Differential expression analysis

Isabelle Dupanloup
BCF - Bioinformatics Core Facility
SIB - Swiss Institute of Bioinformatics
isabelle.dupanloup@sib.swiss

Swiss Institute of Bioinformatics

Teaching material from
Harvard Chan Bioinformatics Core training

Differential gene expression (DGE) analysis

Learning Objectives

- Getting familiar with the differential gene expression analysis workflow
- Exploring different types of normalization methods

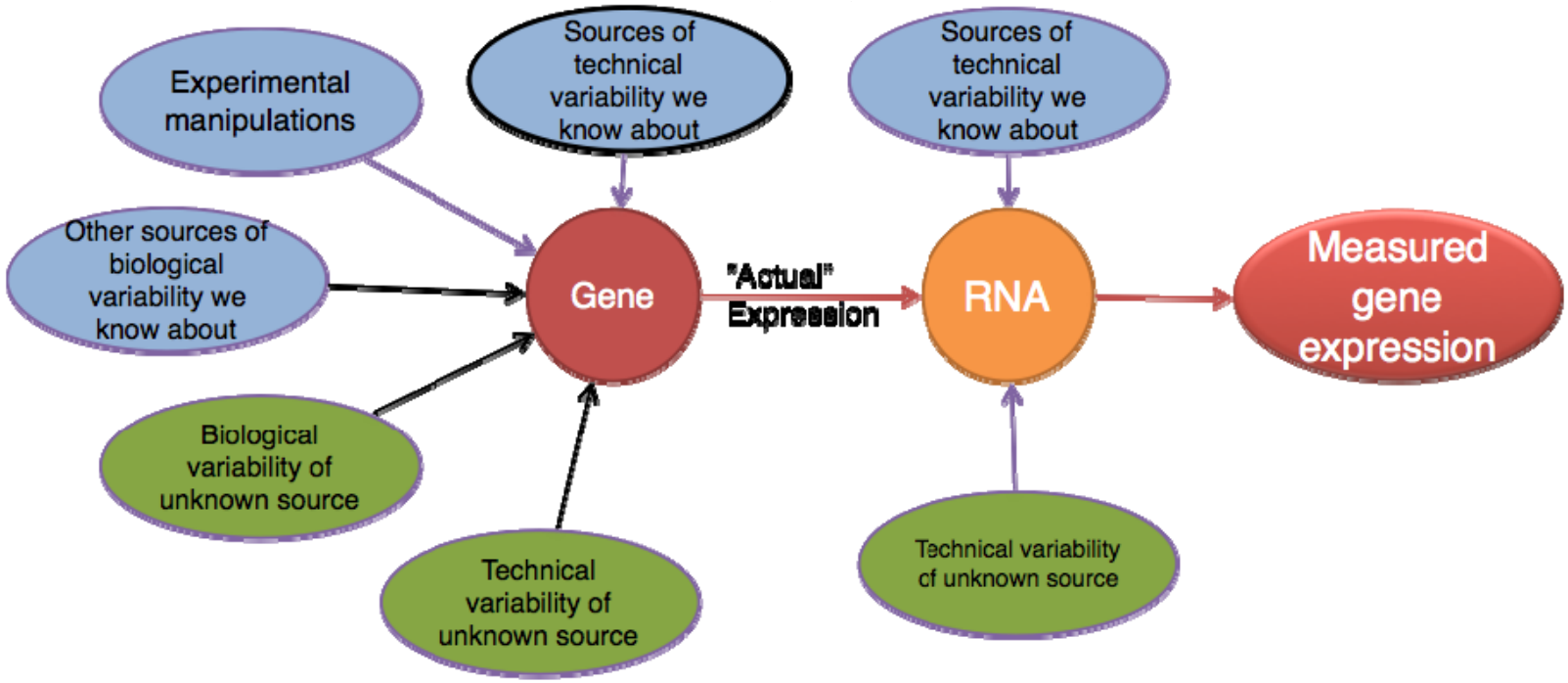
Differential gene expression (DGE) analysis

samples: want to see if differences across condition are significant (w.r.t. biological and technical variation)

features (e.g. genes)

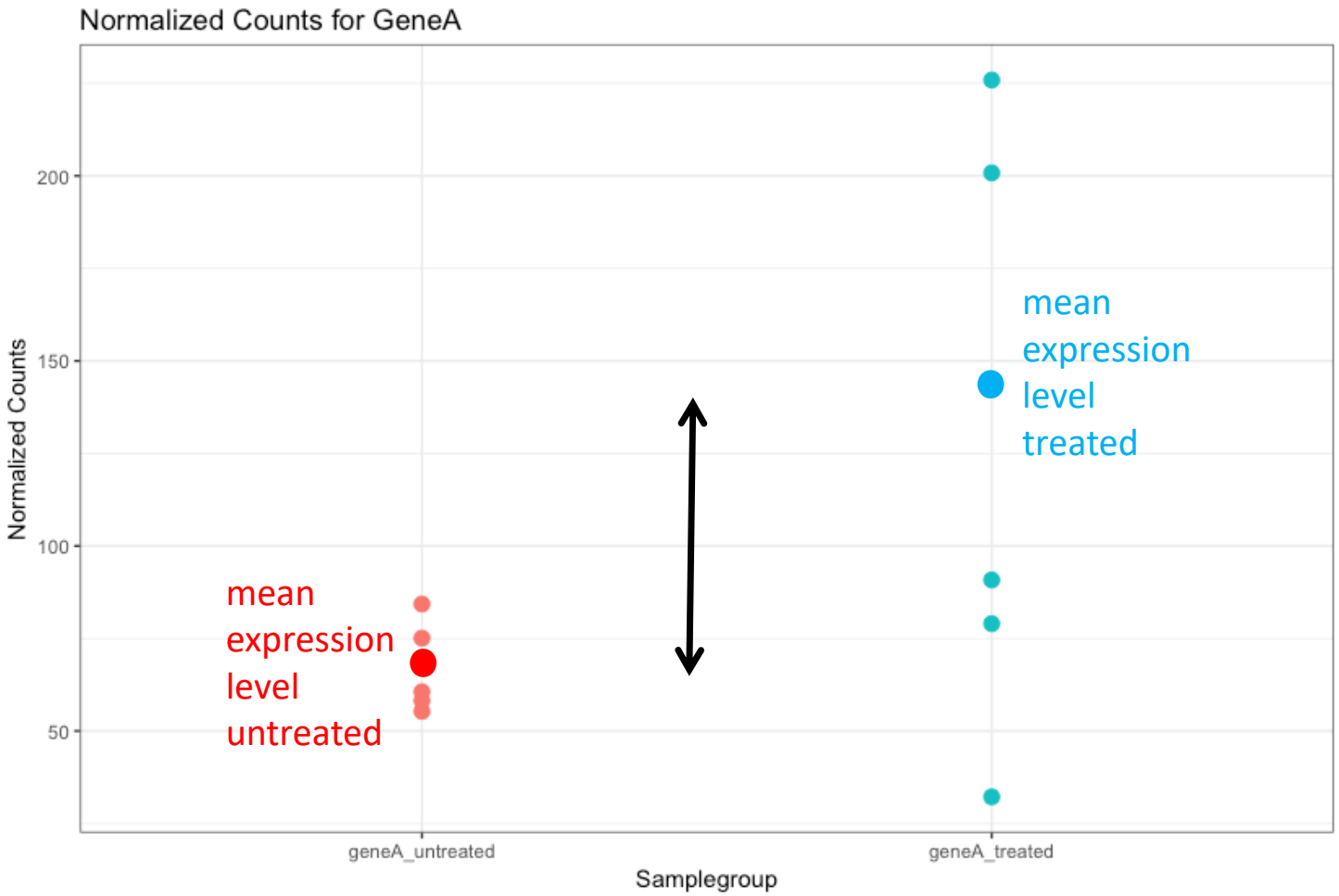
	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

