RNAseq counts

Center for Health Data Science

Health Data Science Sandbox

Overview

1 Rstudio & Rmarkdown

2 Count Matrix & Normalization

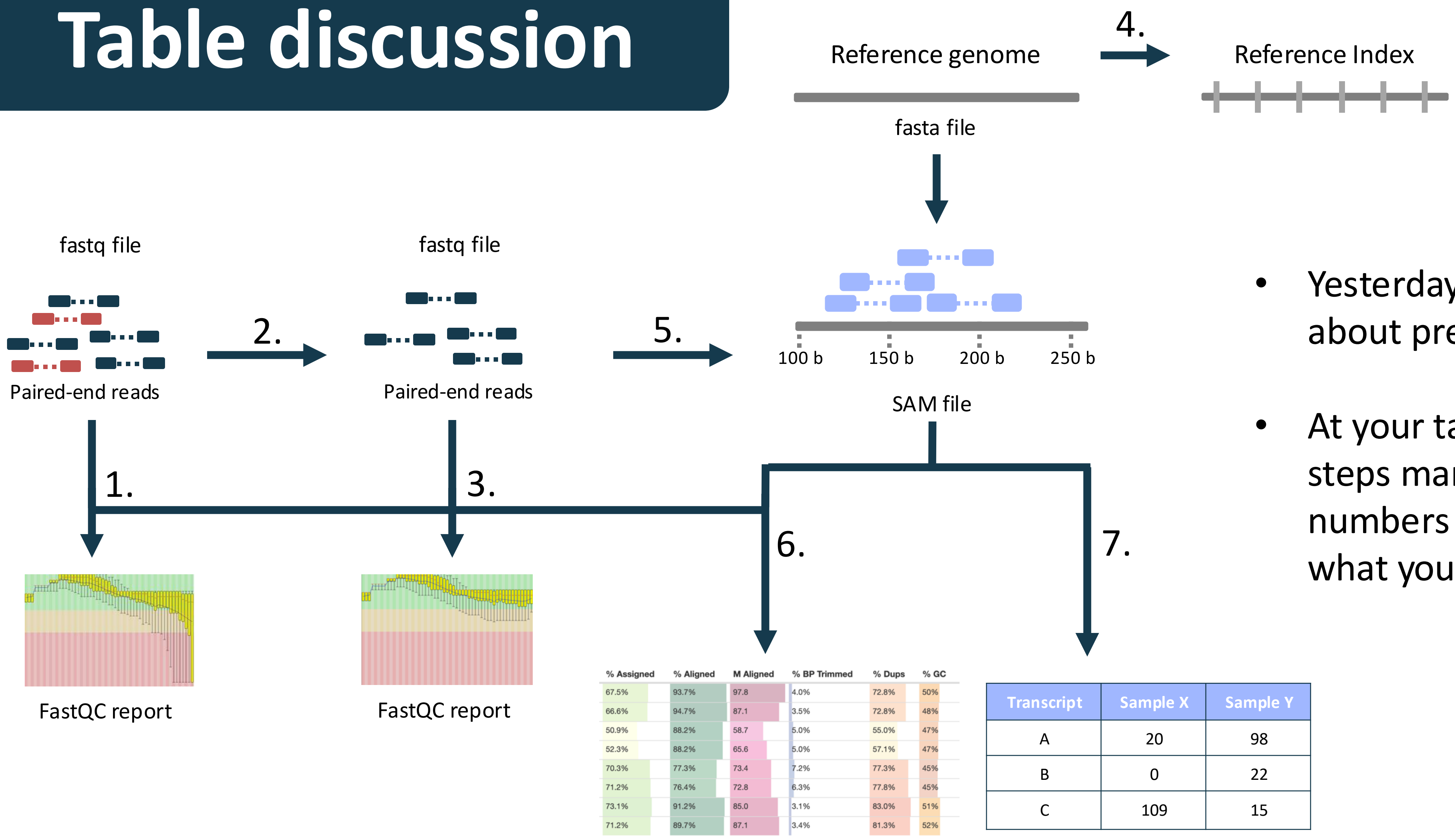
3 Exploratory Analysis

4 Differential Expression

5 Functional Analysis

Time for recap!

Table discussion



% Assigned	% Aligned	M Aligned	% BP Trimmed	% Dups	% GC
67.5%	93.7%	97.8	4.0%	72.8%	50%
66.6%	94.7%	87.1	3.5%	72.8%	48%
50.9%	88.2%	58.7	5.0%	55.0%	47%
52.3%	88.2%	65.6	5.0%	57.1%	47%
70.3%	77.3%	73.4	7.2%	77.3%	45%
71.2%	76.4%	72.8	6.3%	77.8%	45%
73.1%	91.2%	85.0	3.1%	83.0%	51%
71.2%	89.7%	87.1	3.4%	81.3%	52%

MultiQC Report

Transcript	Sample X	Sample Y
A	20	98
B	0	22
C	109	15

Count Matrix

- Yesterday we learned about pre-processing.
- At your table name the steps marked with numbers and summarize what you do.

