# Differential Expression in RNA-Seq

Sonia Tarazona
*starazona@cipf.es*

Genomics of Gene Expression Lab, Bioinformatics Department
**Centro de Investigación Príncipe Felipe, Valencia, Spain**

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Valencia, March 2011

# Outline

# Outline

# Some questions

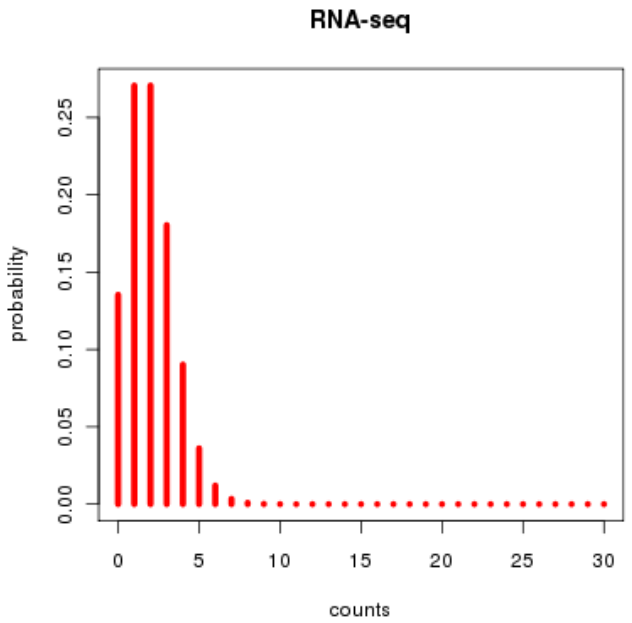
- How do we measure expression?
- What is differential expression?
- Experimental design in RNA-Seq
- Can we used the same statistics as in microarrays?
- Do I need any "normalization"?

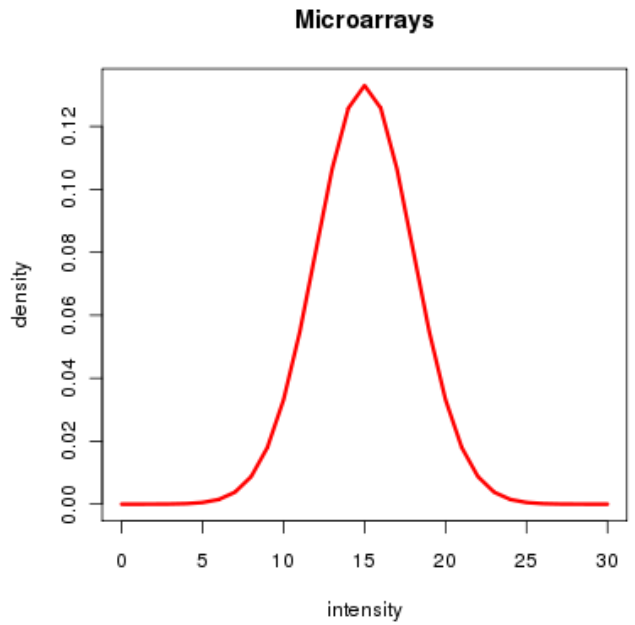
# Some definitions

## Expression level

**RNA-Seq:** The number of reads (**counts**) mapping to the biological feature of interest (gene, transcript, exon, etc.) is considered to be linearly related to the abundance of the target feature.