Differential gene expression (DGE) analysis



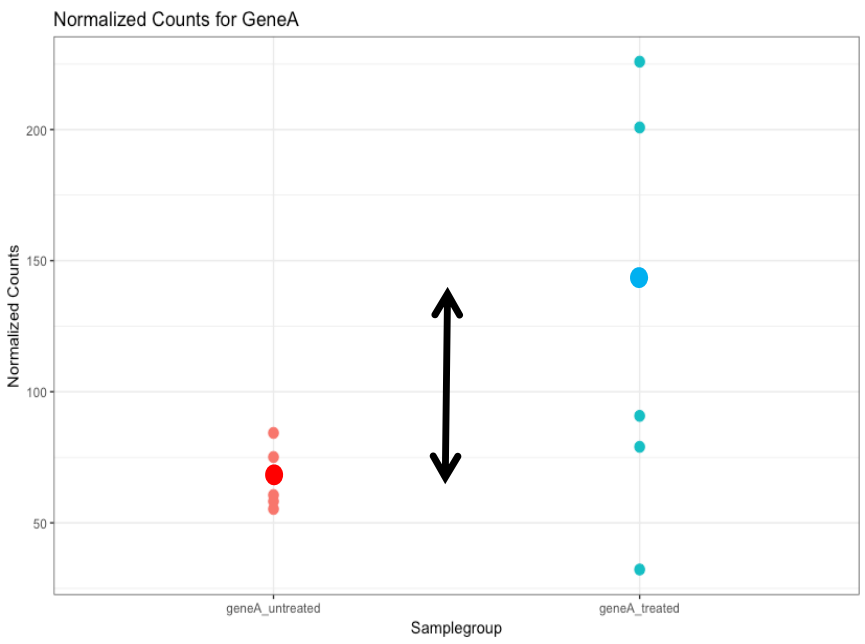
Courtesy of Paul Pavlidis, UBC

Differential gene expression (DGE) analysis



We need to take into account the variation in the data (and where it might be coming from) when determining whether genes are differentially expressed.

Differential gene expression (DGE) analysis



- **Goal** of differential expression analysis : determine, for each gene, whether the differences in expression (counts) **between groups** is significant given the amount of variation observed **within groups** (replicates)
- **Test** for **significance** with an **appropriate statistical model** that accurately performs **normalization** (to account for differences in sequencing depth, etc.) and **variance modeling** (to account for few numbers of replicates and large dynamic expression range)

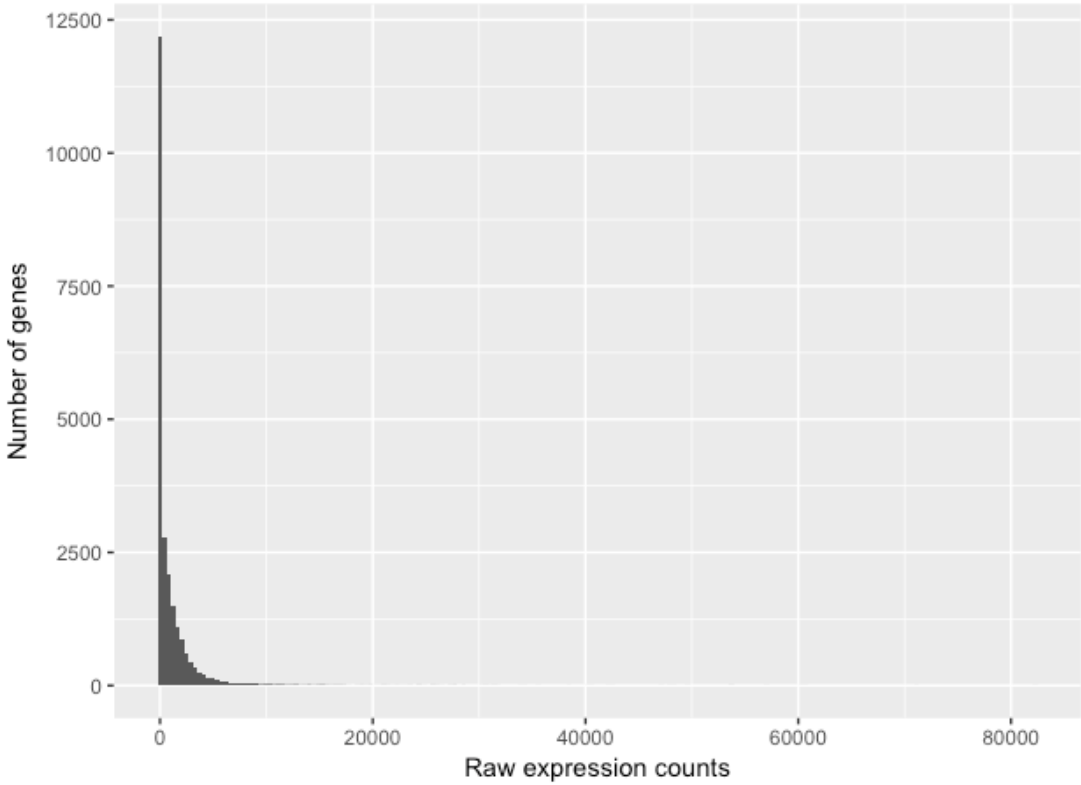
Differential gene expression (DGE) analysis

