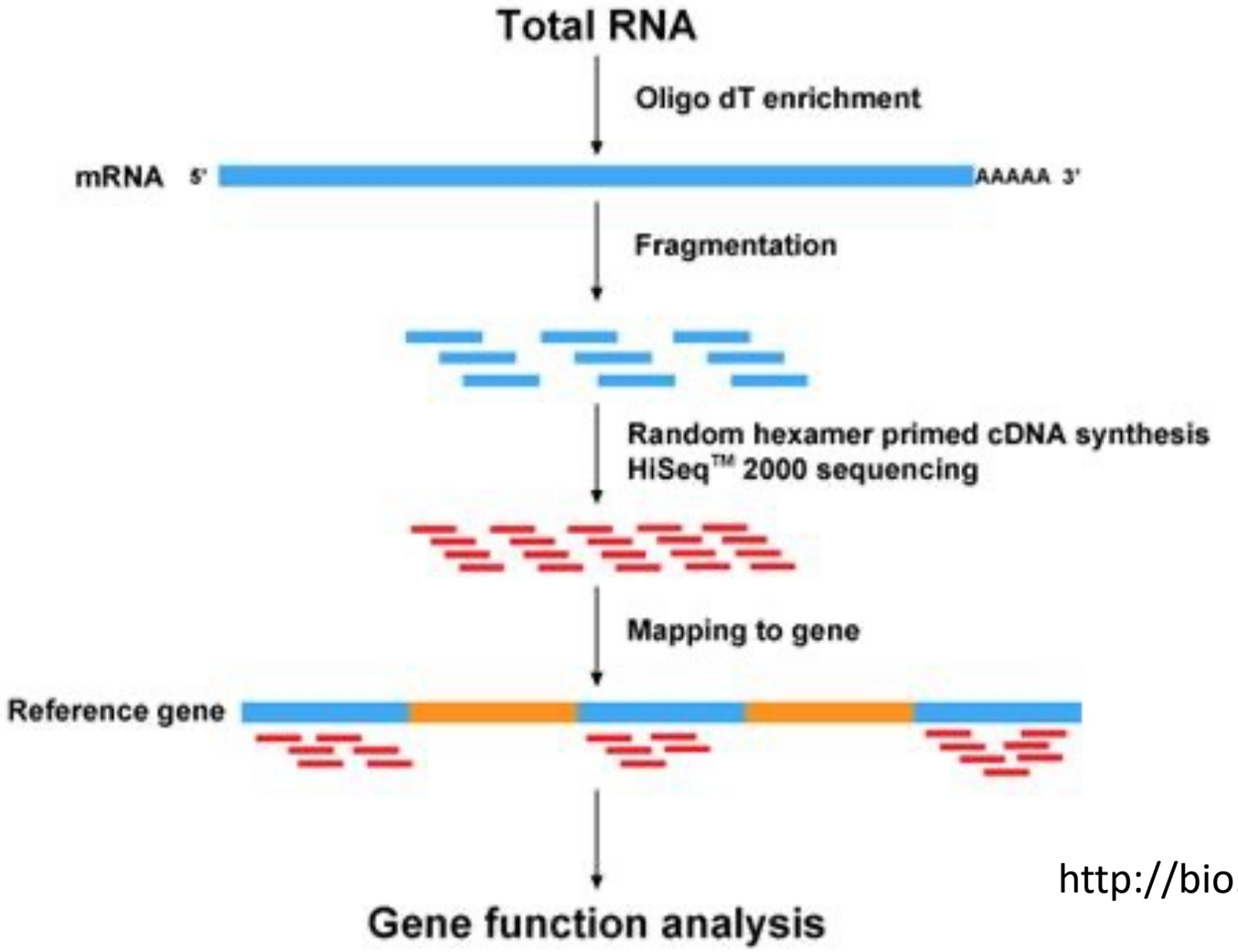
- How to produce a gene expression count matrix from the row sequence reads
- How to analyze differentially expressed genes between tissues

