

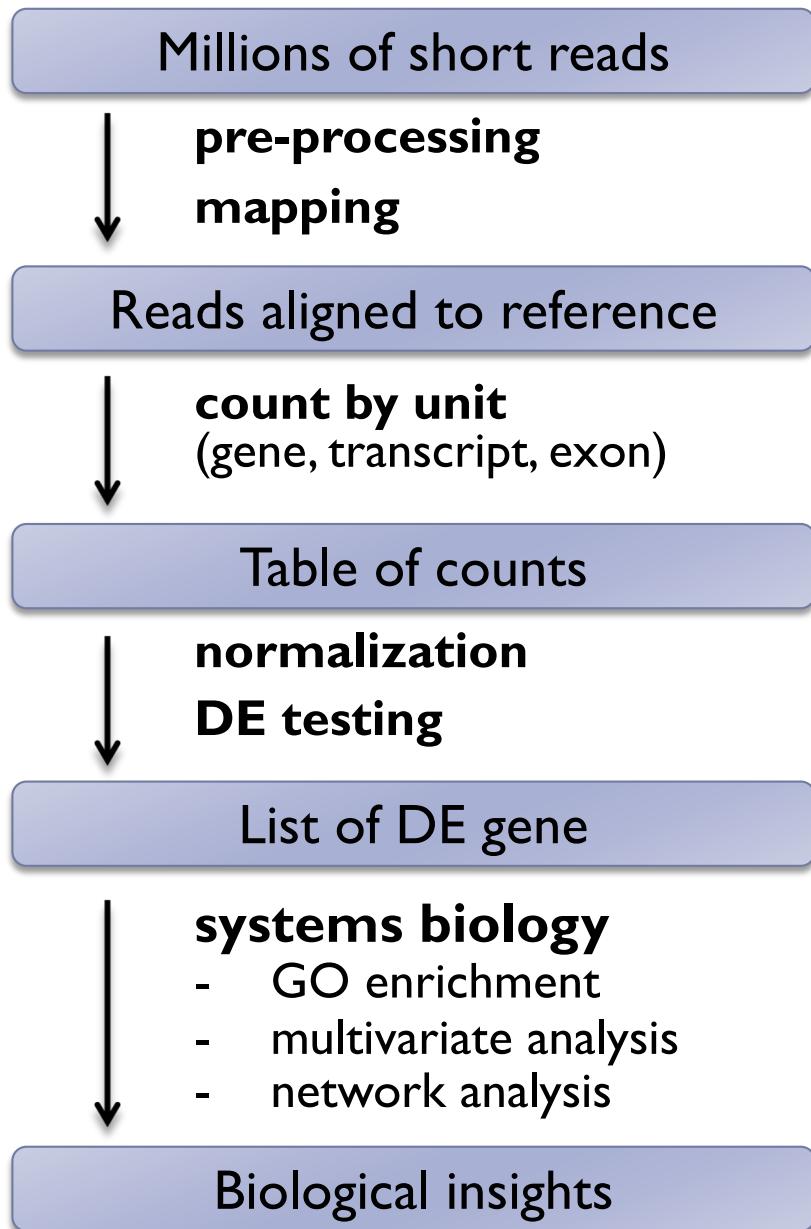
# RNA-seqの解析パイプライン：基礎

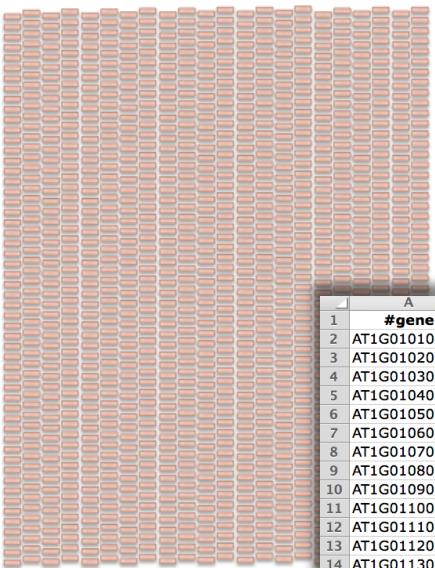
RNA-seq Analysis Pipeline:  
Basics

Shuji Shigenobu  
NIBB, Japan  
<shige@nibb.ac.jp>

# RNA-seq analysis pipeline for DE

Differential Expression analysis





NGS reads

A	B	C	D	E	F	G
#gene	m1	m2	m3	h1	h2	h3
2 AT1G01010	35	77	40	46	64	60
3 AT1G01020	43	45	32	43	39	49
4 AT1G01030	16	24	26	27	35	20
5 AT1G01040	72	43	64	66	25	90
6 AT1G01050	49	78	90	67	45	60
7 AT1G01060	0	15	2	0	21	8
8 AT1G01070	16	34	6	9	20	1
9 AT1G01080	170	191	382	127	98	184
10 AT1G01090	291	346	563	171	116	453
11 AT1G01100	113	125	246	78	27	361
12 AT1G01110	0	1	1	0	0	0
13 AT1G01120	228	189	270	147	83	174
14 AT1G01130	9	11	1	0	2	2
15 AT1G01140	181	120	142	161	73	73
16 AT1G01150	0	2	0	0	0	3
17 AT1G01160	117	125	215	86	46	4
18 AT1G01170	74	57	82	36	22	5
19 AT1G01180	46	7	26	24	18	6
20 AT1G01190	0	3	2	1	2	7
21 AT1G01200	5	0	2	0	0	8
22 AT1G01210	178	203	98	205	83	9
23 AT1G01220	26	49	40	21	15	10
24 AT1G01225	4	10	6	6	11	11
25 AT1G01230	72	51	58	70	18	13
26 AT1G01240	81	89	45	62	24	14
27 AT1G01250	1	1	5	1	2	15
28 AT1G01260	15	52	37	33	27	16
29 AT1G01290	7	16	23	30	5	18
30 AT1G01300	75	115	232	89	109	35

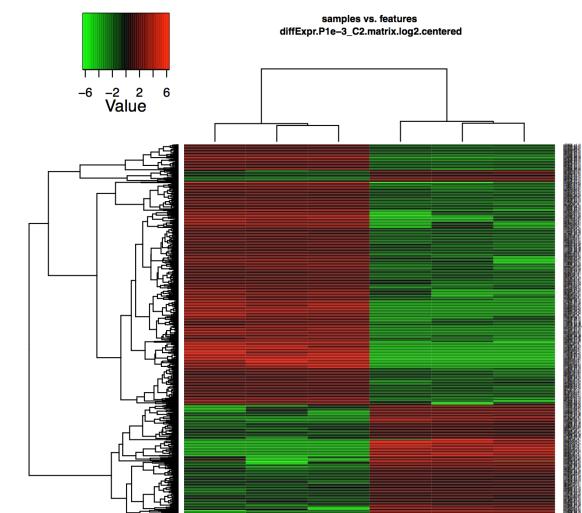
Table of counts

A	B	C	D	E
gene	signal (A)	logFC (M)	p-value	FDR
AT5G48430	-9.17	6.30	3.7E-27	9.8E-23
AT2G08986	-9.15	5.04	2.8E-21	3.6E-17
AT5G31702	-9.74	5.68	5.0E-21	4.3E-17
AT1G51850	-10.07	5.32	2.2E-19	1.4E-15
AT2G39380	-9.82	4.97	2.9E-19	1.5E-15
AT2G44370	-10.23	5.48	1.1E-18	4.7E-15
AT3G46280	-8.19	4.93	1.3E-18	5.0E-15
AT3G55150	-10.44	5.86	1.8E-18	5.8E-15
AT2G19190	-8.70	4.55	8.5E-18	2.5E-14
AT1G51820	-9.40	4.40	1.3E-17	3.3E-14
AT2G39530	-9.16	4.36	1.7E-17	4.1E-14
AT1G40104	-9.58	4.61	2.2E-17	4.9E-14
AT2G30750	-9.97	4.64	1.5E-16	3.0E-13
AT2G17740	-9.12	4.18	1.8E-16	3.3E-13
AT2G07981	-10.07	4.75	1.0E-15	1.8E-12
AT1G30700	-9.21	3.91	3.5E-15	5.8E-12
AT5G31770	-9.12	3.76	1.9E-14	2.7E-11
AT1G51800	-8.47	3.99	1.9E-14	2.7E-11
AT1G30720	-9.06	3.75	4.7E-14	6.5E-11
AT4G12500	-6.58	4.39	1.5E-13	1.9E-10
AT1G07160	-10.76	5.02	2.0E-13	2.5E-10
AT2G39200	-7.57	3.98	6.0E-13	7.2E-10
AT2G10600	-9.30	3.50	8.8E-13	1.0E-09
AT1G65500	-9.60	3.50	9.6E-13	1.0E-09
AT5G39580	-9.24	3.45	1.3E-12	1.3E-09
AT5G44380	-9.51	3.45	1.7E-12	1.7E-09
AT5G13220	-7.20	3.88	2.8E-12	2.7E-09
AT2G45220	-8.45	3.54	3.9E-12	3.6E-09
AT1G30730	-9.80	3.45	6.0E-12	5.4E-09
AT1G64170	-9.91	3.39	1.2E-11	1.1E-08
AT4G20860	-8.10	3.49	1.5E-11	1.2E-08
AT4G23550	-10.08	3.49	1.5E-11	1.2E-08
AT4G12490	-6.75	3.88	1.6E-11	1.2E-08
AT5G64120	-7.09	3.76	2.1E-11	1.6E-08