- **Biological replicates** represent multiple samples (i.e. RNA from different mice) representing the same sample class
- **Technical replicates** represent the same sample (i.e. RNA from the same mouse) but with technical steps replicated
- Usually biological variance is much greater than technical variance, so we do not need to account for technical variance to identify biological differences in expression

Differential gene expression (DGE) analysis

To determine the appropriate statistical model, we need information about the distribution of counts.

low number of counts associated with a large proportion of genes



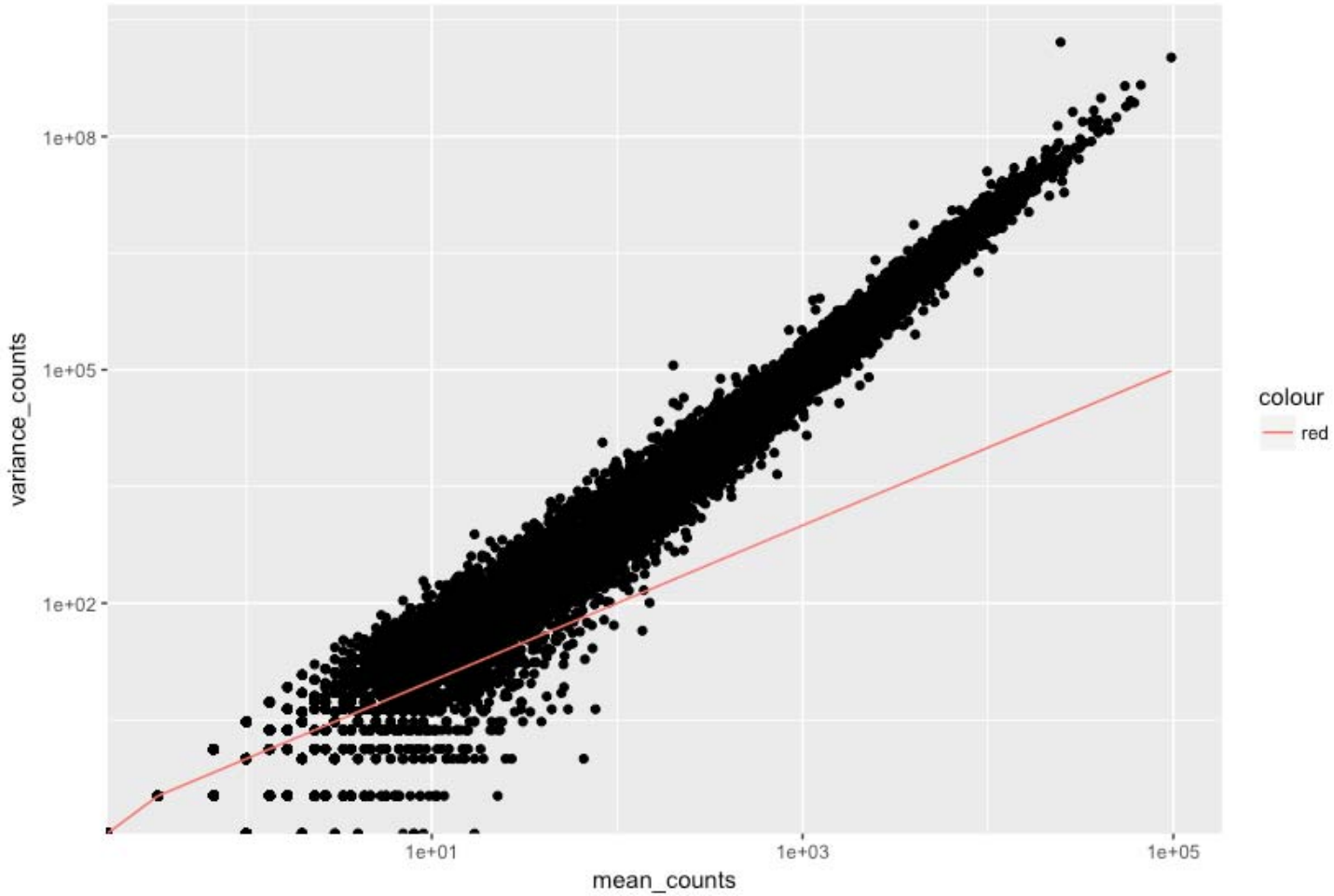
a long right tail due to the **lack of any upper limit for expression**

Differential gene expression (DGE) analysis

Modeling count data

- Count data: often modeled using the binomial distribution (number of possible outcomes = 2)
- RNA-seq data
  - **large number of RNAs are represented** (large number of possible outcomes)
  - **probability of pulling out a particular transcript is very small**
  - **Poisson distribution**
- RNA-seq data
  - **biological variation across biological replicates**
  - genes with larger average expression levels will tend to have larger observed variances across replicates
  - **Negative Binomial (NB) distribution !**

