Practical course in bioinformatics (FOR271:2025)

RNA sequencing: applications and principles

-- an overview

Zilan Wen
15-05-2025

Intended Learning Outcome

- To illustate the work flow of differential gene expression analysis.
- To normalize the reads count data (count normalization).
- Sample-level quality control using Principal Component Analysis (PCA) and hierarchical clustering
- To calculate quantitative changes in gene expression level between experimental groups and explain the results.

Schedule

Friday 16th May

09:15-12:00 RNA-seq tutorial I: data set-up, Count Normalization and PCA

Monday 19th May

13:15 – 16:00:RNA-seq tutorial II: Differential expression analysis with EdgeR

You can get all the data we will use for RNA analysis in Zilan Wen's GitHub

https://github.com/zilanwen/Introduction-of-RNA_seq.

You can find the work on the web page:

https://zilanwen.github.io/Introduction-of-RNA_seq/

Assignment

- A data analysis report with discussion on the group data.

Option 1: to push all the files into GitHub cloud as public repository and create a GitHub page site which links to your knit to html file from R markdown. All the files in your repository include results, data, R markdown file named index.Rmd, the html file named index.html

Option 2: to write data analysis report (.Doc or .PDF format) with the result description and discussion with supplementary figures, tables, code file.

- Deadline: 02.06.2025
- Sent email to zilan.wen@helsinki.fi

Assessment Criteria

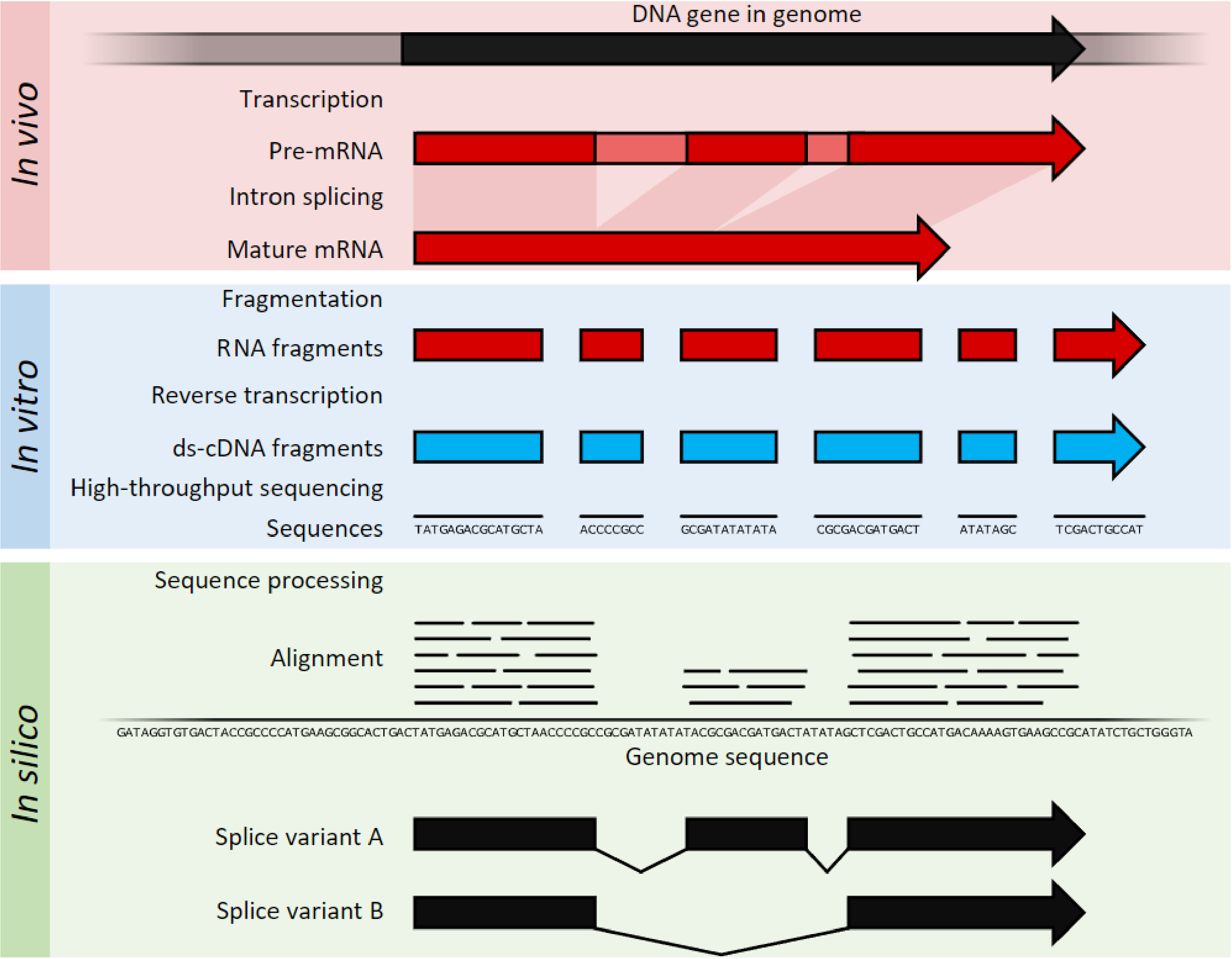
Assessment task	poor	good	excellent
Differential gene expression analysis	The process of differential gene expression analysis is not completed. The application of different count normalization methods is unclear. The quality control at experiment level is not clear. The selection of differentially expressed genes is failed. R code is provided without any instruction. Tables and figures are produced without any explanation. Data analysis report is missing. Group report is missing	The process of differential gene expression analysis is completed. The differences among count normalization methods are well understood. The quality control at experiment level is clear. The selection of differentially expressed genes is successful. R code is provided with simple instructions. Tables and figures are produced with simple descriptions. Group report is provided with figures or tables with simple description.	The process of differential gene expression analysis is completed. The differences among count normalization methods are well understood. The quality control at experiment level is clear. The selection of differentially expressed genes is successful. R code is provided with detailed instructions. Tables and figures are produced with constructive descriptions. Group report is provided with figures or tables. The discussion about the full data set is provided

What is RNA sequencing?

- RNA-Seq is a technique that can examine the quantity and sequences of RNA in a sample using next-generation sequencing (NGS).
- Total cellular content of RNAs including mRNA, rRNA and tRNA.
- It analyses the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent in a specific cell or tissue type at a distinct time.
- RNA-seq data uses short reads of mRNA which is free of intronic non-coding DNA. These reads must then be aligned back to the reference genome.

