

# **Module 7**

## **Transcriptomics**

**Helminth Bioinformatics**  
**Khon Kaen University, 2023**

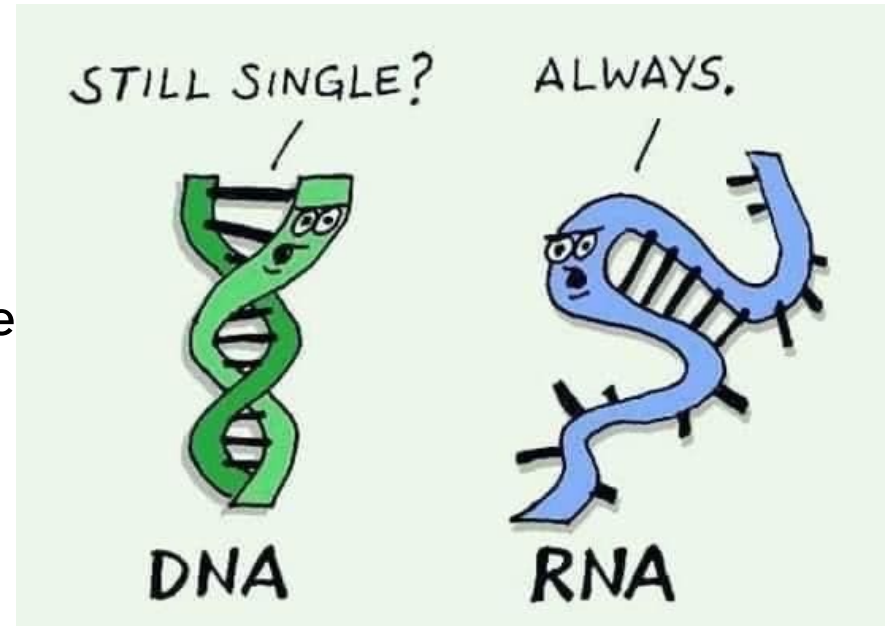
# Module aims

You will learn how to:

- map RNA-seq data to reference genome
- acquire read counting results and import them to R
- visualise transcriptomic profiles in R
- using R packages to identify differentially expressed genes and finding patterns in the data
- performing GO term enrichment and interpret the results

# What is transcriptome?

All RNA being transcribed  
at a certain developmental stage  
in a certain type of cells  
in response to certain stimuli  
...