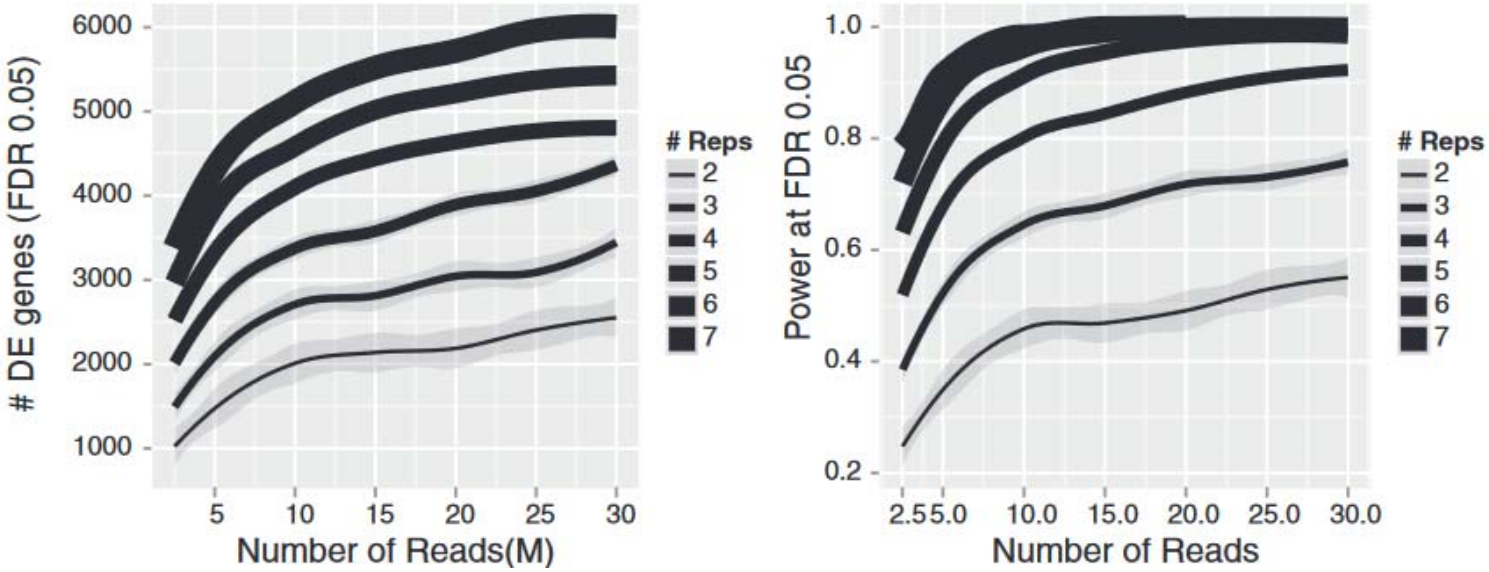
Differential gene expression (DGE) analysis



Differential gene expression (DGE) analysis

more biological replicates

- **more precise estimates of group means**
- **greater confidence in the ability to distinguish differences between sample classes (i.e. more DE genes)**



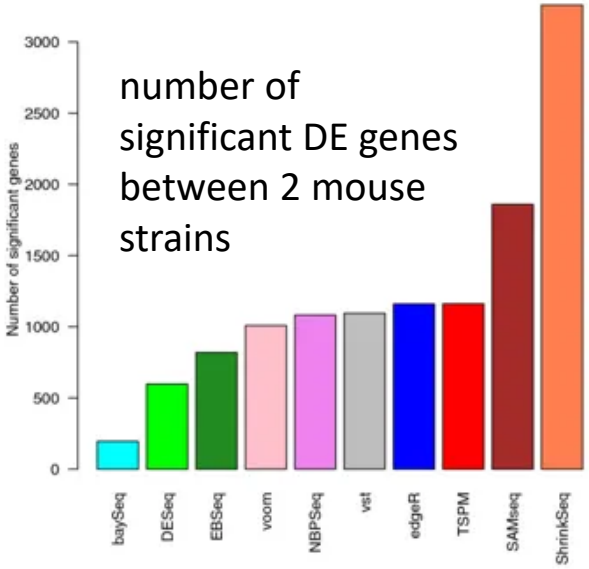
an increase in the number of replicates tends to return more DE genes than increasing the sequencing depth !!!

Differential gene expression (DGE) analysis

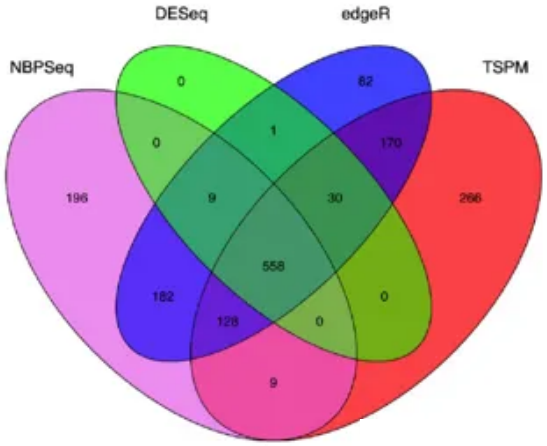
Differential expression analysis workflow

- DESeq2 and EdgeR
  - continuously being developed
  - recommended as best practice
  - use the negative binomial model
  - yield similar results
- Limma-Voom
  - another set of tools often used together for DE analysis
  - less sensitive for small sample sizes
  - method recommended when number of biological replicates per group > 20

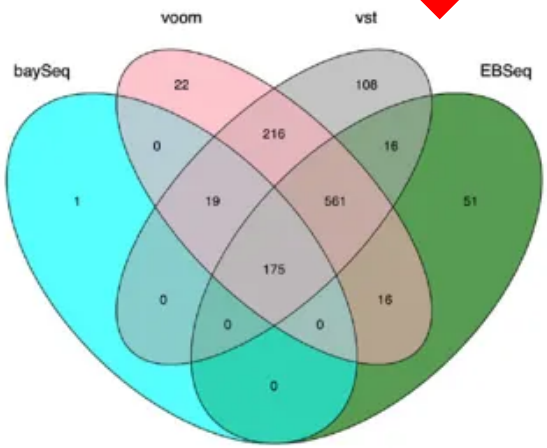
Differential gene expression (DGE) analysis

