Normalization and differential analysis of RNA-seq data

Nathalie Villa-Vialaneix
INRA, Toulouse, MIAT
(Mathématiques et Informatique Appliquées de Toulouse)

nathalie.villa@toulouse.inra.fr
http://www.nathalievilla.org

SPS Summer School 2016

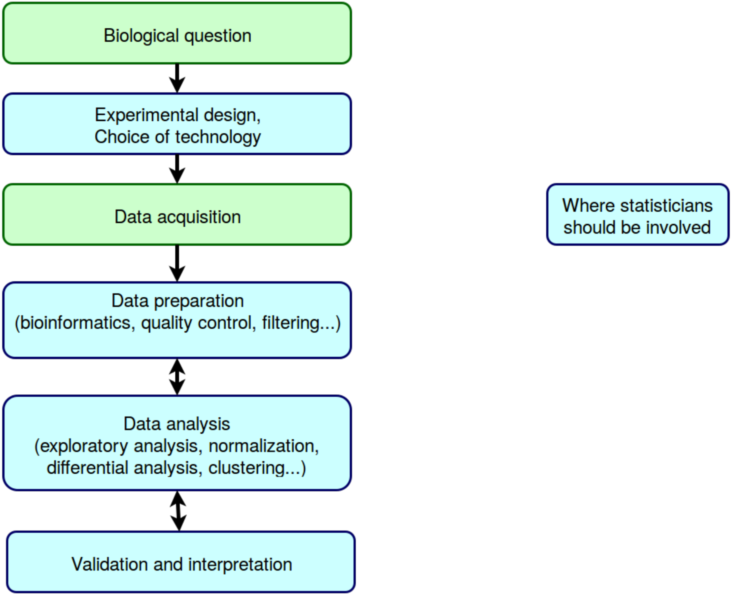
From gene expression to genomic network

Outline

1. Normalization

2. Differential expression analysis
   - Hypothesis testing and correction for multiple tests
   - Differential expression analysis for RNAseq data

A typical transcriptomic experiment



Biological question

Experimental design,
Choice of technology

Data acquisition

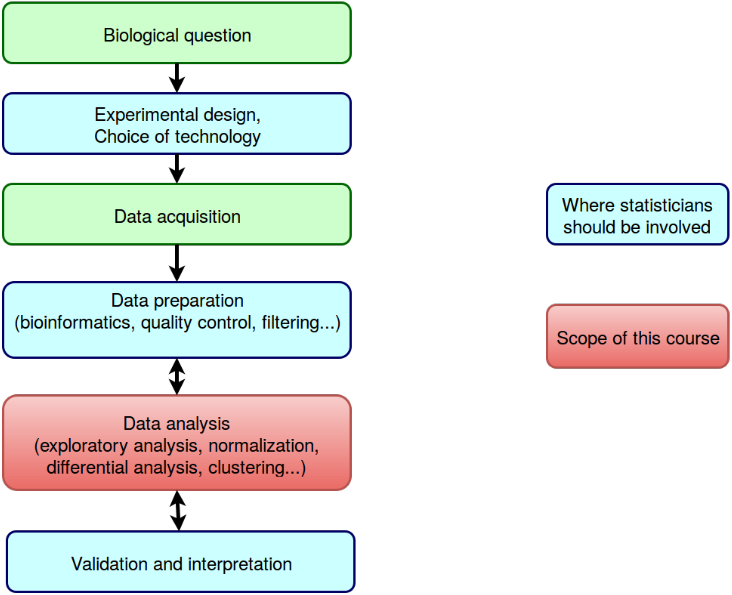
Data preparation
(bioinformatics, quality control, filtering...)

Data analysis
(exploratory analysis, normalization,
differential analysis, clustering...)

Validation and interpretation

Where statisticians
should be involved

A typical transcriptomic experiment



Biological question

Experimental design,
Choice of technology

Data acquisition

Data preparation
(bioinformatics, quality control, filtering...)

Data analysis
(exploratory analysis, normalization,
differential analysis, clustering...)

Validation and interpretation

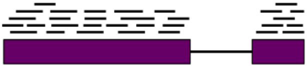
Where statisticians
should be involved

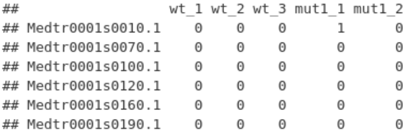
Scope of this course

Some features of RNAseq data

What must be taken into account?

- discrete, non-negative data (total number of aligned reads)



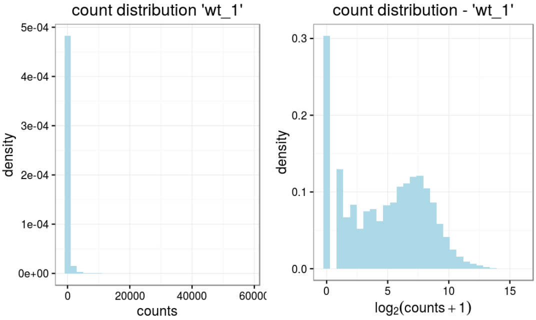


```
##                   wt_1 wt_2 wt_3 mut1_1 mut1_2
## Medtr0001s0010.1    0    0    0      1      0
## Medtr0001s0070.1    0    0    0      0      0
## Medtr0001s0100.1    0    0    0      0      0
## Medtr0001s0120.1    0    0    0      0      0
## Medtr0001s0160.1    0    0    0      0      0
## Medtr0001s0190.1    0    0    0      0      0
```

Some features of RNAseq data

What must be taken into account?

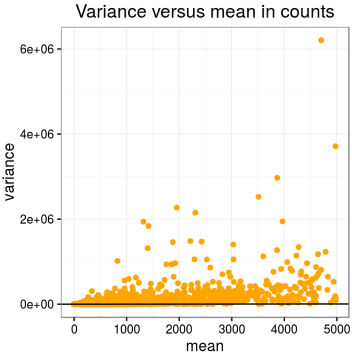
- discrete, non-negative data (total number of aligned reads)
- skewed data



Some features of RNAseq data

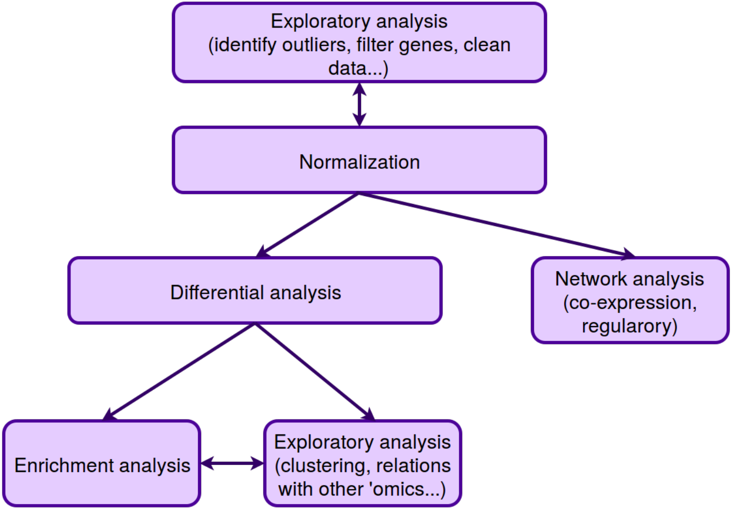
What must be taken into account?