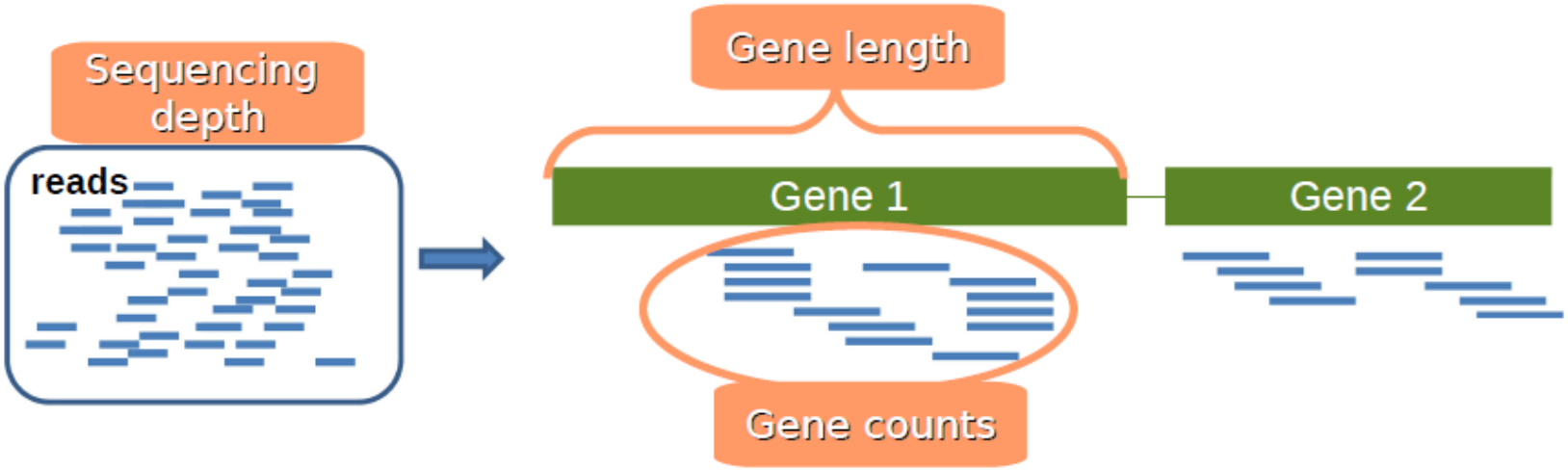
**Microarrays:** The abundance of each sequence is a function of the fluorescence level recovered after the hybridization process.



# Some definitions

- **Sequencing depth:** Total number of reads mapped to the genome. *Library size.*
- **Gene length:** Number of bases.
- **Gene counts:** Number of reads mapping to that gene (expression measurement).



# Some definitions

## What is differential expression?

- A gene is declared differentially expressed if an observed difference or change in read counts between two experimental conditions is statistically significant, i.e. whether it is greater than what would be expected just due to natural random variation.

- Statistical tools are needed to make such a decision by studying counts probability distributions.

# RNA-Seq expression data

## Experimental design

- **Pairwise comparisons:** Only two experimental conditions or groups are to be compared.
- **Multiple comparisons:** More than two conditions or groups.

## Replication

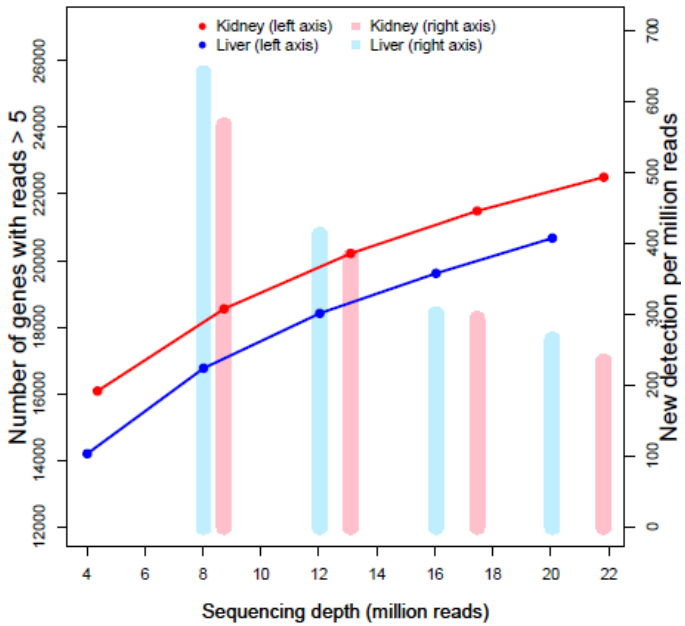
- **Biological replicates.** To draw general conclusions: from samples to population.
- **Technical replicates.** Conclusions are only valid for compared samples.

