

# **Module 7**

## **Transcriptomics**

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

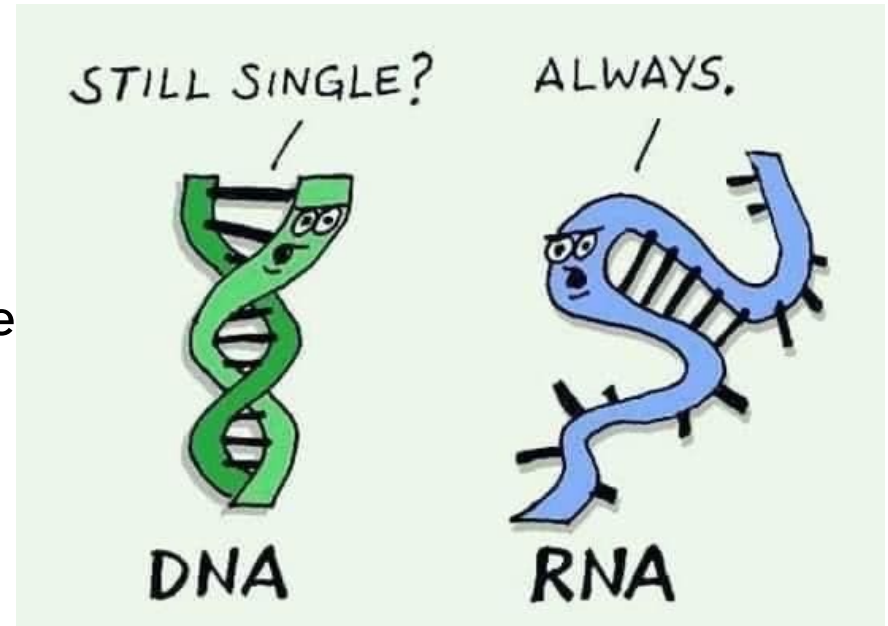
# Module aims

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  
...