- discrete, non-negative data (total number of aligned reads)
- skewed data
- overdispersion (variance $\gg$ mean)



black line is
"variance = mean"

Steps in RNAseq data analysis



Exploratory analysis (identify outliers, filter genes, clean data...)

Normalization

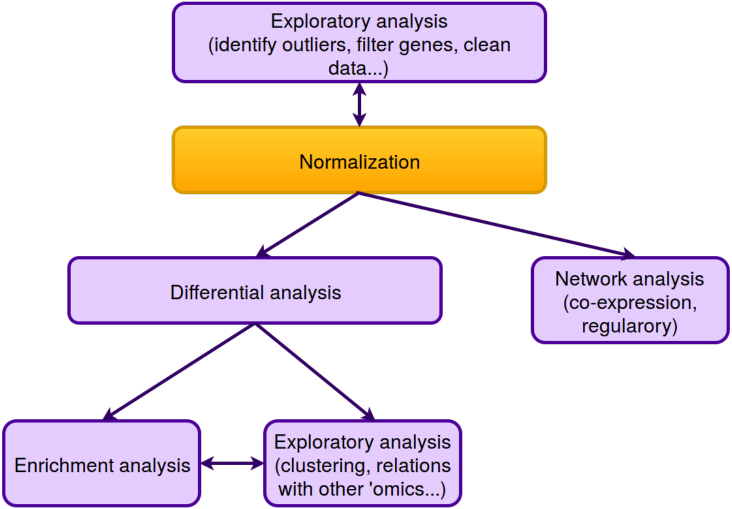
Differential analysis

Network analysis (co-expression, regularory)

Enrichment analysis

Exploratory analysis (clustering, relations with other 'omics...)

Part I: Normalization

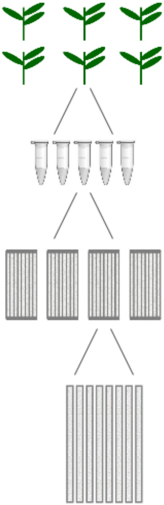

Exploratory analysis (identify outliers, filter genes, clean data...)
Normalization
Differential analysis
Network analysis (co-expression, regularory)
Enrichment analysis
Exploratory analysis (clustering, relations with other 'omics...)

Purpose of normalization

- identify and correct technical biases (due to sequencing process) to make counts comparable

- types of normalization: within sample normalization and between sample normalization

Source of variation in RNA-seq experiments

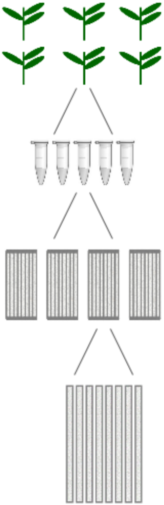


1. at the top layer: biological variations (*i.e.*, individual differences due to *e.g.*, environmental or genetic factors)

2. at the middle layer: technical variations (library effect)

3. at the bottom layer: technical variations (lane and cell flow effects)

Source of variation in RNA-seq experiments



1. at the top layer: biological variations (*i.e.*, individual differences due to *e.g.*, environmental or genetic factors)
2. at the middle layer: technical variations (library effect)
3. at the bottom layer: technical variations (lane and cell flow effects)

lane effect $<$ cell flow effect $<$ library effect $\ll$ biological effect

Within sample normalization

Example: (read counts)

	sample 1	sample 2	sample 3
gene A	752	615	1203
gene B	1507	1225	2455

counts for gene B are twice larger than counts for gene A because:

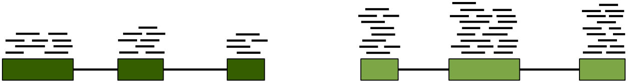
Within sample normalization

Example: (read counts)

	sample 1	sample 2	sample 3
gene A	752	615	1203
gene B	1507	1225	2455

counts for gene B are twice larger than counts for gene A because:

- gene B is expressed with a number of transcripts twice larger than gene A



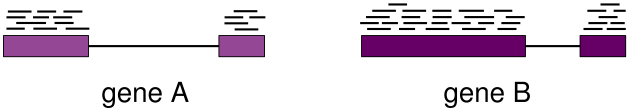
Within sample normalization

Example: (read counts)

	sample 1	sample 2	sample 3
gene A	752	615	1203
gene B	1507	1225	2455

counts for gene B are twice larger than counts for gene A because:

- both genes are expressed with the same number of transcripts but gene B is twice longer than gene A



Within sample normalization

- Purpose of within sample comparison: enabling comparisons of genes from a same sample
- Sources of variability: gene length, sequence composition (GC content)

These differences need not to be corrected for a differential analysis and are not really relevant for data interpretation.

Between sample normalization

Example: (read counts)

	sample 1	sample 2	sample 3
gene A	752	615	1203
gene B	1507	1225	2455

counts in sample 3 are much larger than counts in sample 2 because:

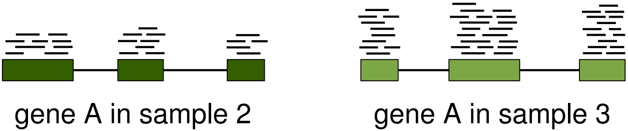
Between sample normalization

Example: (read counts)

	sample 1	sample 2	sample 3
gene A	752	615	1203
gene B	1507	1225	2455

counts in sample 3 are much larger than counts in sample 2 because:

- gene A is more expressed in sample 3 than in sample 2



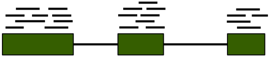
Between sample normalization

Example: (read counts)

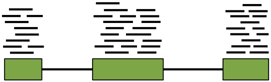
	sample 1	sample 2	sample 3
gene A	752	615	1203
gene B	1507	1225	2455

counts in sample 3 are much larger than counts in sample 2 because:

- gene A is expressed similarly in the two samples but sequencing depth is larger in sample 3 than in sample 2 (*i.e.*, differences in library sizes)



gene A in sample 2



gene A in sample 3

Between sample normalization

- Purpose of between sample comparison: enabling comparisons of a gene in different samples
- Sources of variability: library size, ...

These differences must be corrected for a differential analysis and for data interpretation.