List of DE genes  
Table of significance score



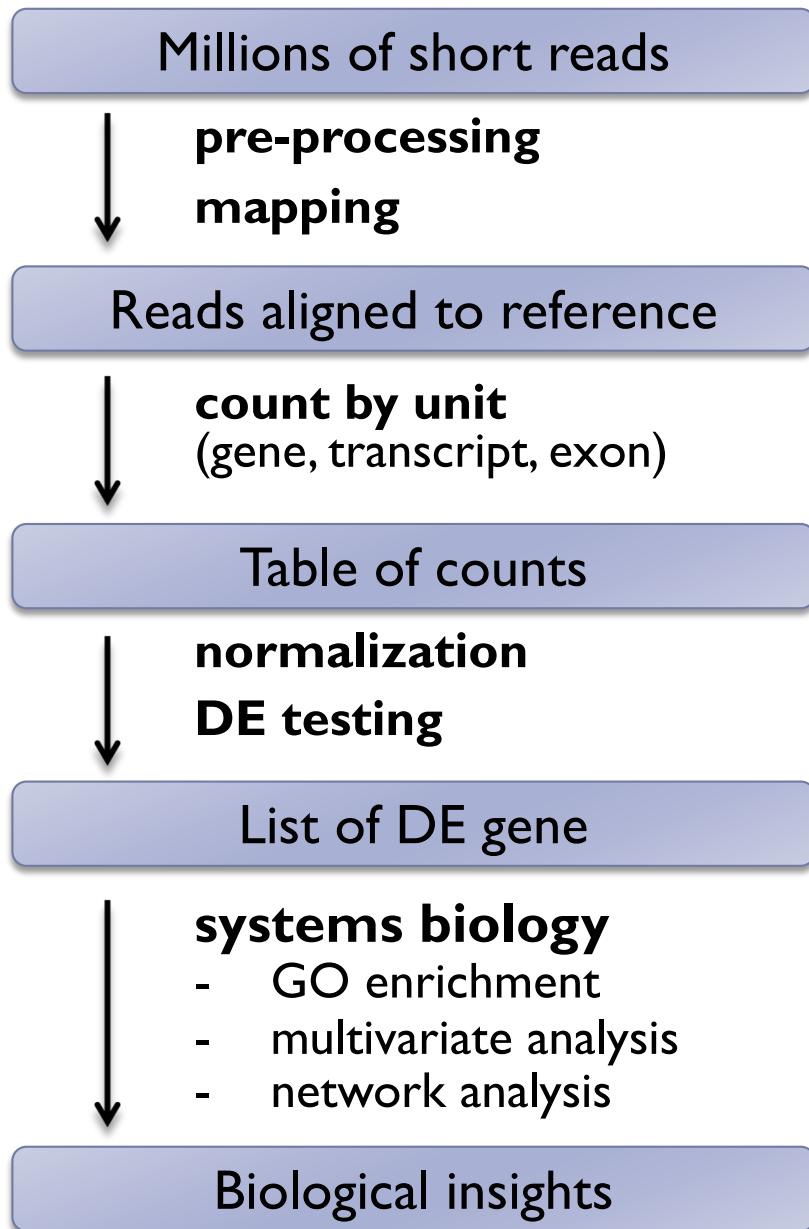
Biological insights



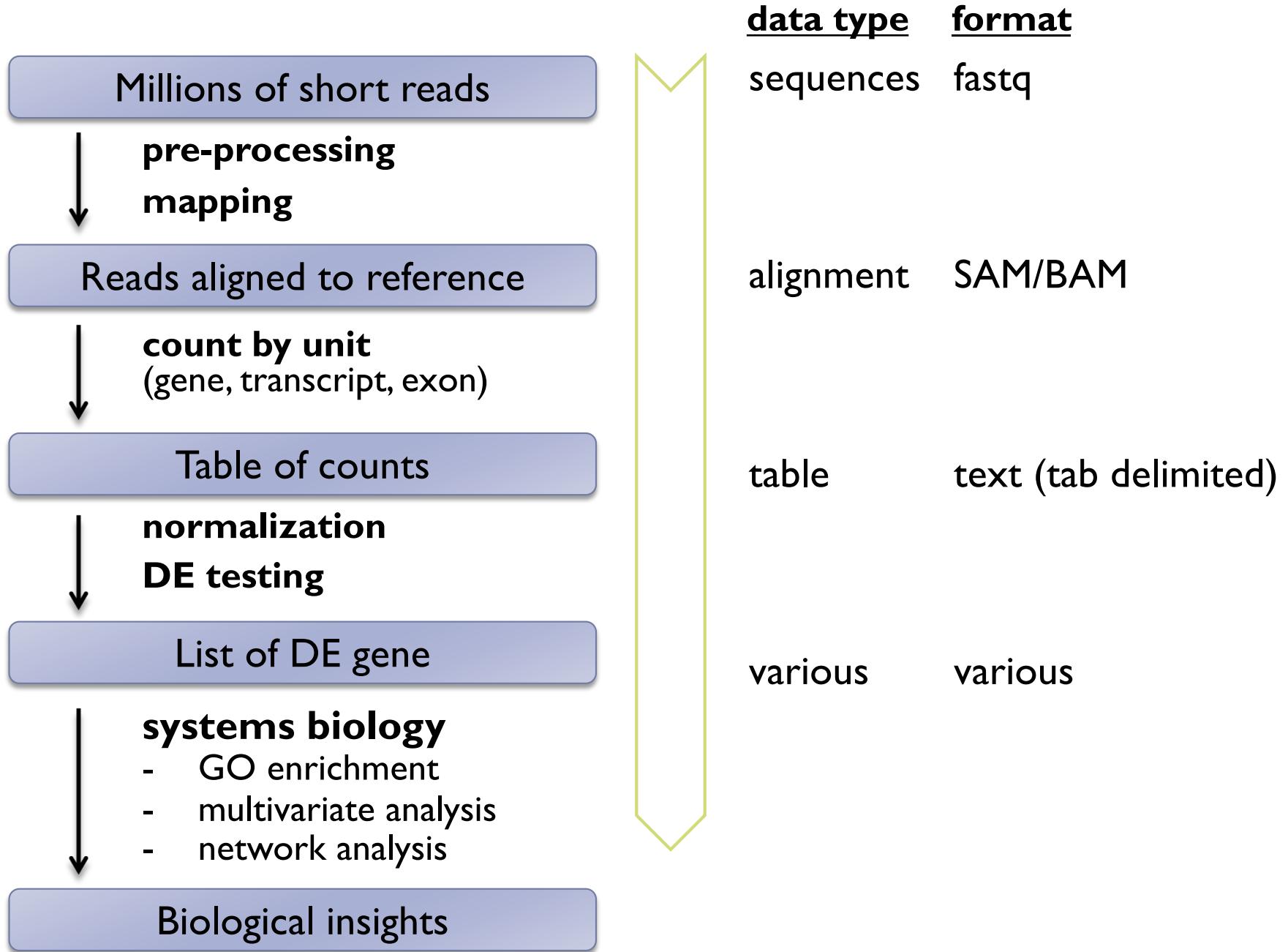
Systems biology:  
clustering and  
network analysis