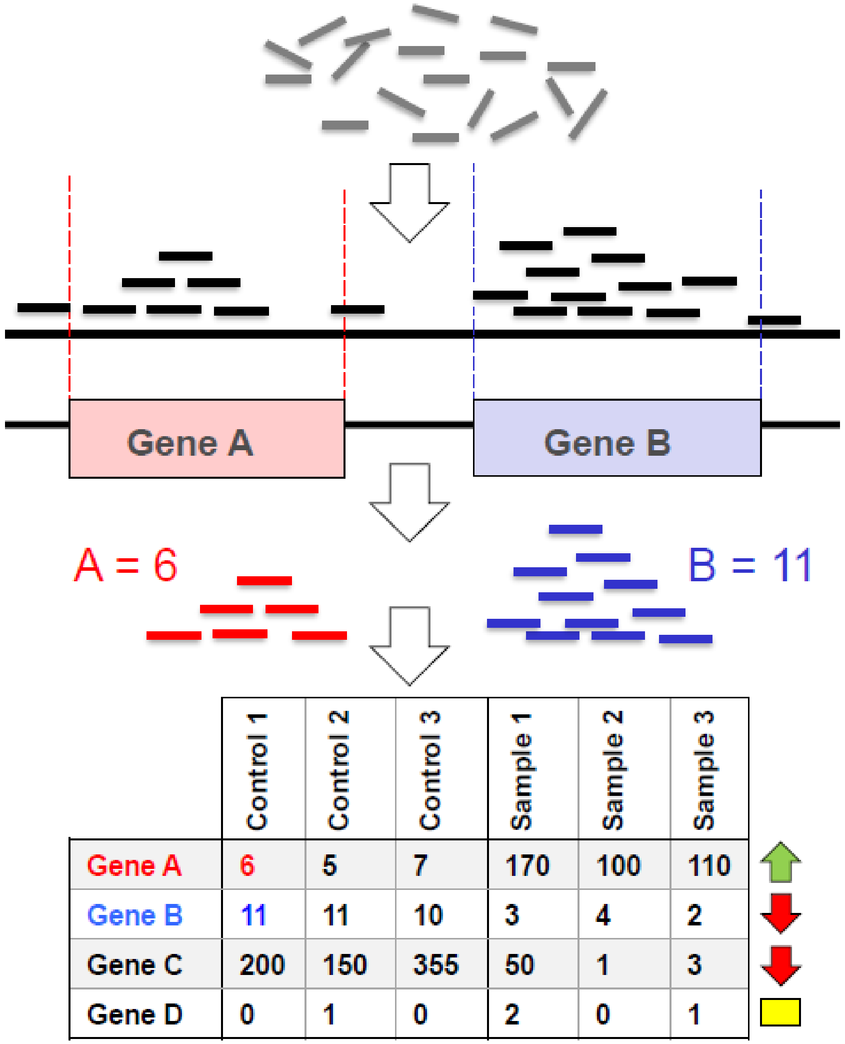
RNA-seq workflow

1. RNA extraction
2. Enrich for RNA by PolyA purification.
3. Fragmentation and size selection.
4. Reverse transcribe RNA into double-stranded cDNA.
5. PCR amplification
6. Sequencing
7. Alignment to reference genome



https://en.wikipedia.org/wiki/RNA-Seq

Differential Gene Expression (DGE) analysis workflow



	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3	
Gene A	6	5	7	170	100	110	⬆
Gene B	11	11	10	3	4	2	⬇
Gene C	200	150	355	50	1	3	⬇
Gene D	0	1	0	2	0	1	🟨

1.Raw data (reads)

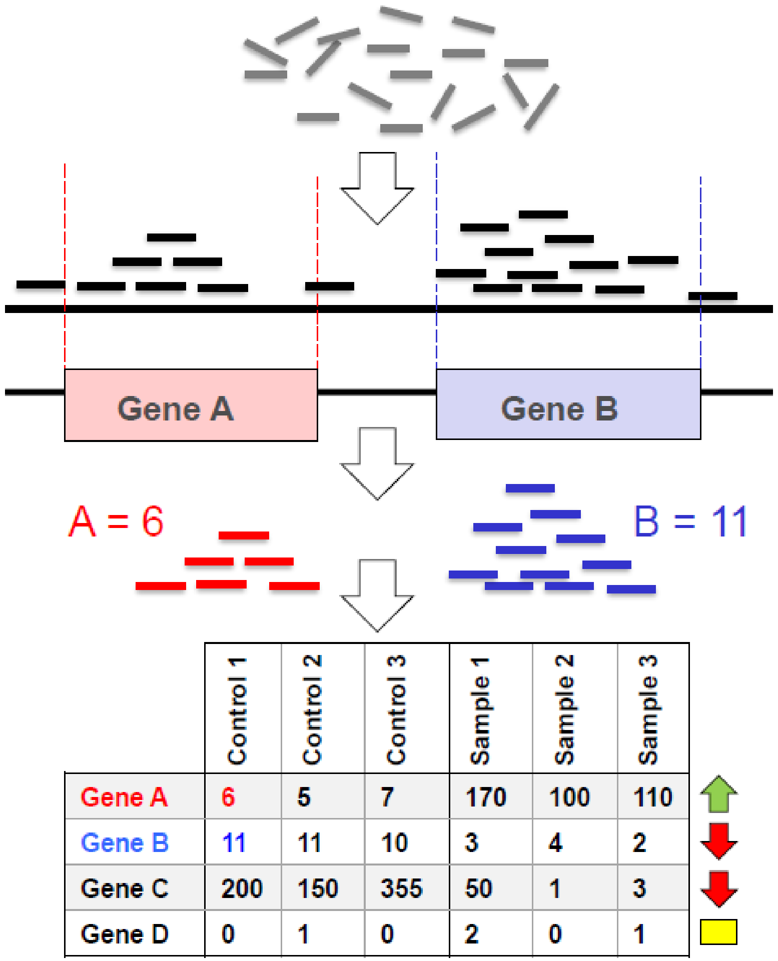
2. Align reads to reference genome

3. Match alignment positions with known gene positions

4. Count how many reads each gene has

5. Compare sample groups: differential expression analysis

DGE analysis: steps, tools and files



	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3	
Gene A	6	5	7	170	100	110	⬆
Gene B	11	11	10	3	4	2	⬇
Gene C	200	150	355	50	1	3	⬇
Gene D	0	1	0	2	0	1	🟨