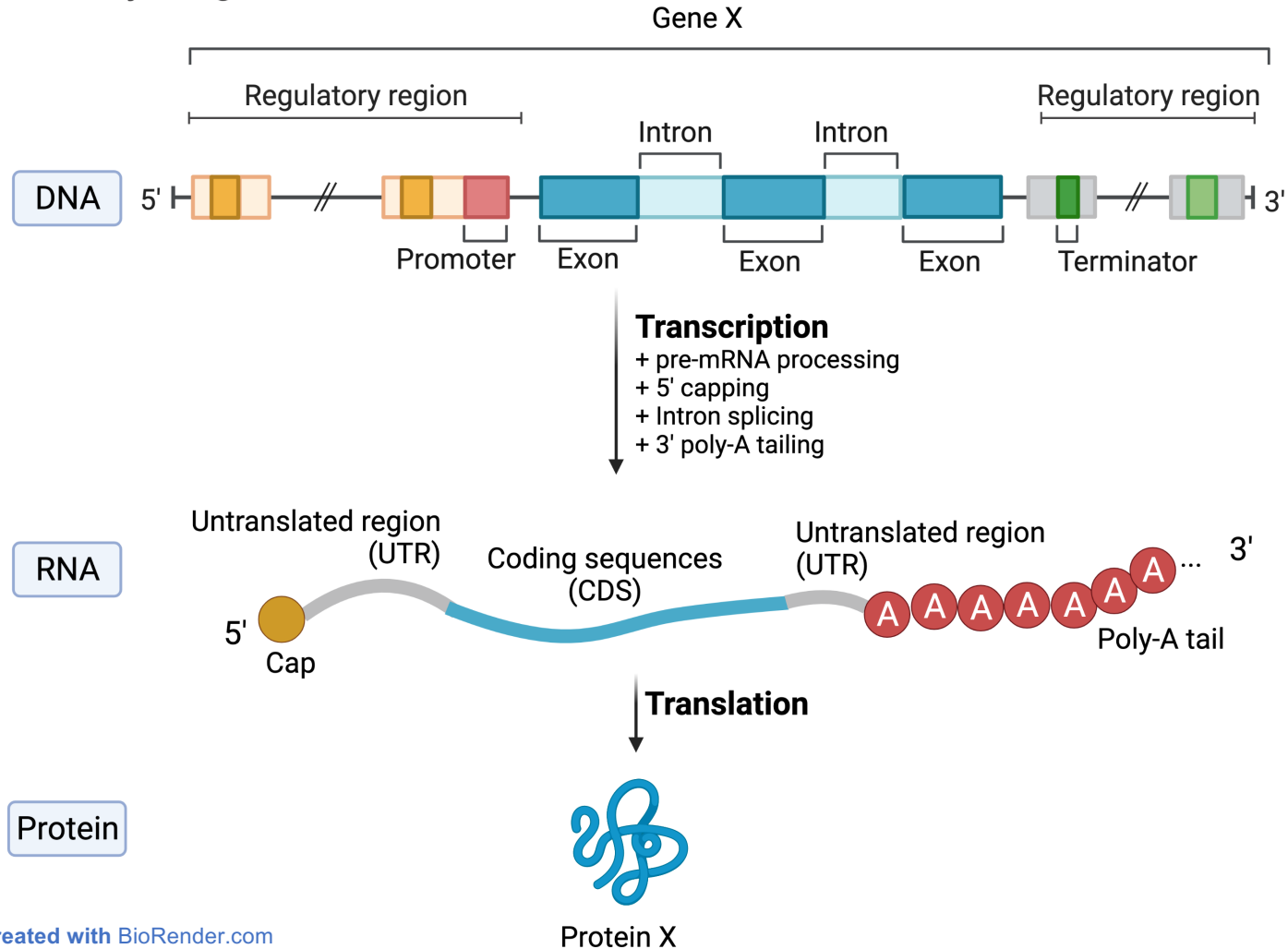


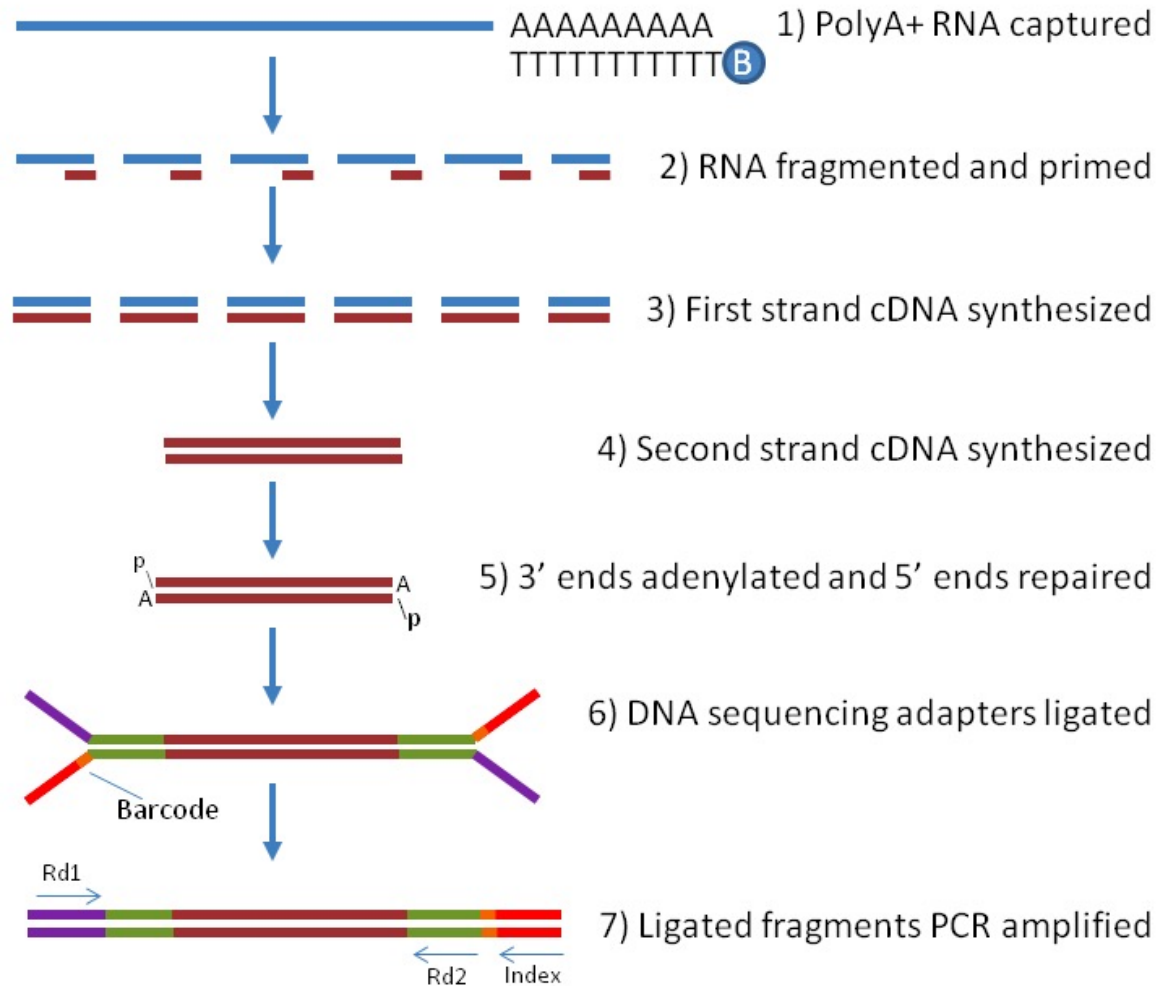
**Tyler Gable** siRNA, miRNA, ceRNA, piRNA, piRNA-like RNA, pesRNA, many viral RNAs ALL DISAGREE.

Like · Reply · 4d



# Eukaryotic gene structure





The diagram illustrates the RNA-seq workflow, divided into three main stages: *In vivo*, *In vitro*, and *In silico*.

- In vivo* (Pink background):**
  - DNA gene in genome:** Represented by a black arrow.
  - Transcription:** The DNA is transcribed into **Pre-mRNA**, shown as a red arrow with internal segments.
  - Intron splicing:** The pre-mRNA is processed into **Mature mRNA**, shown as a continuous red arrow.
- In vitro* (Blue background):**
  - Fragmentation:** The mature mRNA is fragmented into **RNA fragments**, shown as red rectangles.
  - Reverse transcription:** The RNA fragments are converted into **ds-cDNA fragments**, shown as blue rectangles.
  - High-throughput sequencing:** The ds-cDNA fragments are sequenced to produce **Sequences**, shown as short DNA reads.
- In silico* (Green background):**
  - Sequence processing:** The raw sequences are processed and aligned to the **Genome sequence**.
  - Alignment:** The sequences are mapped to the genome, with some reads spanning across exons and introns.
  - Splice variant A and B:** The aligned reads are used to identify different **Splice variants** (isoforms) of the gene, shown as black arrows with different internal structures.