# RNA-seq analysis pipeline for DE

Differential Expression analysis



# RNA-seq analysis pipeline for DE



# Two Basic Pipelines

- ▶ Choice of reference

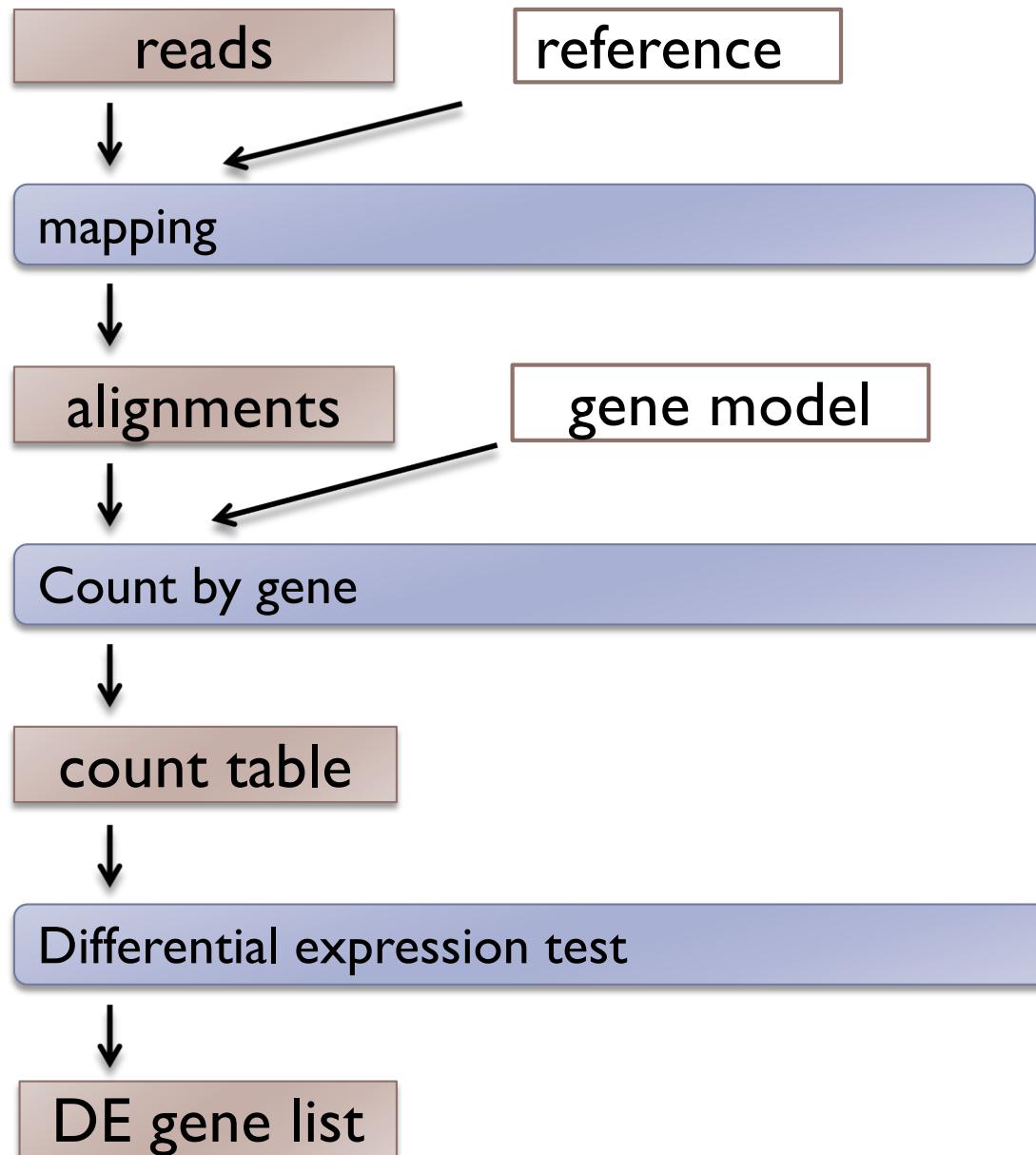
- ▶ **Genome**

- ▶ **Transcript**

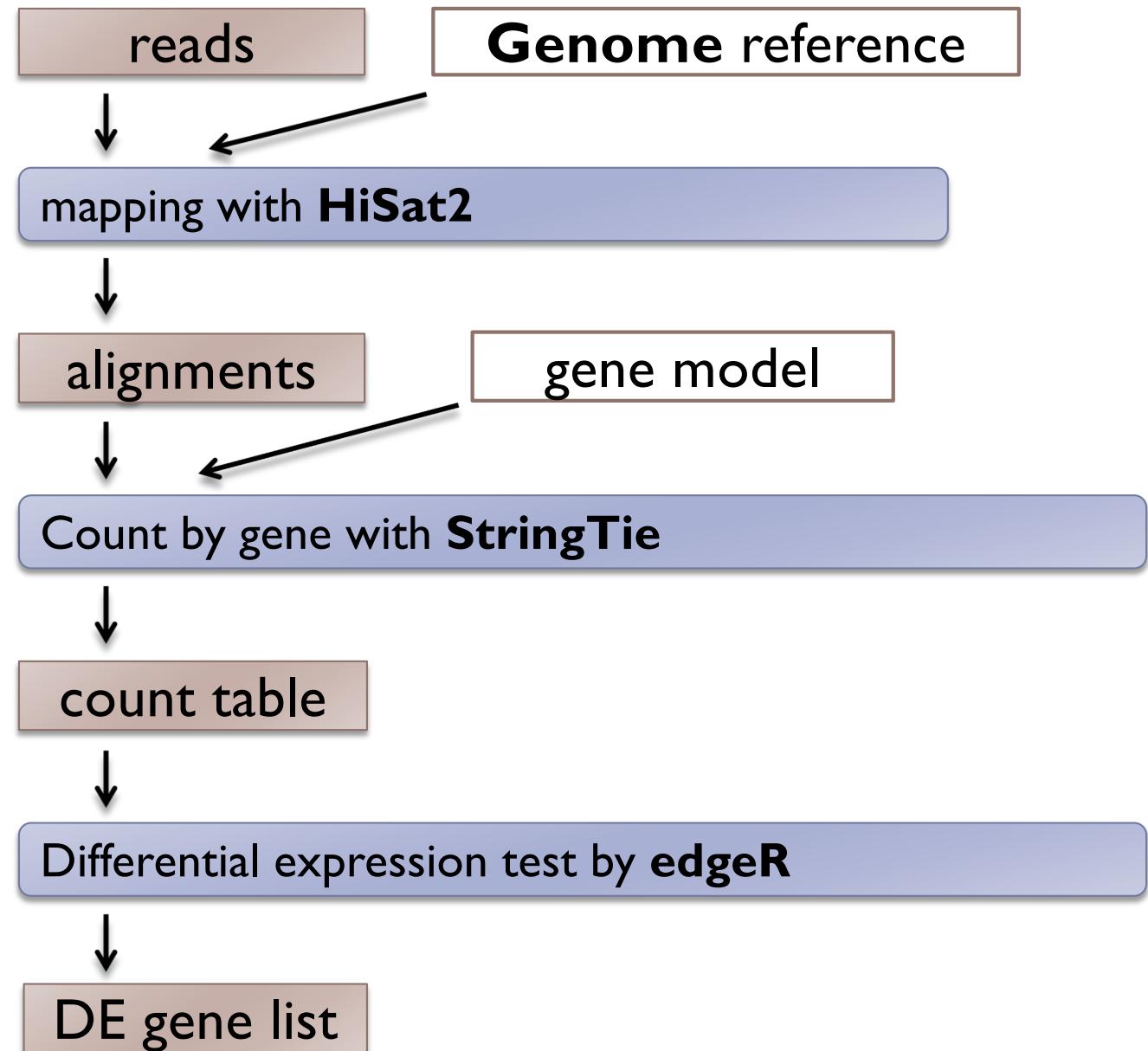
# Two Basic Pipelines

- ▶ **Choice of reference**
  - ▶ **Genome** – standard for genome-known species
  - ▶ **Transcript** – the only way for genome-unknown species
    - can be used for genome-known species