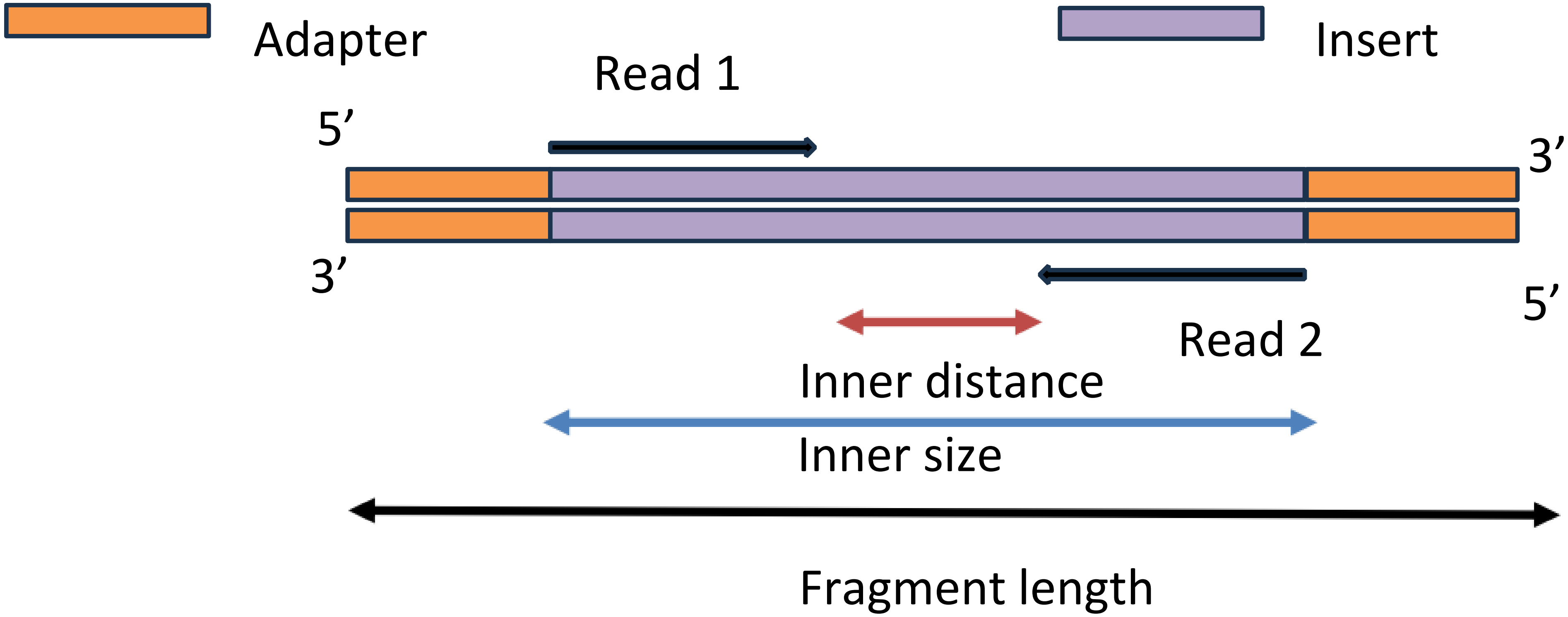
CSV file: comma-separated value” file

Step	Tools	File
Quality control	FastQC	FASTQ
Pre-processing	Trimmomatic	FASTQ
Alignment	HISAT2, BOWTIE2	BAM
Alignment quality control	RseQc	
Count quantitation	HTSeq	Read count table (.tsv, .txt. .csv)
Combine counts files to table	Define NGS experiemnt	Read count table
Quality control	PCA, clustering	
Differential expression analysis	DESeq2; EdgeR	Gene list (.csv)

How was the FASTQ files produced

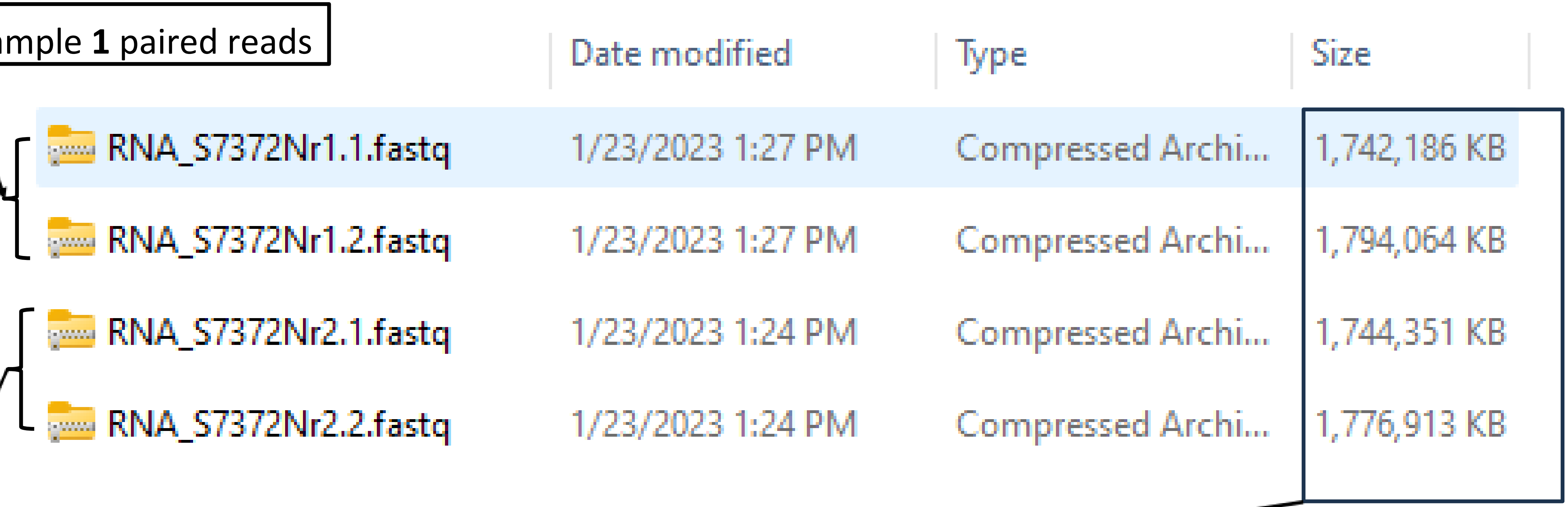
Sequencing Illumina

- Illumina reads are always of same length
- SE-Single ende dataset => Only read 1
- PE-Paried end dataset => Read 1 + Read 2



Raw reads – stored as FASTQ files

Sample **1** paired reads



	Date modified	Type	Size
RNA_S7372Nr1.1.fastq	1/23/2023 1:27 PM	Compressed Archi...	1,742,186 KB
RNA_S7372Nr1.2.fastq	1/23/2023 1:27 PM	Compressed Archi...	1,794,064 KB
RNA_S7372Nr2.1.fastq	1/23/2023 1:24 PM	Compressed Archi...	1,744,351 KB
RNA_S7372Nr2.2.fastq	1/23/2023 1:24 PM	Compressed Archi...	1,776,913 KB

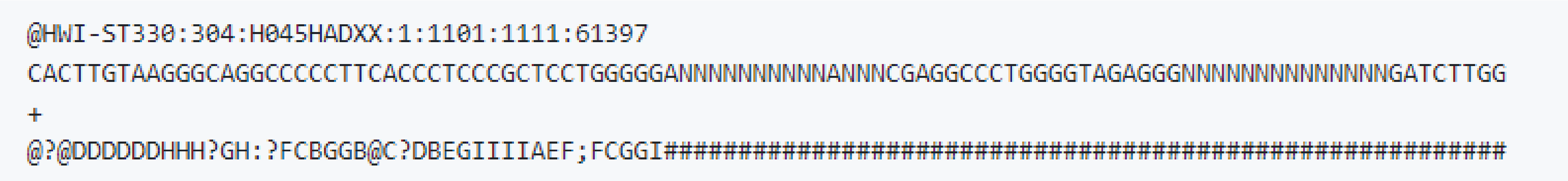
Sample **2** paired reads

Large size: Do not unzip FASTQ files

FASTQ file format

Raw reads: FASTQ file format

Each read is represented by 4 lines as shown below:



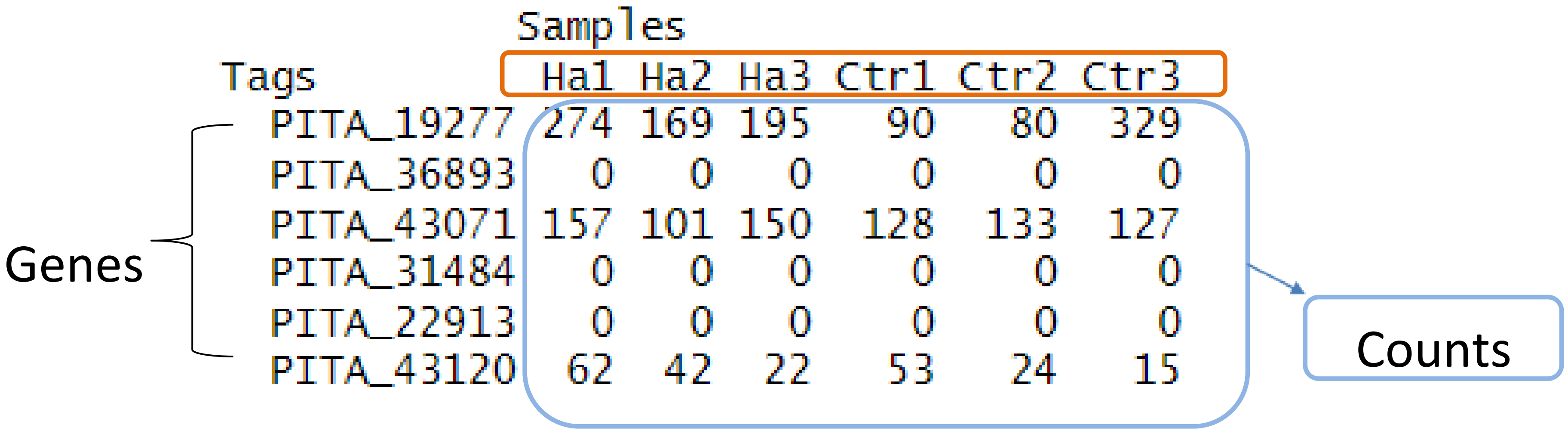
```
@HWI-ST330:304:H045HADXX:1:1101:1111:61397
CACTTGTAAGGGCAGGCCCCCTTCACCCTCCCGCTCCTGGGGGANNNNNNNNNNANNNCGAGGCCCTGGGGTAGAGGGNNNNNNNNNNNNNNGATCTTGG
+
@?@DDDDDDHHH?GH:?FCBGGB@C?DBEGIIIIAEF;FCGGI#########################################################
```

Line	Description
1	Always begin with '@' followed by sequence identifier and an optional description
2	The actual raw sequence letters
3	Always begins with a '+' and sometimes the same info as in line 1
4	Has a string of characters which represent the quality scores; must have same number of characters as line 2

More details about FASTQ format: https://en.wikipedia.org/wiki/FASTQ_format

Count matrix

Count matrix: a matrix summarizing the gene-level expression in each sample of your dataset.



```
             Samples
Tags          Ha1 Ha2 Ha3 Ctr1 Ctr2 Ctr3
  PITA_19277 274 169 195   90   80  329
  PITA_36893   0   0   0    0    0    0
  PITA_43071 157 101 150  128  133  127
  PITA_31484   0   0   0    0    0    0
  PITA_22913   0   0   0    0    0    0
  PITA_43120  62  42  22   53   24   15
```

Count Normalization

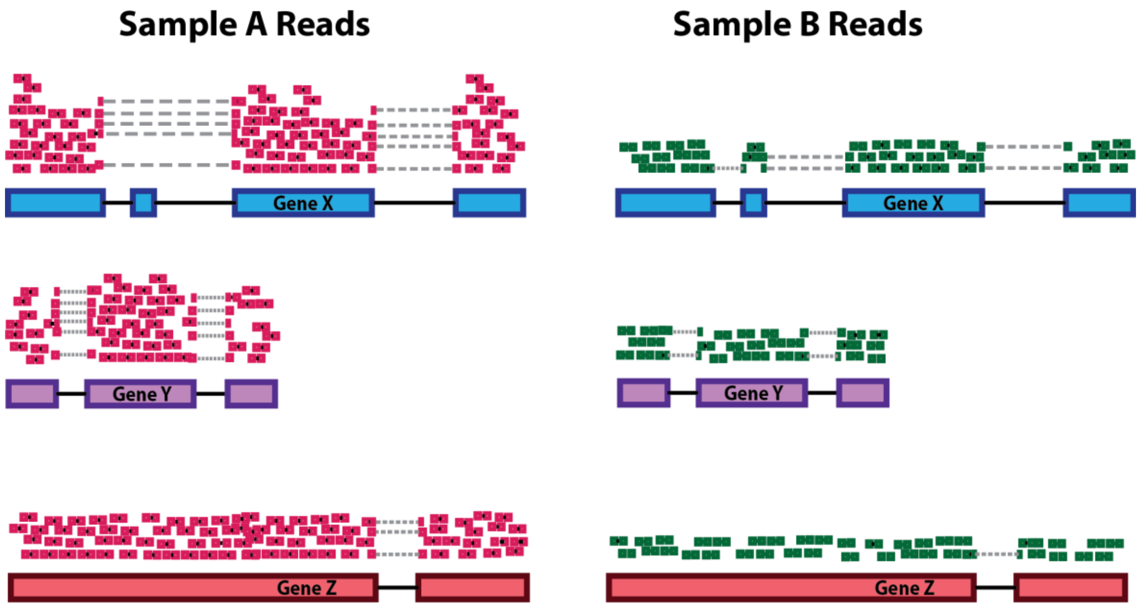
Normalization is the process of adjusting raw count values to account for the "uninteresting" factors. In this way the expression levels are more comparable between and/or within samples.

The main "uninteresting" factors often considered during normalization are:

- Sequencing depth
- Gene length
- RNA composition

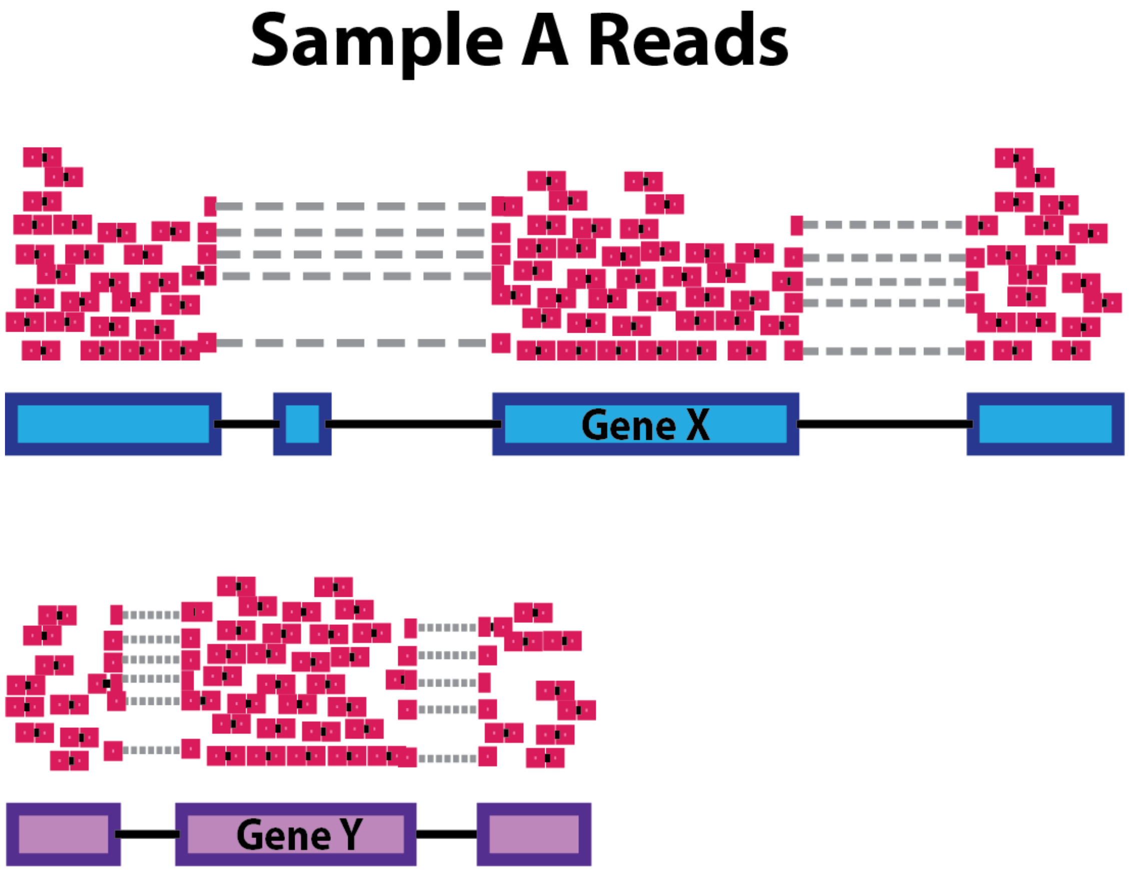
Sequencing depth

- Accounting for sequencing depth is necessary for comparison of gene expression between samples.
- In the example right, each gene appears to have doubled in expression in Sample A relative to Sample B, however this is a consequence of Sample A having double the sequencing depth.