# Outline

# Why Normalization?

## RNA-seq biases

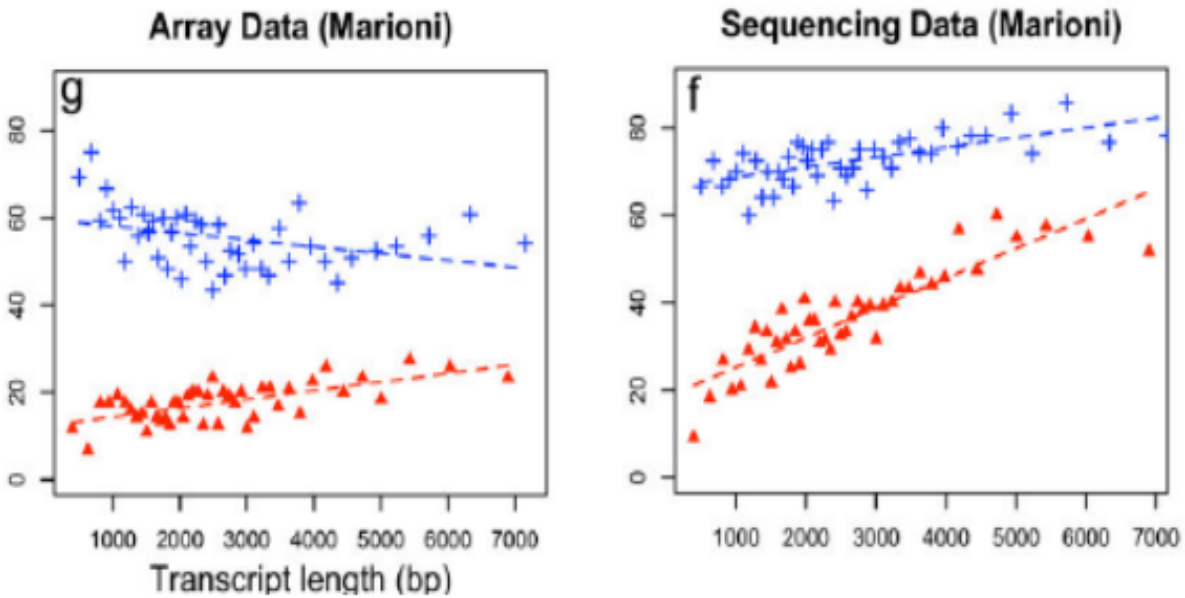
- Influence of sequencing depth: The higher sequencing depth, the higher counts.



# Why Normalization?

## RNA-seq biases

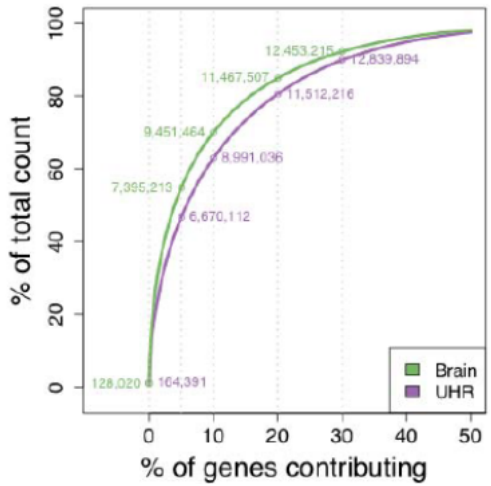
- Dependence on gene length: Counts are proportional to the transcript length times the mRNA expression level.



# Why Normalization?

## RNA-seq biases

- Differences on the counts distribution among samples.



# Why Normalization?

## RNA-seq biases

- Influence of sequencing depth: The higher sequencing depth, the higher counts.
- Dependence on gene length: Counts are proportional to the transcript length times the mRNA expression level.
- Differences on the counts distribution among samples.

## Options

1. Normalization: Counts should be previously corrected in order to minimize these biases.
2. Statistical model should take them into account.

# Some Normalization Methods

- RPKM (Mortazavi et al., 2008): Counts are divided by the transcript length (kb) times the total number of millions of mapped reads.

$$RPKM = \frac{\text{number of reads of the region}}{\frac{\text{total reads}}{1000000} \times \frac{\text{region length}}{1000}}$$

- Upper-quartile (Bullard et al., 2010): Counts are divided by upper-quartile of counts for transcripts with at least one read.
- TMM (Robinson and Oshlack, 2010): Trimmed Mean of M values.
- Quantiles, as in microarray normalization (Irizarry et al., 2003).
- FPKM (Trapnell et al., 2010): Instead of counts, Cufflinks software generates FPKM values (Fragments Per Kilobase of exon per Million fragments mapped) to estimate gene expression, which are analogous to RPKM.