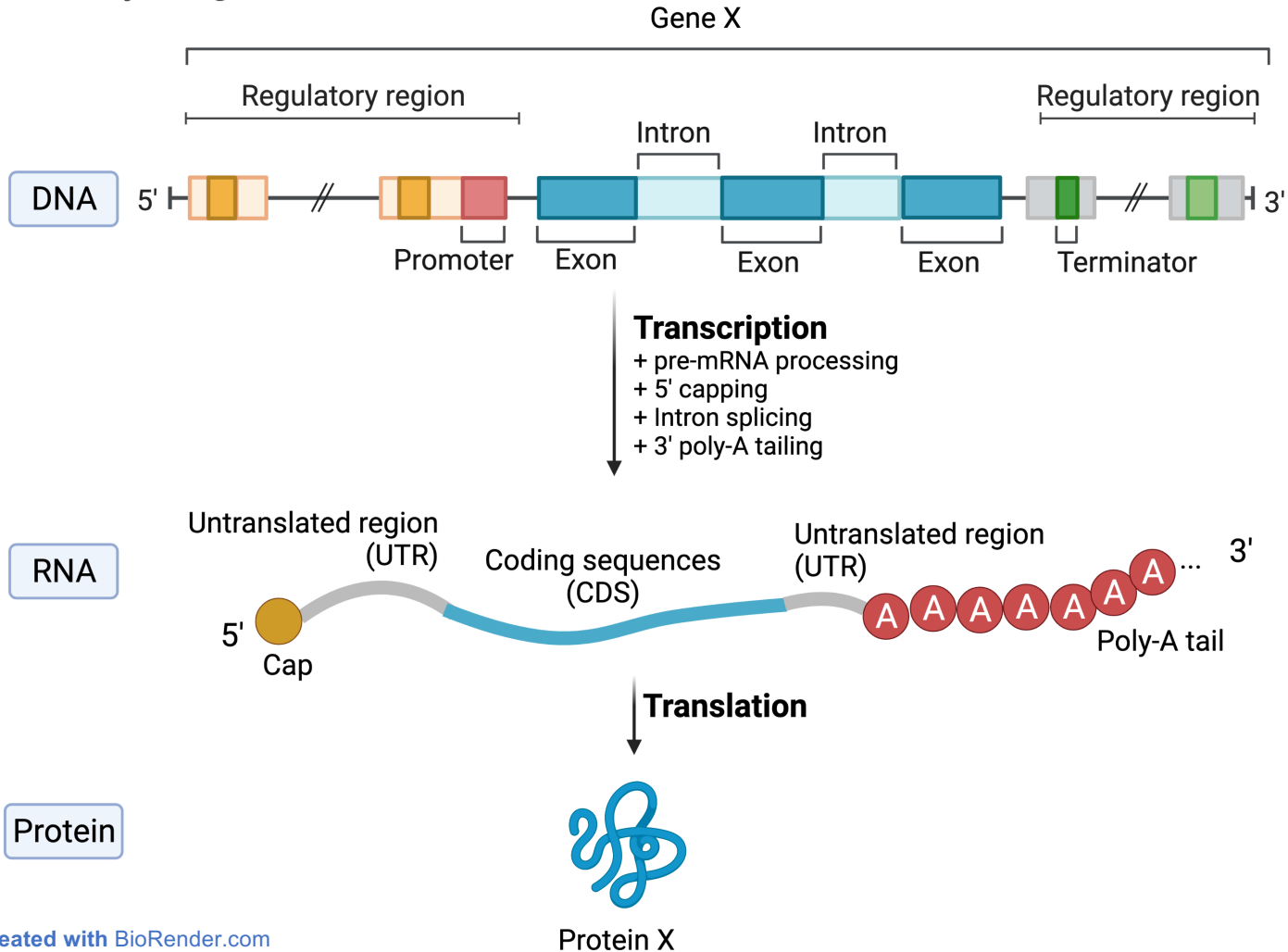


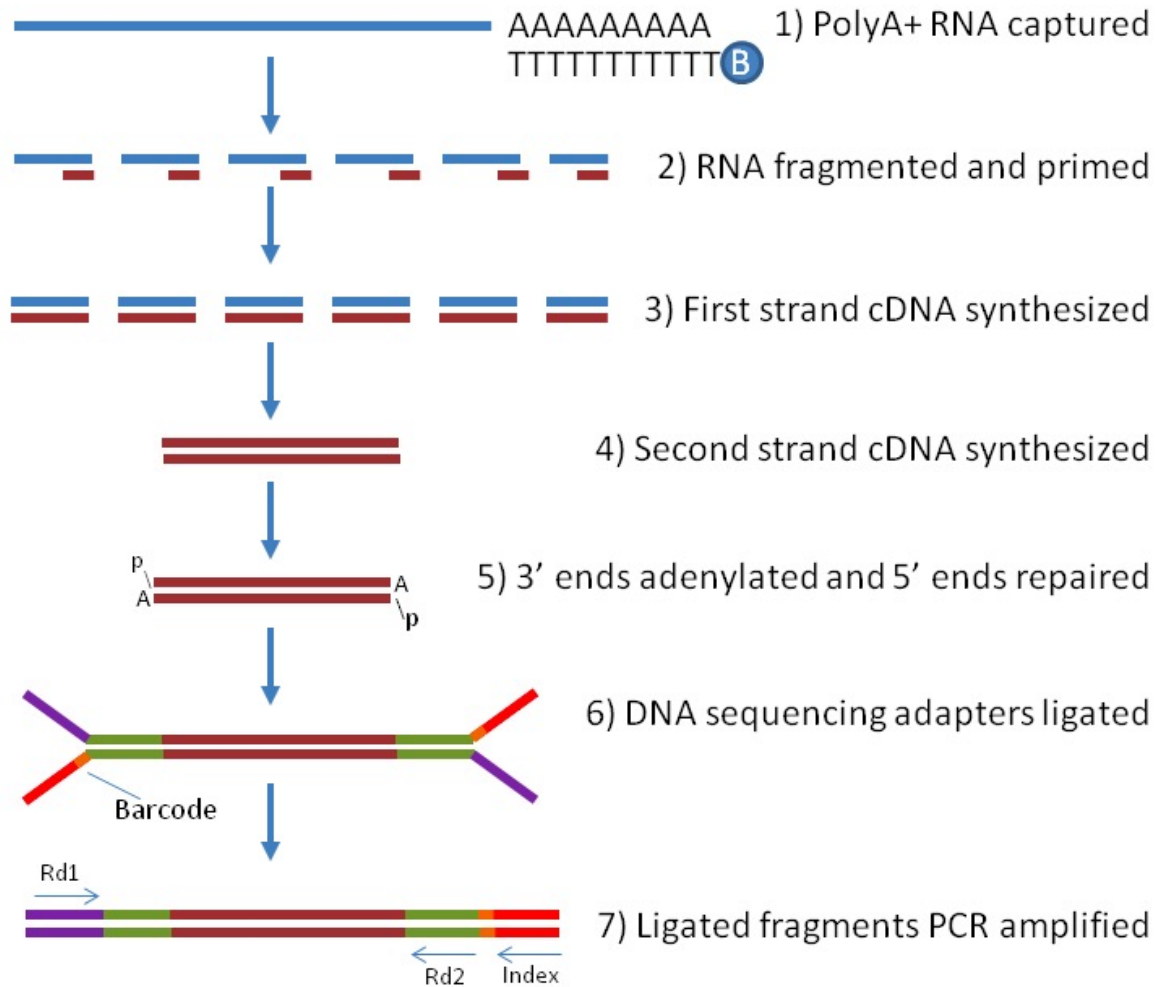
**Tyler Gable** siRNA, miRNA, ceRNA, piRNA, piRNA-like RNA, pesRNA, many viral RNAs ALL DISAGREE.