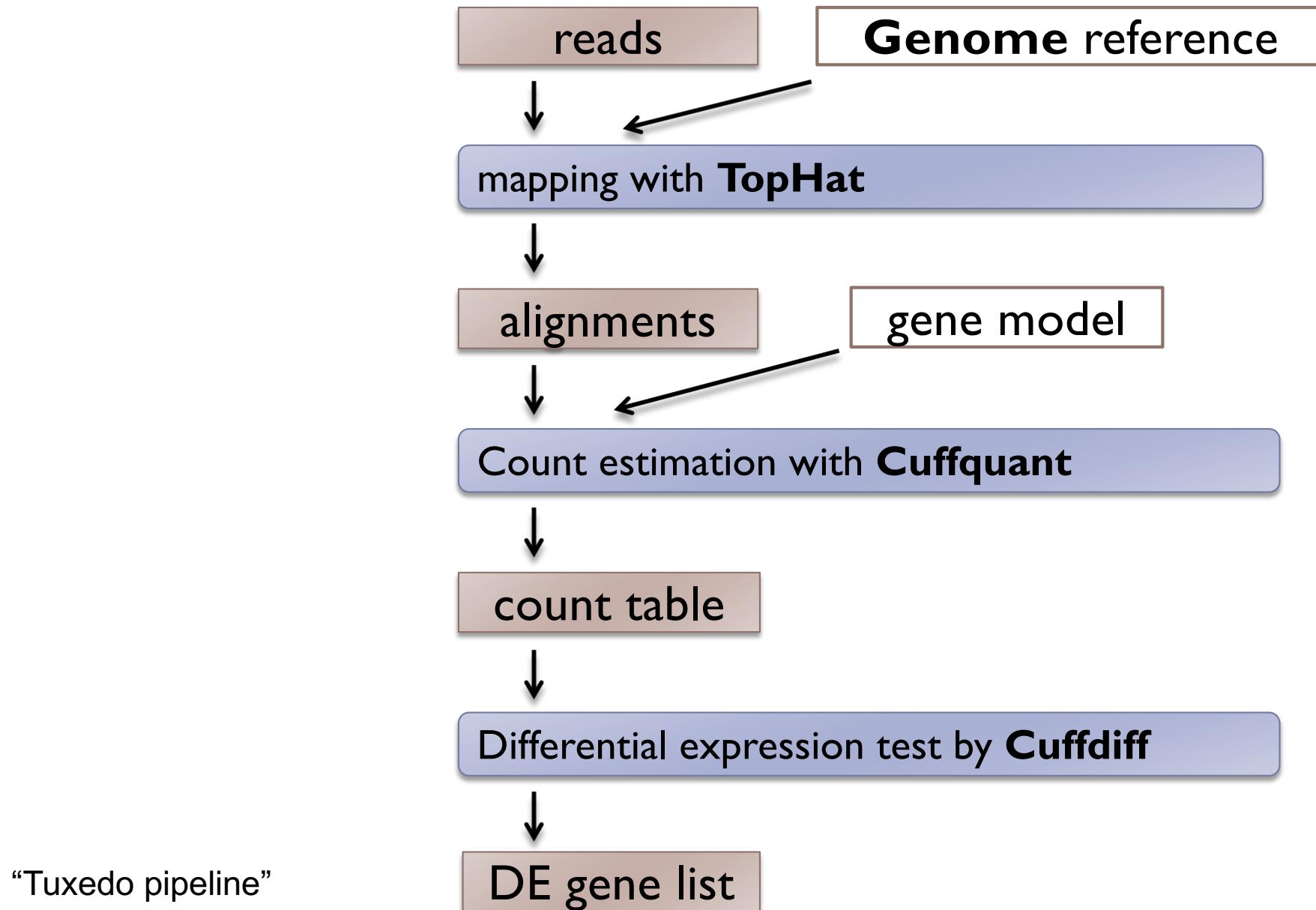
# Common workflow



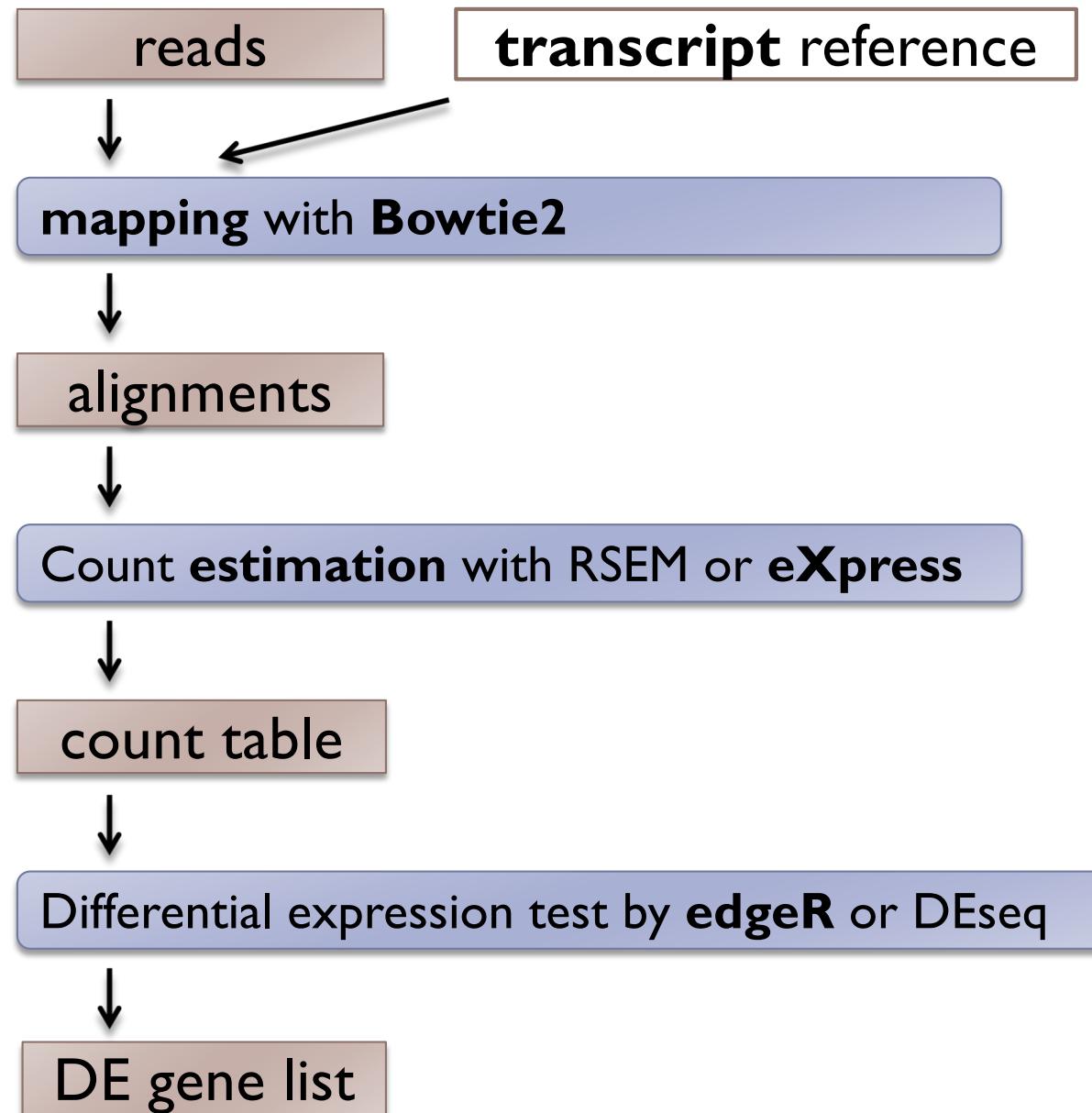
# A genome-based pipeline



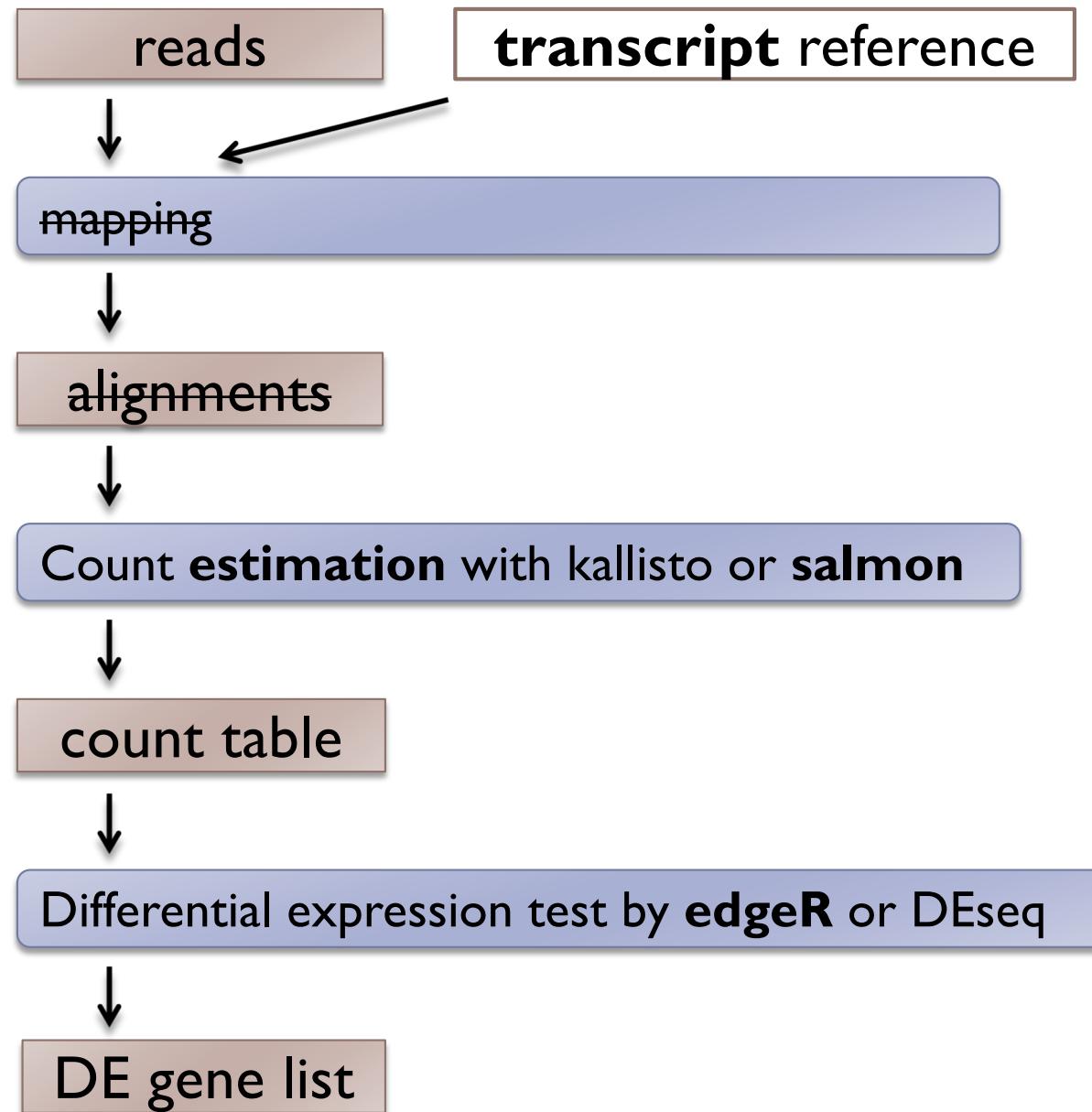
# A genome-based pipeline



# A transcript-based pipeline

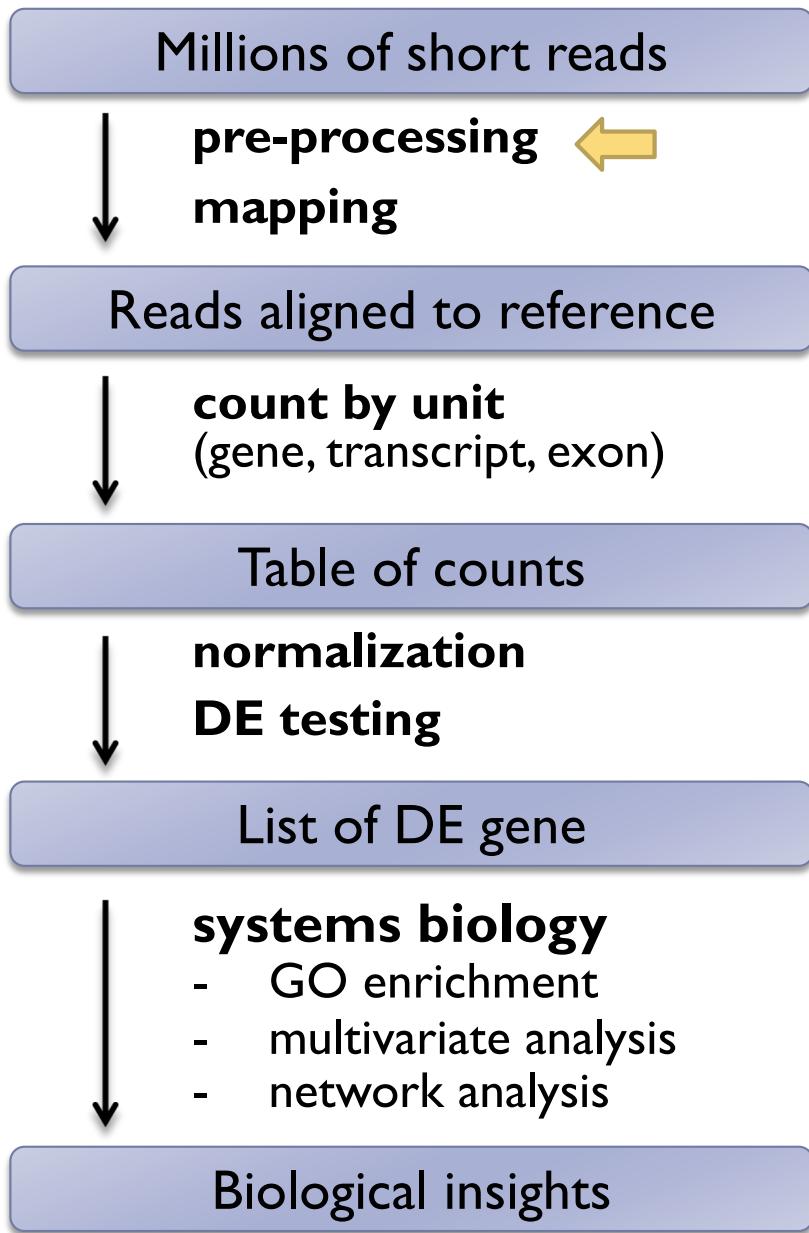


# A transcript-based pipeline (alignment-free method)



# RNA-seq analysis pipeline for DE

Differential Expression analysis



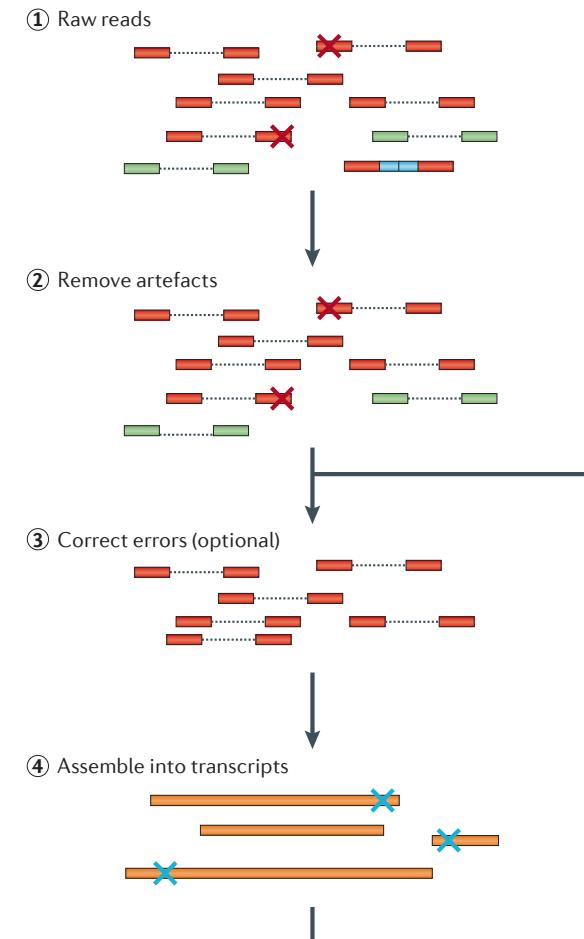
# Read QC and Pre-processing

## ▶ Read QC

- ▶ Tools: FastQC etc.