# Example on RPKM normalization

| EnsemblGeneID | kidney | liver |
|---|---|---|
| ENSG00000146556 | 0 | 0 |
| ENSG00000197194 | 0 | 0 |
| ENSG00000197490 | 0 | 0 |
| ENSG00000205292 | 0 | 0 |
| ENSG00000177693 | 0 | 0 |
| ENSG00000209338 | 0 | 0 |
| ENSG00000196573 | 0 | 0 |
| ENSG00000177799 | 0 | 0 |
| ENSG00000209341 | 0 | 0 |
| ENSG00000209342 | 9 | 1 |
| ENSG00000209343 | 0 | 0 |
| ENSG00000209344 | 0 | 0 |
| ENSG00000209346 | 0 | 0 |
| ENSG00000209349 | 0 | 0 |
| ENSG00000209350 | 27 | 161 |
| ENSG00000209351 | 0 | 0 |
| ENSG00000209352 | 2 | 2 |
| ENSG00000212679 | 620 | 746 |
| ENSG00000212678 | 64591 | 44870 |
| ENSG00000185097 | 0 | 0 |
| ENSG00000209353 | 0 | 0 |

## Sequencing depth

Marioni *et al.*, 2008

- Kidney: 9.293.530 reads
- Liver: 8.361.601 reads

## RPKM normalization

- Length: 1500 bases.
- Kidney: RPKM $= \frac{620}{\frac{9293530}{10^6} \times \frac{1500}{1000}} = 44{,}48$
- Liver: RPKM $= \frac{746}{\frac{8361601}{10^6} \times \frac{1500}{1000}} = 59{,}48$

# Example

| EnsemblGeneID | Length | kidney | liver | RPKM kidney | RPKM liver | UQUA kidney | UQUA liver |
|---|---|---|---|---|---|---|---|
| ENSG00000187642 | 3035 | 39 | 7 | 1.38 | 0.28 | 382.37 | 74.66 |
| ENSG00000188290 | 877 | 54 | 9 | 6.63 | 1.23 | 981.59 | 182.39 |
| ENSG00000187608 | 634 | 59 | 63 | 10.01 | 11.88 | 1252.58 | 1484.31 |
| ENSG00000188157 | 7353.5 | 2108 | 259 | 30.85 | 4.21 | 13193.25 | 1796.14 |
| ENSG00000131591 | 2039.83 | 54 | 34 | 2.85 | 1.99 | 637.77 | 445.46 |
| ENSG00000215916 | 2008 | 57 | 34 | 3.05 | 2.03 | 678.67 | 448.97 |
| ENSG00000207730 | 95 | 0 | 0 | 0 | 0 | 0 | 0 |
| ENSG00000207607 | 90 | 0 | 0 | 0 | 0 | 0 | 0 |
| ENSG00000198976 | 83 | 2 | 0 | 2.59 | 0 | 113.62 | 0 |
| ENSG00000205231 | 3532 | 4 | 0 | 0.12 | 0 | 35.71 | 0 |
| ENSG00000162571 | 2060.25 | 4 | 0 | 0.21 | 0 | 47.82 | 0 |
| ENSG00000186891 | 964 | 0 | 3 | 0 | 0.37 | 0 | 57.78 |
| ENSG00000186827 | 987 | 5 | 1 | 0.55 | 0.12 | 86.6 | 19.42 |
| ENSG00000078808 | 1870.67 | 1136 | 883 | 65.34 | 56.45 | 14095 | 12172.11 |
| ENSG00000176022 | 2793 | 143 | 165 | 5.51 | 7.07 | 1453.42 | 1853.37 |
| ENSG00000184163 | 1036 | 16 | 14 | 1.66 | 1.62 | 268.45 | 258.3 |
| ENSG00000160087 | 1614.14 | 315 | 290 | 21 | 21.49 | 4220.61 | 4320 |
| ENSG00000162572 | 2635.86 | 47 | 28 | 1.92 | 1.27 | 489.42 | 323.8 |
| ENSG00000131584 | 3861.75 | 379 | 216 | 10.56 | 6.69 | 3281.75 | 2076.27 |
| ENSG00000169972 | 1239 | 105 | 143 | 9.12 | 13.8 | 1611.32 | 2423.83 |
| ENSG00000127054 | 1813.5 | 496 | 330 | 29.43 | 21.76 | 6250.88 | 4618.39 |
| ENSG00000187488 | 2193 | 17 | 6 | 0.83 | 0.33 | 194.62 | 76.17 |
| ENSG00000215792 | 2193 | 17 | 6 | 0.83 | 0.33 | 194.62 | 76.17 |
| ENSG00000169962 | 3402 | 2 | 0 | 0.06 | 0 | 18.2 | 0 |
| ENSG00000107404 | 2554 | 51 | 16 | 2.15 | 0.75 | 542.14 | 188.61 |
| ENSG00000162576 | 2169.67 | 56 | 11 | 2.78 | 0.61 | 645.79 | 138.29 |
| ENSG00000175756 | 874.25 | 0 | 0 | 0 | 0 | 0 | 0 |
| ENSG00000131586 | 676 | 9 | 0 | 1.43 | 0 | 187.59 | 0 |
| ENSG00000205116 | 492 | 7 | 0 | 1.53 | 0 | 175.4 | 0 |
| ENSG00000179403 | 2442.5 | 53 | 46 | 2.33 | 2.25 | 578.29 | 553.48 |
| ENSG00000215915 | 2800.5 | 0 | 4 | 0 | 0.17 | 0 | 46.38 |