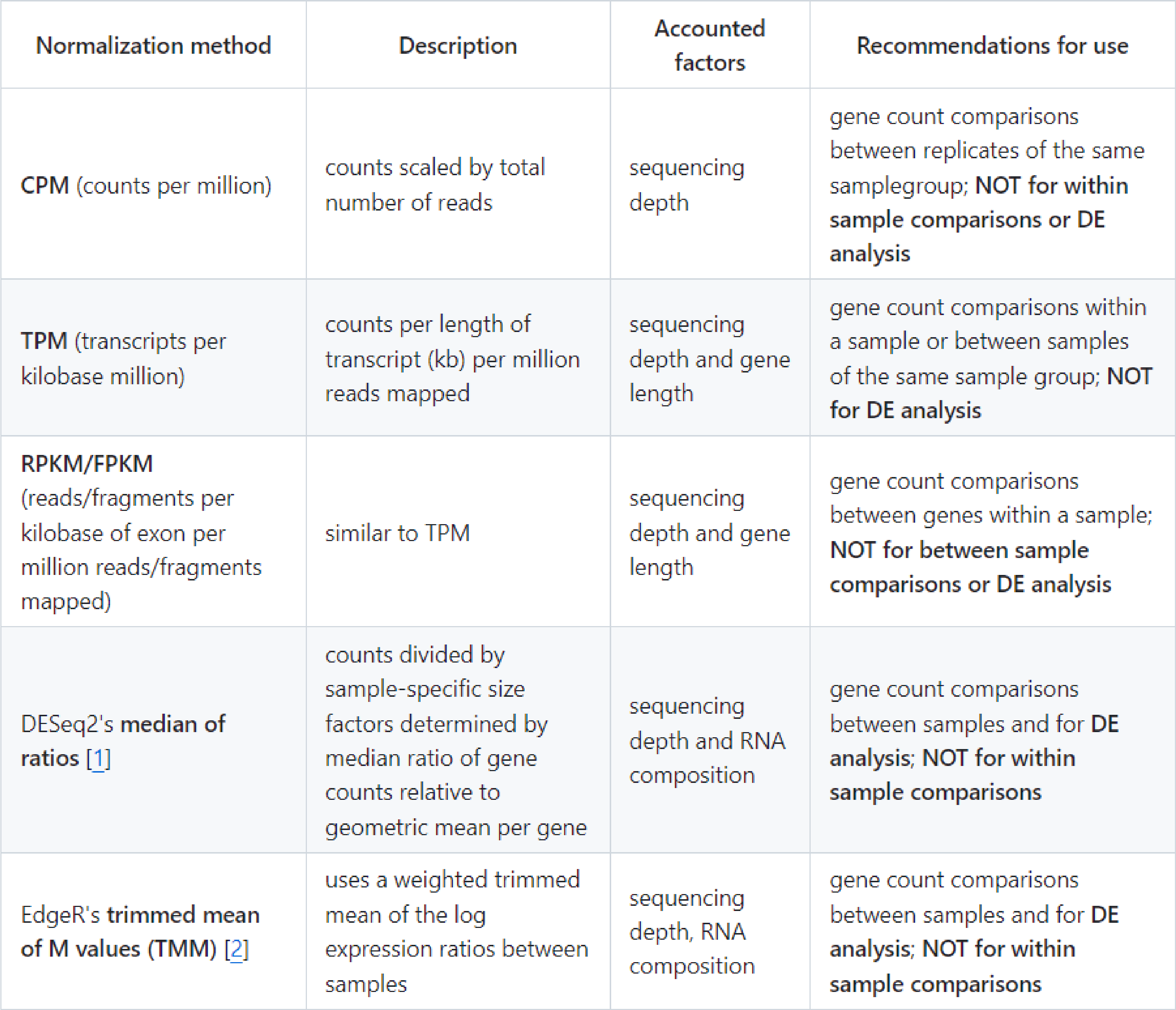


Gene length

- Accounting for gene length is necessary for comparing expression between different genes within the same sample.
- In the example, Gene X and Gene Y have similar levels of expression, but the number of reads mapped to Gene X would be many more than the number mapped to Gene Y because Gene X is longer.



Common normalization methods



Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis; NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition	gene count comparisons between samples and for **DE analysis; NOT for within sample comparisons**

RPKM, FPKM and TPM

- RPKM: Reads per kilobase Million
- FRKM: Fragments Per Kilobase Million
- TPM: Transcripts per Million

The normalized reads counts for:

1) Sequencing depth: Sequencing runs with more depth will get more reads mapping to each gene.
2) Gene length: longer genes will have more reads mapping to them

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	10	12	30
B (4 KB)	20	25	60
C (1 KB)	5	8	15
D (10KB)	0	0	1

- Rep3 has more reads than the other replicates regardless of genes which mean it has higher sequencing depth.
- GeneB is twice as long as Gene A, resulting in getting twice as many as reads regardless of replicates.

RPKM-Step1: normalize for sequencing depth

Raw read counts

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	10	12	30
B (4 KB)	20	25	60
C (1 KB)	5	8	15
D (10KB)	0	0	1
Total reads	35	45	106
Tens of reads	3.5	4.5	10.6

Reads per million

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	2.86	2.67	2.83
B (4 KB)	5.71	5.56	5.66
C (1 KB)	1.43	1.78	1.42
D (10KB)	0	0	0.09

Regard as ´per million’ scaling factors, GeneA reads per million: 10/3.5=2.86

Here we only have 4 genes as an example, we scale the total reas counts by 10 instead of by 1 million. If we have ten thousands genes, we will scale by 1 million.