## ▶ Pre-processing

- ▶ Filter or trim by base quality
- ▶ Remove artifacts
  - ▶ adaptors
  - ▶ low complexity reads
  - ▶ PCR duplications (optional)
- ▶ Remove rRNA and other contaminations (optional)
- ▶ Sequence error correction (optional)
- ▶ Tools: **cutadapt**, **trimmmomatic**



Martin et al (2011) *Nat Rev Genet*

# RNA-seq analysis pipeline for DE

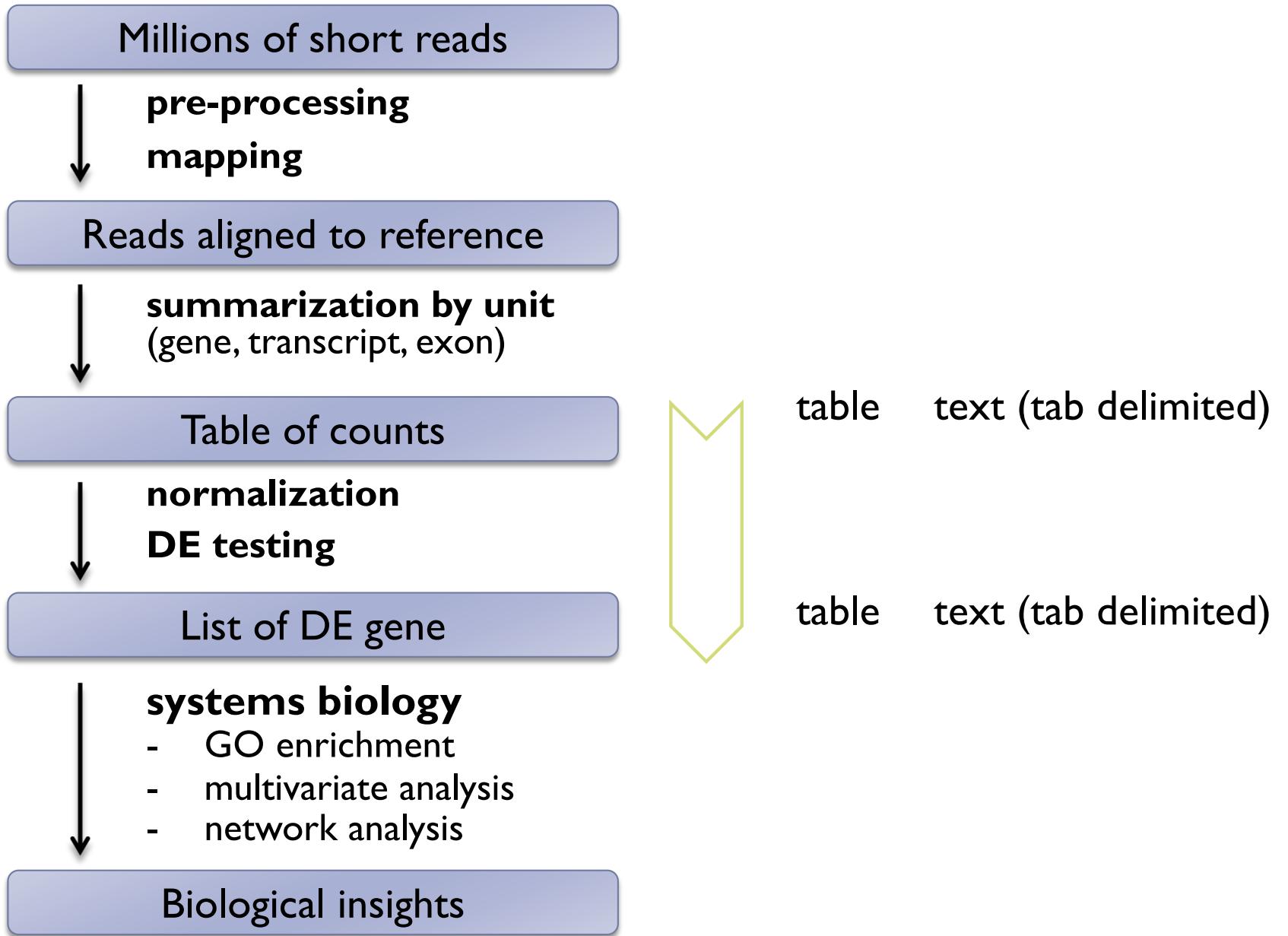


Table of counts

**data import**

**diagnostics**

**normalization**

**DE testing**

**evaluation**

List of DE gene

# Input

	A	B	C	D	E	F	G
1	#gene	m1	m2	m3	h1	h2	h3
2	AT1G01010	35	77	40	46	64	60
3	AT1G01020	43	45	32	43	39	49
4	AT1G01030	16	24	26	27	35	20
5	AT1G01040	72	43	64	66	25	90
6	AT1G01050	49	78	90	67	45	60
7	AT1G01060	0	15	2	0	21	8
8	AT1G01070	16	34	6	9	20	1
9	AT1G01080	170	191	382	127	98	184
10	AT1G01090	291	346	563	171	116	453
11	AT1G01100	113	125	246	78	27	361
12	AT1G01110	0	1	1	0	0	0
13	AT1G01120	228	189	270	147	83	174
14	AT1G01130	9	11	1	0	2	9
15	AT1G01140	181	120	142	161	73	134
16	AT1G01150	0	2	0	0	0	0
17	AT1G01160	117	125	215	86	46	212
18	AT1G01170	74	57	82	36	22	29
19	AT1G01180	46	7	26	24	18	58
20	AT1G01190	0	3	2	1	2	2
21	AT1G01200	5	0	2	0	0	0
22	AT1G01210	178	203	98	205	83	143
23	AT1G01220	26	49	40	21	15	34
24	AT1G01225	4	10	6	6	0	3
25	AT1G01230	72	51	58	70	18	77
26	AT1G01240	81	89	45	62	24	33
27	AT1G01250	1	1	5	1	2	2
28	AT1G01260	15	52	37	33	27	54
29	AT1G01290	7	16	23	30	5	19
30	AT1G01300	75	115	232	89	109	224

Table of counts

# Output

	A	B	C	D	E
1	gene	signal (A)	logFC (M)	p-value	FDR
2	AT5G48430	-9.17	6.30	3.7E-27	9.8E-23
3	AT2G08986	-9.15	5.04	2.8E-21	3.6E-17
4	AT5G31702	-9.74	5.68	5.0E-21	4.3E-17
5	AT1G51850	-10.07	5.32	2.2E-19	1.4E-15
6	AT2G39380	-9.82	4.97	2.9E-19	1.5E-15
7	AT2G44370	-10.23	5.48	1.1E-18	4.7E-15
8	AT3G46280	-8.19	4.93	1.3E-18	5.0E-15
9	AT3G55150	-10.44	5.86	1.8E-18	5.8E-15
10	AT2G19190	-8.70	4.55	8.5E-18	2.5E-14
11	AT1G51820	-9.40	4.40	1.3E-17	3.3E-14
12	AT2G39530	-9.16	4.36	1.7E-17	4.1E-14
13	AT1G40104	-9.58	4.61	2.2E-17	4.9E-14
14	AT2G30750	-9.97	4.64	1.5E-16	3.0E-13
15	AT2G17740	-9.12	4.18	1.8E-16	3.3E-13
16	AT2G07981	-10.07	4.75	1.0E-15	1.8E-12
17	AT1G30700	-9.21	3.91	3.5E-15	5.8E-12
18	AT5G31770	-9.12	3.76	1.9E-14	2.7E-11
19	AT1G51800	-8.47	3.99	1.9E-14	2.7E-11
20	AT1G30720	-9.06	3.75	4.7E-14	6.5E-11
21	AT4G12500	-6.58	4.39	1.5E-13	1.9E-10
22	AT1G07160	-10.76	5.02	2.0E-13	2.5E-10
23	AT2G39200	-7.57	3.98	6.0E-13	7.2E-10
24	AT2G10600	-9.30	3.50	8.8E-13	1.0E-09
25	AT1G65500	-9.60	3.50	9.6E-13	1.0E-09
26	AT5G39580	-9.24	3.45	1.3E-12	1.3E-09
27	AT5G44380	-9.51	3.45	1.7E-12	1.7E-09
28	AT5G13220	-7.20	3.88	2.8E-12	2.7E-09
29	AT2G45220	-8.45	3.54	3.9E-12	3.6E-09
30	AT1G30730	-9.80	3.45	6.0E-12	5.4E-09
31	AT1G64170	-9.91	3.39	1.2E-11	1.1E-08
32	AT4G20860	-8.10	3.49	1.5E-11	1.2E-08
33	AT4G23550	-10.08	3.49	1.5E-11	1.2E-08
34	AT4G12490	-6.75	3.88	1.6E-11	1.2E-08
35	AT5G64120	-7.09	3.76	2.1E-11	1.6E-08

List of DE genes  
Table of significance score

# Identify differentially expressed genes (DEG)

Question: Which are differentially expressed genes (DEG)?

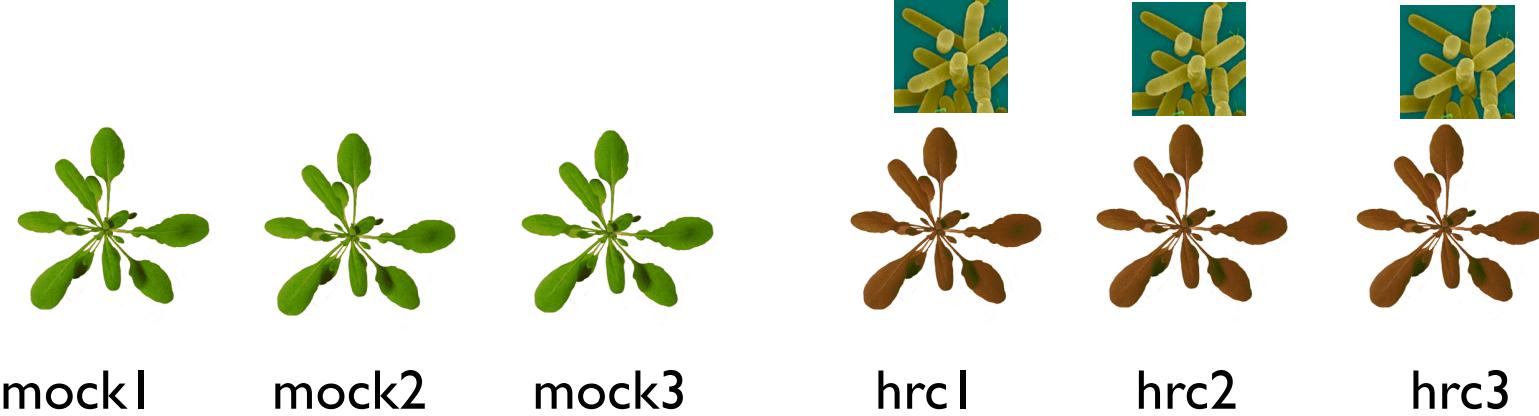
[simple examples (pairwise comparison)]

- mutant v.s. WT
- tissue A v.s. tissue B
- developmental time point A (ex. Early) v.s. B (ex. Late)

Goal:

- Find DE genes
- Rank by significance

# Example: Arabidopsis RNA-seq



mock inoculation (treated w/  
10mM MgCl<sub>2</sub>)

Challenged by defense-eliciting delta-hrcC mutant of *Pseudomonas syringae* pathovar *tmato* DC3000.