Like · Reply · 4d



# Eukaryotic gene structure





# What RNA-seq sequences represent

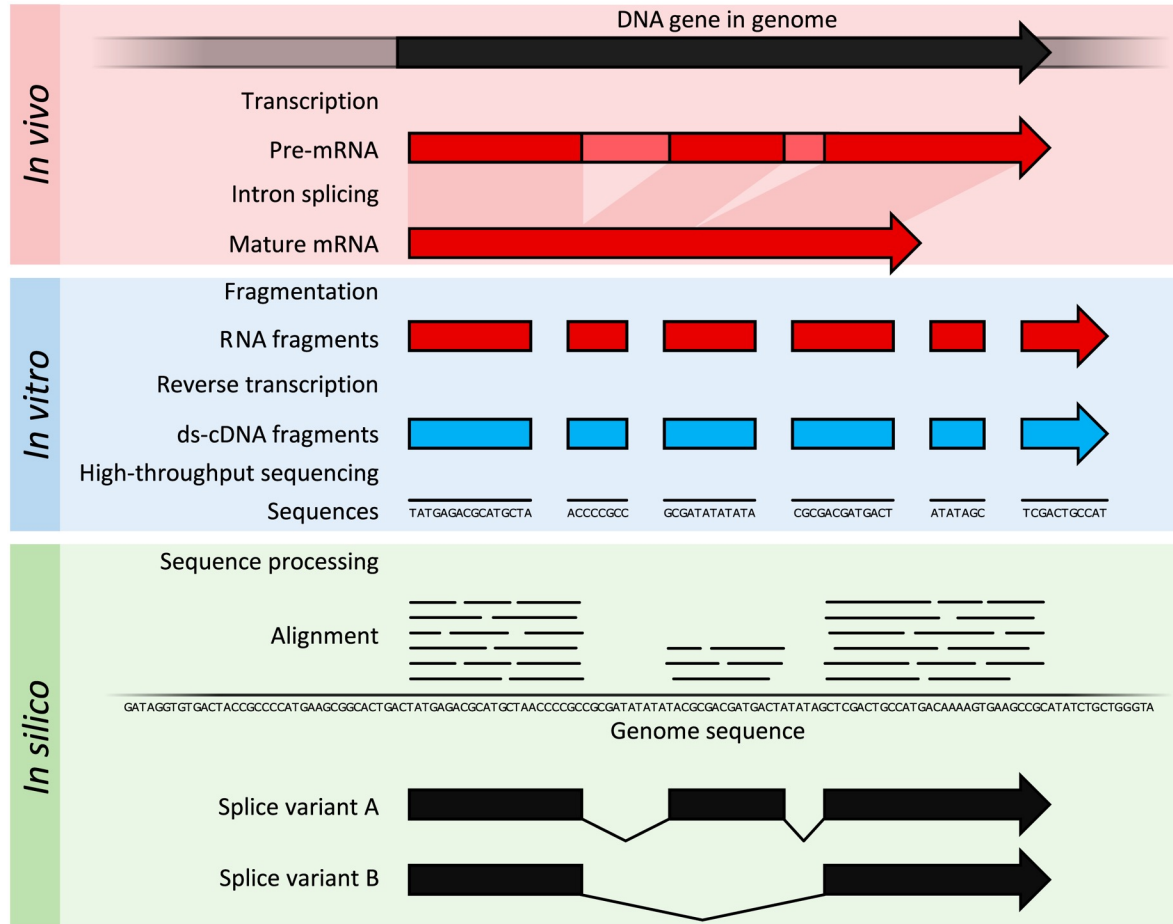


Image from:

<https://www.youtube.com/watch?v=14N111111111>

# 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

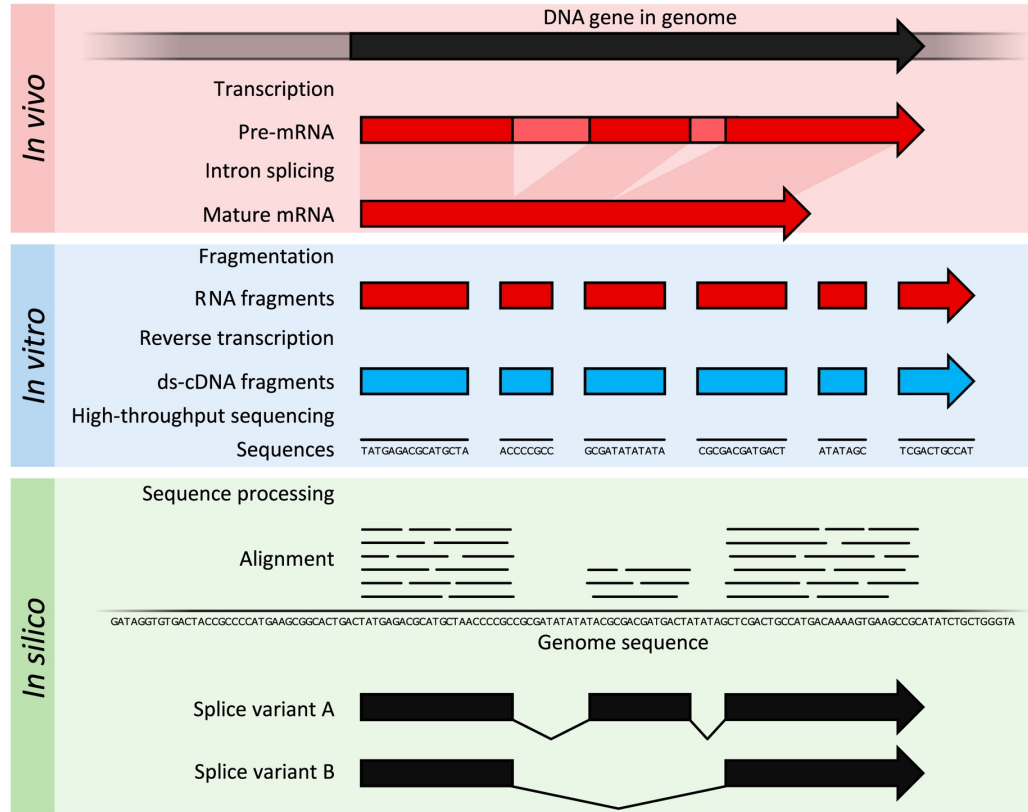
## Profiling genes in an organism

e.g. for gene annotation, refining gene model

# Terms you might come across

number of reads strand-specific

single-end/pair-end

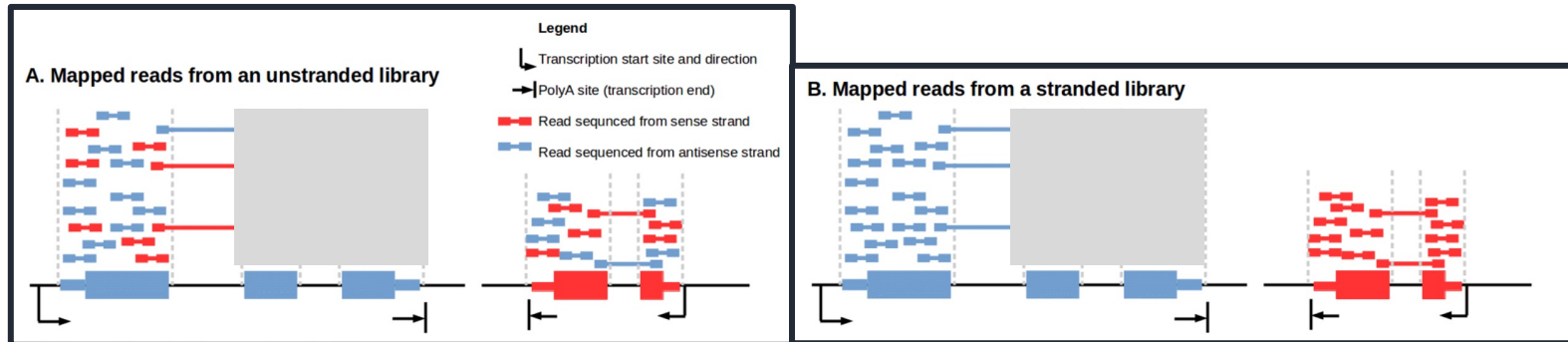
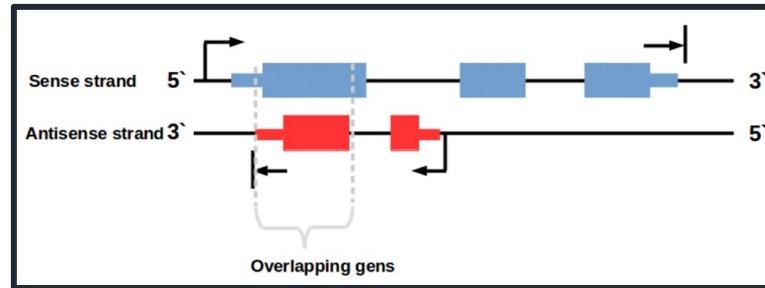




# Terms you might come across

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

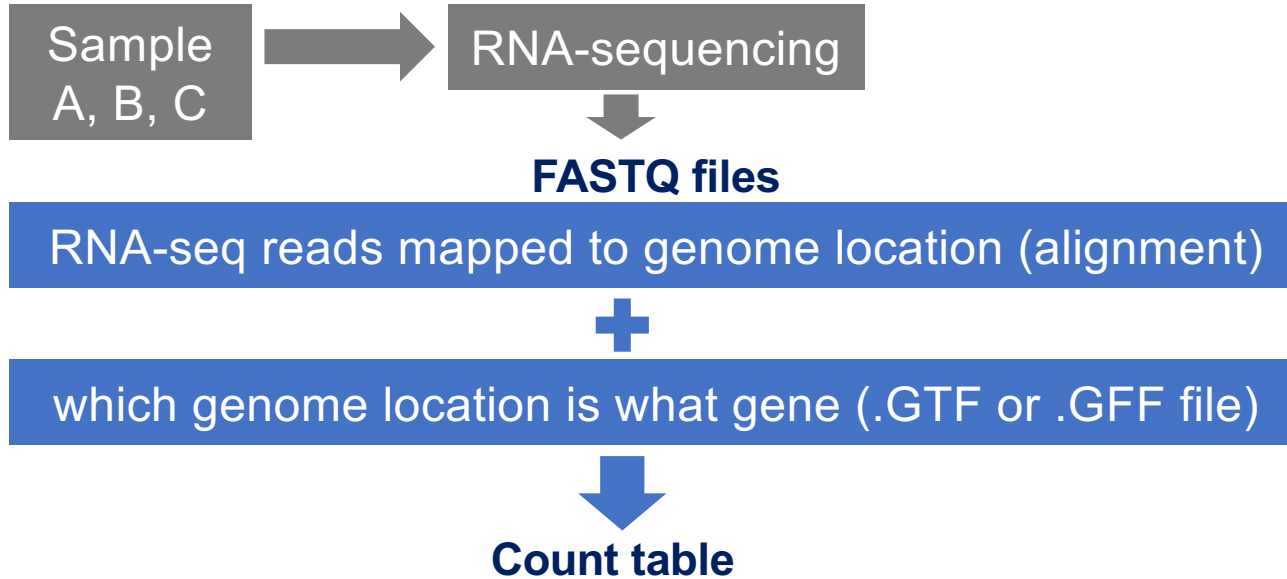
Read from both ends of the fragment

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



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

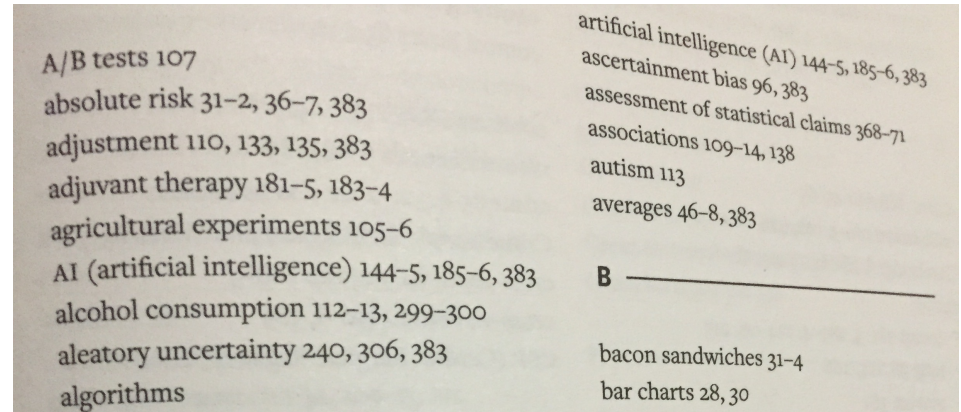
# Almost hands-on time: genome indexing – why?