# Common uses of RNA-seq data

## Gene expression study

e.g. differential expression, time course profile

## Profiling total RNA (e.g. miRNA and mRNA)

e.g. in exosomes and other secretory products

## Splice isoform

only useful for organism with polished reference genomes

## SNP calling

use transcriptome as a reduced subset of genomic variation study

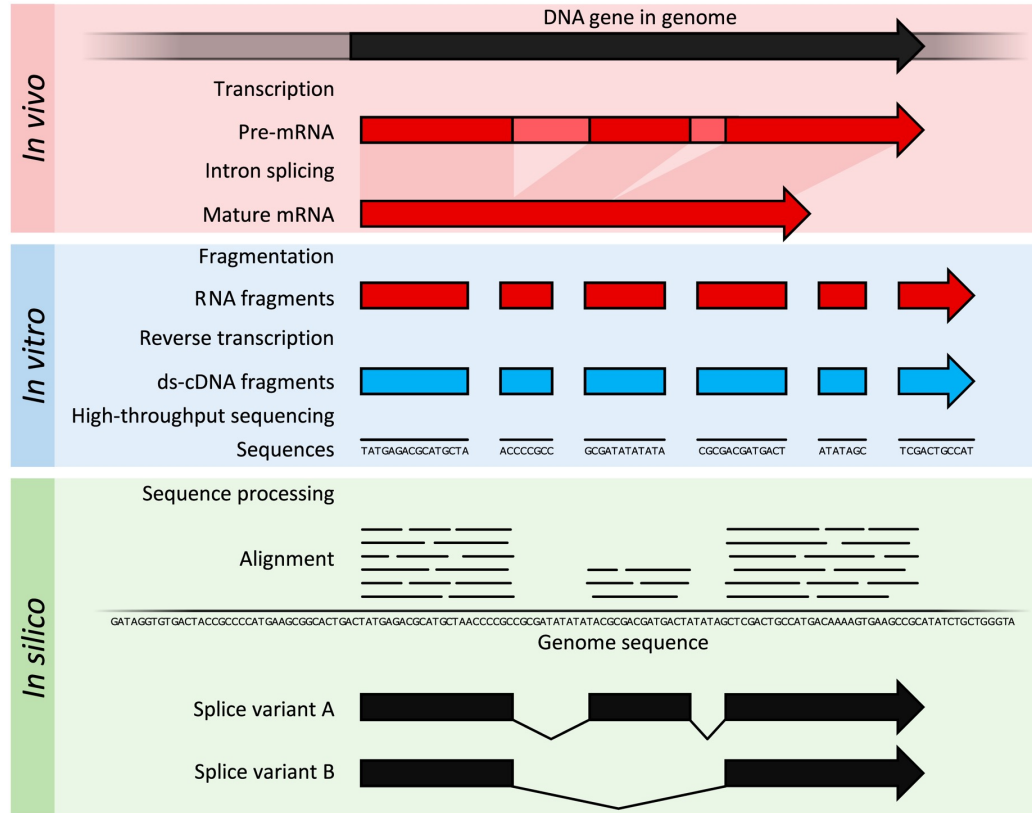
## Profiling genes in an organism

e.g. for gene annotation, refining gene model

# Terms you might come across

number of reads strand-specific

single-end/pair-end

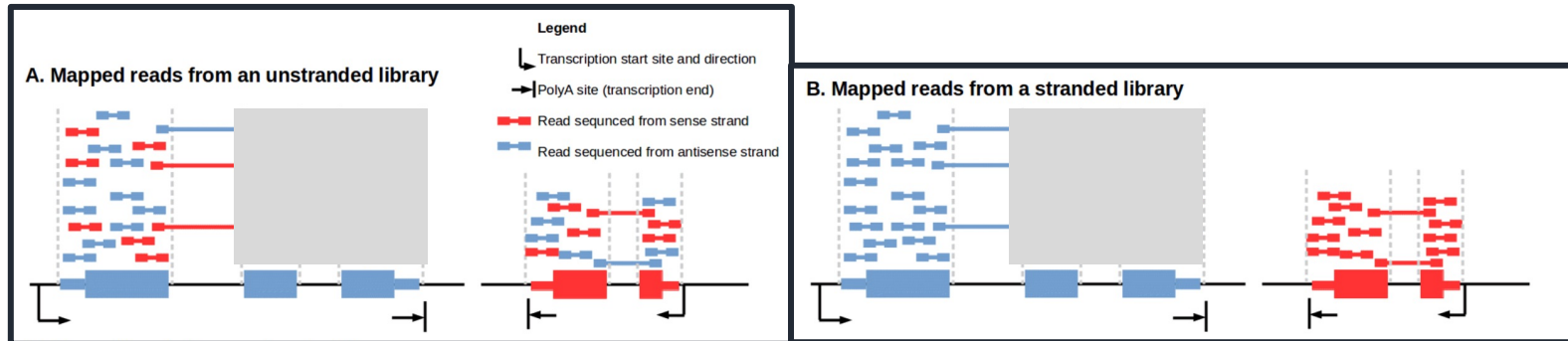
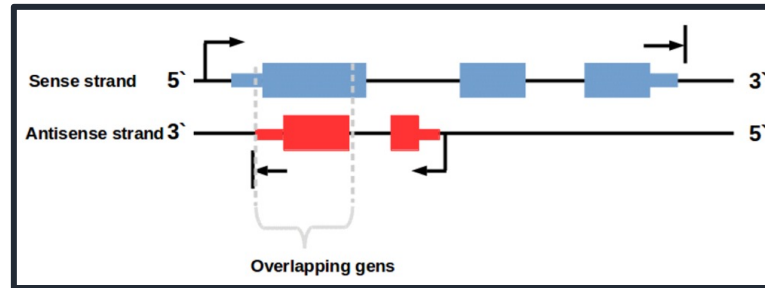