- ▶ 6 libraries = 2 groups × 3 biological replicates

Di, Y. et al. *Stat Appl Genet Mol* (2011).  
Cumbie, J. S. et al. *PLoS ONE* (2011).

# Input

- ▶ Typical primary data = matrix of #genes x #samples

column x number of samples (libraries)

row x number of genes (probes)

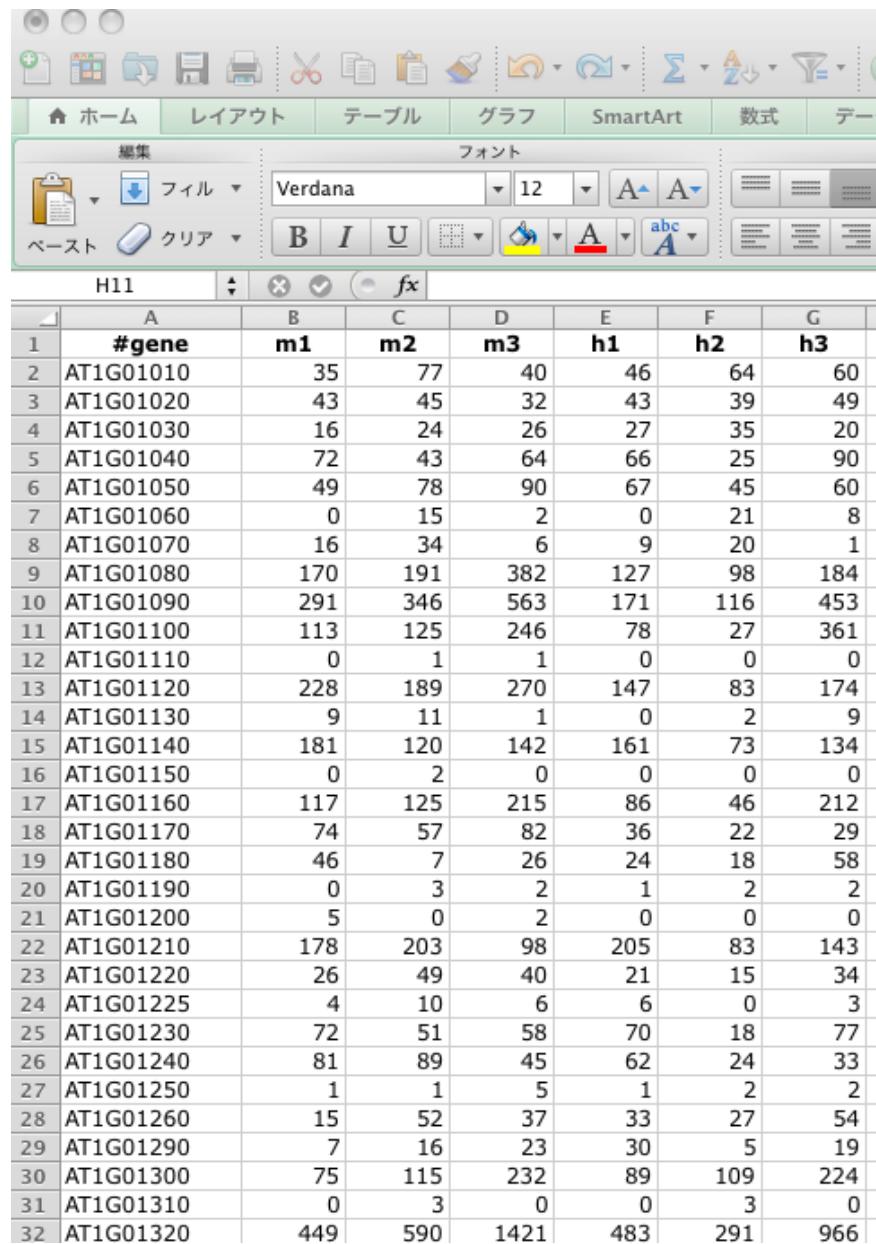
	A	B	C	D	E	F	G
1	#gene	m1	m2	m3	h1	h2	h3
2	AT1G01010	35	77	40	46	64	60
3	AT1G01020	43	45	32	43	39	49
4	AT1G01030	16	24	26	27	35	20
5	AT1G01040	72	43	64	66	25	90
6	AT1G01050	49	78	90	67	45	60
7	AT1G01060	0	15	2	0	21	8
8	AT1G01070	16	34	6	9	20	1
9	AT1G01080	170	191	382	127	98	184
10	AT1G01090	291	346	563	171	116	453
11	AT1G01100	113	125	246	78	27	361
12	AT1G01110	0	1	1	0	0	0
13	AT1G01120	228	189	270	147	83	174
14	AT1G01130	9	11	1	0	2	9
15	AT1G01140	181	120	142	161	73	134
16	AT1G01150	0	2	0	0	0	0
17	AT1G01160	117	125	215	86	46	212
18	AT1G01170	74	57	82	36	22	29
19	AT1G01180	46	7	26	24	18	58
20	AT1G01190	0	3	2	1	2	2
21	AT1G01200	5	0	2	0	0	0
22	AT1G01210	178	203	98	205	83	143
23	AT1G01220	26	49	40	21	15	34
24	AT1G01225	4	10	6	6	0	3
25	AT1G01230	72	51	58	70	18	77
26	AT1G01240	81	89	45	62	24	33
27	AT1G01250	1	1	5	1	2	2
28	AT1G01260	15	52	37	33	27	54
29	AT1G01290	7	16	23	30	5	19
30	AT1G01300	75	115	232	89	109	224

# Import count table / diagnostics

## Look into the input data first

- ▶ Quick view of the table (tools: R, MS Excel etc.)
  - ▶ Check: Format, data structure, data size etc.
- ▶ Scatter plot, MA plot (tools: R, MS Excel etc.)

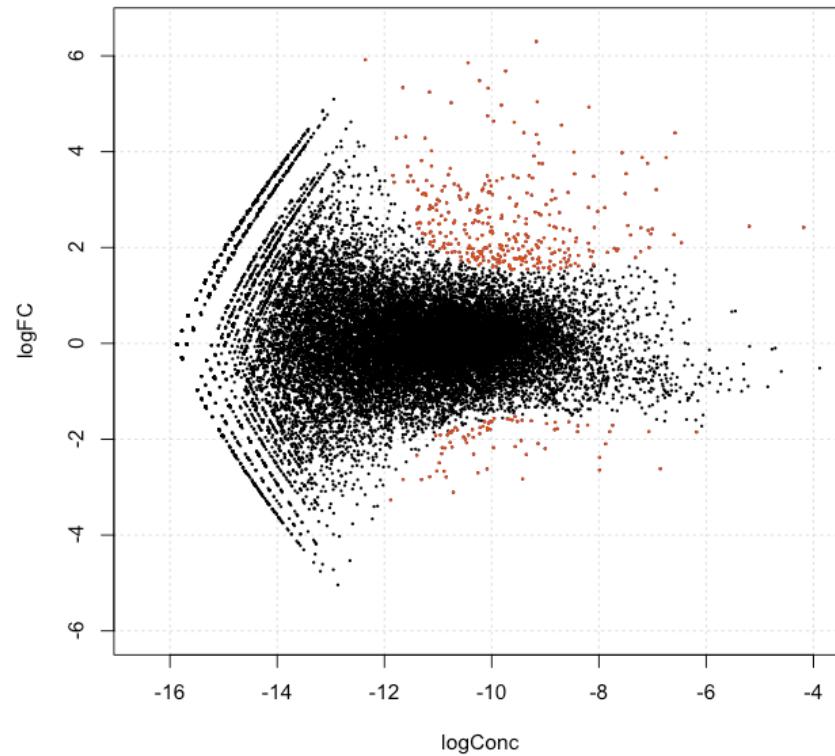
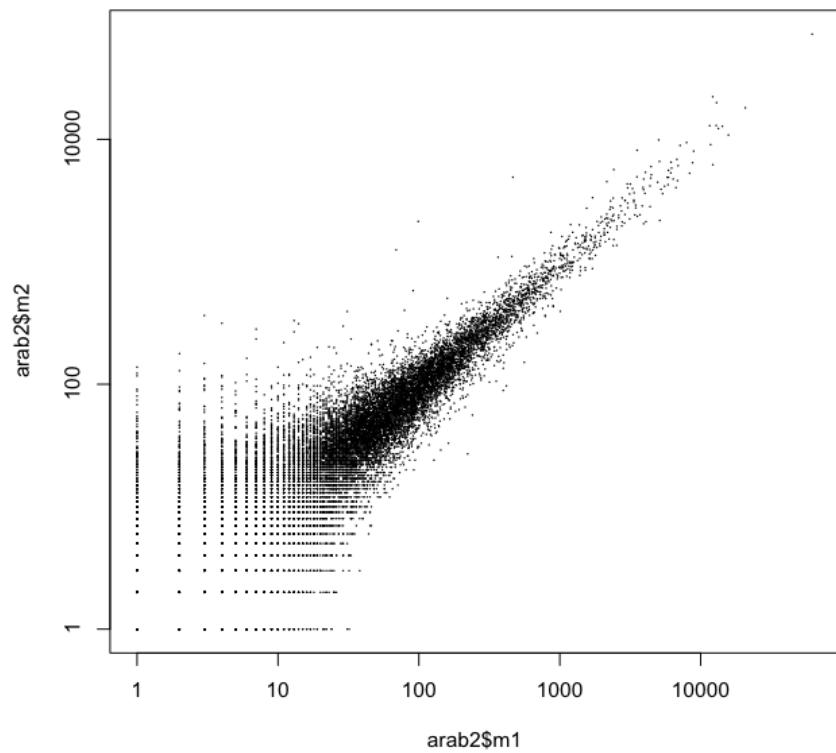
# MS Excel



A screenshot of Microsoft Excel showing a gene expression dataset. The interface includes the ribbon with tabs like Home, Layout, Tables, Graphs, SmartArt, Styles, and Data. The font ribbon shows Verdana at size 12. The table below contains 32 rows of data, starting with a header row.