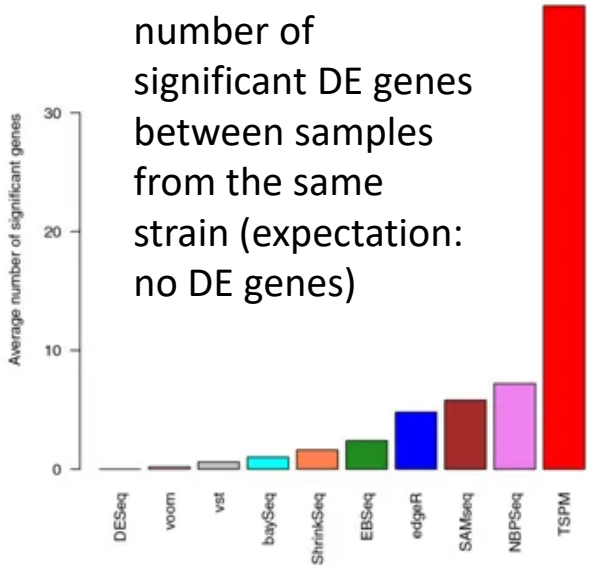


number of significant DE genes between 2 mouse strains



overlap among the set of DE genes found by different methods





number of significant DE genes between samples from the same strain (expectation: no DE genes)

there is no one method that performs optimally under all conditions !!

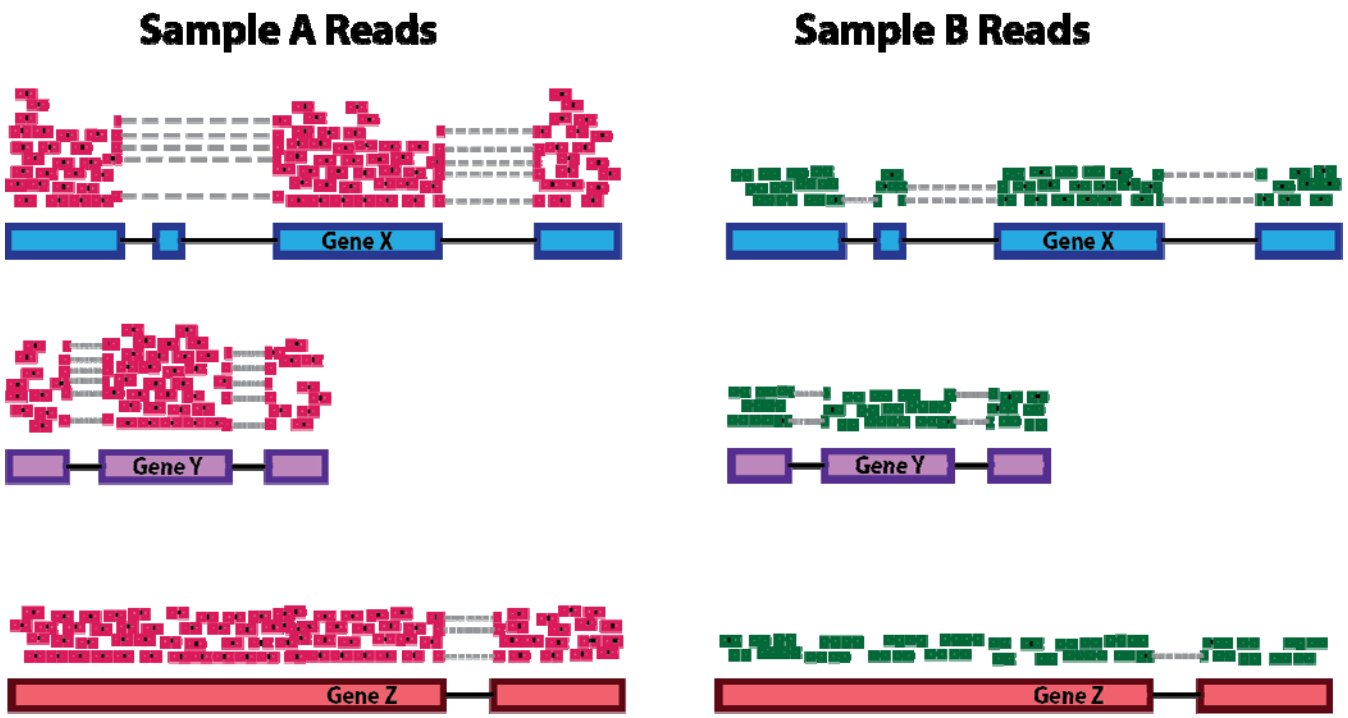
Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14, 91.

Differential gene expression (DGE) analysis

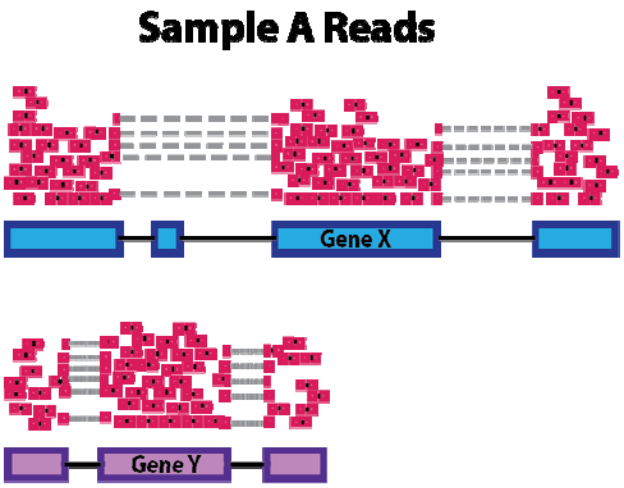
Normalization

- necessary to make accurate comparisons of gene expression between samples
- ***essential for differential expression analyses***
- ***also necessary for exploratory data analysis, visualization of data***
- main factors often considered during normalization
  - ❑ Sequencing depth
  - ❑ Gene length
  - ❑ RNA composition