Principles for sequencing depth normalization

Basics

1. choose an appropriate baseline for each sample
2. for a given gene, compare counts relative to the baseline rather than raw counts

Principles for sequencing depth normalization

Basics

1. choose an appropriate baseline for each sample
2. for a given gene, compare counts relative to the baseline rather than raw counts

In practice: Raw counts correspond to different sequencing depths

	control				treated		
Gene 1	5	1	0	0	**4**	**0**	**0**
Gene 2	0	2	1	2	**1**	**0**	**0**
Gene 3	92	161	76	70	**140**	**88**	**70**
:		:				:	
:		:				:	
:		:				:	
Gene G	15	25	9	5	**20**	**14**	**17**

Principles for sequencing depth normalization

Basics

1. choose an appropriate baseline for each sample
2. for a given gene, compare counts relative to the baseline rather than raw counts

In practice: A correction multiplicative factor is calculated for every sample

	control				treated		
Gene 1	5	1	0	0	**4**	**0**	**0**
Gene 2	0	2	1	2	**1**	**0**	**0**
Gene 3	92	161	76	70	**140**	**88**	**70**
:		:				:	
:		:				:	
:		:				:	
Gene G	15	25	9	5	**20**	**14**	**17**
C_j	1.1	1.6	0.6	0.7	**1.4**	**0.7**	**0.8**

Principles for sequencing depth normalization

Basics

1. choose an appropriate baseline for each sample
2. for a given gene, compare counts relative to the baseline rather than raw counts

In practice: Every counts is multiplied by the correction factor corresponding to its sample

Gene 3	92	161	76	70	**140**	**88**	**70**	
C_j	1.1	1.6	0.6	0.7	**1.4**	**0.7**	**0.8**	x
Gene 3	101.2	257.6	45.6	49	**196**	**61.6**	**56**	

Principles for sequencing depth normalization

Basics

1. choose an appropriate baseline for each sample
2. for a given gene, compare counts relative to the baseline rather than raw counts

Consequences: Library sizes for normalized counts are roughly equal.

	control				treated			
Gene 1	5.5	1.6	0	0	**5.6**	**0**	**0**	
Gene 2	0	3.2	0.6	1.4	**1.4**	**0**	**0**	
Gene 3	101.2	257.6	45.6	49	**196**	**61.6**	**56**	
:		:				:		
:		:				:		
:		:				:		
Gene G	16.5	40	5.4	5.5	**28**	**9.8**	**13.6**	+
Lib. size	13.1	13.0	13.2	13.1	**13.2**	**13.0**	**13.1**	x 10^5

Principles for sequencing depth normalization

Definition

If K_{gj} is the raw count for gene g in sample j then, the normalized counts is defined as:

$$\widetilde{K}_{gj} = \frac{K_{gj}}{s_j}$$

in which $s_j = C_j^{-1}$ is the scaling factor for sample j.

Principles for sequencing depth normalization

Definition

If K_{gj} is the raw count for gene g in sample j then, the normalized counts is defined as:

$$\widetilde{K}_{gj} = \frac{K_{gj}}{s_j}$$

in which $s_j = C_j^{-1}$ is the scaling factor for sample j.

Three types of methods:

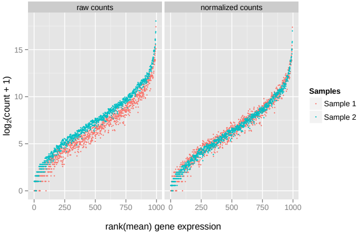
- distribution adjustment
- method taking length into account
- the “effective library size” concept

Distribution adjustment

- Total read count adjustment [Mortazavi et al., 2008]

$$s_j = \frac{D_j}{\frac{1}{N}\sum_{l=1}^{N} D_l}$$

in which N is the number of samples and $D_j = \sum_g K_{gj}$.



edgeR:

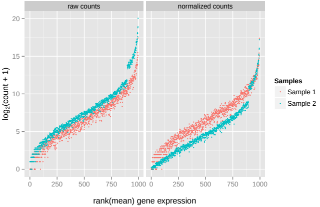
```
cpm(...,
    normalized.lib.sizes=TRUE)
```

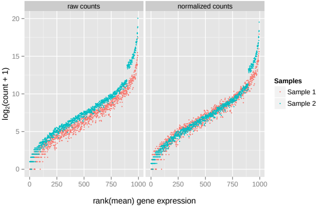
Distribution adjustment

- Total read count adjustment [Mortazavi et al., 2008]
- (Upper) Quartile normalization [Bullard et al., 2010]

$$s_j = \frac{Q_j^{(p)}}{\frac{1}{} \sum_{l=1}^{N} Q_l^{(p)}}$$

in which $Q_j^{(p)}$ is a given quantile (generally 3rd quartile) of the count distribution in sample j.





edgeR:

```
calcNormFactors(..., method = "upperquartile",
                     p = 0.75)
```

Method using gene lengths (intra & inter sample normalization)

RPKM: Reads Per Kilobase per Million mapped Reads

Assumptions: read counts are proportional to expression level, transcript length and sequencing depth

$$s_j = \frac{D_j L_g}{10^3 \times 10^6}$$

in which L_g is gene length (bp).

edgeR:

```
rpkm(..., gene.length = ...)
```

Unbiaised estimation of number of reads but affect variability [Oshlack and Wakefield, 2009].

Relative Log Expression (RLE)

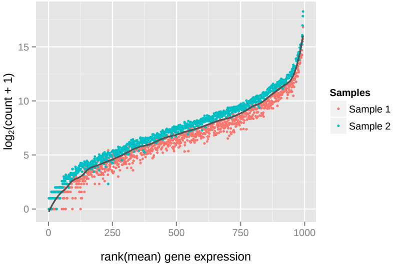
[Anders and Huber, 2010], **edgeR** - **DESeq** - **DESeq2**

Method:

1. compute a pseudo-reference sample: geometric mean across samples

$$R_g = \left(\prod_{j=1}^{N} K_{gj}\right)^{1/N}$$

(geometric mean is less sensitive to extreme values than standard mean)



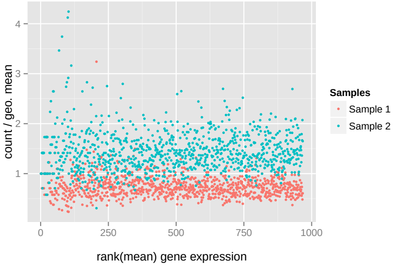
Relative Log Expression (RLE)

[Anders and Huber, 2010], **edgeR** - **DESeq** - **DESeq2**

Method:

1. compute a pseudo-reference sample
2. center samples compared to the reference

$$\tilde{K}_{gj} = \frac{K_{gj}}{R_g} \qquad \text{with} \qquad R_g = \left(\prod_{j=1}^{N} K_{gj} \right)^{1/N}$$



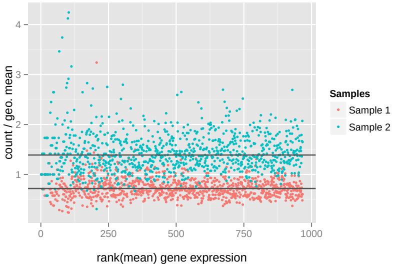
Relative Log Expression (RLE)

[Anders and Huber, 2010], **edgeR** - **DESeq** - **DESeq2**

Method:

1. compute a pseudo-reference sample
2. center samples compared to the reference
3. calculate normalization factor: median of centered counts over the genes

$$\tilde{s}_j = \operatorname*{median}_g \left\{ \tilde{K}_{gj} \right\} \quad \text{factors multiply to 1:} \quad s_j = \frac{\tilde{s}_j}{\exp\left(\frac{1}{N}\sum_{l=1}^N \log(\tilde{s}_l)\right)}$$



with

$$\tilde{K}_{gj} = \frac{K_{gj}}{R_g}$$

and

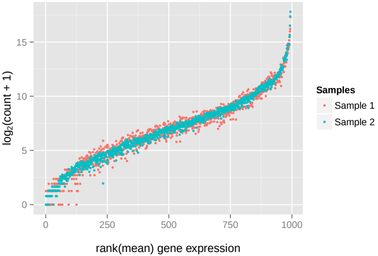
$$R_g = \left(\prod_{j=1}^N K_{gj} \right)^{1/N}$$

Relative Log Expression (RLE)

[Anders and Huber, 2010], **edgeR** - **DESeq** - **DESeq2**

Method:

1. compute a pseudo-reference sample
2. center samples compared to the reference
3. calculate normalization factor: median of centered counts over the genes



```
## with edgeR
calcNormFactors(...,
  method="RLE")

## with DESeq
estimateSizeFactors(...)
```

Trimmed Mean of M-values (TMM)

[Robinson and Oshlack, 2010], **edgeR**

Assumptions behind the method

- the total read count strongly depends on a few highly expressed genes
- most genes are not differentially expressed

Trimmed Mean of M-values (TMM)

[Robinson and Oshlack, 2010], **edgeR**

Assumptions behind the method

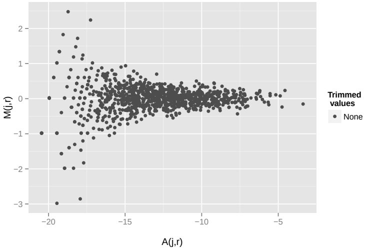
- the total read count strongly depends on a few highly expressed genes
- most genes are not differentially expressed

⇒ remove extreme data for fold-changed (M) and average intensity (A)

$$M_g(j,r) = \log_2\left(\frac{K_{gj}}{D_j}\right) - \log_2\left(\frac{K_{gr}}{D_r}\right) \qquad A_g(j,r) = \frac{1}{2}\left[\log_2\left(\frac{K_{gj}}{D_j}\right) + \log_2\left(\frac{K_{gr}}{D_r}\right)\right]$$

select as a reference sample, the sample r with the upper quartile closest to the average upper quartile

M- vs A-values



Trimmed Mean of M-values (TMM)

[Robinson and Oshlack, 2010], **edgeR**

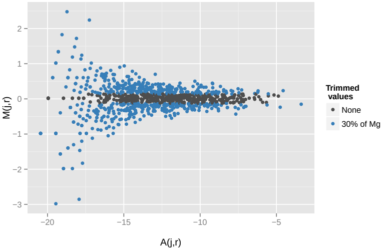
Assumptions behind the method

- the total read count strongly depends on a few highly expressed genes
- most genes are not differentially expressed

⇒ remove extreme data for fold-changed (M) and average intensity (A)

$$M_g(j,r) = \log_2\left(\frac{K_{gj}}{D_j}\right) - \log_2\left(\frac{K_{gr}}{D_r}\right) \qquad A_g(j,r) = \frac{1}{2}\left[\log_2\left(\frac{K_{gj}}{D_j}\right) + \log_2\left(\frac{K_{gr}}{D_r}\right)\right]$$

Trim 30% on M-values



Trimmed Mean of M-values (TMM)

[Robinson and Oshlack, 2010], **edgeR**

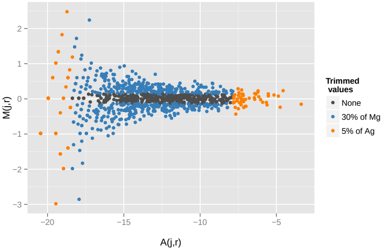
Assumptions behind the method

- the total read count strongly depends on a few highly expressed genes
- most genes are not differentially expressed

⇒ remove extreme data for fold-changed (M) and average intensity (A)

$$M_g(j,r) = \log_2\left(\frac{K_{gj}}{D_j}\right) - \log_2\left(\frac{K_{gr}}{D_r}\right) \qquad A_g(j,r) = \frac{1}{2}\left[\log_2\left(\frac{K_{gj}}{D_j}\right) + \log_2\left(\frac{K_{gr}}{D_r}\right)\right]$$

Trim 5% on A-values

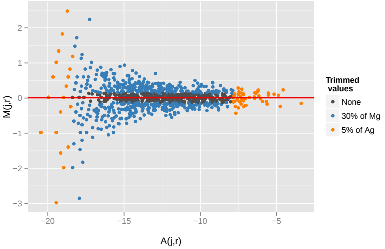


Trimmed Mean of M-values (TMM)

[Robinson and Oshlack, 2010], **edgeR**

Assumptions behind the method

- the total read count strongly depends on a few highly expressed genes
- most genes are not differentially expressed



On remaining data, calculate the weighted mean of M-values:

$$\mathrm{TMM}(j,r) = \frac{\sum_{g:\text{not trimmed}} w_g(j,r) M_g(j,r)}{\sum_{g:\text{not trimmed}} w_g(j,r)}$$

with $w_g(j,r) = \left(\frac{D_j - K_{gj}}{D_j K_{gj}} + \frac{D_r - K_{gr}}{D_r K_{gr}}\right)$.

Trimmed Mean of M-values (TMM)

[Robinson and Oshlack, 2010], **edgeR**

Assumptions behind the method

- the total read count strongly depends on a few highly expressed genes
- most genes are not differentially expressed

Correction factors:

$$\tilde{s}_j = 2^{\text{TMM}(j,r)} \quad \text{factors multiply to 1:} \quad s_j = \frac{\tilde{s}_j}{\exp\left(\frac{1}{N}\sum_{l=1}^{N}\log(\tilde{s}_l)\right)}$$

```
calcNormFactors(..., method="TMM")
```

Comparison of the different approaches

[Dillies et al., 2013], (6 simulated datasets)

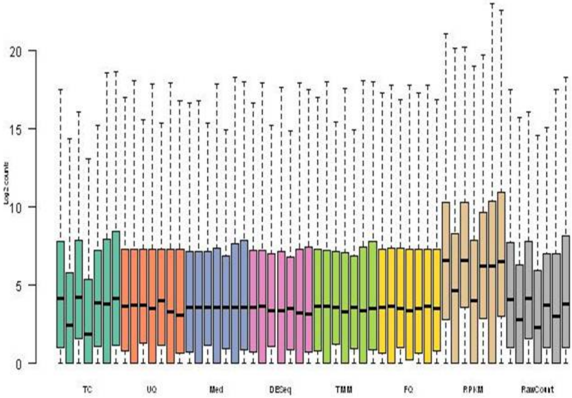
Purpose of the comparison:

- finding the “best” method for all cases is not a realistic purpose
- find an approach which is robust enough to provide relevant results in all cases
- Method: comparison based on several criteria to select a method which is valid for multiple objectives

Comparison of the different approaches

[Dillies et al., 2013], (6 simulated datasets)

Effect on count distribution:

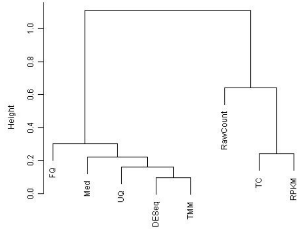


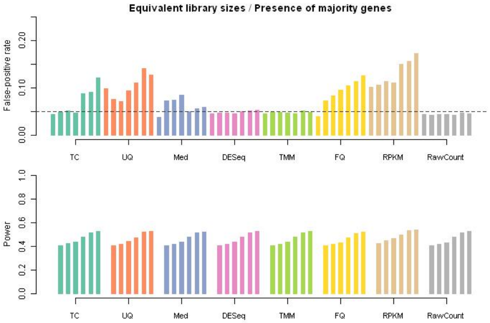
RPKM and TC are very similar to raw data.