# Terms you might come across

number of reads    **strand-specific**    single-end/pair-end

- More reliable quantification of genes on opposite strand
- Allow discovery of anti-sense transcription



# Terms you might come across

number of reads    strand-specific

single-end/pair-end

## Single-end

Read fragment from only one end

Can be good enough for gene expression study, if there is a good reference genome

## Pair-end

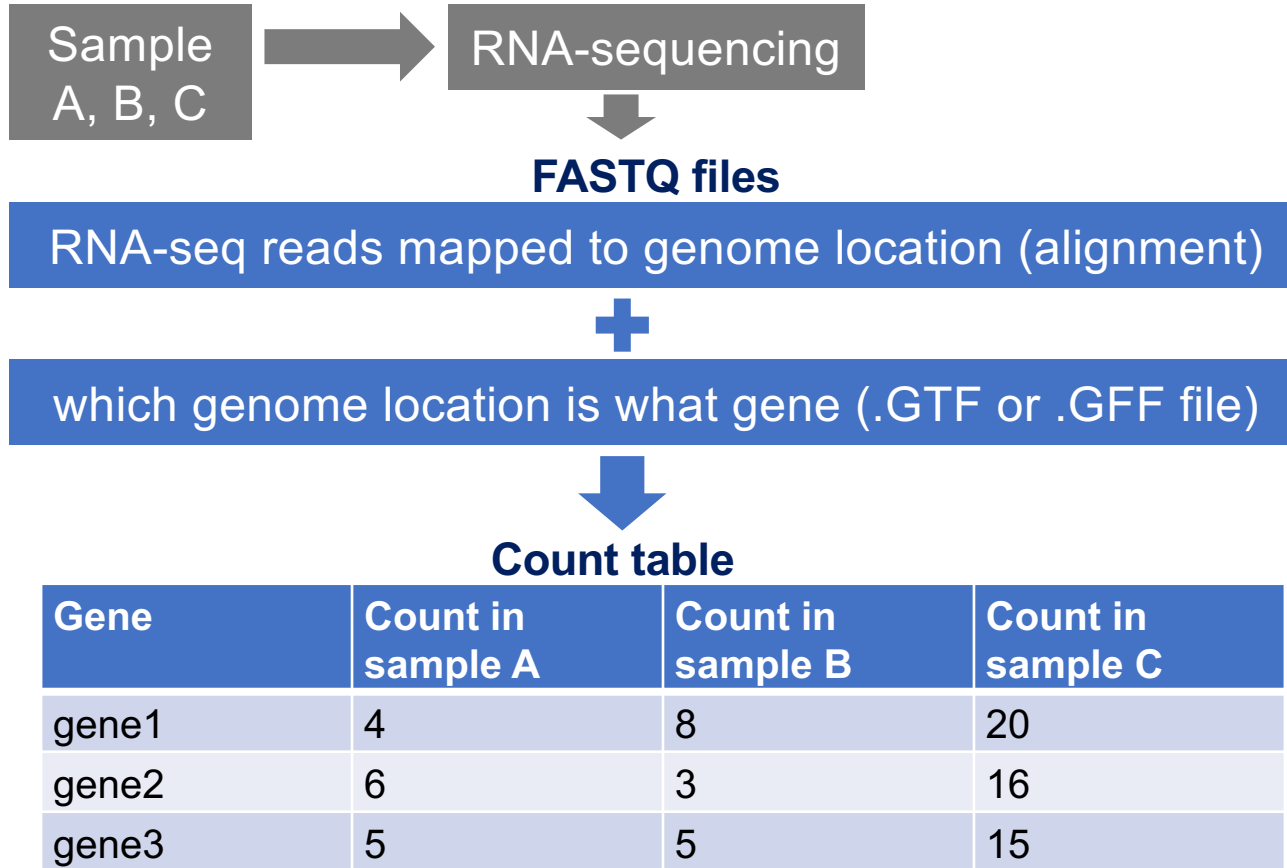
Read from both ends of the fragment

Provide more information which can help with mapping

Highly recommend for organism with only draft reference genome, or without a genome



# From sequencing data to read count



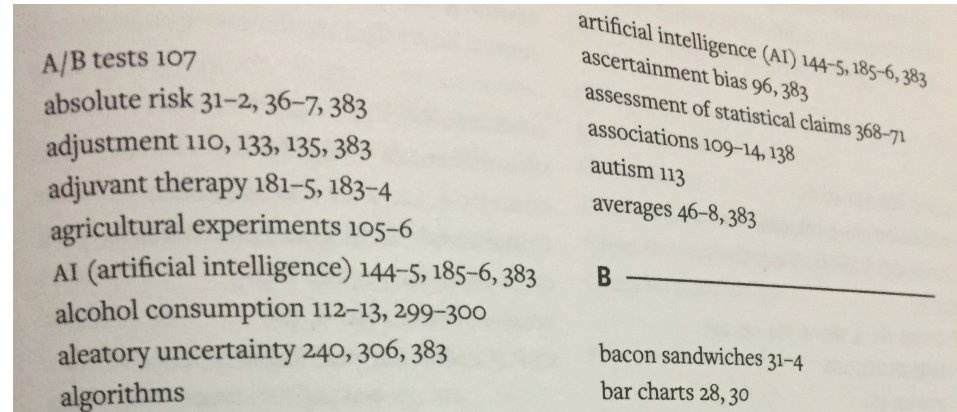
# Almost hands-on time: genome indexing – why?

Mapping reads to a genome as approximate pattern matching

Finding your sequences (short texts) in a genome (large book)

## Choices

- A) Scan the whole genome (large book) for the sequence
- B) Pre-process the genome – then searching through book index instead of page by page



# Hands-on time!

Index genome using hisat2 (this will take a few minutes)

`/location/of/your/data/`

replace text inside with information related to your situation e.g. location of your files