RPKM-Step2: normalize for gene length

Raw read counts

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	10	12	30
B (4 KB)	20	25	60
C (1 KB)	5	8	15
D (10KB)	0	0	1

Reads per million

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	2.86	2.67	2.83
B (4 KB)	5.71	5.56	5.66
C (1 KB)	1.43	1.78	1.42
D (10KB)	0	0	0.09

→

Scale per Kb

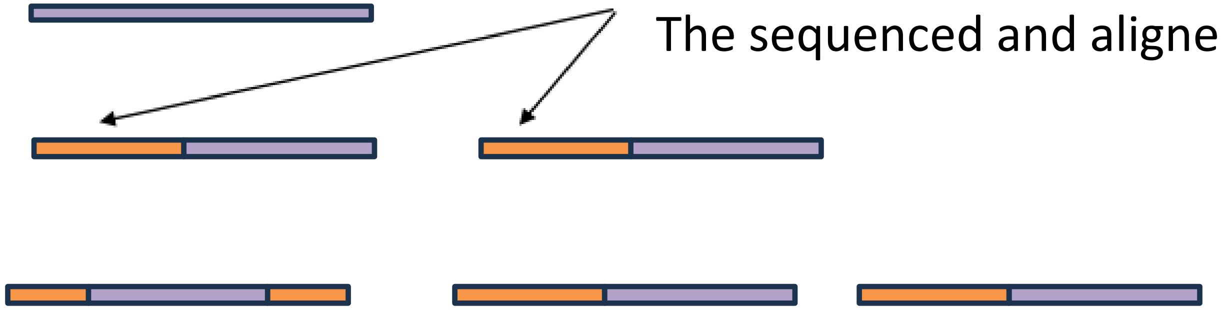
Reads per million per KB

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	1.43	1.33	1.42
B (4 KB)	1.43	1.39	1.42
C (1 KB)	1.43	1.78	1.42
D (10KB)	0	0	0.009

RPKM and FPKM – two very closely related

RPKM: Reads per kilobase Million ---- >> for single end RNA-Seq

FRKM: Fragments Per Kilobase Million ---->> for paired end RNA-Seq



A fragment to be sequenced:

The sequenced and aligned *reads*

Single end sequencing:

Paried end sequencing:

Both ends can map, giving you
two reads per fragment, or ...

Sometiones only one end of the
paried end has a quality tead
and maps

TPM-Step1: normalize for gene length

Raw read counts

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	10	12	30
B (4 KB)	20	25	60
C (1 KB)	5	8	15
D (10KB)	0	0	1

Reads scaled by gene length

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	5	6	15
B (4 KB)	5	6.25	15
C (1 KB)	5	8	15
D (10KB)	0	0	0.1

Total RPK:	15	20.25	45.1
Tens of RPK:	1.5	2.025	4.51

Reads scaled by gene length and sequence depth

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	3.33	2.96	3.326
B (4 KB)	3.33	3.09	3.326
C (1 KB)	3.33	3.95	3.326
D (10KB)	0	0	0.02

RPKM VS TPM

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

Gene name	Rep1 counts	Rep2 counts	Rep3 counts
A (2KB)	3.33	2.96	3.326
B (4 KB)	3.33	3.09	3.326
C (1 KB)	3.33	3.95	3.326
D (10KB)	0	0	0.02
Total:	10	10	10

With RPKM, we get a **different value** for each sample column. While we get a **same value** for each sample column with TPM.

Why is rthe same value is important?
You can get answer from the video (https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/). This is the reason why RPKM/FPKM is not recommended for between sample comparisons

DESeq2-normalized counts: Median of ratios method

- **Gene length** does not need to be accounted for, because differential expression analysis are comparing the counts of the same gene between sample groups.
- However, **sequencing depth** and **RNA composition** do need to be taken into account.
- To normalize for sequencing depth and RNA composition, DESeq2 uses the median of ratios method.

DESeq2-normalized counts: Median of ratios method

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

For each gene, a pseudo-reference sample is created that is equal to the geometric mean across all samples.

Gene	Ha1	Ha2	Ha3	pseudo-reference sample
PITA_19277	274	169	195	POWER(274*169*195,1/3) = 208.24
PITA_43071	157	101	150	133.49
....	...	..	..	..

Step 2: calculates ratio of each sample to the reference

=274/208.24

=169/208.24

	Gene	Ha1	Ha2	Ha3	pseudo-reference sample	ratio of Ha1/ref	ratio of Ha2/ref	ratio of Ha3/ref
1	PITA_19277	274	169	195	208.24	1.32	0.81	0.94
2	PITA_43071	157	101	150	133.49	1.18	0.76	1.12
3	PITA_43120	62	42	22	38.55	1.61	1.09	0.57
4	PITA_19387	901	582	778	741.67	1.21	0.78	1.05
5	PITA_38886	42	48	67	51.31	0.82	0.94	1.31
6	PITA_15089	469	289	347	360.97	1.30	0.80	0.96
7	PITA_43425	2580	1697	2337	2170.97	1.19	0.78	1.08

Step 3: calculate the normalization factor for each sample (size factor)

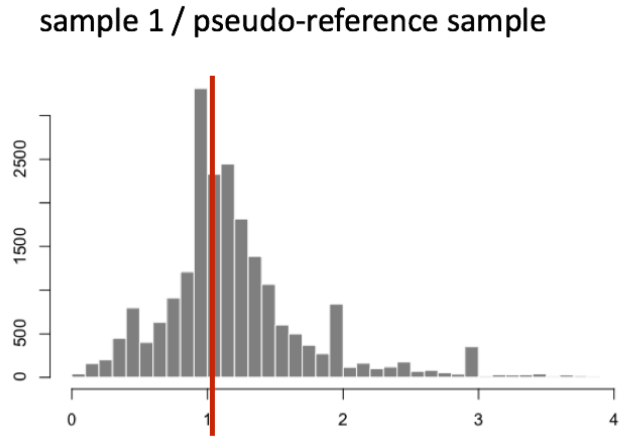
The median value (column-wise for the above table) of all ratios for a given sample is taken as the normalization factor (size factor) for that sample, as calculated below. Notice that the differentially expressed genes should not affect the median value:

```
normalization_factor_Ha1 <- median(c(1.61, 1.32, 1.30, 1.21, 1.19, 1.18, 0.82))
normalization_factor_Ha2<- median(c(1.09, 0.94, 0.81, 0.80, 0.78, 0.78, 0.76))
normalization_factor_Ha3<- median(c(1.31, 1.12, 1.08, 1.05, 0.96, 0.94, 0.57))
```

Ha1 median ratio = 1.21

Ha2 median ratio = 0.80

Ha3 median ratio = 1.05



Step 3: calculate the normalized count values using the normalization factor

Ha1 median ratio = 1.21

Ha2 median ratio = 0.80

Ha3 median ratio = 1.05

Raw Counts

Gene	Ha1	Ha2	Ha3
PITA_19277	274	169	195
PITA_43071	157	101	150
…	…	…	…

Normalized Counts

Gene	Ha1	Ha2	Ha3
PITA_19277	274/1.21=**226.44**	169/0.80=**211.25**	195/1.05=**185.71**
PITA_43071	**129.75**	**126.25**	**142.86**
…	…	…	…

Experiment-level Quality Control

- Getting an overview of similarities and dissimilarities between samples allows you to check

1) How well our replicates cluster together.
2) whether our experimental condition represents the major source of variation in the data.
3) Are there sample outliers that should be removed.