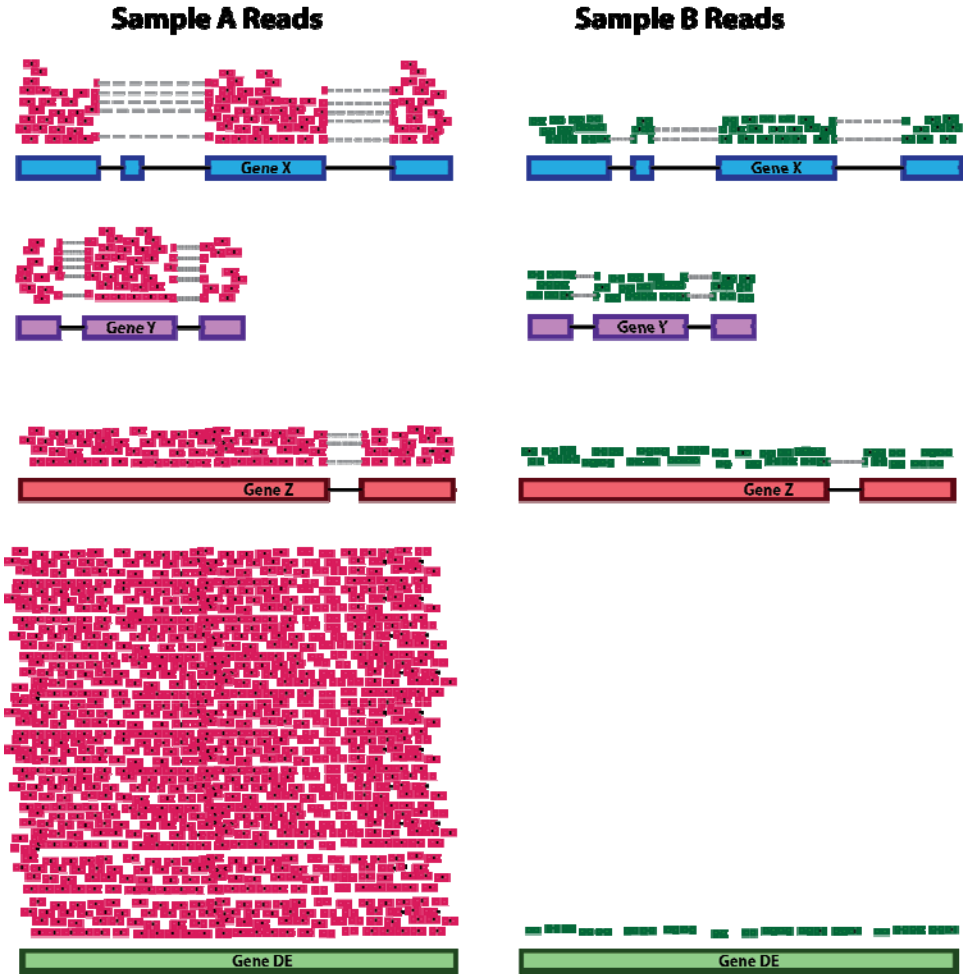
Differential gene expression (DGE) analysis



Differential gene expression (DGE) analysis



Differential gene expression (DGE) analysis



Differential gene expression (DGE) analysis

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**	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)**	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

Differential gene expression (DGE) analysis

RPKM/FPKM (not recommended)

- normalized count values output by the RPKM/FPKM method are not comparable between samples
- RPKM: single-end reads
- FPKM: paired-end reads

gene name	read counts S1	read counts S2	read counts S3
A (2 Kb)	10	12	30
B (4 Kb)	20	25	60
C (1 Kb)	5	8	15
D (10 Kb)	0	0	1
total	35	45	106
total/10(6)	0.000035	0.000045	0.000106

gene name	RPM S1	RPM S2	RPM S3
A (2 Kb)	285714.29	266666.67	283018.87
B (4 Kb)	571428.57	555555.56	566037.74
C (1 Kb)	142857.14	177777.78	141509.43
D (10 Kb)	0.00	0.00	9433.96
total	1000000.00	1000000.00	1000000.00

gene name	RPKM S1	RPKM S2	RPKM S3
A (2 Kb)	142857.14	133333.33	141509.43
B (4 Kb)	142857.14	138888.89	141509.43
C (1 Kb)	142857.14	177777.78	141509.43
D (10 Kb)	0.00	0.00	943.40
	428571.43	450000.00	425471.70

sum of normalized reads in each sample is different !!!

Differential gene expression (DGE) analysis

DESeq2-normalized counts: Median of ratios method

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

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**

- Step 2: calculates ratio of each sample to the reference

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**

since the majority of genes are not differentially expressed, the majority of genes in each sample should have similar ratios within the sample