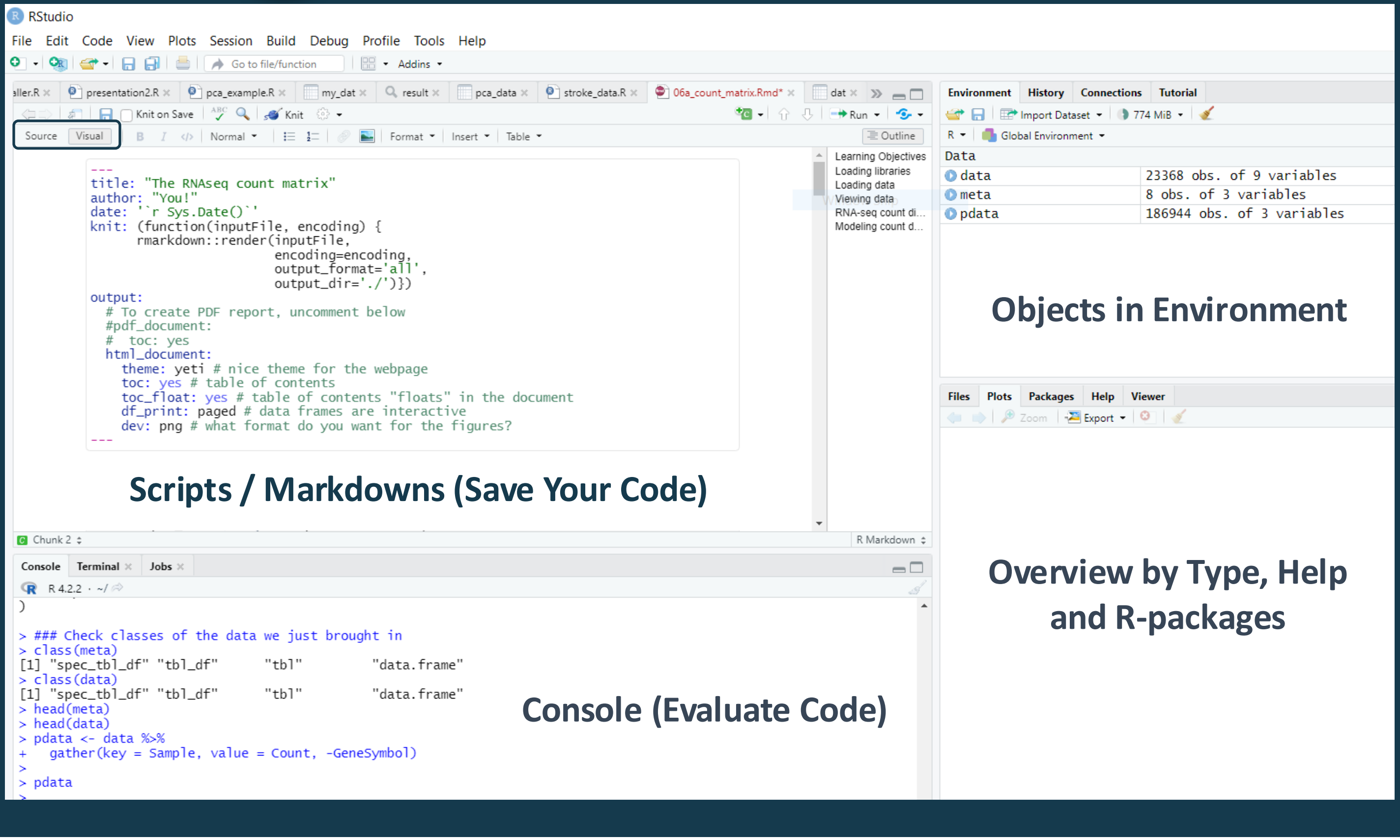
HeaDS

Rstudio and Markdown

Scripting / Programming Language

Reports (html, pdf, docx)

Studio - Code Interpreter and Editor



```
---
title: "The RNAseq count matrix"
author: "You!"
date: '`r Sys.Date()`'
knit: (function(inputFile, encoding) {
      rmarkdown::render(inputFile,
                        encoding=encoding,
                        output_format='all',
                        output_dir='./')})
output:
  # To create PDF report, uncomment below
  #pdf_document:
  #  toc: yes
  html_document:
    theme: yeti # nice theme for the webpage
    toc: yes # table of contents
    toc_float: yes # table of contents "floats" in the document
    df_print: paged # data frames are interactive
    dev: png # what format do you want for the figures?
---
```

Scripts / Markdowns (Save Your Code)

```
)

> ### Check classes of the data we just brought in
> class(meta)
[1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
> class(data)
[1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
> head(meta)
> head(data)
> pdata <- data %>%
+   gather(key = Sample, value = Count, -GeneSymbol)
>
> pdata
>
```

Console (Evaluate Code)