Today’s schedule:

- [**Lecture 10:00-10:45**] Intro to the RNA-seq analysis (theory/lab)
  Intro to Public sequence repository
- [**Hands-on 10:45-12:00**] Do the analysis by yourself, Q and A

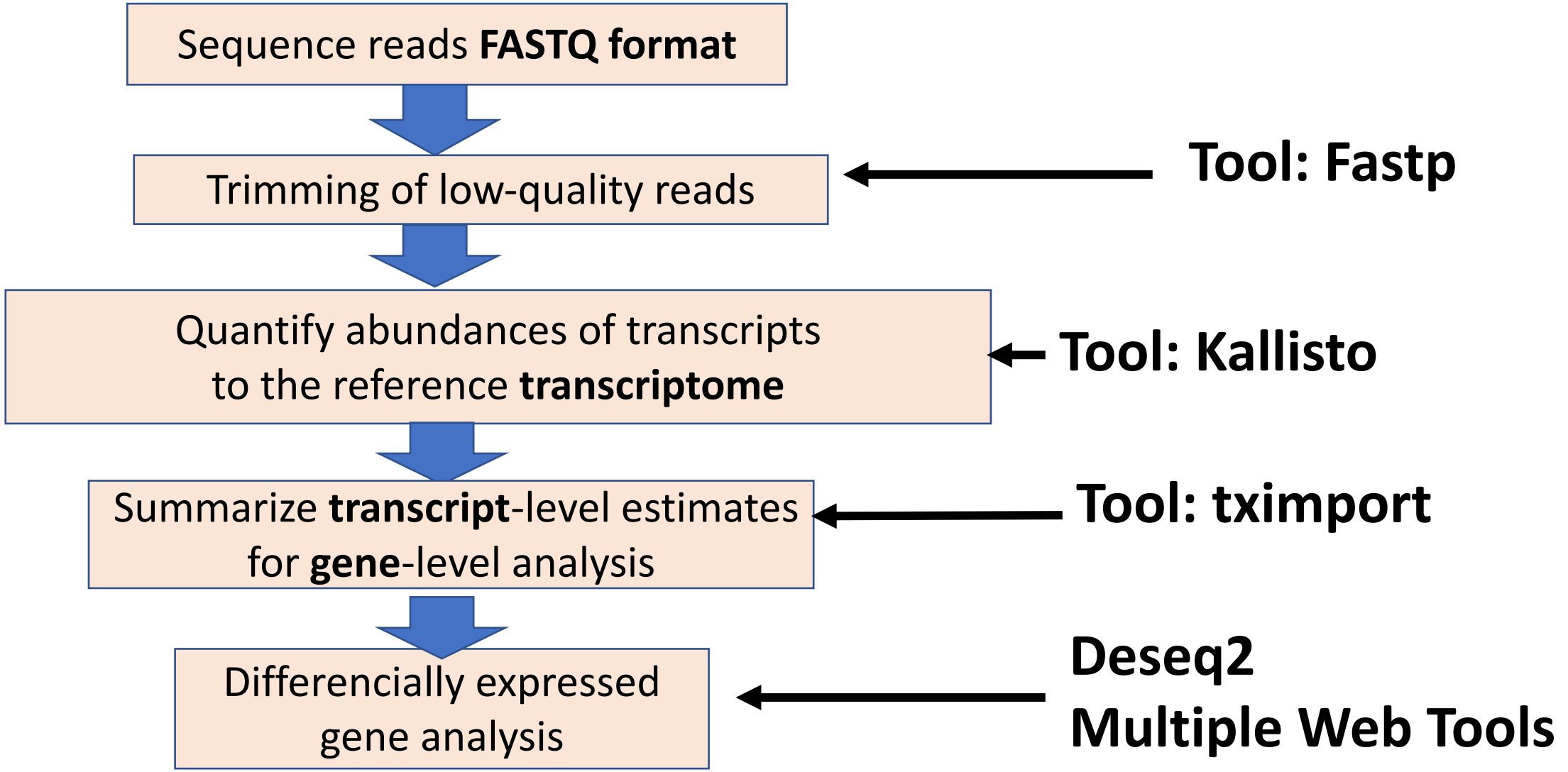
RNA-sequencing

Examine the gene expression pattern in a genome-wide manner



http://bio.lundberg.gu.se/courses/vt13/rnaseq.html

RNA sequencing workflow



With RNAseq, we can investigate

Gene expression change between:

- Conditions
  (different diet, treatment, infected vs healthy...)
- Developmental stages/cell replicative age
- Tissues, organs

When you submit a paper with RNAseq Analysis

You should:

- Submit the raw sequences to a public repository (such as ENA)
- Describe

  Sample preparation protocol/RNAseq experimental design/Software versions and parameters used

Example (Saitou et al., 2020, Cell Reports)
RNA isolation and sequencing
Human adult and fetal tissue samples were mechanically homogenized using a hand-held homogenizer (Thermo Fisher Scientific) and lysed in 500 µl RNA lysis buffer (Ambion) by sonication (1 × 2-4 s pulse, Branson SFX150). RNA was isolated from 3 × 30 µm sections of human adult and fetal tissue using the RNAqueous Micro Kit (Ambion), and total RNA samples were DNase-treated (Ambion). Sample yield and integrity was analyzed using a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). RNA sequencing was performed by standard operating procedure by GENEWIZ (https://www.genewiz.com/en) using Illumina HiSeq with a 2 × 150 bp configuration. Quality control of the obtained sequences was performed using FastQC (Wingett and Andrews, 2018). Adaptor sequences, low-quality bases from both sides of the read (3 bases or smaller), and reads with a length smaller than 36 bp were discarded by Trimmomatic (Bolger et al., 2014). [...]

Key Concepts

- Transcripts Per Million (TPM)
- p-value vs fold change
- Gene Ontology analysis