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)

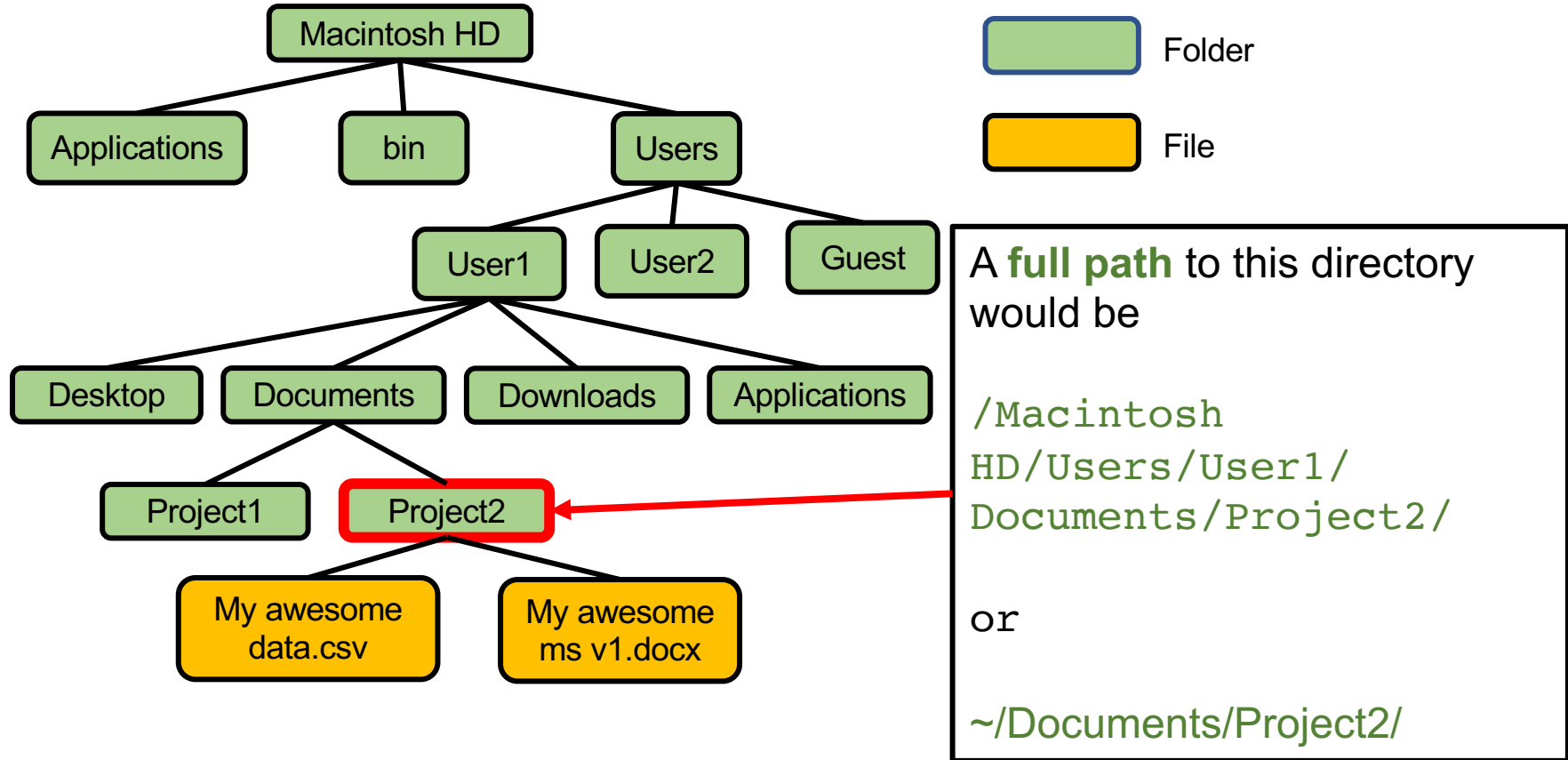
`/location/of/your/data/`

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

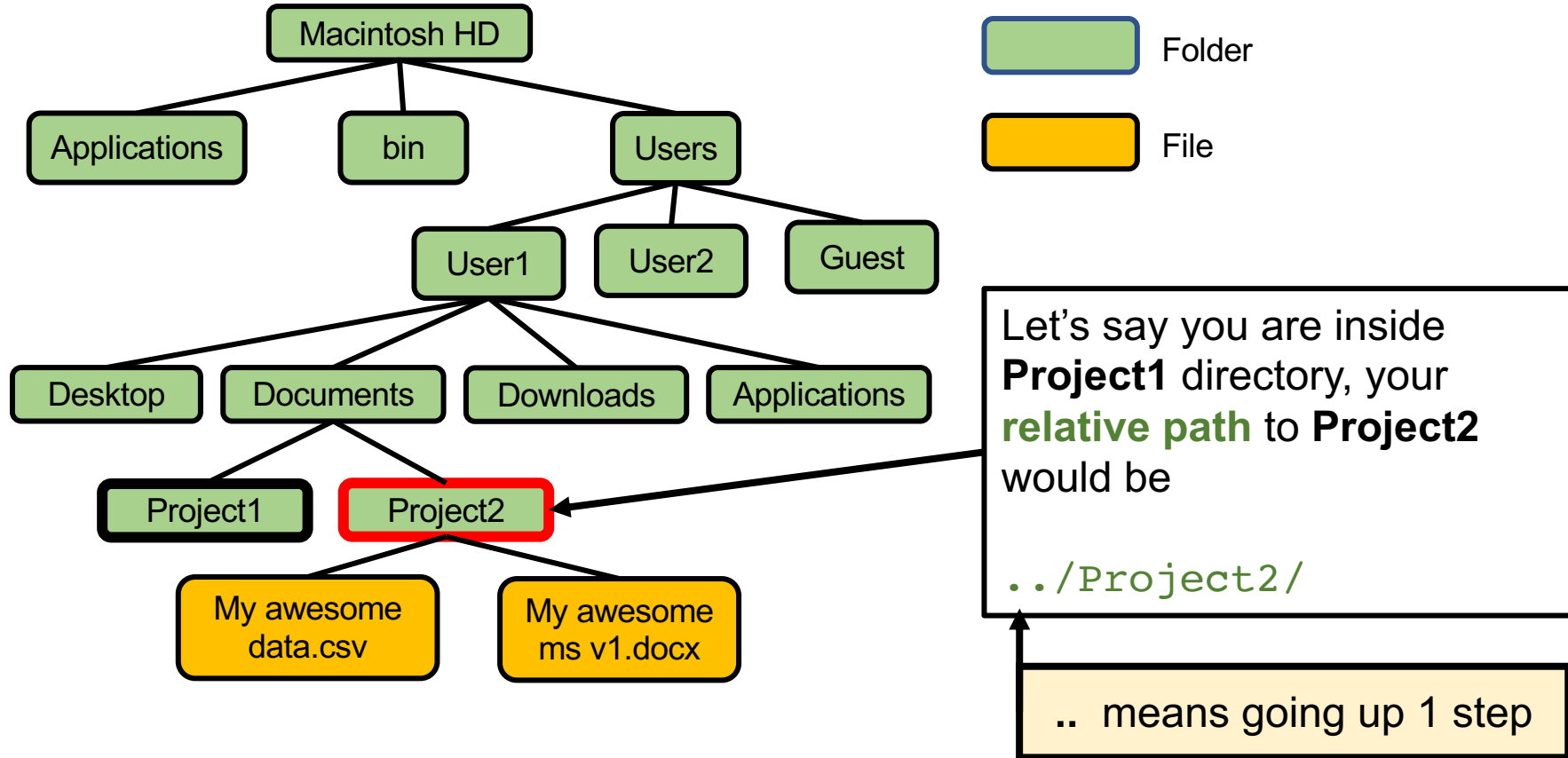
**USE TAB** (also try double tab)