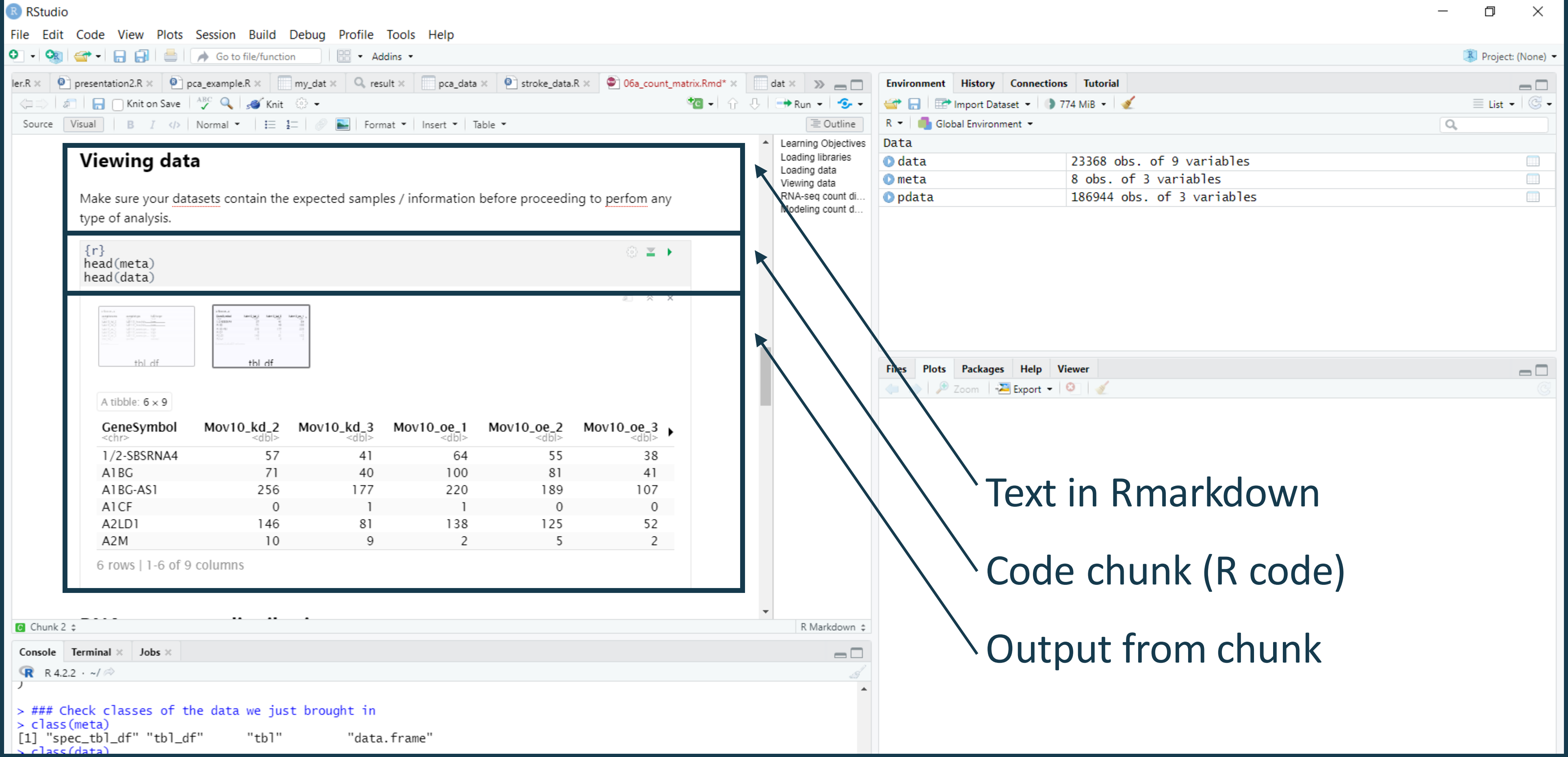
Data

data	23368 obs. of 9 variables
meta	8 obs. of 3 variables
pdata	186944 obs. of 3 variables

Objects in Environment

Overview by Type, Help and R-packages

Rstudio and Markdown



Viewing data

Make sure your datasets contain the expected samples / information before proceeding to perfom any type of analysis.

```
{r}
head(meta)
head(data)
```

tbl_df tbl_df

A tibble: 6 × 9

GeneSymbol <chr>	Mov10_kd_2 <dbl>	Mov10_kd_3 <dbl>	Mov10_oe_1 <dbl>	Mov10_oe_2 <dbl>	Mov10_oe_3 <dbl>
1/2-SBSRNA4	57	41	64	55	38
A1BG	71	40	100	81	41
A1BG-AS1	256	177	220	189	107
A1CF	0	1	1	0	0
A2LD1	146	81	138	125	52
A2M	10	9	2	5	2

6 rows | 1-6 of 9 columns

Data

data	23368 obs. of 9 variables
meta	8 obs. of 3 variables
pdata	186944 obs. of 3 variables