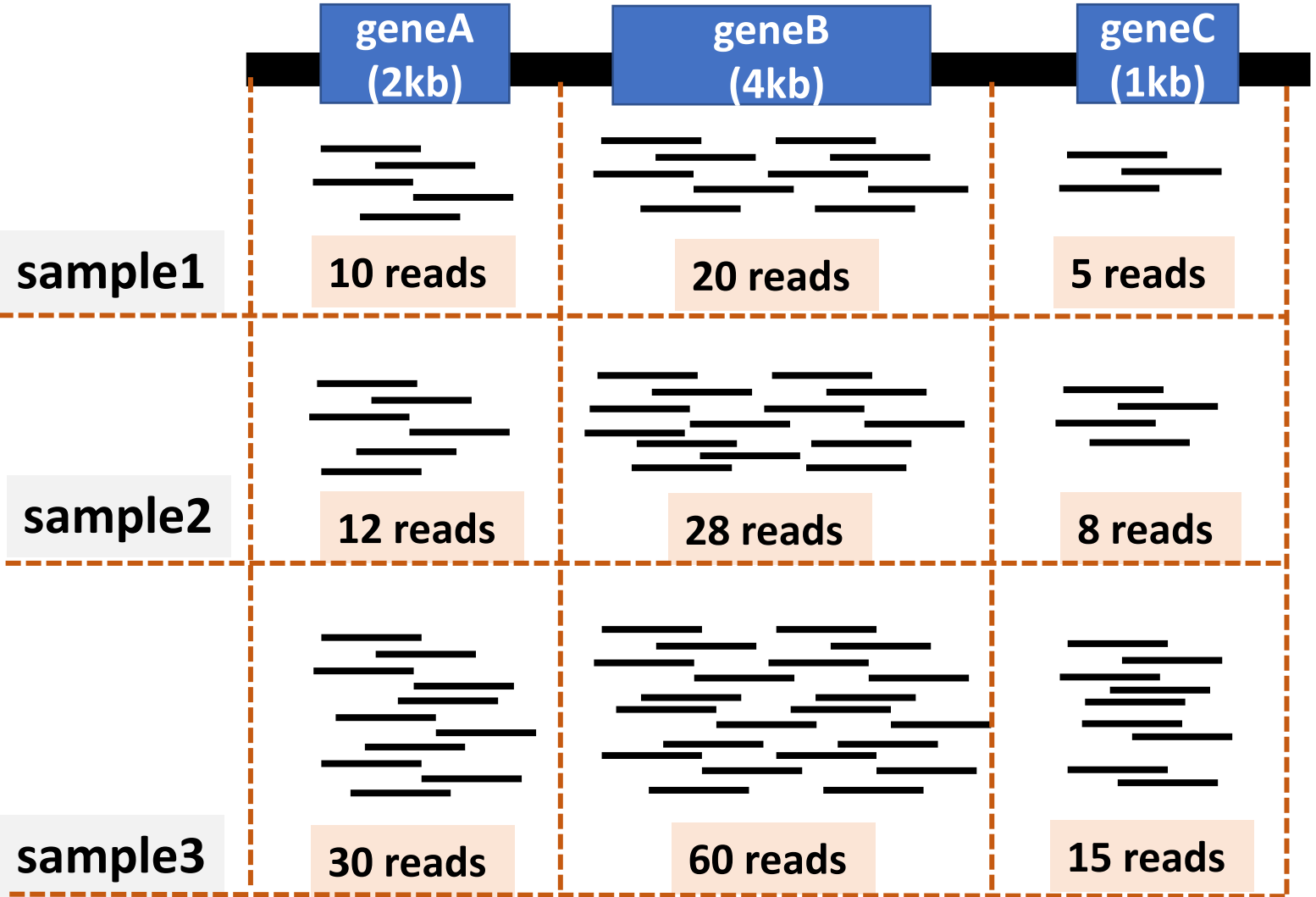
Transcripts Per Million (TPM)
Unit of gene expression

Let’s assume that we got the following read count table.

gene	geneA (2kb)	geneB (4kb)	geneC (1kb)
sample1	10	20	5
sample2	12	25	8
sample3	30	60	15