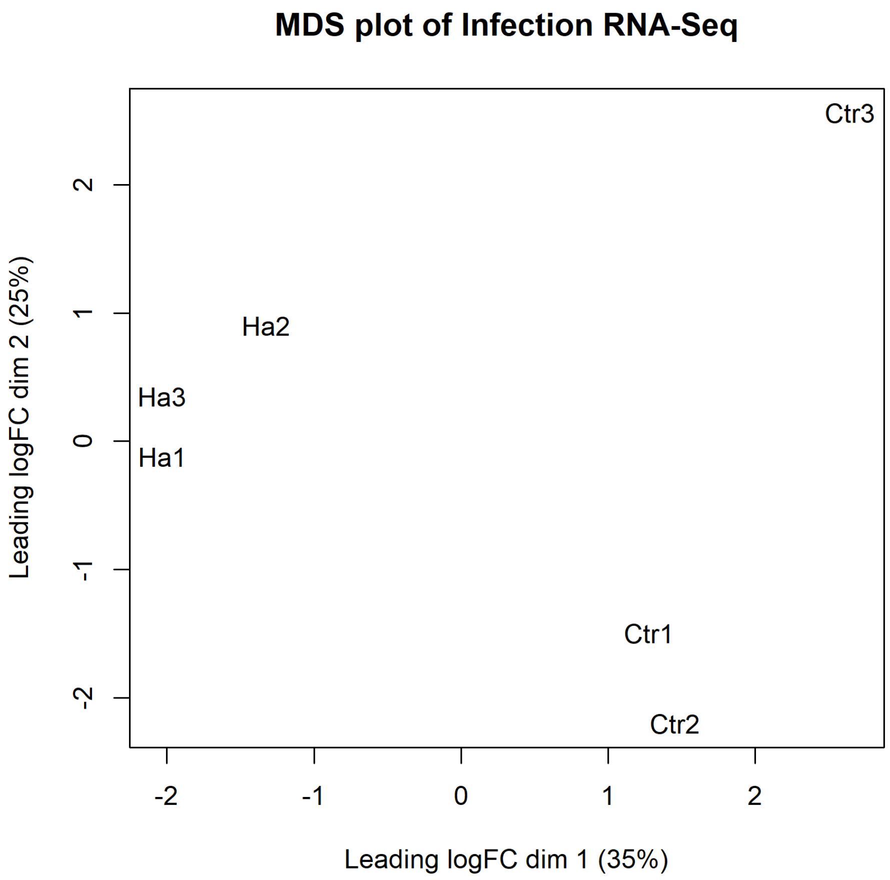
- Several methods available

1) MDS (multidimensional scaling)
2) PCA (principal component analysis
3) Clustering

MDS plot by edgeR

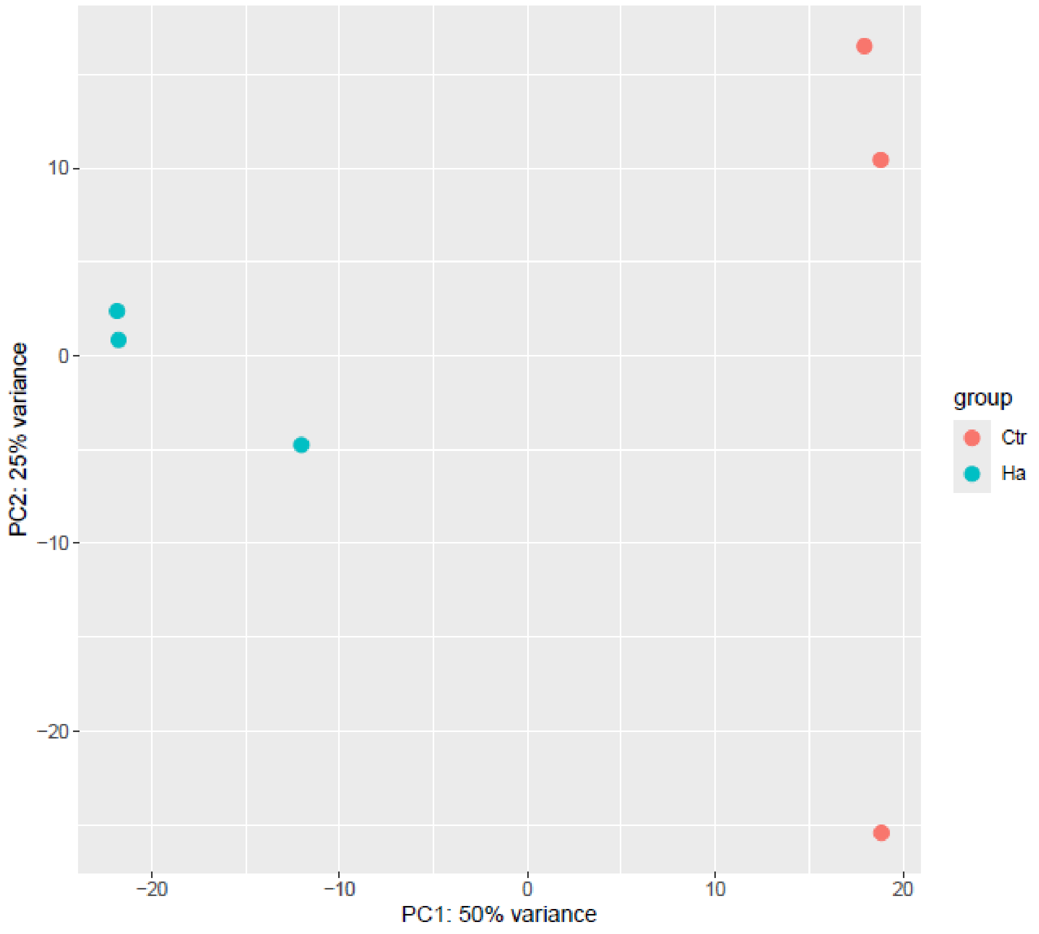
- Distances correspond to the logFC or biological coefficient of variation (BCV) between each pair of samples

- Calculated using 500 most heterogenous genes (that have largest dispersion when treating all samples as one group)



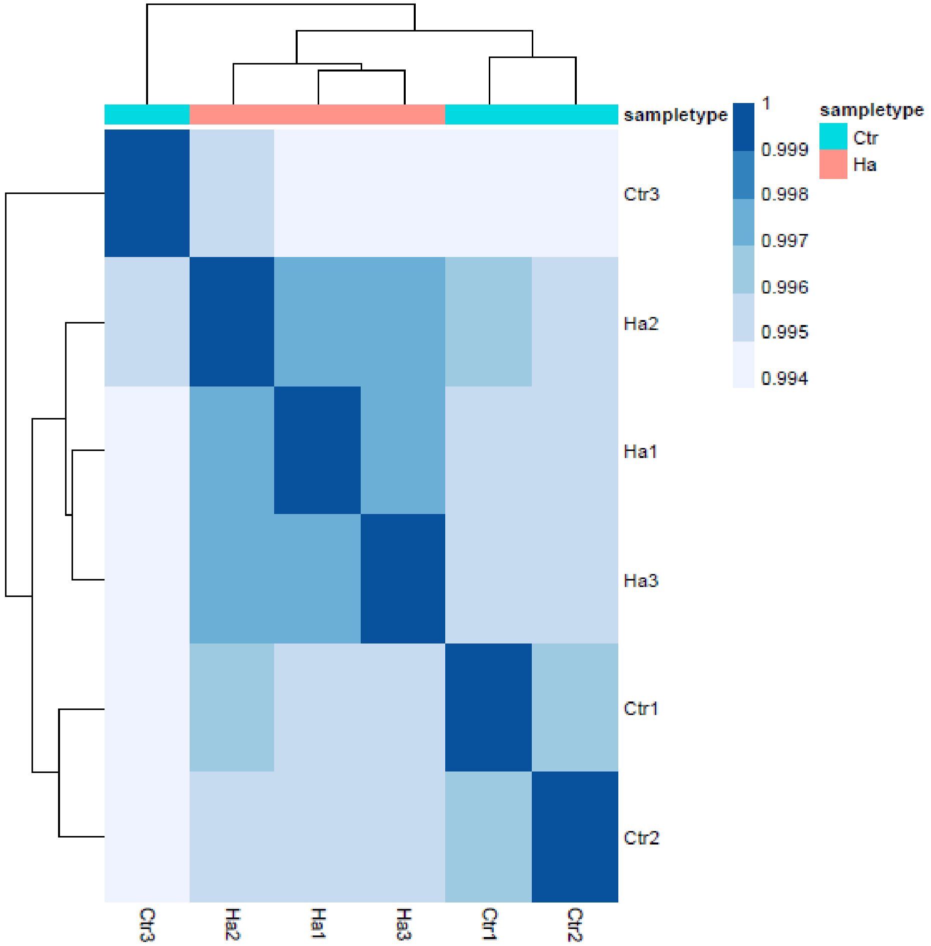
PCA plot by DESeq2

- The first two principal components , calculated after variance stabilizing transformation
- Indicates the proportion of variance explained by each component
- If PC2 explains only a small percentage of variance, it can be ignored