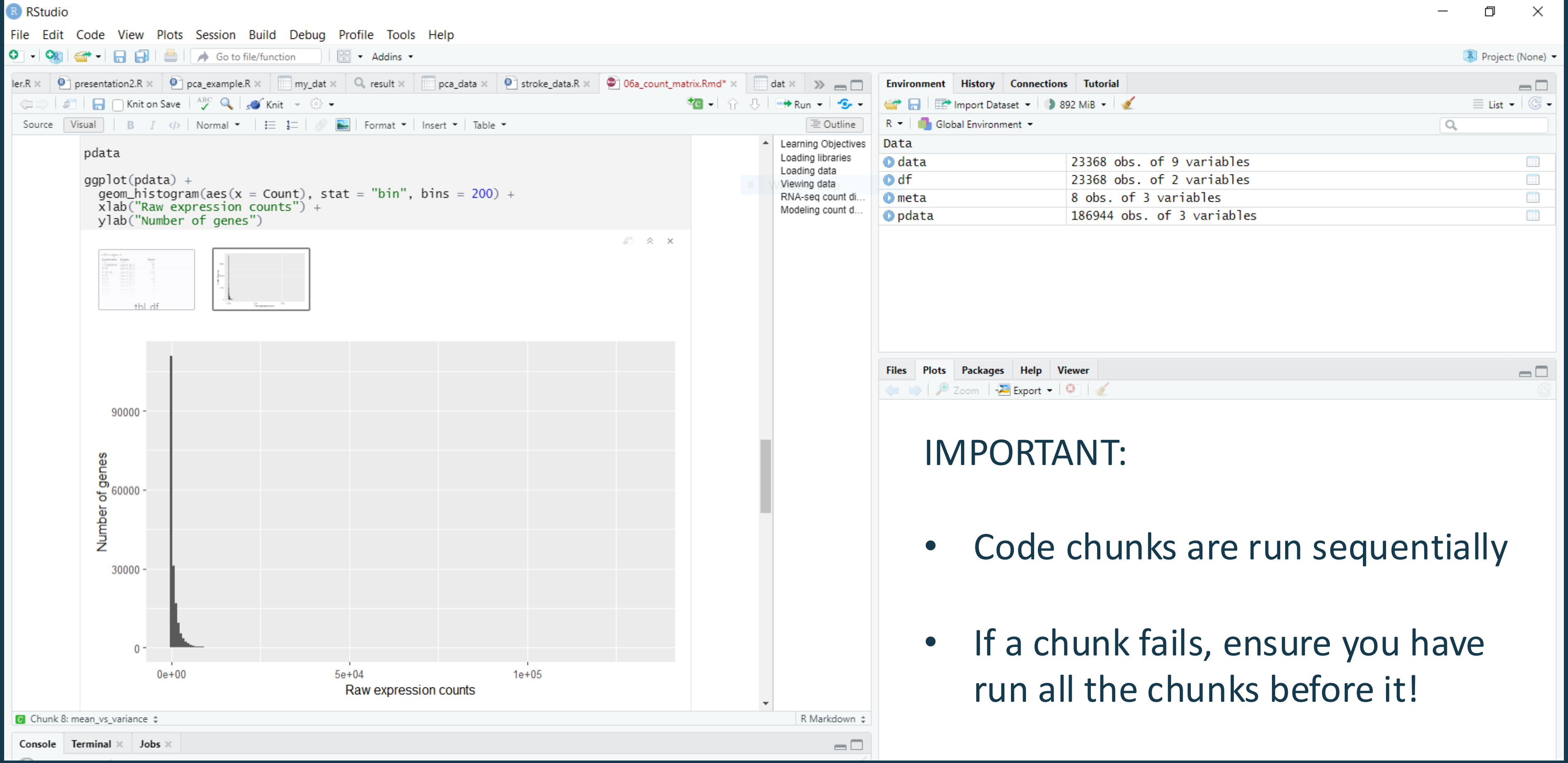
```
> ### Check classes of the data we just brought in
> class(meta)
[1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
> class(data)
```

Text in Rmarkdown

Code chunk (R code)

Output from chunk

Rstudio and Markdown



```
pdata

ggplot(pdata) +
  geom_histogram(aes(x = Count), stat = "bin", bins = 200) +
  xlab("Raw expression counts") +
  ylab("Number of genes")
```

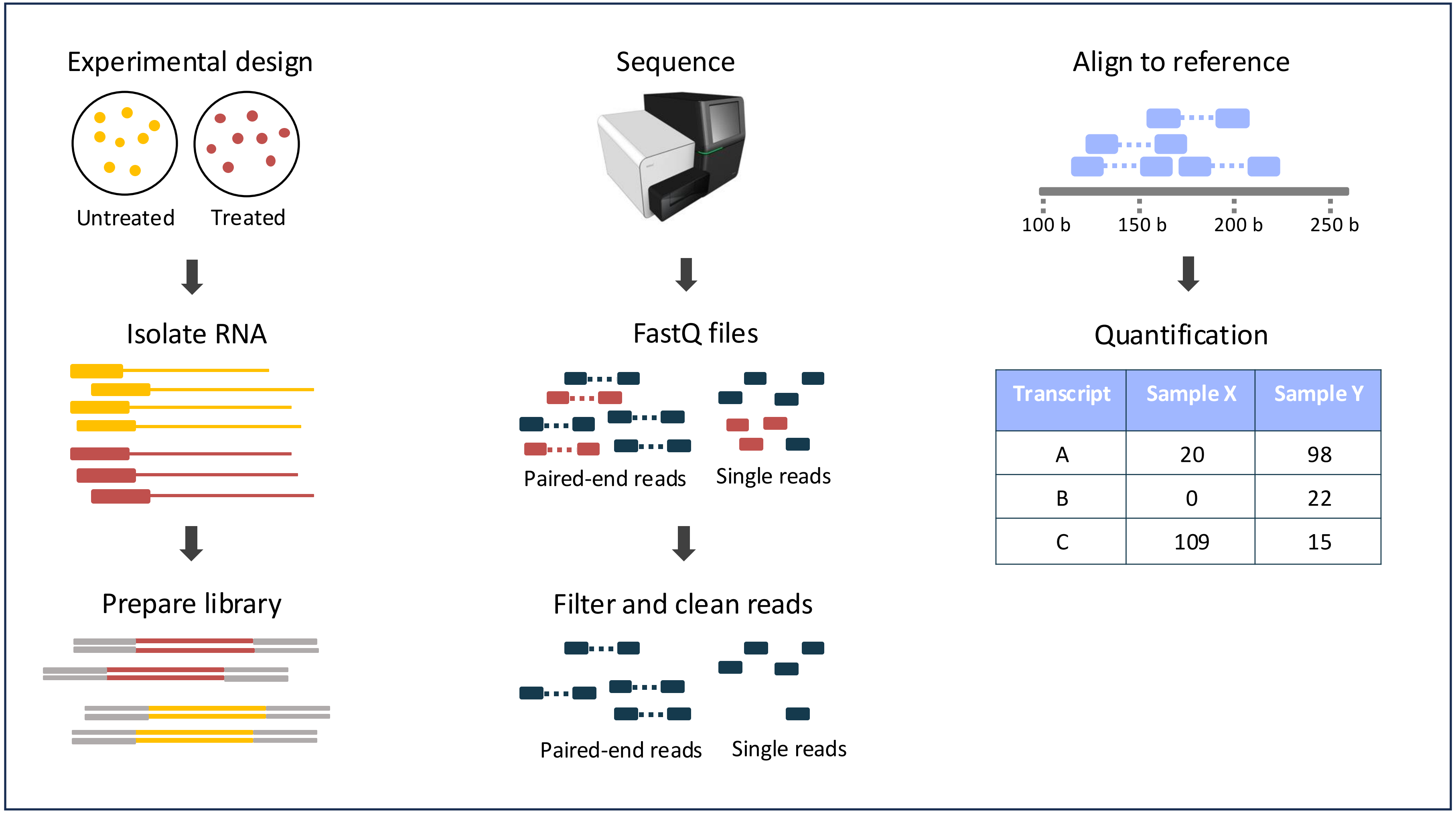
Data

```
data    23368 obs. of 9 variables
df      23368 obs. of 2 variables
meta    8 obs. of 3 variables
pdata   186944 obs. of 3 variables
```