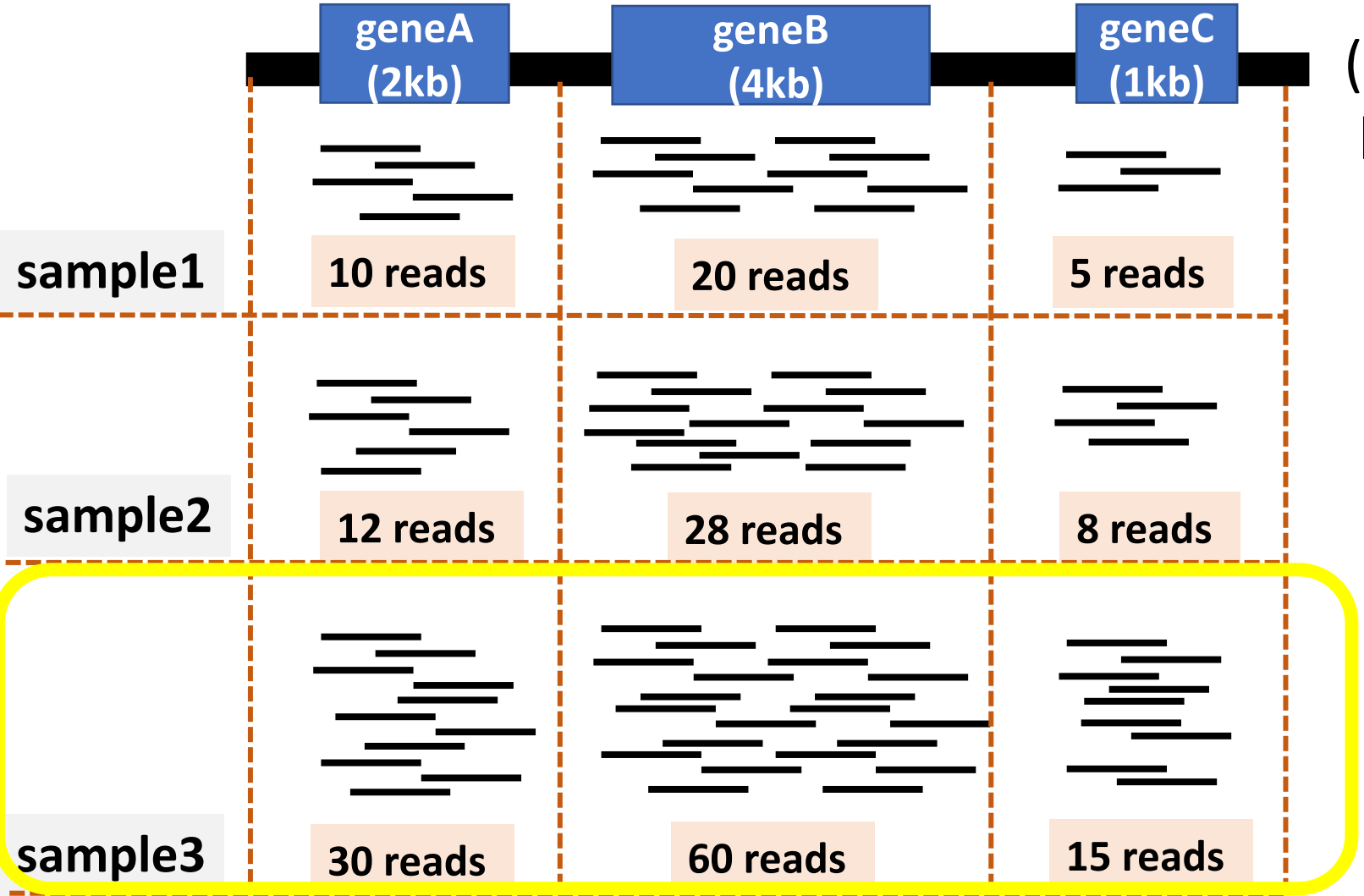
https://www.youtube.com/watch?v=TTUrtCY2k-w&t=7s

gene	geneA (2kb)	geneB (4kb)	geneC (1kb)
sample1	10	20	5
sample2	12	25	8
sample3	30	60	15