Sample Heatmap by DESeq2

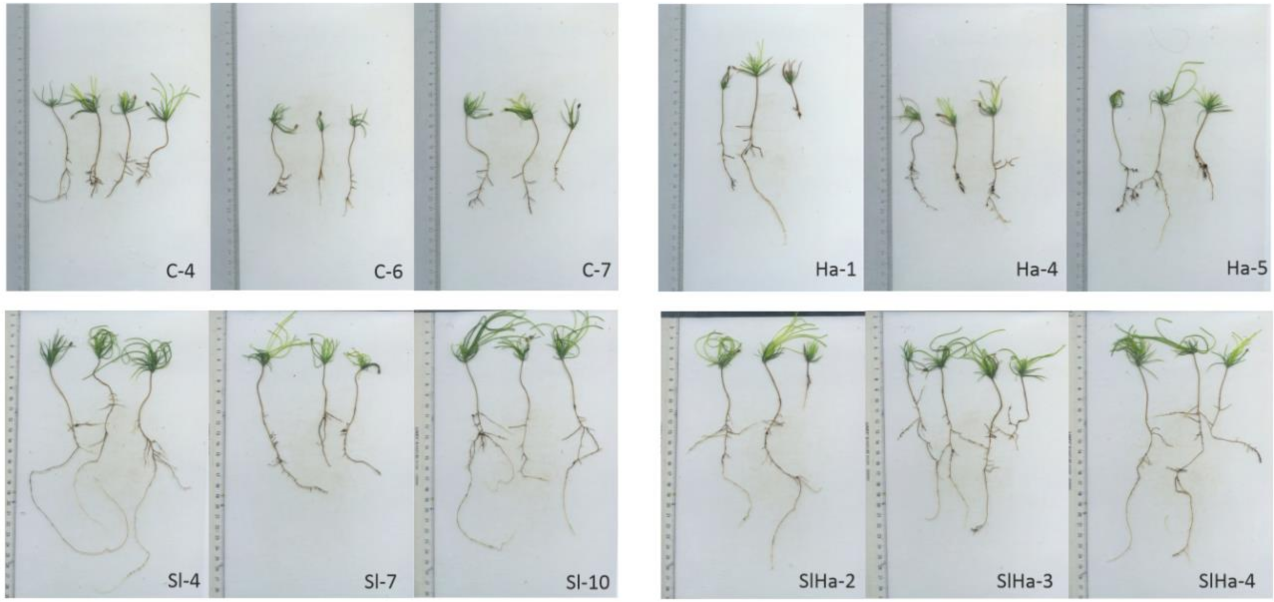
- Euclidean distances between the samples , calculated after variance stabilizing transformation



Now let’s play RNA-seq data

- *Heterobasidion annousm* is pathogen, *Suillus luteus* is beneficial mutualistic fungus.
- We have three treatment groups and one control group.

(1) Control Scots pine seedlings without any inoculum (C).



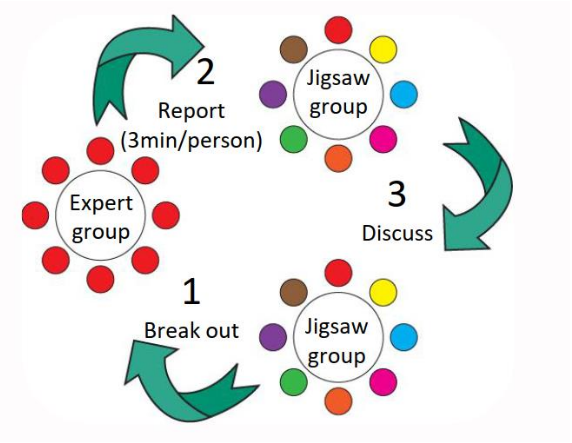
(2) Mutualistic fungus-inoculated seedlings (Sl)

(3) Pathogen-infected seedlings (Ha)

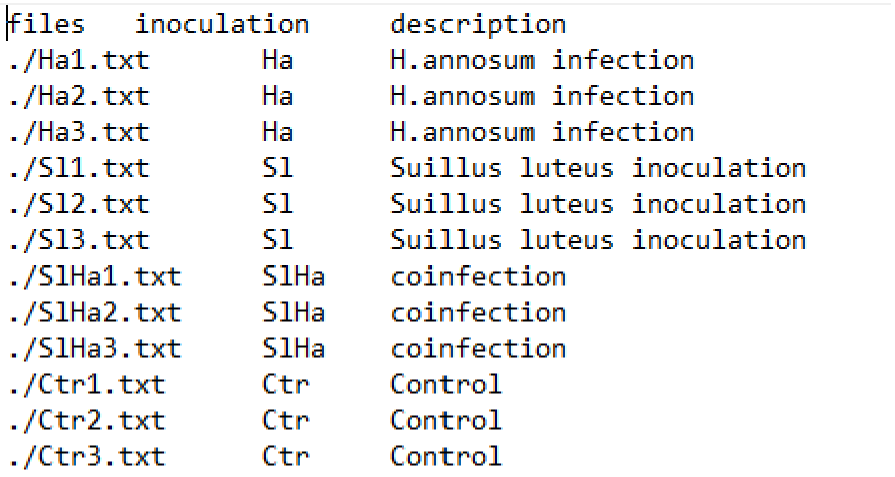
(4) Co-inoculated seedlings with both the ECM fungus and the pathogen

How to play the RNA-seq data

- Each .txt file has two columns with gene IDs and raw read counts.
- We will have two student groups (A, B) and each group have three students. Each student choose different treatment group (Ha, Sl, Coinfection).
- Students work on same dataset can work together.
- After finish the data analysis, back to home group and compare the data together.
- You can get all the data we will use for RNA analysis in my GitHubhttps://github.com/zilanwen/Introduction-of-RNA_seq.



- You can find the work on the web page:https://zilanwen.github.io/Introduction-of-RNA_seq/



```
files   inoculation     description
./Ha1.txt       Ha      H.annosum infection
./Ha2.txt       Ha      H.annosum infection
./Ha3.txt       Ha      H.annosum infection
./Sl1.txt       Sl      Suillus luteus inoculation
./Sl2.txt       Sl      Suillus luteus inoculation
./Sl3.txt       Sl      Suillus luteus inoculation
./SlHa1.txt     SlHa    coinfection
./SlHa2.txt     SlHa    coinfection
./SlHa3.txt     SlHa    coinfection
./Ctr1.txt      Ctr     Control
./Ctr2.txt      Ctr     Control
./Ctr3.txt      Ctr     Control
```