IMPORTANT:

- Code chunks are run sequentially
- If a chunk fails, ensure you have run all the chunks before it!

Workflow



Experimental design

Untreated

Treated

Isolate RNA

Prepare library

Sequence

FastQ files

Paired-end reads

Single reads

Filter and clean reads

Paired-end reads

Single reads

Align to reference

Quantification

Transcript	Sample X	Sample Y
A	20	98
B	0	22
C	109	15

Count matrix

- Understand the output of your pre-processing
- How can we model gene counts?
- Gene expression biases
- How do we normalize our count matrix?

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A	10	12	30
B	20	25	60
C	5	8	15
D	0	0	1
Total counts	**35**	**45**	**106**

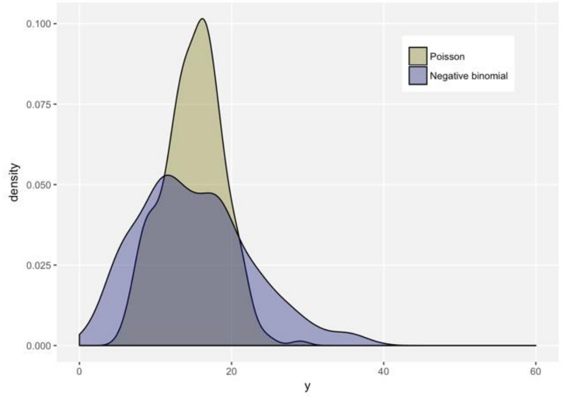
Raw count matrix

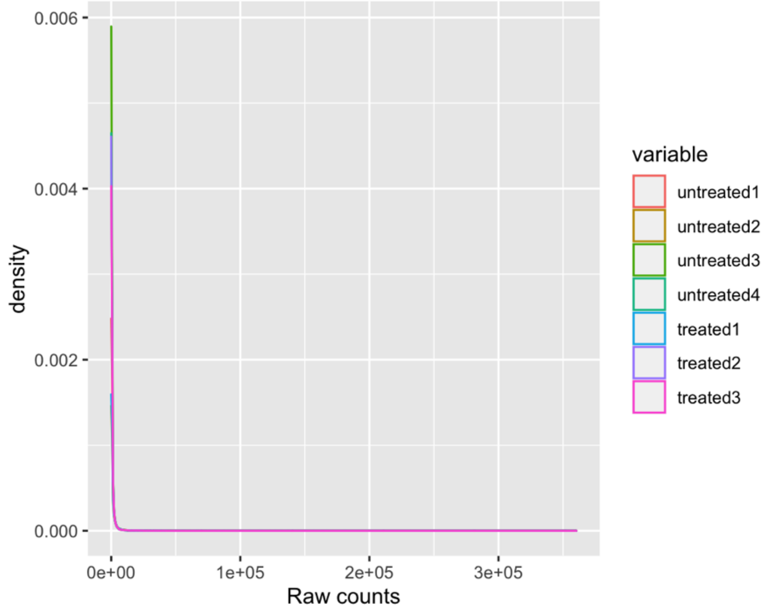
Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A	10	12	30
B	20	25	60
C	5	8	15
D	0	0	1
Total counts	**35**	**45**	**106**