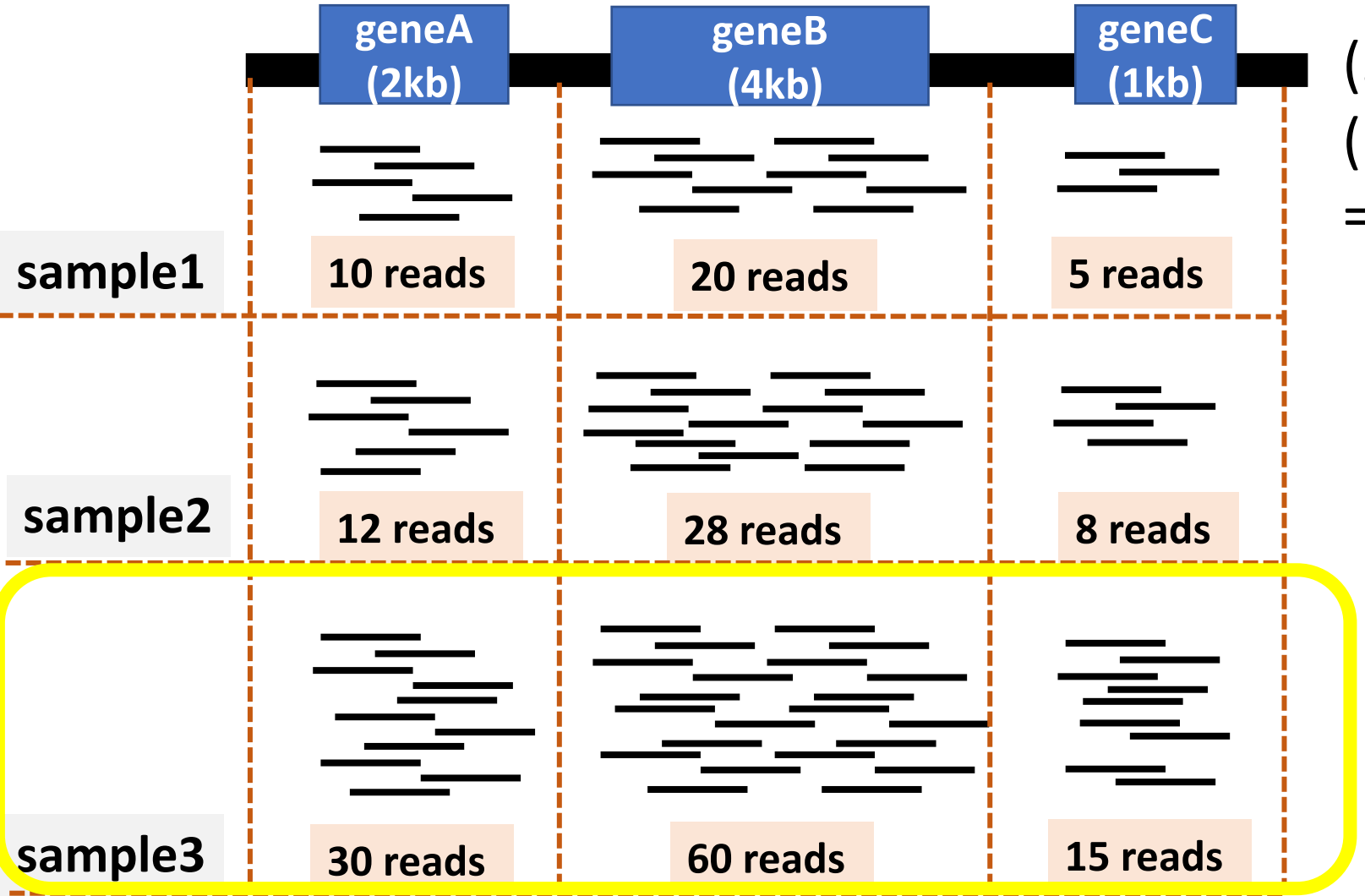
(1) Normalize the read counts by gene lentgh

gene	geneA (2kb)	geneB (4kb)	geneC (1kb)	sum
sample1	5	5	5	15
sample2	6	6.25	8	20.25
sample3	15	15	15	45



(2) Normalize the read counts by total read counts

gene	geneA (2kb)	geneB (4kb)	geneC (1kb)	sum
sample1	0.33	0.33	0.33	1
sample2	0.30	0.31	0.40	1
sample3	0.33	0.33	0.33	1



(3) total read counts will be one
(In reality, 1 Million)
= transcripts per million