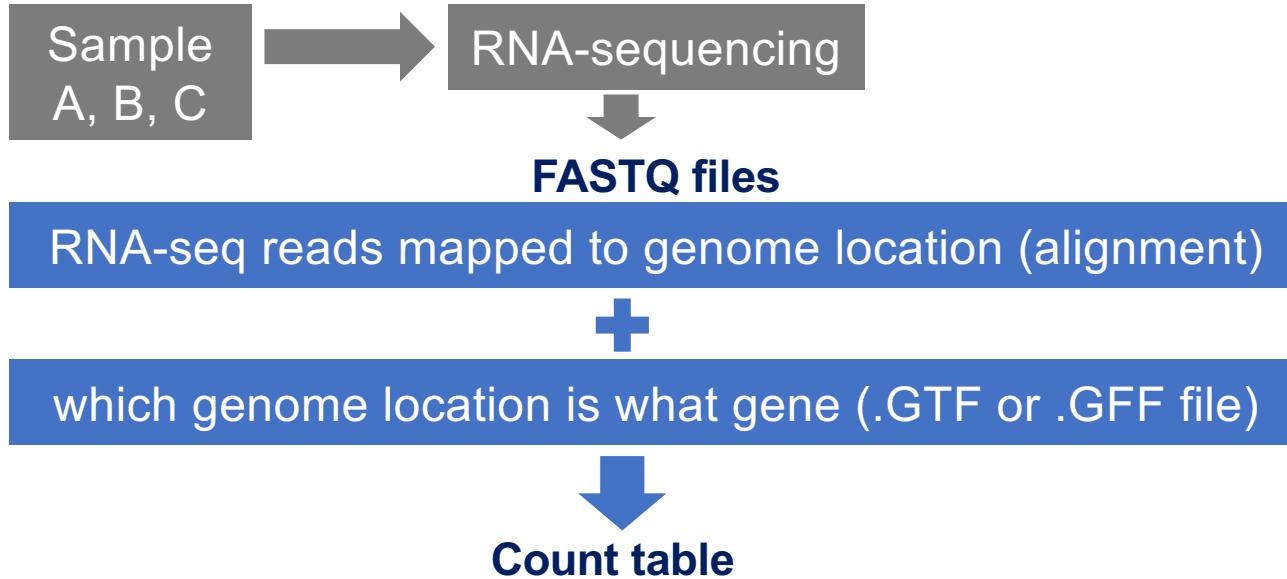
USE TAB (also try double tab)

When copy-paste, check this symbol `-` and this `"`

# What we did in unix

- Genome indexing
- Map (align) reads to genome
  - SAM & BAM files
- Get read counts per gene
  - (\*\_v10.count)

# From sequencing data to read count



Gene	Count in sample A	Count in sample B	Count in sample C
gene1	4	8	20
gene2	6	3	16
gene3	5	5	15

```
$ head *.count
```

```
==> D06_1_v10.count <==
```

```
Smp_000020.1 299
```

```
Smp_000030.1 1071
```

```
Smp_000040.1 425
```

```
Smp_000050.1 190
```

```
Smp_000070.1 156
```

```
==> D06_2_v10.count <==
```

```
Smp_000020.1 76
```

```
Smp_000030.1 310
```

```
Smp_000040.1 134
```

```
Smp_000050.1 67
```

```
Smp_000070.1 46
```



# Next.. R

- Prepare data for analysis in R
- Identify differentially expressed (**DE**) genes
- Create plots
- Functional analysis

# Fold change

**A** (D13)

---

**B** (D06)

**Log<sub>2</sub>FC (log<sub>2</sub> of fold change)**

$$\log_2 \left( \frac{A(D13)}{B(D06)} \right)$$

**Log<sub>2</sub>FC (log<sub>2</sub> of fold change)**

$$\log_2 \left( \frac{8}{2} \right)$$

**Log<sub>2</sub>FC (log<sub>2</sub> of fold change)**

$$\log_2 \left( \frac{A(D13)}{B(D06)} \right)$$

**Log<sub>2</sub>FC (log<sub>2</sub> of fold change)**

$$\log_2 \left( \frac{2}{8} \right)$$

# Functional analysis

- Rather than going through the list of differentially expressed genes to find genes that you expect to see changes
  - Do functional analysis
  - Let data guide the way
- Possibly the most common = GO enrichment

# GO term enrichment

Genes often have associated GO terms (Gene Ontology terms).

**WormBase ParaSite** Version: WBPS9 (WS258)

Search WormBase ParaSite...