Differential gene expression (DGE) analysis

DESeq2-normalized counts: Median of ratios method

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

```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))
```

```
normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```

- Step 4: calculate the normalized count values using the normalization factor

Normalized Counts

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...

Differential gene expression (DGE) analysis

DESeq2-normalized counts: Median of ratios method

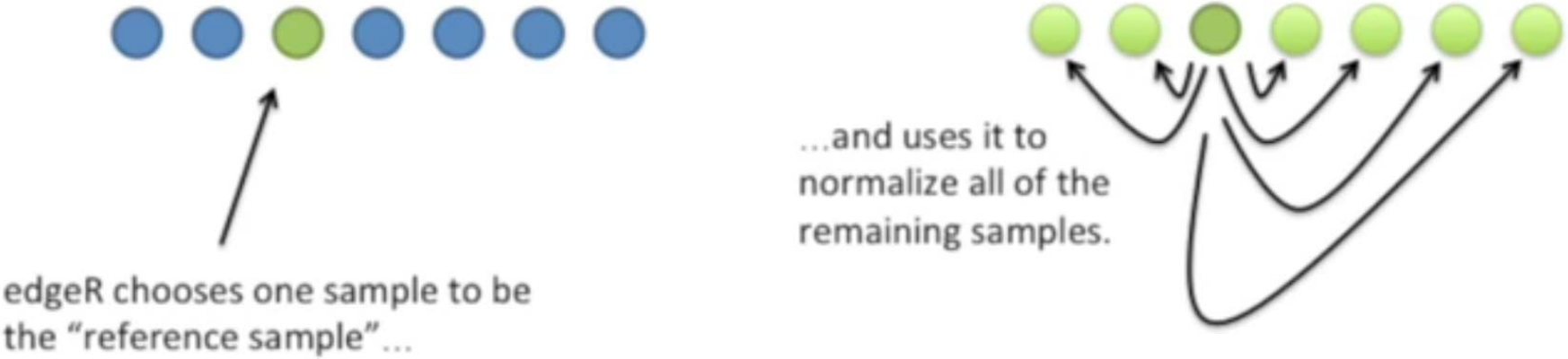
DESeq2 doesn't actually use normalized counts, rather it uses the raw counts and models the normalization inside the Generalized Linear Model (GLM).

These normalized counts will be useful for downstream visualization of results, but cannot be used as input to DESeq2 or any other tools that perform differential expression analysis which use the negative binomial model.

Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

- Step 1: picks one sample as a reference sample



The illustrations of the EdgeR normalizations were taken from Josh Starmer (StatQuest)

Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

An example of an extremely bad "reference sample"

To avoid choosing extreme samples, edgeR attempts to identify the most "average" sample.

Let's see how it does this!

	Sample #1	Sample #2	Sample #3
Gene1	0	10	0
Gene2	2	6	0
Gene3	33	55	200
Gene4	12	40	0
Gene5	117	187	0
Gene6	86	123	0
Gene7	34	91	0
Gene8	10	72	0
Gene9	217	250	0

Sample #3 would be a terrible reference sample.

Scaling would be based on a single, potentially very noisy, measurement.

Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

Original read counts

	Sample #1	Sample #2	Sample #3
Gene1	0	10	4
Gene2	2	6	12
Gene3	33	55	200
Gene4	12	40	33
Total reads:	47	111	249

Scaled read counts

	Sample #1	Sample #2	Sample #3
Gene1	0/47	10/111	4/249
Gene2	2/47	5/111	12/249
Gene3	33/47	55/111	200/249
Gene4	12/47	40/111	33/249

Original read counts... ...divided by the total

Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

Part a: Scale each sample by its total read counts.

Part b: For each sample, determine the value such that 75% of the scaled data are equal to or smaller than it.

Scaled read counts

	Sample #1	Sample #2	Sample #3
Gene1	0	0.09	0.02
Gene2	0.04	0.05	0.05
Gene3	0.70	0.50	0.80
Gene4	0.26	0.36	0.13

In Sample #1, 3 of the 4 values (75%) are less than or equal to 0.26

Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

Part a: Scale each sample by its total read counts.