	#gene	m1	m2	m3	h1	h2	h3
1	AT1G01010	35	77	40	46	64	60
2	AT1G01020	43	45	32	43	39	49
3	AT1G01030	16	24	26	27	35	20
4	AT1G01040	72	43	64	66	25	90
5	AT1G01050	49	78	90	67	45	60
6	AT1G01060	0	15	2	0	21	8
7	AT1G01070	16	34	6	9	20	1
8	AT1G01080	170	191	382	127	98	184
9	AT1G01090	291	346	563	171	116	453
10	AT1G01100	113	125	246	78	27	361
11	AT1G01110	0	1	1	0	0	0
12	AT1G01120	228	189	270	147	83	174
13	AT1G01130	9	11	1	0	2	9
14	AT1G01140	181	120	142	161	73	134
15	AT1G01150	0	2	0	0	0	0
16	AT1G01160	117	125	215	86	46	212
17	AT1G01170	74	57	82	36	22	29
18	AT1G01180	46	7	26	24	18	58
19	AT1G01190	0	3	2	1	2	2
20	AT1G01200	5	0	2	0	0	0
21	AT1G01210	178	203	98	205	83	143
22	AT1G01220	26	49	40	21	15	34
23	AT1G01225	4	10	6	6	0	3
24	AT1G01230	72	51	58	70	18	77
25	AT1G01240	81	89	45	62	24	33
26	AT1G01250	1	1	5	1	2	2
27	AT1G01260	15	52	37	33	27	54
28	AT1G01290	7	16	23	30	5	19
29	AT1G01300	75	115	232	89	109	224
30	AT1G01310	0	3	0	0	3	0
31	AT1G01320	449	590	1421	483	291	966

# Diagnostics: Scatter plot & MA plot



# Let's try: data import and quick check

```
> dat <- read.delim("~/data/SS/arab2.txt", row.names=1)
> head(arab2)                                # look at the first several lines
                                                # for checking
  m1  m2  m3  h1  h2  h3
AT1G01010 35  77  40  46  64  60
AT1G01020 43  45  32  43  39  49
AT1G01030 16  24  26  27  35  20
AT1G01040 72  43  64  66  25  90
AT1G01050 49  78  90  67  45  60
AT1G01060  0  15   2   0  21   8

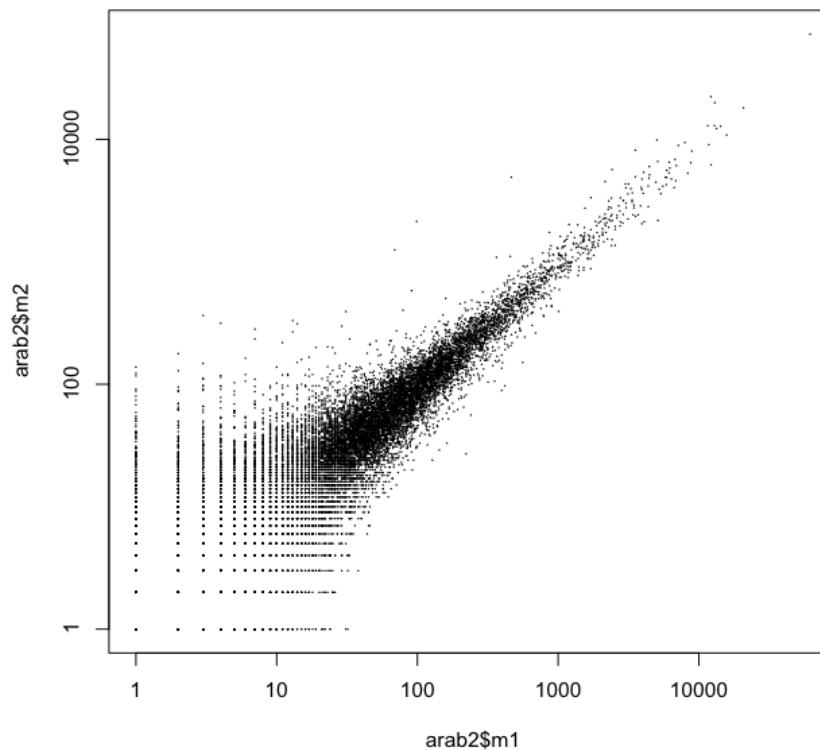
> dim(dat)                                    # get numbers of rows and columns
[1] 26221      6

> colSums(dat)                               # get column sums
  m1        m2        m3        h1        h2        h3
1902032 1934029 3259705 2129854 1295304 3526579
```

演習問題 ex3

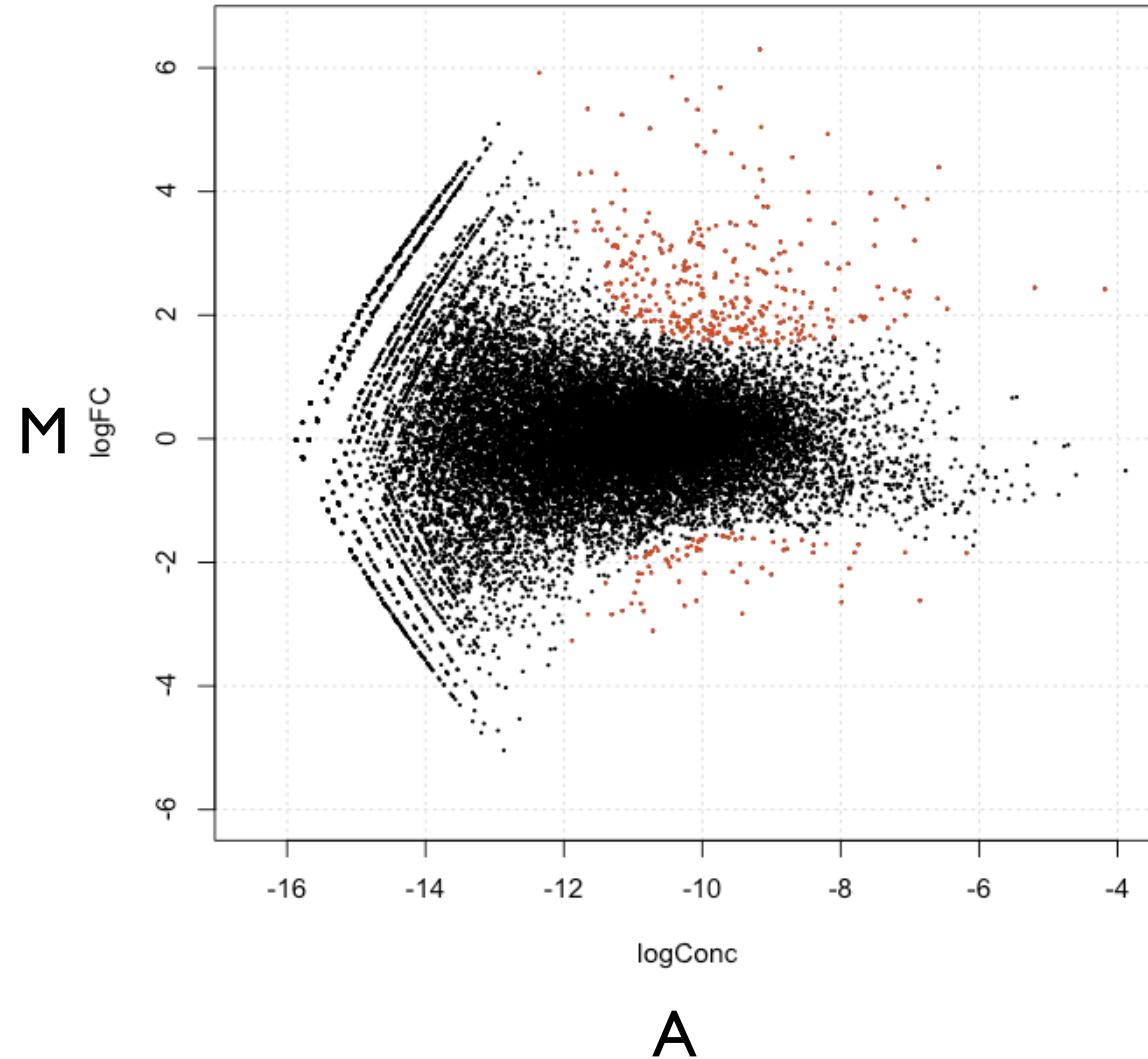
# Let's try: Scatter plot

```
> plot(dat$m1 + 1, dat$m2 + 1, log="xy")
```



演習問題 ex3

# MA plot



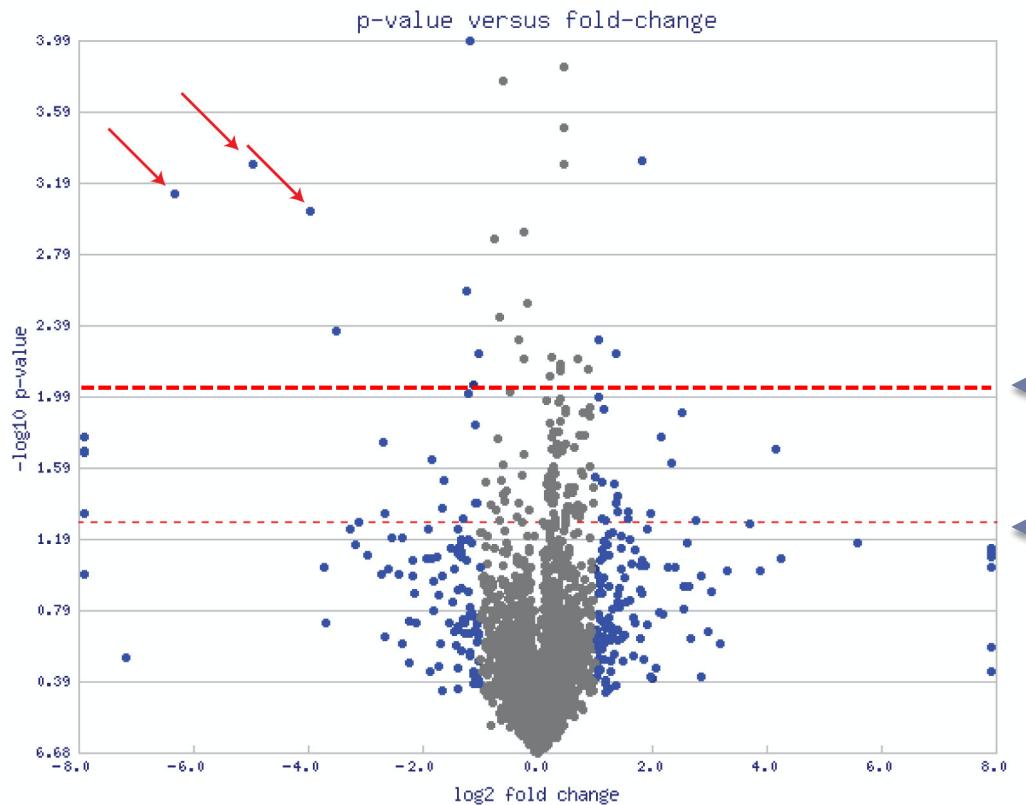
**M:** log fold-change  
**A:** log intensity average

$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$
$$A = \frac{1}{2}\log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))$$

R: expression level of sample 1  
G: expression level of sample 2

演習問題 ex4

# Volcano Plot



$$P = 0.01 \\ (-\log_{10}(0.01)) = 2$$

$$P = 0.05 \\ (-\log_{10}(0.05)) = 1.301$$

Fig: Wikipedia

- ▶ X axis: log fold change – 発現比
- ▶ Y axis: log p-value -- significance

演習問題 ex16

DEG解析を学んだ  
後でtry

Table of counts

**data import**

**diagnostics**

**normalization**

**DE testing**

**evaluation**

List of DE gene

# Normalization

- ▶ What is normalization? Why it is required?
- ▶ Types of normalization.
- ▶ RNAseq specific issue.