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

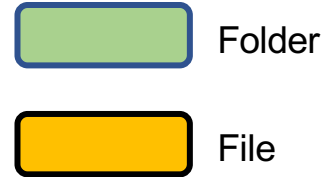
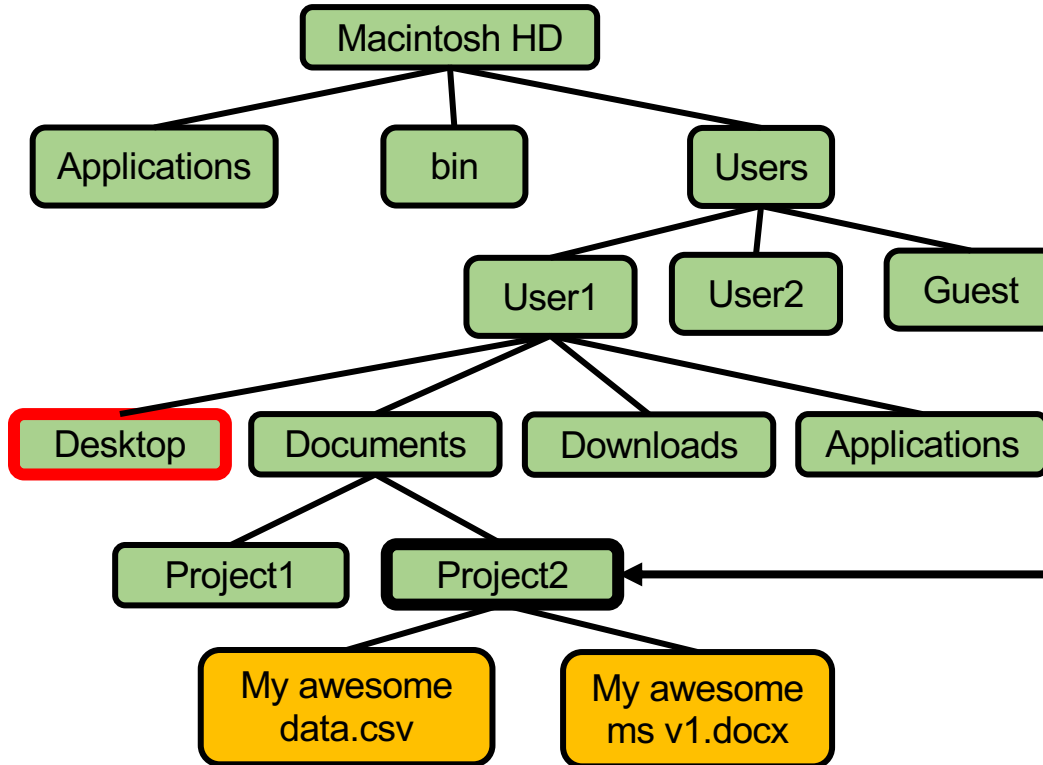
# Mac file system (simplified)