Comparison of the different approaches

[Dillies et al., 2013], (6 simulated datasets)

Effect on differential analysis (DESeq v. 1.6):



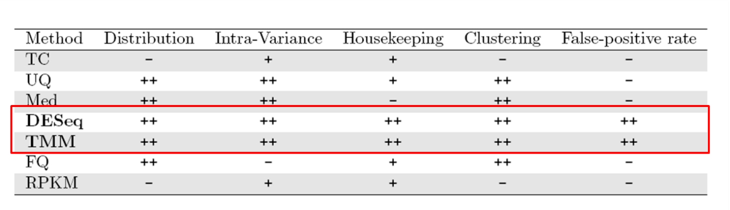


Inflated FPR for all methods except for TMM and DESeq (RLE).

Comparison of the different approaches

[Dillies et al., 2013], (6 simulated datasets)

Conclusion: Differences appear based on data characteristics

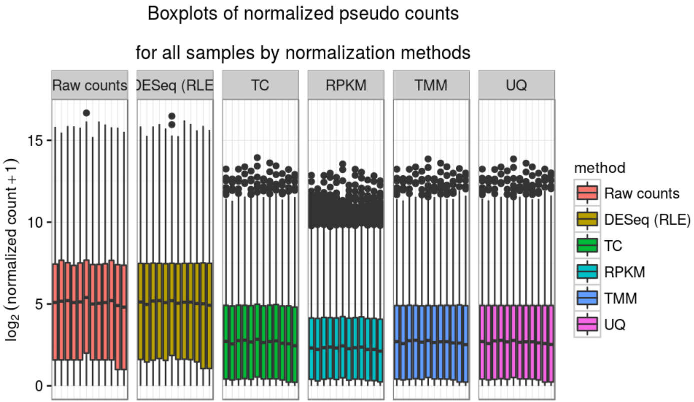


Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
FQ	++	–	+	++	–
RPKM	–	+	+	–	–

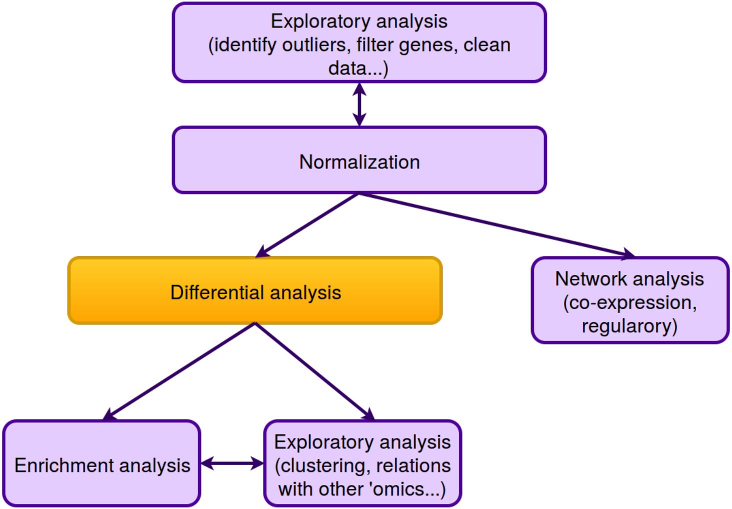
TMM and DESeq (RLE) are performant in a differential analysis context.

Practical session

- import and understand data;
- run different types of normalization;
- compare the results...



Part II: Differential expression analysis



Exploratory analysis (identify outliers, filter genes, clean data...)

Normalization

Differential analysis

Network analysis (co-expression, regularory)

Enrichment analysis

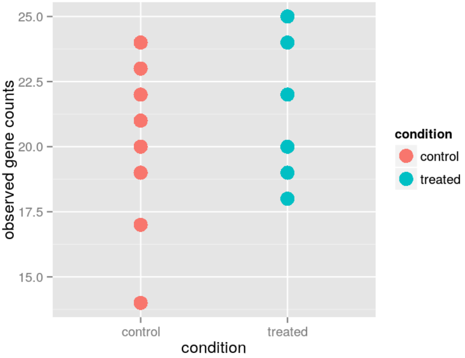
Exploratory analysis (clustering, relations with other 'omics...)

Different steps in hypothesis testing

1. formulate an hypothesis H_0:

   H_0: the average count for gene *g* in the control samples is the same that the average count in the treated samples

   which is tested against an alternative H_1: the average count for gene *g* in the control samples is different from the average count in the treated samples

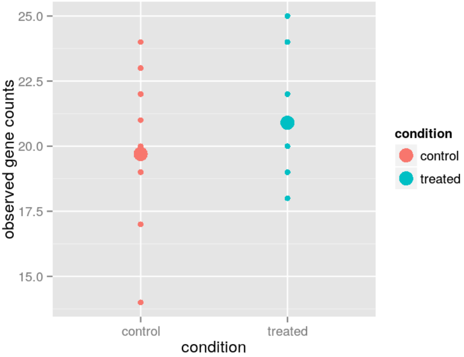


Different steps in hypothesis testing

1. formulate an hypothesis H_0:

   H_0: the average count for gene *g* in the control samples is the same that the average count in the treated samples

2. from observations, calculate a test statistics (*e.g.*, the mean in the two samples)



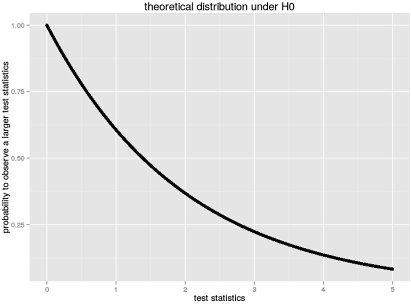
Different steps in hypothesis testing

1. formulate an hypothesis H_0:

   H_0: the average count for gene *g* in the control samples is the same that the average count in the treated samples

2. from observations, calculate a test statistics (*e.g.*, the mean in the two samples)

3. find the theoretical distribution of the test statistics under H_0

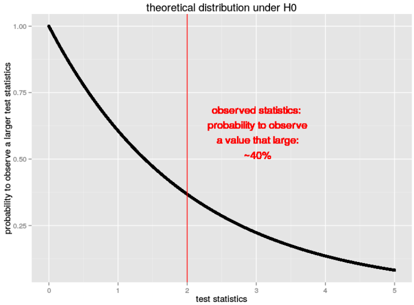


Different steps in hypothesis testing

1. formulate an hypothesis H_0:

   H_0: the average count for gene *g* in the control samples is the same that the average count in the treated samples

2. from observations, calculate a test statistics (*e.g.*, the mean in the two samples)
3. find the theoretical distribution of the test statistics under H_0
4. deduce the probability that the observations occur under H_0: this is called the p-value



Different steps in hypothesis testing

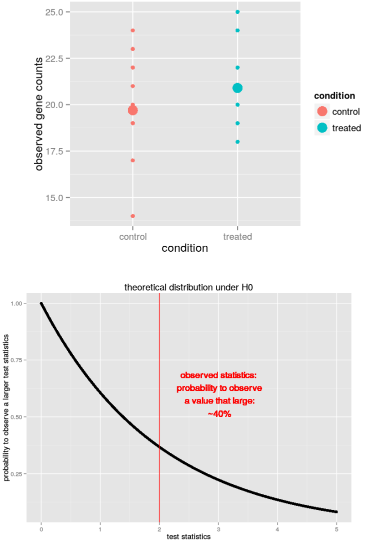
1. formulate an hypothesis H_0:

   H_0: the average count for gene *g* in the control samples is the same that the average count in the treated samples

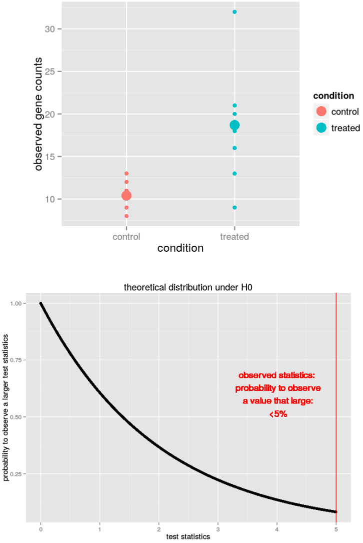
2. from observations, calculate a test statistics (*e.g.*, the mean in the two samples)
3. find the theoretical distribution of the test statistics under H_0
4. deduce the probability that the observations occur under H_0: this is called the p-value
5. conclude: if the p-value is low (usually below $\alpha = 5\%$ as a convention), H_0 is unlikely: we say that "H_0 is rejected".

   We have that: $\alpha = \mathbb{P}_{H_0}(H_0 \text{ is rejected})$.

Summary of the possible decisions



Not reject H_0



Reject H_0

Types of errors in tests

		Reality	
		H_0 is true	H_0 is false
Decision	Do not reject H_0	Correct decision ☺ **(True Negative)**	Type II error ☹ **(False Negative)**
	Reject H_0	Type I error ☹ **(False Positive)**	Correct decision ☺ **(True Positive)**

$$\mathbb{P}(\text{Type I error}) = \alpha \text{ (risk)}$$

$$\mathbb{P}(\text{Type II error}) = 1 - \beta \text{ (}\beta\text{: power)}$$

Why performing a large number of tests might be a problem?

Framework: Suppose you are performing G tests at level α.

$$\mathbb{P}(\text{at least one FP if H}_0\text{ is always true}) = 1 - (1 - \alpha)^G$$

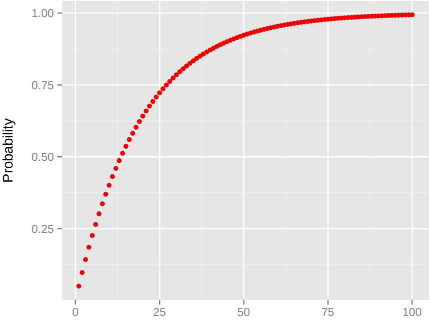
Ex: for $\alpha = 5\%$ and $G = 20$,
$\mathbb{P}(\text{at least one FP if H}_0\text{ is always true}) \simeq 64\%$!!!

Why performing a large number of tests might be a problem?

Framework: Suppose you are performing G tests at level α.

$$\mathbb{P}(\text{at least one FP if H}_0\text{ is always true}) = 1 - (1-\alpha)^G$$

Ex: for $\alpha = 5\%$ and $G = 20$,
$\mathbb{P}(\text{at least one FP if H}_0\text{ is always true}) \simeq 64\%$!!!
Probability to have at least one false positive versus the number of tests performed when H_0 is true for all G tests



For more than 75 tests and if H_0 is always true, the probability to have at least one false positive is very close to 100%!

Notation for multiple tests

Number of decisions for G independent tests:

	True null hypotheses	False null hypotheses	Total
Rejected	U	V	R
Not rejected	$G_0 - U$	$G_1 - V$	$G - R$
Total	G_0	G_1	G

Notation for multiple tests

Number of decisions for G independent tests:

	True null hypotheses	False null hypotheses	Total
Rejected	U	V	R
Not rejected	$G_0 - U$	$G_1 - V$	$G - R$
Total	G_0	G_1	G

Instead of the risk α, control:

- familywise error rate (FWER): FWER $= \mathbb{P}(U > 0)$ (*i.e.*, probability to have at least one false positive decision)
- false discovery rate (FDR): FDR $= \mathbb{E}(Q)$ with

$$Q = \begin{cases} U/R & \text{if } R > 0 \\ 0 & \text{otherwise} \end{cases}$$

Adjusted p-values

Settings: p-values $p_1, \dots, p_G$ (*e.g.*, corresponding to G tests on G different genes)

Adjusted p-values

adjusted p-values are $\tilde{p}_1, \dots, \tilde{p}_G$ such that

$$\text{Rejecting tests such that } \tilde{p}_g < \alpha \quad \Longleftrightarrow \quad \mathbb{P}(U > 0) \leq \alpha \text{ or } \mathbb{E}(Q) \leq \alpha$$

Adjusted p-values

Settings: p-values $p_1, \dots, p_G$ (*e.g.*, corresponding to G tests on G different genes)

Adjusted p-values

adjusted p-values are $\tilde{p}_1, \dots, \tilde{p}_G$ such that

$$\text{Rejecting tests such that } \tilde{p}_g < \alpha \quad \Longleftrightarrow \quad \mathbb{P}(U > 0) \leq \alpha \text{ or } \mathbb{E}(Q) \leq \alpha$$

Calculating p-values

1. order the p-values $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(G)}$

Adjusted p-values

Settings: p-values $p_1, \dots, p_G$ (*e.g.*, corresponding to G tests on G different genes)

Adjusted p-values

adjusted p-values are $\tilde{p}_1, \dots, \tilde{p}_G$ such that

$$\text{Rejecting tests such that } \tilde{p}_g < \alpha \quad \Longleftrightarrow \quad \mathbb{P}(U > 0) \leq \alpha \ \text{ or } \ \mathbb{E}(Q) \leq \alpha$$

Calculating p-values

1. order the p-values $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(G)}$
2. calculate $\tilde{p}_{(g)} = a_g p_{(g)}$
   - with Bonferroni method: $a_g = G$ (FWER)
   - with Benjamini & Hochberg method: $a_g = G/g$ (FDR)