# Outline

# Differential Expression

## Parametric approaches

Counts are modeled using known probability distributions such as Binomial, Poisson, Negative Binomial, etc.

R packages in Bioconductor:

- **edgeR** (Robinson et al., 2010): Exact test based on Negative Binomial distribution.
- **DESeq** (Anders and Huber, 2010): Exact test based on Negative Binomial distribution.
- **DEGseq** (Wang et al., 2010): MA-plots based methods (MATR and MARS), assuming Normal distribution for $M|A$.
- **baySeq** (Hardcastle et al., 2010): Estimation of the posterior likelihood of differential expression (or more complex hypotheses) via empirical Bayesian methods using Poisson or NB distributions.

# Differential Expression

## Non-parametric approaches

No assumptions about data distribution are made.

- **Fisher's exact test** (better with normalized counts).

| | KIDNEY | LIVER | Total |
|---|---|---|---|
| **ENSG00000188157** | 31 | 4 | 35 |
| **Remaining genes** | 809347 | 799468 | 1608815 |
| Total | 809378 | 799472 | |

- **cuffdiff** (Trapnell et al., 2010): Based on entropy divergence for relative transcript abundances. Divergence is a measurement of the "distance" between the relative abundances of transcripts in two difference conditions.
- NOISeq (Tarazona et al., coming soon)

# Differential Expression

## Drawbacks of differential expression methods

- Parametric assumptions: Are they fulfilled?
- Need of replicates.
- Problems to detect differential expression in genes with low counts.

# Differential Expression

## Drawbacks of differential expression methods

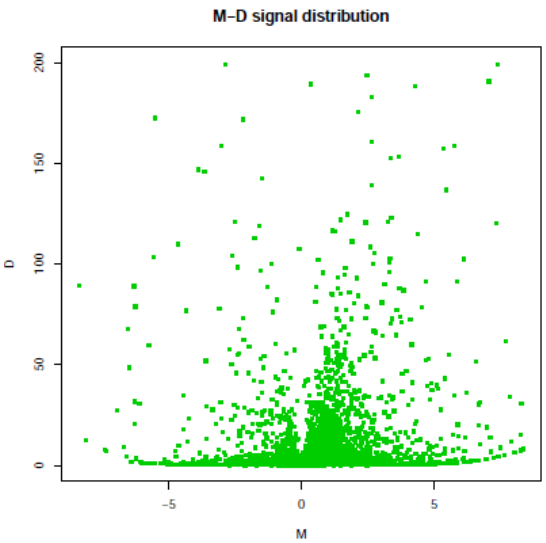
- Parametric assumptions: Are they fulfilled?
- Need of replicates.
- Problems to detect differential expression in genes with low counts.