# Mac file system (simplified)



# Mac file system (simplified)

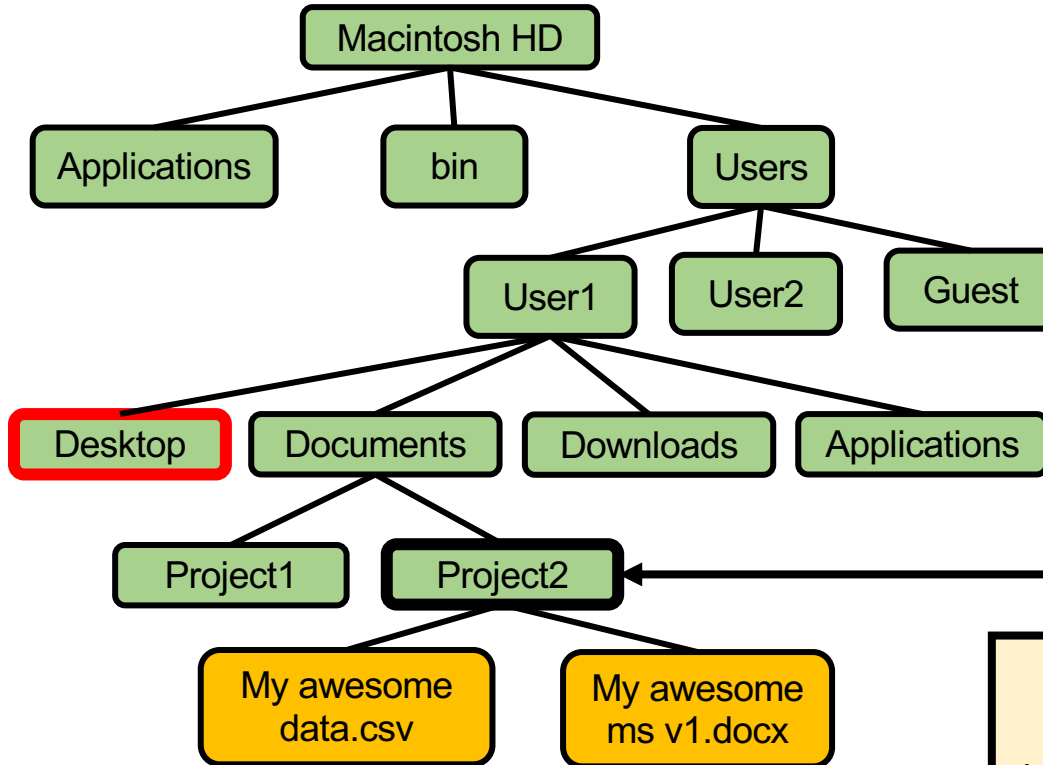


Say, you are inside Project2 directory, what would be your **relative path** to Desktop?

???



# Mac file system (simplified)



 Folder

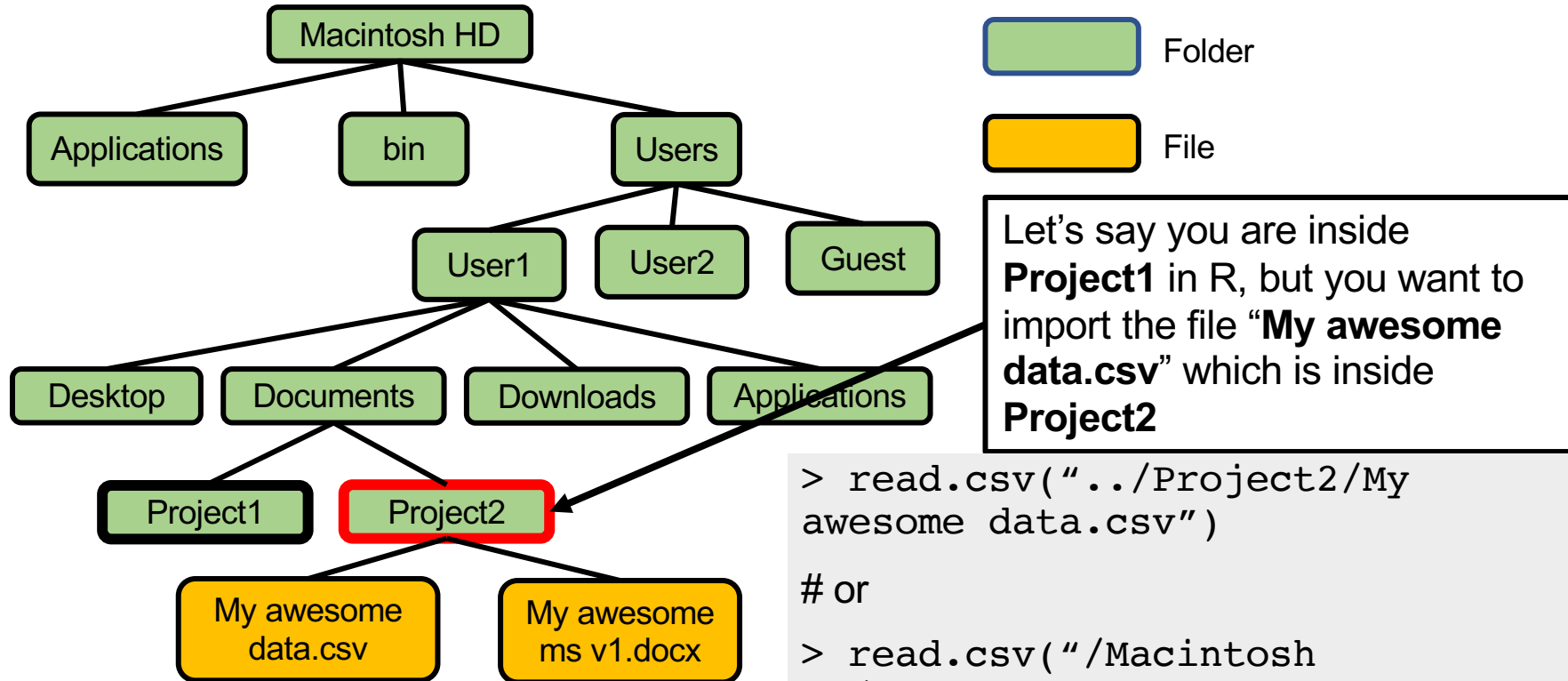
 File

Say, you are inside Project2 directory, what would be your **relative path** to Desktop?