Adjusted p-values

Settings: p-values $p_1, \ldots, p_G$ (*e.g.*, corresponding to G tests on G different genes)

Adjusted p-values

adjusted p-values are $\tilde{p}_1, \ldots, \tilde{p}_G$ such that

$$\text{Rejecting tests such that } \tilde{p}_g < \alpha \quad \Longleftrightarrow \quad \mathbb{P}(U > 0) \leq \alpha \text{ or } \mathbb{E}(Q) \leq \alpha$$

Calculating p-values

1. order the p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(G)}$
2. calculate $\tilde{p}_{(g)} = a_g p_{(g)}$
   - with Bonferroni method: $a_g = G$ (FWER)
   - with Benjamini & Hochberg method: $a_g = G/g$ (FDR)
3. if adjusted p-values $\tilde{p}_{(g)}$ are larger than 1, correct $\tilde{p}_{(g)} \leftarrow \min\{\tilde{p}_{(g)}, 1\}$

Fisher's exact test for contingency tables

After normalization, one may build a contingency table like this one:

	treated	control	Total
gene g	n_{gA}	n_{gB}	n_g
other genes	$N_A - n_{gA}$	$N_B - n_{gB}$	$N - n_g$
Total	N_A	N_B	N

Question: is the number of reads of gene g in the treated sample significatively different than in the control sample?

Fisher's exact test for contingency tables

After normalization, one may build a contingency table like this one:

	treated	control	Total
gene g	n_{gA}	n_{gB}	n_g
other genes	$N_A - n_{gA}$	$N_B - n_{gB}$	$N - n_g$
Total	N_A	N_B	N

Question: is the number of reads of gene g in the treated sample significatively different than in the control sample?

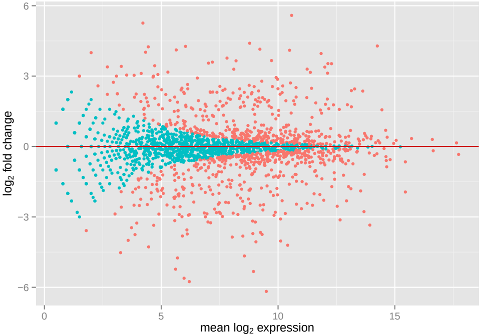
Method

Direct calculation of the probability to obtain such a contingency table (or a "more extreme" contingency table) with:

- independency between the two columns of the contingency tables;
- the same marginals ("Total").

Example of results obtained with the Fisher test

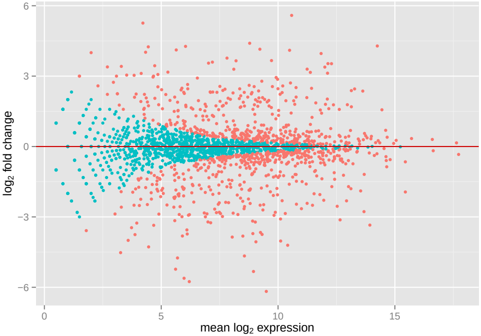
Genes declared significantly differentially expressed are in pink:



Main remark: more conservative for genes with a low expression

Example of results obtained with the Fisher test

Genes declared significantly differentially expressed are in pink:



Main remark: more conservative for genes with a low expression

Limitation of Fisher test

Highly expressed genes have a very large variance! As Fisher test does not estimate variance, it tends to detect false positives among highly expressed genes ⇒ do not use it!

Basic principles of tests for count data: 2 conditions and replicates

Notations: for gene g, K^1_{g1}, ..., $K^1_{gn_1}$ (condition 1) and K^2_{g1}, ..., $K^2_{gn_2}$ (condition 2)

- choose an appropriate distribution to model count data (discrete data, overdispersion)
- estimate its parameters for both conditions
- conclude by calculating p-value

Basic principles of tests for count data: 2 conditions and replicates

Notations: for gene g, K^1_{g1}, ..., $K^1_{gn_1}$ (condition 1) and K^2_{g1}, ..., $K^2_{gn_2}$ (condition 2)

- choose an appropriate distribution to model count data (discrete data, overdispersion)

  $$K^k_{gj} \sim \text{NB}(s^k_j \lambda_{gk}, \phi_g)$$

  in which:

  - s^k_j is library size of sample j in condition k
  - λ_{gk} is the proportion of counts for gene g in condition k
  - ϕ_g is the dispersion of gene g (supposed to be identical for all samples)

- estimate its parameters for both conditions

- conclude by calculating p-value

Basic principles of tests for count data: 2 conditions and replicates

Notations: for gene g, K^1_{g1}, ..., $K^1_{gn_1}$ (condition 1) and K^2_{g1}, ..., $K^2_{gn_2}$ (condition 2)

- choose an appropriate distribution to model count data (discrete data, overdispersion)

  $$K^k_{gj} \sim \text{NB}(s^k_j \lambda_{gk}, \phi_g)$$

  in which:

  - s^k_j is library size of sample j in condition k
  - λ_{gk} is the proportion of counts for gene g in condition k
  - ϕ_g is the dispersion of gene g (supposed to be identical for all samples)

- estimate its parameters for both conditions

  λ_{g1} λ_{g2} ϕ_g

- conclude by calculating p-value

Basic principles of tests for count data: 2 conditions and replicates

Notations: for gene g, K_{g1}^1, ..., $K_{gn_1}^1$ (condition 1) and K_{g1}^2, ..., $K_{gn_2}^2$ (condition 2)

- choose an appropriate distribution to model count data (discrete data, overdispersion)

  $$K_{gj}^k \sim \mathrm{NB}(s_j^k \lambda_{gk}, \phi_g)$$

  in which:

  - s_j^k is library size of sample j in condition k
  - λ_{gk} is the proportion of counts for gene g in condition k
  - ϕ_g is the dispersion of gene g (supposed to be identical for all samples)

- estimate its parameters for both conditions

  λ_{g1} $\qquad$ λ_{g2} $\qquad$ ϕ_g

- conclude by calculating p-value ⇒ Test

  $$\mathrm{H0} : \{\lambda_{g1} = \lambda_{g2}\}$$

First method: Exact Negative Binomial test

[Robinson and Smyth, 2008]

Normalization is performed to get equal size librairies $\Rightarrow s$

$K^1_{g1} + \ldots + K^1_{gn_1} \sim \mathrm{NB}(s\lambda_{g1}, \phi_g/n_1)$ (and similarly for the second condition)

First method: Exact Negative Binomial test

[Robinson and Smyth, 2008]

Normalization is performed to get equal size librairies ⇒ s

$K_{g1}^1 + \ldots + K_{gn_1}^1 \sim \text{NB}(s\lambda_{g1}, \phi_g/n_1)$ (and similarly for the second condition)

1. λ_{g1} and λ_{g2} are estimated (mean of the distributions)

First method: Exact Negative Binomial test

[Robinson and Smyth, 2008]

Normalization is performed to get equal size librairies ⇒ s

$K_{g1}^1 + \ldots + K_{gn_1}^1 \sim \mathrm{NB}(s\lambda_{g1}, \phi_g/n_1)$ (and similarly for the second condition)