# Normalization

## What is normalization? Why it is required?

- ▶ Normalization means to adjust transcriptome data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between genes.
- ▶ Normalization is an essential step in the analysis of DE from RNA-seq data to make them really comparable.

# Normalization: two types

- ▶ **Between-libraries**
  - ▶ Comparing expression (counts) of genes between libraries
- ▶ **Within-library**
  - ▶ Comparing expression (counts) of genes within a library  
(should be possible with NGS – in contrast to microarray)

# Normalization

- ▶ **Between-library:**  
gene vs gene **between** libraries/sample

Adjust by the total number of reads

- ▶ CPM (Counts Per Million mapped reads)

$$\text{CPM}_i = \frac{X_i}{N} = \frac{X_i}{N} \cdot 10^6$$

**$X_i$ :** count of gene

**N:** number of fragments sequenced

# Normalization

- ▶ **Within-library:**  
gene vs gene **within** sample

Longer transcripts gets higher counts. => Needs normalization by length

- ▶ RPKM/FPKM (Reads/Fragments Per Kb per Million mapped reads)

$$\text{FPKM}_i = \frac{X_i}{\left(\frac{\tilde{l}_i}{10^3}\right) \left(\frac{N}{10^6}\right)} = \frac{X_i}{\tilde{l}_i N} \cdot 10^9$$

- ▶ TPM (Transcript per million)

$$\text{TPM}_i = \frac{X_i}{\tilde{l}_i} \cdot \left( \frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6$$

***l<sub>i</sub>*:** effective length of gene  
***N*:** number of fragments sequenced  
***X<sub>i</sub>*:** count of gene

## ► Relationship between TPM and FPKM

$$\text{TPM}_i = \left( \frac{\text{FPKM}_i}{\sum_j \text{FPKM}_j} \right) \cdot 10^6$$

```
countToTpm <- function(counts, effLen){  
  rate <- log(counts) - log(effLen)  
  denom <- log(sum(exp(rate)))  
  exp(rate - denom + log(1e6))  
}  
  
countToFpkm <- function(counts, effLen){  
  N <- sum(counts)  
  exp( log(counts) + log(1e9) - log(effLen) - log(N) )  
}  
  
fpkmToTpm <- function(fpkm){  
  exp(log(fpkm) - log(sum(fpkm)) + log(1e6))  
}  
  
countToEffCounts <- function(counts, len, effLen){  
  counts * (len / effLen)  
}
```

Table of counts

**data import**

**diagnostics**

**normalization**

**DE testing**

**evaluation**

sophisticated way specialized  
for RNA-seq

List of DE gene

# DEG: RNA-seq specific issues

- ▶ RNA-seq count data is Non-Gaussian
- ▶ Normalization: composition effects
- ▶  $N$  (biological replicates) is so small
- ▶ Multiple comparisons (多重検定の問題)

# RNA-seq data is Non-Gaussian

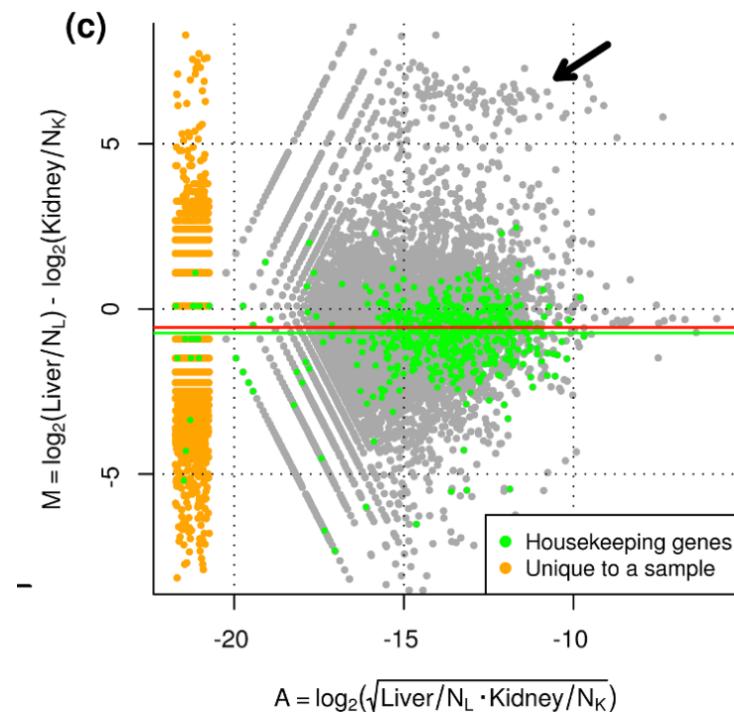
- ▶ RNA-seq data
  - ▶ Discrete-valued data (離散値)
  - ▶ Not normally distributed random variables
  - ▶ **Poisson distribution** for technical replicates
  - ▶ **Negative binomial distribution** for biological replicates.  
(負の二項分布)

# RNA-seq issue: Normalization

- ▶ **Simple normalization**
  - ▶ Simple CPM, TPM or RPKM/FPKM works well in many cases.  
But RNA comparison differences requires further normalization.
- ▶ **Composition effects**
  - ▶ A small number of highly expressed genes can consume a significant amount of the total sequence.
  - ▶ Strategies to control composition effects
    - ▶ estimate scaling factors from data and statistical models
    - ▶ quantile normalization
    - ▶ ...

# TMM Method

- ▶ Trimmed Mean of M values method (Robinson et al., 2010)



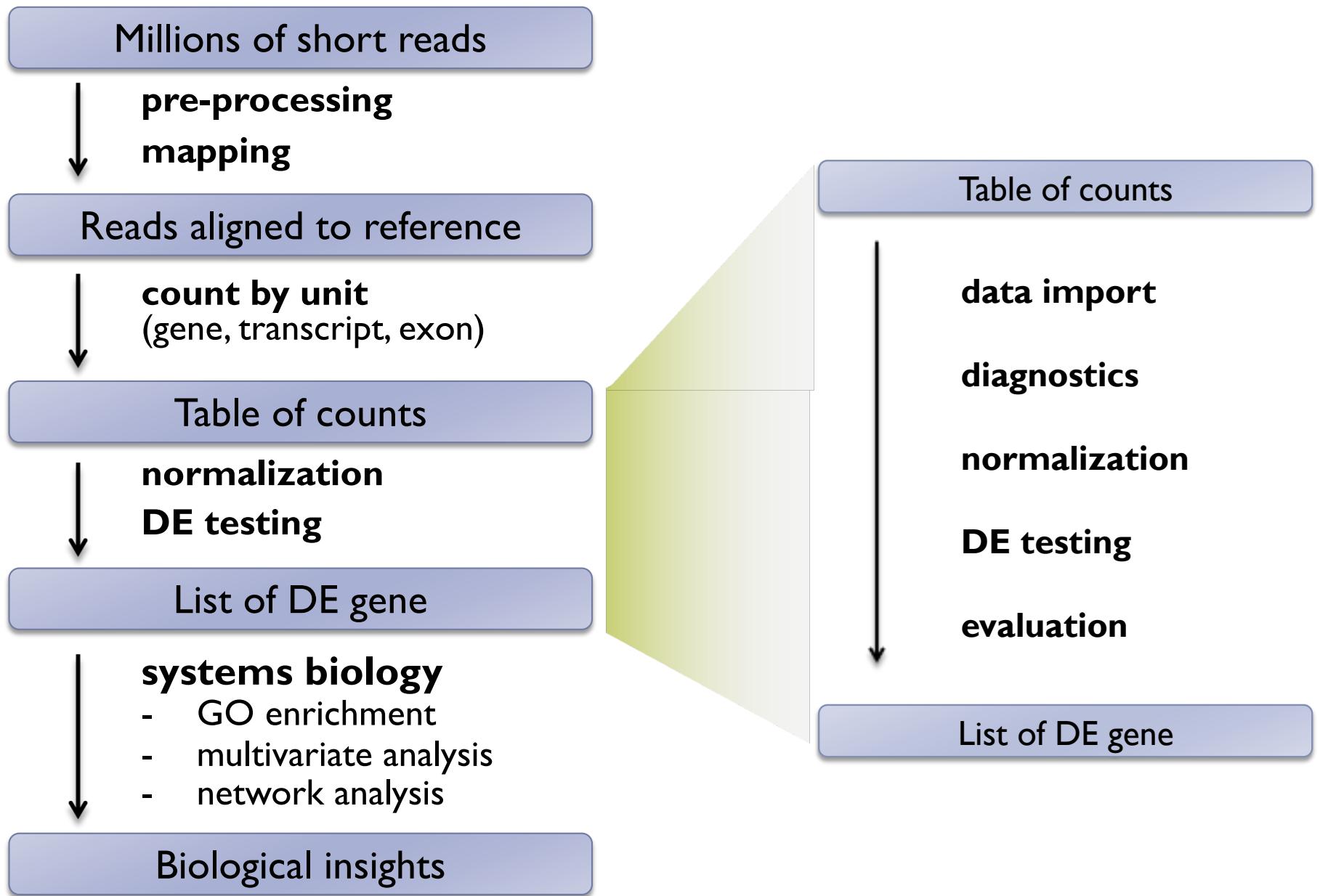
1. Calculate  $M$  and  $A$  for all genes
2. Remove genes with  $M$  of top and bottom 30%
3. Remove genes with  $A$  of top and bottom 5%
4. Calculate scaling factor using remained genes.

# Implementation

## edgeR

- ▶ **Model:** An over dispersed Poisson model, **negative binomial (NB) model** is used
- ▶ **Normalization:** **TMM method** (trimmed mean of M values; Robinson et al., 2010), **RLE** (Anders et al., 2010) and **upperquantile** (Bullard et al., 2010)

# RNA-seq analysis pipeline for DE



- ▶ Let's try 演習問題
  - ▶ ex3: Count data import and scatter plot
  - ▶ ex4: MA plot
  - ▶ ex16:Volcano plot (after learning DEG analysis)