**../../Desktop**

Up to **Documents**, up to **User1**, then to **Desktop**  
**\*\*If in doubt, always use full path\*\***

# Mac file system (simplified)

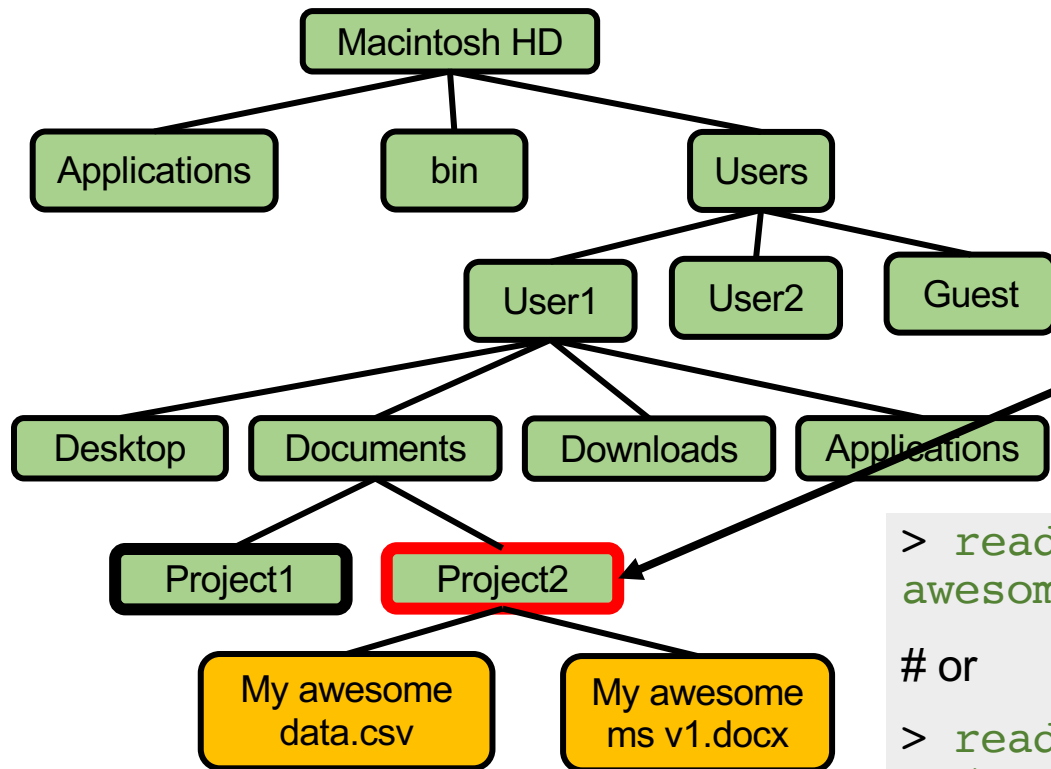


```
> read.csv("../Project2/My  
awesome data.csv")
```

# or

```
> read.csv("/Macintosh  
HD/  
_____/My awesome data.csv")
```

# Mac file system (simplified)



Folder

File

Let's say you are inside **Project1** in R, but you want to import the file "**My awesome data.csv**" which is inside **Project2**

```
> read.csv("../Project2/My  
awesome data.csv")
```

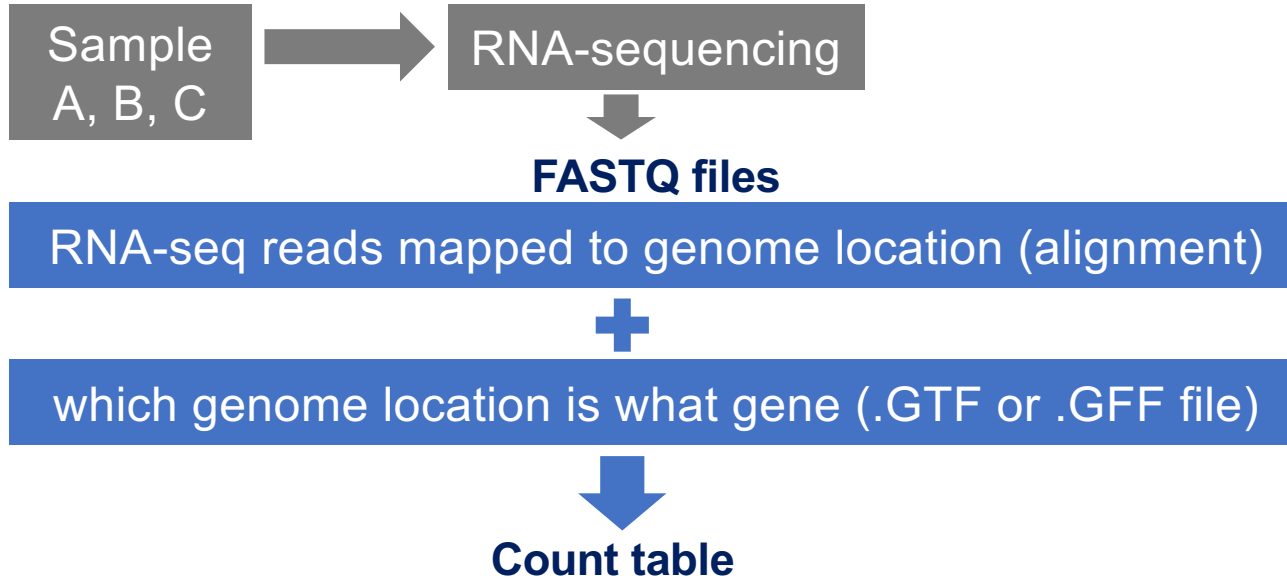
# or

```
> read.csv("/Macintosh  
HD/Users/User1/Documents/Project  
2/My awesome data.csv")
```

# 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).

The screenshot displays the WormBase ParaSite interface for the gene **Smp\_013040**. The left sidebar shows a navigation menu with 'Gene Ontology' expanded to show 'Molecular function', 'Cellular component', and 'Biological process'. The main content area shows the gene's description, location, and a table of associated GO terms.

**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](#)

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

**Table of associated GO terms:**

Accession	Term	Evidence	Annotation source	Transcript IDs
<a href="#">GO:0004190</a>	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="#">GO:0008233</a>	peptidase 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>

• [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?”