Genome List BLAST BioMart REST API VEP Downloads WormBase

Schistosoma mansoni (PRJEA36577) Location: Smp.Chr\_3:12,709,526-12,722,895 Gene: Smp\_013040

**Gene-based displays**

- Summary
  - Splice variants
  - Transcript comparison
- Sequence
  - Literature
- Comparative genomics
  - Gene tree
  - Orthologues
  - Paralogues
- Gene Ontology
  - Molecular function**
  - Cellular component
  - Biological process
- External references
  - Expression
- Variation
  - Variation Table

**Gene: Smp\_013040**

**Description** Cathepsin D (A01 family) [Source:UniProtKB/TrEMBL;Acc:[G4VEV6](#)]

**Location** [Scaffold Smp.Chr\\_3:12,709,526-12,722,895](#) reverse strand.

**INSDC Sequence ID** [HE601626.1](#)

**Gene Overview** This gene has 2 transcripts ([splice variants](#)), [1048 orthologues](#) and [1 paralogue](#).

**Gene Type** Protein coding

**Annotation Method** Gene models from Wellcome Trust Sanger Institute [Reference Helminth Genomes project](#)

**Transcripts** [Show transcript table](#)

**Molecular function**

GO Term	Evidence	Annotation source	Transcript IDs	Actions
Aspartic-type endopeptidase activity	IEA	<a href="#">UniProtKB/TrEMBL:G4VEV6</a> , <a href="#">UniProtKB/TrEMBL:P91802</a> , <a href="#">InterPro:Aspartic_peptidase_AS</a> , <a href="#">InterPro:Aspartic_peptidase_A1</a> , <a href="#">InterPro:Cathepsin_D</a> , <a href="#">InterPro:Aspartic_peptidase_N</a>	<a href="#">Smp_013040.1</a> <a href="#">Smp_013040.2</a>	<a href="#">Search BioMart</a> <a href="#">View associated genes</a>
Aspartic-type endopeptidase activity	IEA	<a href="#">UniProtKB/TrEMBL:G4VEV6</a> , <a href="#">UniProtKB/TrEMBL:P91802</a>	<a href="#">Smp_013040.1</a> <a href="#">Smp_013040.2</a>	<a href="#">Search BioMart</a> <a href="#">View associated genes</a>



# GO term enrichment

Genes often have associated GO terms (Gene Ontology terms).  
GO terms describe functions of a gene, and can be derived from  
sequence similarity, experiment, homology etc.

**ID number**   **Description**

**WormBase ParaSite** Version: W

Genome List BLAST BioMart REST API VEP

*Schistosoma mansoni* (PRJEA36577) Location: Smp.Chr\_3:12

**Gene-based displays**

- Summary
  - Splice variants
  - Transcript comparison
- Sequence
  - Literature
- Comparative genomics
  - Gene tree
  - Orthologues
  - Paralogues
- Gene Ontology
  - Molecular function**
  - Cellular component
  - Biological process
- External references
- Expression
- Variation
  - Variation Table
  - Variation Image

**Gene: Smp\_01304**

Description

Location

INSDC Sequence ID

Gene Overview

Gene Type

Annotation Method

Transcripts

**Molecular function**

Accession	Term
<a href="#">GO:0004190</a>	aspartic-type endopeptidase activity
<a href="#">GO:0008233</a>	peptidase activity

**Export data**

**Share this page**

Show/hide columns (1 hidden) Filter

Accession	Term	Evidence	Annotation source	Transcript IDs
<a href="#">GO:0004190</a>	aspartic-type endopeptidase activity	IEA	UniProtKB/TrEMBL:G4VEV6, UniProtKB/TrEMBL:P91802, InterPro:Aspartic_peptidase_AS, InterPro:Aspartic_peptidase_A1, InterPro:Cathepsin_D, InterPro:Aspartic_peptidase_N	<a href="#">Smp_013040.1</a> <a href="#">Smp_013040.2</a>
<a href="#">GO:0008233</a>	peptidase activity	IEA	UniProtKB/TrEMBL:G4VEV6, UniProtKB/TrEMBL:P91802	<a href="#">Smp_013040.1</a> <a href="#">Smp_013040.2</a>

- [Search BioMart](#)
- [View associated genes](#)

# GO term enrichment

Genes often have associated GO terms (Gene Ontology terms).

GO terms describe functions of a gene, and can be derived from sequence similarity, experiment, homology etc.

**GO term enrichment:** “Are there any GO terms present in my data more frequently than expected by chance alone?”