1. λ_{g1} and λ_{g2} are estimated (mean of the distributions)
2. ϕ_g is estimated independently of λ_{g1} and λ_{g2}, using different approaches to account for small sample size

First method: Exact Negative Binomial test

[Robinson and Smyth, 2008]

Normalization is performed to get equal size librairies ⇒ s

$K^1_{g1} + \ldots + K^1_{gn_1} \sim \text{NB}(s\lambda_{g1}, \phi_g / n_1)$ (and similarly for the second condition)

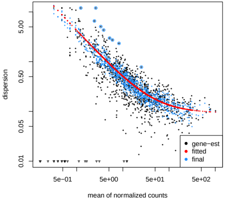
1. λ_{g1} and λ_{g2} are estimated (mean of the distributions)
2. ϕ_g is estimated independently of λ_{g1} and λ_{g2}, using different approaches to account for small sample size
3. The test is performed similarly as for Fisher test (exact probability calculation according to estimated paramters)

Estimating the dispersion parameter ϕ_g

Some methods:

- **DESeq**, **DESeq2**: ϕ_g is a smooth function of $\lambda_g = \lambda_{g1} = \lambda_{g2}$

```
dge <- estimateDispersion(dge)
```



- **edgeR**: estimate a common dispersion parameter for all genes and use it as a prior in a Bayesian approach to estimate a gene specific dispersion parameter

```
dge <- estimateCommonDisp(dge)
dge <- estimateTagwiseDisp(dge)
```

Perform the test

Some methods:

- **DESeq**, **DESeq2**: exact (**DESeq**) or approximate (Wald and LR in **DESeq2**) tests

```
res <- nbinomWaldTest(dge)
results(res)
```

```
res <- nbinomLR(dge)
results(res)
```

- **edgeR**: exact tests

```
res <- exactTest(dge)
topTags(res)
```

(comparison between methods in [Zhang et al., 2014])

More complex experiments: GLM

Framework:

$$K_{gj} \sim \text{NB}(\mu_{gj}, \phi_g) \qquad \text{with} \qquad \log(\mu_{gj}) = \log(s_j) + \log(\lambda_{gj})$$

in which:

- s_j is the library size for sample j;

More complex experiments: GLM

Framework:

$$K_{gj} \sim \text{NB}(\mu_{gj}, \phi_g) \qquad \text{with} \qquad \log(\mu_{gj}) = \log(s_j) + \log(\lambda_{gj})$$

in which:

- s_j is the library size for sample j;
- $\log(\lambda_{gj})$ is estimated (for instance) by a Generalized Linear Model (GLM):

  $$\log(\lambda_{gj}) = \lambda_0 + \mathbf{x}_j^\top \beta_g$$

  in which $\mathbf{x}_i$ is a vector of covariates.

More complex experiments: GLM

Framework:

$$K_{gj} \sim \mathrm{NB}(\mu_{gj}, \phi_g) \qquad \text{with} \qquad \log(\mu_{gj}) = \log(s_j) + \log(\lambda_{gj})$$

in which:

- s_j is the library size for sample j;
- $\log(\lambda_{gj})$ is estimated (for instance) by a Generalized Linear Model (GLM):

  $$\log(\lambda_{gj}) = \lambda_0 + \mathbf{x}_j^\top \beta_g$$

  in which $\mathbf{x}_i$ is a vector of covariates.

GLM allows to decompose the effects on the mean of

- different factors
- their interactions

More complex experiments: GLM in practice

edgeR

```
dge <- estimateDisp(dge, design)
fit <- glmFit(dge, design)
res <- glmRT(fit, ...)
topTags(res)
```

DESeq, DESeq2

```
dge <- newCountDataSet(counts, design)
dge <- estimateSizeFactors(dge)
dge <- estimateDispersions(dge)
fit <- fitNbinomGLMs(dge, count ~ ...)
fit0 <- fitNbinomGLMs(dge, count ~ 1)
res <- nbinomGLMTest(fit, fit0)
p.adjust(res, method = "BH")
```

Alternative approach: linear model for count data

[Law et al., 2014], **limma**

Basic idea:

1. data are transformed so that they are approximately normally distributed

   ```r
   tcount <- voom(counts, design)
   ```

2. a linear (Gaussian) model is fitted (with a Bayesian approach to improve FDR [McCarthy and Smyth, 2009]):

   $$\widetilde{K}_{gj} \sim \mathcal{N}(\mu_{gj}, \sigma_g^2)$$

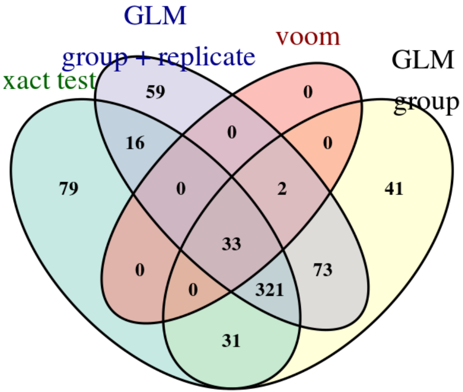
   with

   $$\mathbb{E}(\widetilde{K}_{gj}) = \beta_0 + \mathbf{x}_j^\top \beta_g$$

   ```r
   fit <- lmFit(tcount, design)
   fit <- eBayes(fit)
   topTables(fit, ...)
   ```

Practical session

- use the same data as before;
- run the analysis with different approaches (using exact test or GLM or voom + LM);
- compare the results...



References

Note: Some images in this slide have been used as a courtesy of **Ignacio Gonzàles.**

Anders, S. and Huber, W. (2010).
Differential expression analysis for sequence count data.
Genome Biology, 11:R106.

Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010).
Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments.
BMC Bioinformatics, 11(1):94.

Dillies, M., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., and Jaffrézic, F. (2013).
A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis.
Briefings in Bioinformatics, 14(6):671–683.
on behalf of The French StatOmique Consortium.

Law, C., Chen, Y., Shi, W., and Smyth, G. (2014).
Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.
Genome Biology, 15(R29).

McCarthy, D. and Smyth, G. (2009).
Testing significance relative to a fold-change threshold is a TREAT.
Bioinformatics, 25:765–771.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. (2008).
Mapping and quantifying mammalian transcriptomes by RNA-Seq.
Nature Methods, 5:621–628.

Oshlack, A. and Wakefield, M. (2009).
Transcript length bias in RNA-seq data confounds systems biology.
Biology Direct, 4(14).

Robinson, M. and Oshlack, A. (2010).
A scaling normalization method for differential expression analysis of RNA-seq data.
Genome Biology, 11:R25.

Robinson, M. and Smyth, G. (2008).
Small-sample estimation of negative binomial dispersion, with applications to SAGE data.
Bioinformatics, 9(2):321–332.

Zhang, Z., Jhaveri, D., Marshall, V., Bauer, D., Edson, J., Narayanan, R., Robinson, G., Lundberg, A., Bartlett, P., Wray, N., and Zhao, Q. (2014).
A comparative study of techniques for differential expression analysis on RNA-seq data.
PLoS ONE, 9(8):e103207.