Fold change vs p-value

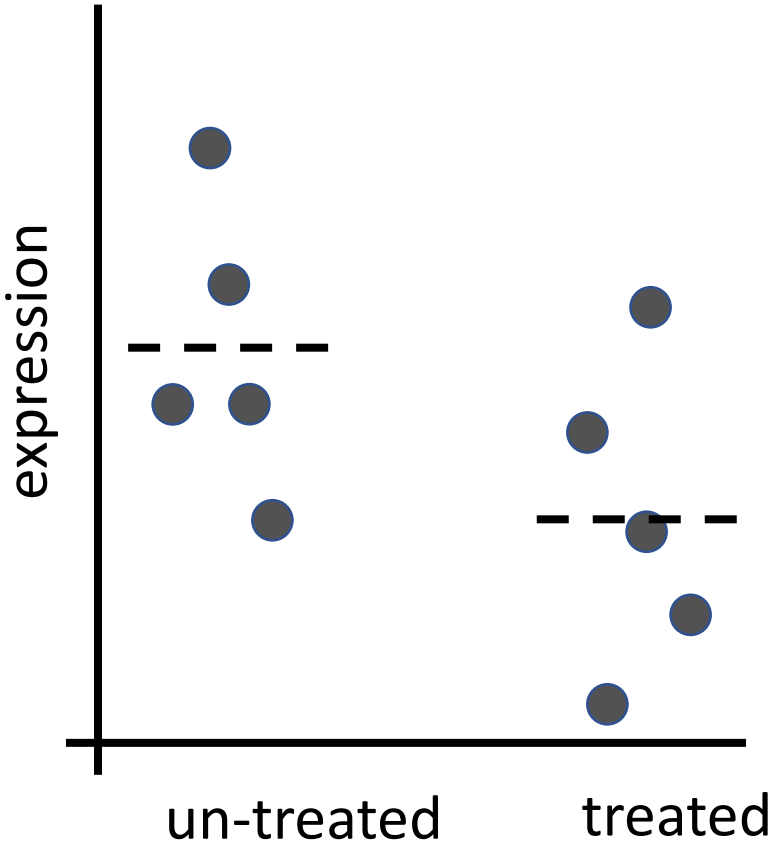
Fold change is $$\frac{(expression_{treated} - expression_{non\text{-}treated})}{expression_{non\text{-}treated}}$$

p-value is *the probability of obtaining results as extreme as the observed results of a hypothesis test, assuming that the null hypothesis is correct.*

...in today's case, the probability of observing our results when the RNAi of the target gene does not affect gene expression.

Case study:

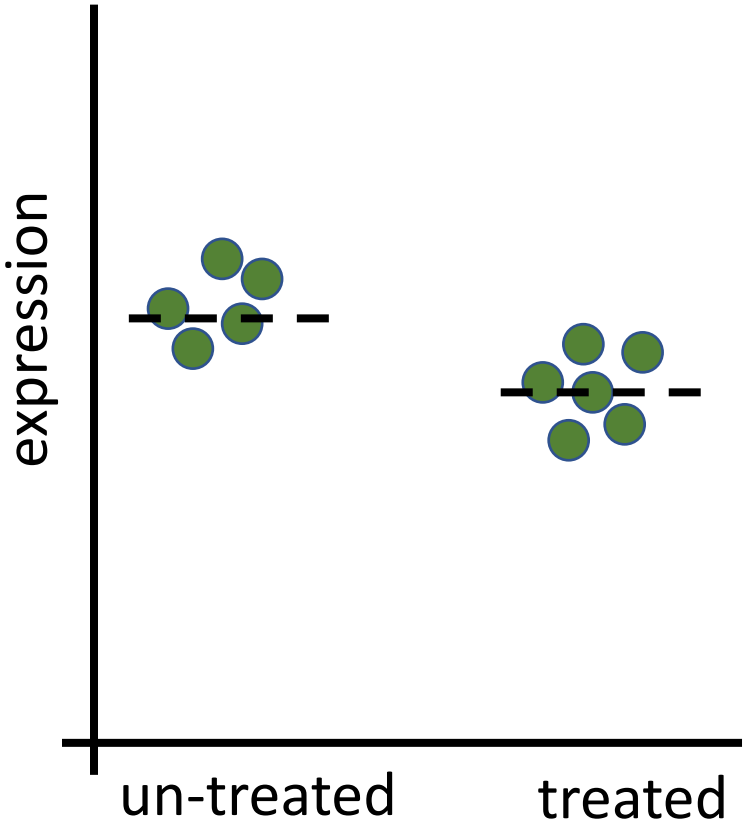
Observed expression of gene A



High fold change

Less significant p-value

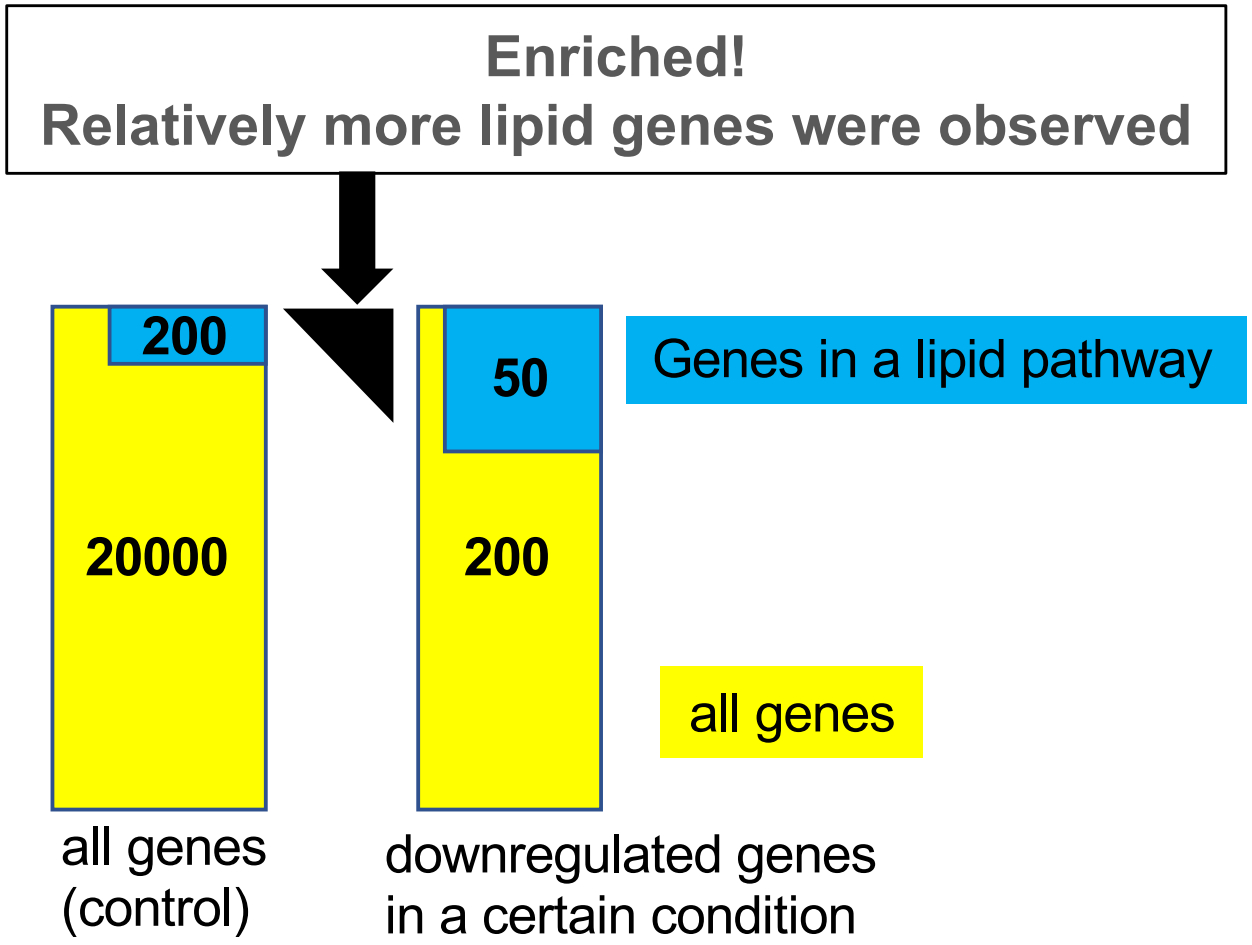
Observed expression of gene B



Low fold change

More significant p-value

Functional categorization of genes (Gene Ontology Analysis)



Are you ready?

https://mariesaitou.github.io/Bio326/RNAseq_for_lab.html