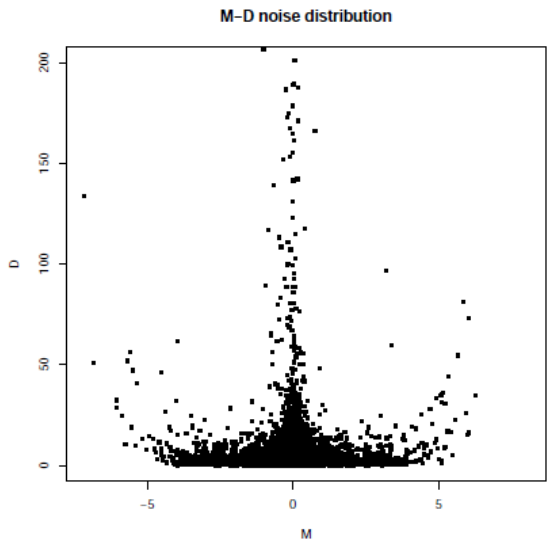
## NOISeq

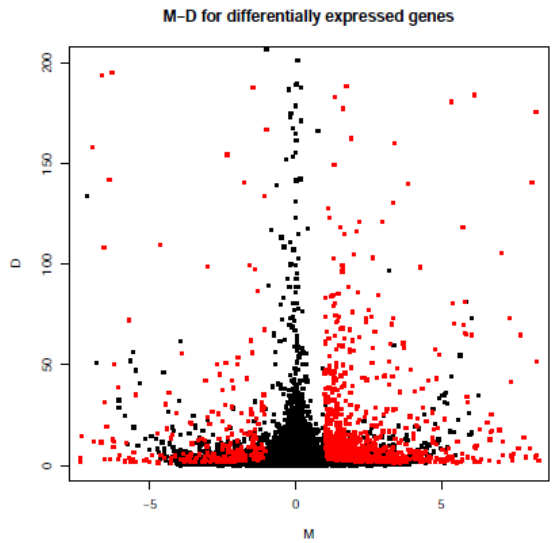
- Non-parametric method.
- No need of replicates.
- Less influenced by sequencing depth or number of counts.

# NOISeq

1. Compute for each gene:
   $M = \log_2 \frac{expression_1}{expression_2}$
   $D = |expression_1 - expression_2|$
2. Compute M-D in noise by comparing each pair of replicates within the same condition.
3. Probability for a gene of being differentially expressed: Obtained by comparing M-D values of that gene against noise distribution.
4. A gene is declared as differentially expressed if this probability is higher than $q$.



# NOISeq

## NOISeq-real

- Replicates are available for each condition.
- Compute M-D in noise by comparing each pair of replicates within the same condition.

# NOISeq

## NOISeq-real

- Replicates are available for each condition.
- Compute M-D in noise by comparing each pair of replicates within the same condition.

## NOISeq-sim

- No replicates are available at all.
- NOISeq simulates technical replicates for each condition. The replicates are generated from a multinomial distribution taking the counts in the only sample as the probabilities for the distribution.
- Compute M-D in noise by comparing each pair of simulated replicates within the same condition.

# NOISeq

## Input

- Data: *datos*1, *datos*2
- Features length: *long* (only if length correction is to be applied)
- Normalization: *norm* = {"rpkm", "uqua", "tmm", "none"}; *lc* = "length correction"
- Simulation: *nss* = "number of replicates to be simulated"; *pnr* = "total counts in each simulated replicate"; *v* = variability for *pnr*
- Probability cutoff: $q$ ($\geq 0{,}8$)
- Others: $k = 0{,}5$

# NOISeq

## Output

- Differential expression probability. For each feature, probability of being differentially expressed.
- Differentially expressed features. List of features names which are differentially expressed according to $q$ cutoff.
- M-D values. For signal (between conditions and for each feature) and for noise (among replicates within the same condition, pooled).

# Exercises

Execute in an R terminal the code provided in *exerciseDEngs.r*

## Reading data

```
simCount <- read.delim("simCount.txt", row.names = 1, header = TRUE)
head(simCount)
depth <- colSums(simCount)
```

## Differential Expression by NOISeq

NOISeq-real

```
res1noiseq <- noiseq(simCount[,1:5], simCount[,6:10], nss = 0, q = 0.8)
```

NOISeq-sim

```
res2noiseq <- noiseq(simCount[,1:5], simCount[,6:10], q=0.8, pnr=0.2,
nss=5)
```

# Some remarks

- Experimental design is decisive to answer correctly your biological questions.
- Differential expression methods for RNA-Seq data must be different to microarray methods.
- Normalization should be applied to raw counts, at least a library size correction.

# For Further Reading


Oshlack, A., Robinson, M., and Young, M. (2010) From RNA-seq reads to differential expression results. *Genome Biology*, **11**, 220+.
*Review on RNA-Seq, including differential expression*.


Auer, P.L., and Doerge R.W. (2010) Statistical Design and Analysis of RNA Sequencing Data. *Genetics*, **185**, 405-416.


Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94+.
*Normalization methods (including Upper Quartile) and differential expression*.


Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer A., and Conesa, A. (in preparation) Differential expression in RNA-seq: a matter of depth.