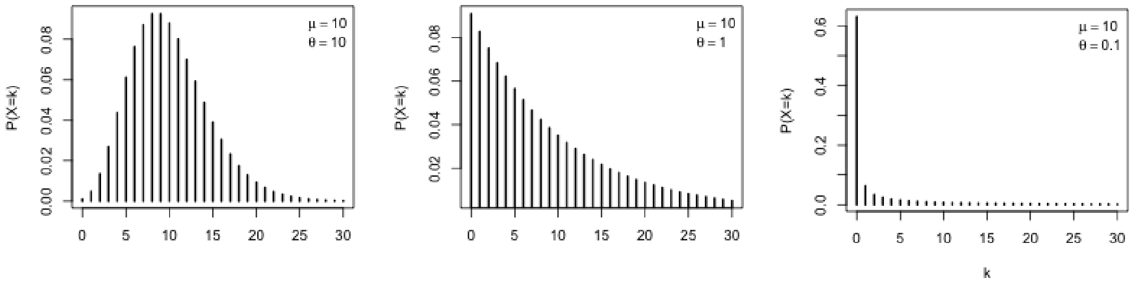
- Is the expression of gene A higher in Rep3 than Rep1?
- Is the expression of gene A higher than gene B in Rep1?
- Can you directly compare genes and reps in this table? Why / why not?

Raw count matrix

- Distribution of RNAseq count data:
  - Model with a Poisson distribution (PD)?
  - PD assumes *mean == variance*, count distributions are overdispersed! **Negative binomial distribution.**







Raw count matrix

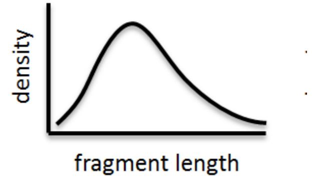
Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A	10	12	30
B	20	25	60
C	5	8	15
D	0	0	1
Total counts	**35**	**45**	**106**

- The **raw count matrix cannot be used** as input for statistical tests, etc.
- There are several biases that affect the count matrix
- Before our analysis, we need to correct for these

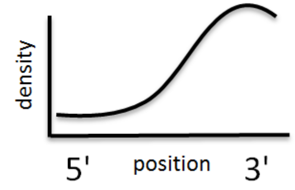
Raw Count Biases

- ❗ Library size bias (total counts):
  - Deeper runs will have more reads mapping to each gene
- ⊖ Gene length bias (Kb):
  - Longer genes will have more reads mapping to them
- ⊖ GC-rich and AT-rich fragment bias:
  - Genes rich in these are underrepresented in the sequencing results
- ❗ RNA composition:
  - Few highly expressed genes can skew normalization

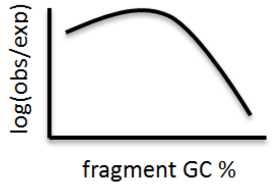
Fragment length
(size selection)



Positional bias
(degradation)



Fragment sequence bias
(PCR amplification)

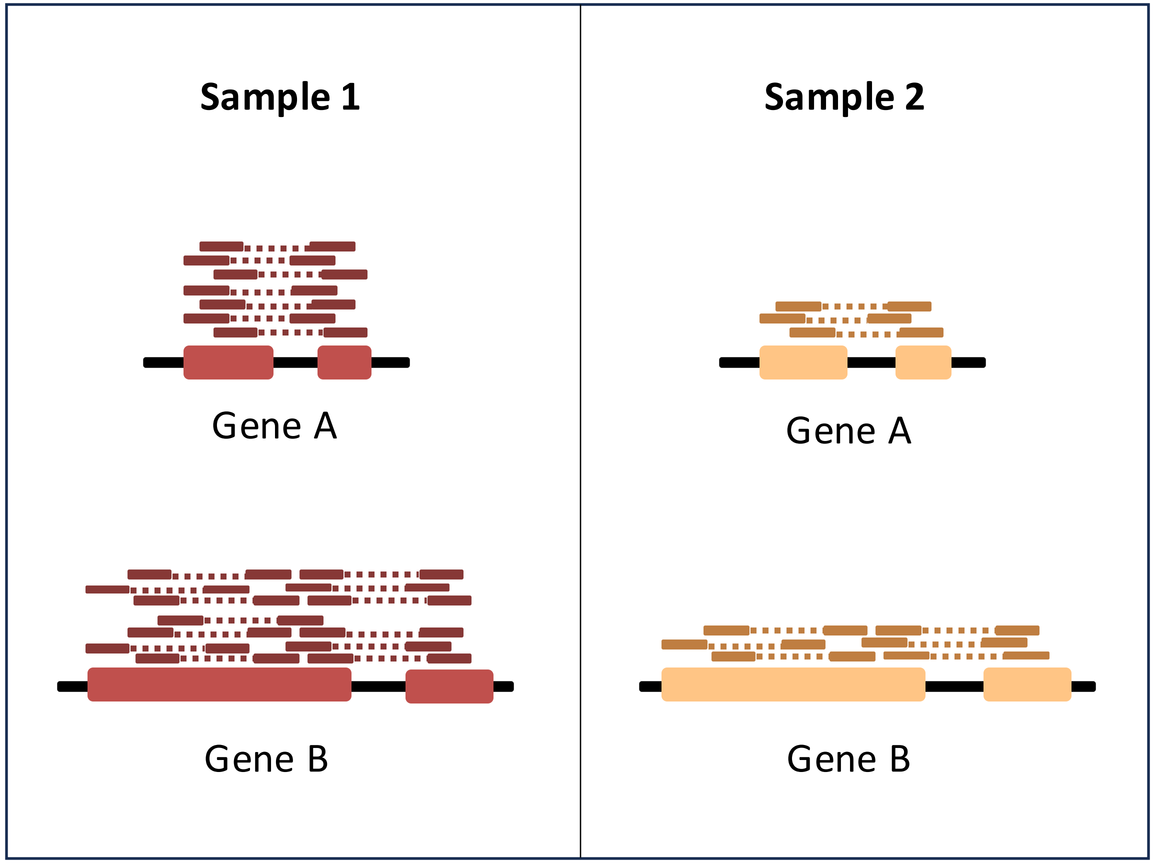


Raw Count Biases

Library size bias:

- Deeper runs will have more reads mapping to each gene
- Lab protocol variability
- Biological variability

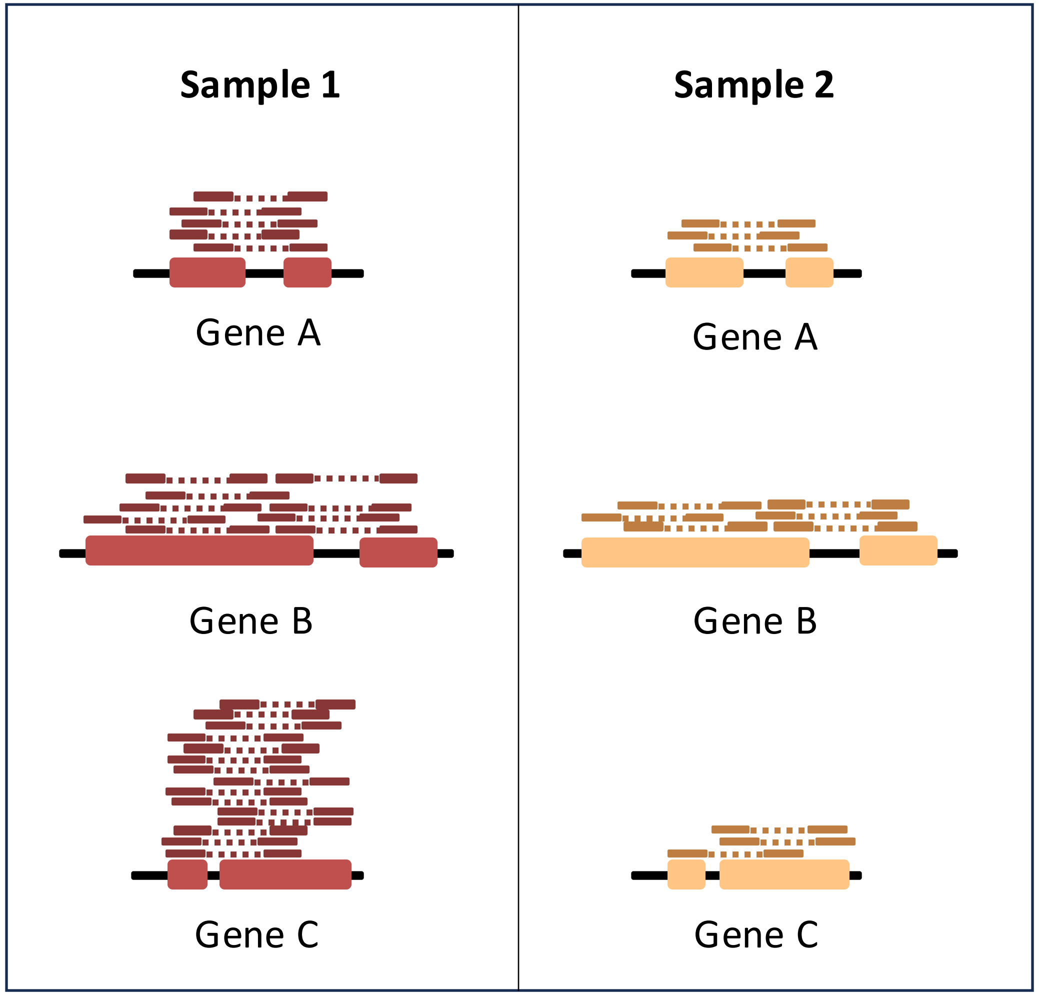
We must correct for this before differential expression analysis!



Raw Counts

RNA composition:

- Few highly expressed genes can skew normalization
- Especially important to consider for differential expression



Normalization

RPKM = Reads per Kilobase Million (single end)

FPKM = Fragments per Kilobase Million (paired end)

$$RPKM_{gX} = \frac{\left(\frac{read\ count_{gX}}{\sum_g read\ count_X\ /\ 10^6}\right)}{gene\ length\ Kb_g}$$

TPM = Transcripts per Million

$$transcript_{gX} = \left(\frac{read\ count_{gX}}{gene\ length\ Kb_g}\right)$$

$$TPM_{gX} = 10^6 \left(\frac{transcript_{gX}}{\sum_g transcript_X}\right)$$

Original

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A	10	12	30
B	20	25	60
C	5	8	15
D	0	0	1
Total counts	**35**	**45**	**106**

TPM

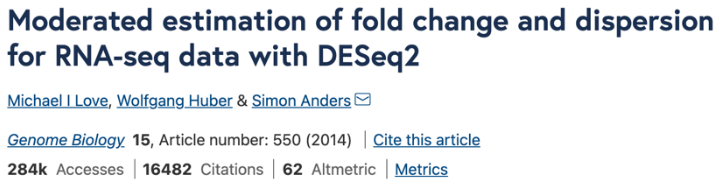
Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A	3.33	2.96	3.326
B	3.33	3.09	3.326
C	3.33	3.95	3.326
D	0	0	0.02
Total counts	**~10**	**~10**	**~10**