Part b: For each sample, determine the value such that 75% of the scaled data are equal to or smaller than it.

Sample#1 will be the reference sample.

Scaled read counts

	Reference	Sample #2	Sample #3
Gene1	0	0.09	0.02
Gene2	0.04	0.05	0.05
Gene3	0.70	0.50	0.80
Gene4	0.26	0.36	0.13
75th Quantiles:	0.26	0.36	0.13

Average = 0.25

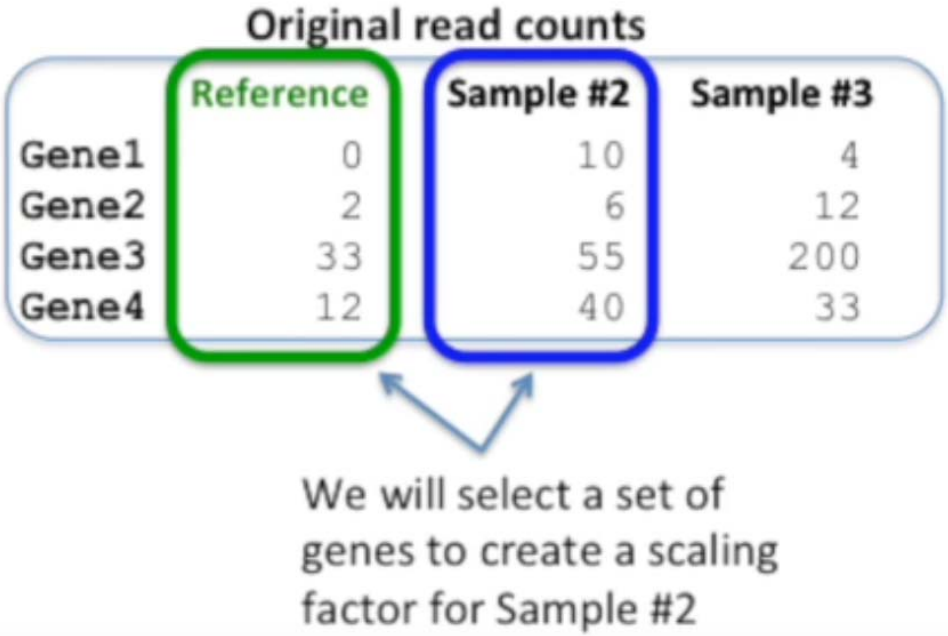
Part c: Average the 75th quantiles.

Part d: The "reference sample" is the one who's 75th quintile is closest to the average.

Differential gene expression (DGE) analysis

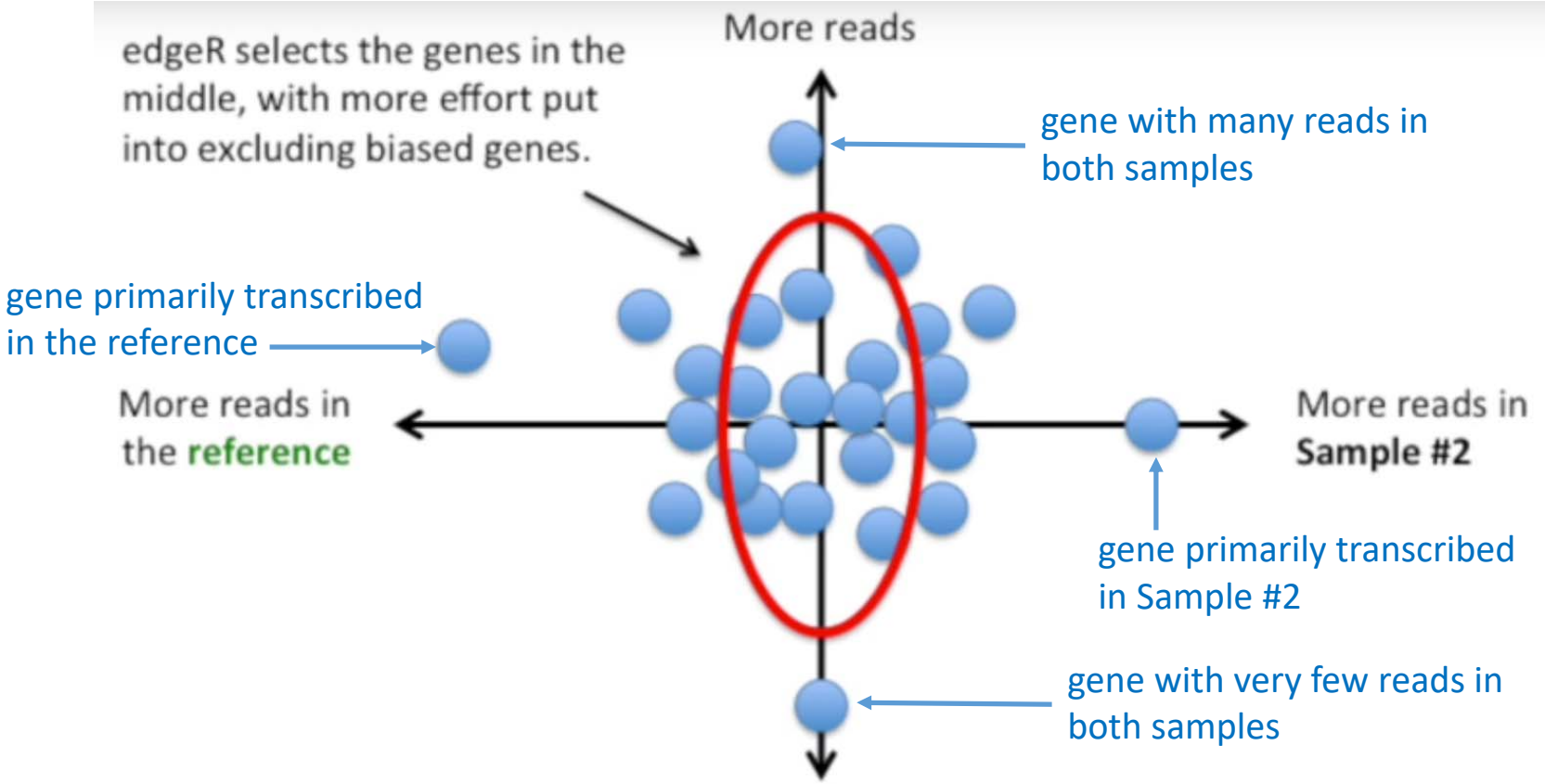
EdgeR-normalized counts: Trimmed mean of M values (TMM)

- Step 2: selects the genes for calculating the scaling factors. This is done separately for each sample relative to the "reference sample".



Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)



Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

One to identify biased genes…

…and one to identify genes that are highly and lowly transcribed in both samples.

$$\frac{\log_2(\text{Reference}) + \log_2(\text{Sample \#2})}{2}$$

$\log_2$(Reference / Sample #2)

Gene2	-0.32
Gene3	0.49
Gene4	-0.47
…	
GeneN	-0.21

Mean of logs

Gene2	-4.48
Gene3	-0.76
Gene4	-1.71
…	
GeneN	-2.84

Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

Sort both tables from low to high.

Filter out the top 30% and the bottom 30% biased genes.

Filter out the top 5% and the bottom 5% of the highly and lowly transcribed genes.

$\log_2$(Reference / Sample #2)

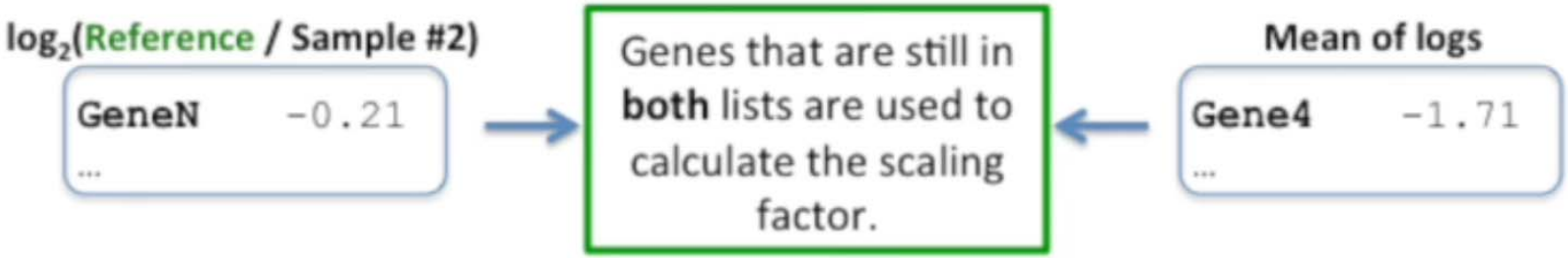
Gene4	-0.47
Gene2	-0.32
GeneN	-0.21
…	
Gene3	0.49

Mean of logs

Gene2	-4.48
GeneN	-2.84
Gene4	-1.71
…	
Gene3	-0.76

Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)



Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

- Step 3: calculates the weighted average of the remaining $\log_2$ ratios

$\log_2$(Reference / Sample #2)

GeneA	-0.07
GeneB	-0.02
GeneC	0.21
...	
GeneZ	0.49

Once you have selected which genes will be used to calculate the scaling factor, just calculate the **weighted average** of their $\log_2$ ratios.

Differential gene expression (DGE) analysis

EdgeR-normalized counts: Trimmed mean of M values (TMM)

- Step 3: calculates the weighted average of the remaining $\log_2$ ratios

Read Counts

	Sample #1	Sample #2	$\log_2$(ratio)
Gene #1	202	101	1
Gene #2	204	101	1.01
Gene #3	206	101	1.02
Gene #4	2	1	1
Gene #5	4	1	2
Gene #6	6	1	2.6

Genes with more reads mapped to them get more weight.

This is because log ratios have more variance with low read counts.

- Step 4: compute a scaling factor $_2$weighted average of $\log_2$ ratios