HeaDS

DESeq2 R-package

- One of the most used R-packages for RNAseq analysis is **DESeq2**

- Normalizing the data with DESeq2:
  - Does not use RPKM/TPM
  - Uses median of ratios and size factor calculation
  - One of most popular normalization methods for DEA

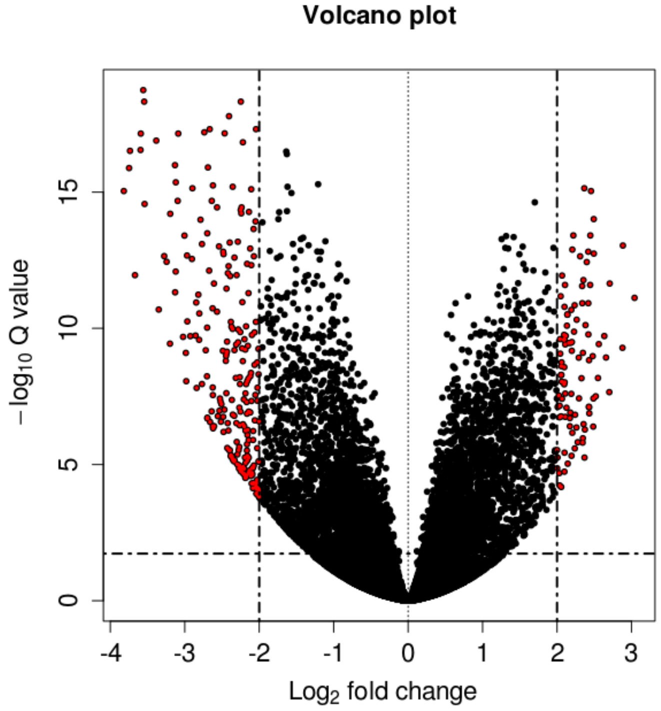


Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber & Simon Anders

Genome Biology **15**, Article number: 550 (2014) | Cite this article

284k Accesses | **16482** Citations | **62** Altmetric | Metrics



Genes differentially expressed (DE) between sample groups

DESeq2 Normalization

Median of ratios

- Accounts for **sequencing depth and RNA composition**

- **Steps:**
  1. Create pseudo-reference sample
  2. Calculate ratio of each sample to the reference
  3. Calculate normalization factor for each sample
  4. Calculate normalized count values using normalization factor

DESeq2 Normalization

1. Create pseudo-reference sample:
   - Geometric mean across all samples
     - Exponential growth data, less sensitive to outliers

2. Calculate ratio of each sample to the reference:
   - Ratios of each gene in a sample compared to the ref.

Gene	Sample 1	Sample 2	Pseudo-ref. sample	Ration Sample 1 / ref.	Ration Sample 2 / ref.
EF2A	1489	906	^{n}sqrt(1489 * 906) = **1161.5**	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ACBD1	22	13	^{n}sqrt(22 * 13) = **17.7**	22/16.9 = **1.30**	13/16.9 = **0.77**
…	…	…	…	…	…

DESeq2 Normalization

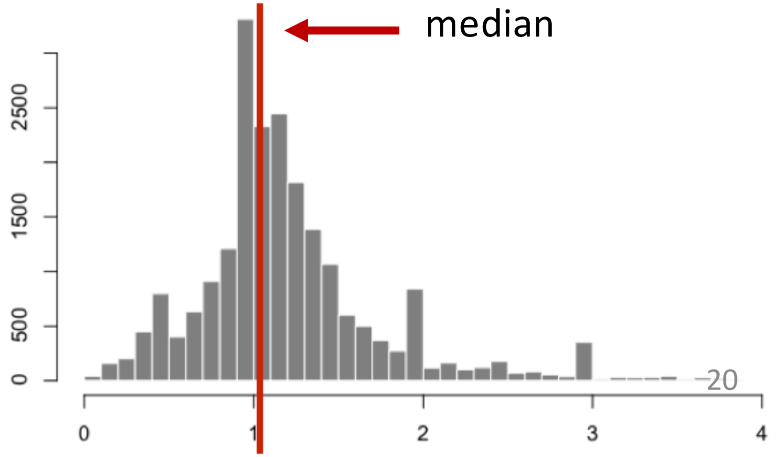
3. Calculate normalization factor for each sample:

- Median value of the ratios for each sample is used as normalization factor

Normalization factor approach:

- Robust to **imbalance in up/down regulation**
- Robust to **large number of DE genes**
- If size factor >> or << 1 = extreme outlier!

sample 1 / pseudo-reference sample



median

Distribution of ratios for a sample

DESeq2 Normalization

4. Calculate normalized count values

Divide raw count value in a sample by sample's normalization factor

Sample1 median ratio = 1.29

Sample2 median ratio = 0.78

Gene	Sample 1	Sample 2
EF2A	1489/1.29 = **1154.26**	906/0.78 = **1161.53**
ACBD1	22/1.29 = **17.905**	13/0.78 = **16.66**
…	…	…

DESeq2 Normalization

Let’s normalize the counts for our dataset

Notebooks:

- *05b_count_matrix.Rmd*
- *05c_count_normalization.Rmd*

Summary Slide

1. RNA counts follow a negative binominal distribution

   Data distribution guides normalization strategies and model choice

2. RNASeq data inherently contain biases which must be taken into account

   DESeq2 performs median of ratios normalization and size factor scaling