Module 16: Analyse statistique de données RNA-Seq

Recherche des régions d'intérêt différentiellement exprimées

(R, RStudio)

J. Aubert et C. Hennequet-Antier

Plateforme Migale

30-31 mai 2022

`https://migale.inrae.fr/trainings`

Outline

Presentation

- My name is...
- I’m working...
- My skills are...
- My interests are...
- I hope to be able to...

Objectifs

- Connaître le vocabulaire et les concepts statistiques utiles pour analyser des données type RNA-Seq
- Savoir effectuer une analyse différentielle dans quelques cas standards à l'aide de logiciel R
- Comprendre le matériel et méthode d'un article du domaine
- Evaluer la pertinence d'une analyse RNA-Seq en identifiant les éléments clefs et comprendre les particularités liées à la nature des données

Modalités pratiques

Horaires

- lundi : 9h30-17h
- mardi : 9h30-17h

RStudio

- RStudio de migale `https://rstudio.migale.inrae.fr`
  un login/mdp par apprenant
- Vous garderez le même compte (et poste) pour les deux jours de la formation.
- N'hésitez pas à demander des pauses en cas de besoin.

Programme : alternance Cours / TP

- Explorer les données
- Normaliser les données de comptage
- Identifier les transcrits différentiellement exprimés
- Se sensibiliser aux tests multiples
- S'initier à la visualisation et l'analyse de voies métaboliques

Le tout avec R dans l'environnement RStudio.

R and RStudio environment

R Packages

- DESeq2
- edgeR

also available in Galaxy environment.

DESeq2 package

Reference

citation("DESeq2")

Michael I Love, Wolfgang Huber and Simon Anders (2014): **Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2**, *Genome Biology*

Tutorial

Love, MI and Anders, S and Kim, V and Huber, W (2016), **RNA-Seq workflow: gene-level exploratory analysis and differential expression**, http://openr.es/7pm, *10.12688/f1000research.7035.2*

rnaseqGene package

DESeq2 package

Installation of the package

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager") BiocManager::install("DESeq2")
```

To use an installed package, we have to load it into the current session

```
library(DESeq2)
```

edgeR package

Reference

citation("edgeR")

Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, 26(1), 139-140.

McCarthy, J. D, Chen, Yunshun, Smyth, K. G (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” Nucleic Acids Research, 40(10), 4288-4297.

Tutorial

Chen, Y, Lun, ATL and Smyth, GK. (2016) **From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline** *[version 2; peer review: 5 approved]. F1000Research*, `https://doi.org/10.12688/f1000research.8987.2`.

edgeR package

Installation of the package

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager") BiocManager::install("edgeR")
```

To use an installed package, we have to load it into the current session

```
library(edgeR)
```

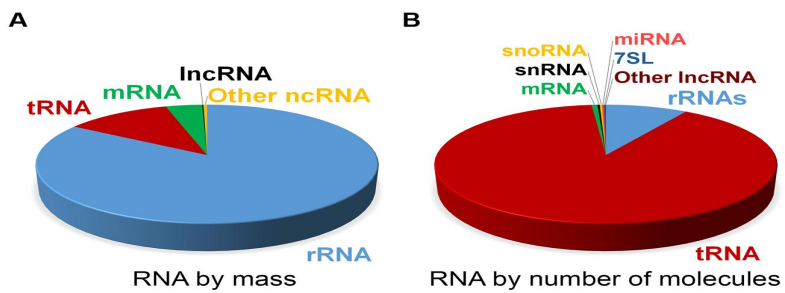
Transcriptomics

Transcriptome: Complete set of transcripts and their level of expression, for a defined population of cells. Unlike the genome, the transcriptome is dynamic and can be modulated by both internal and external factors. (Velculescu et al, 1997)

The aims of transcriptomics:

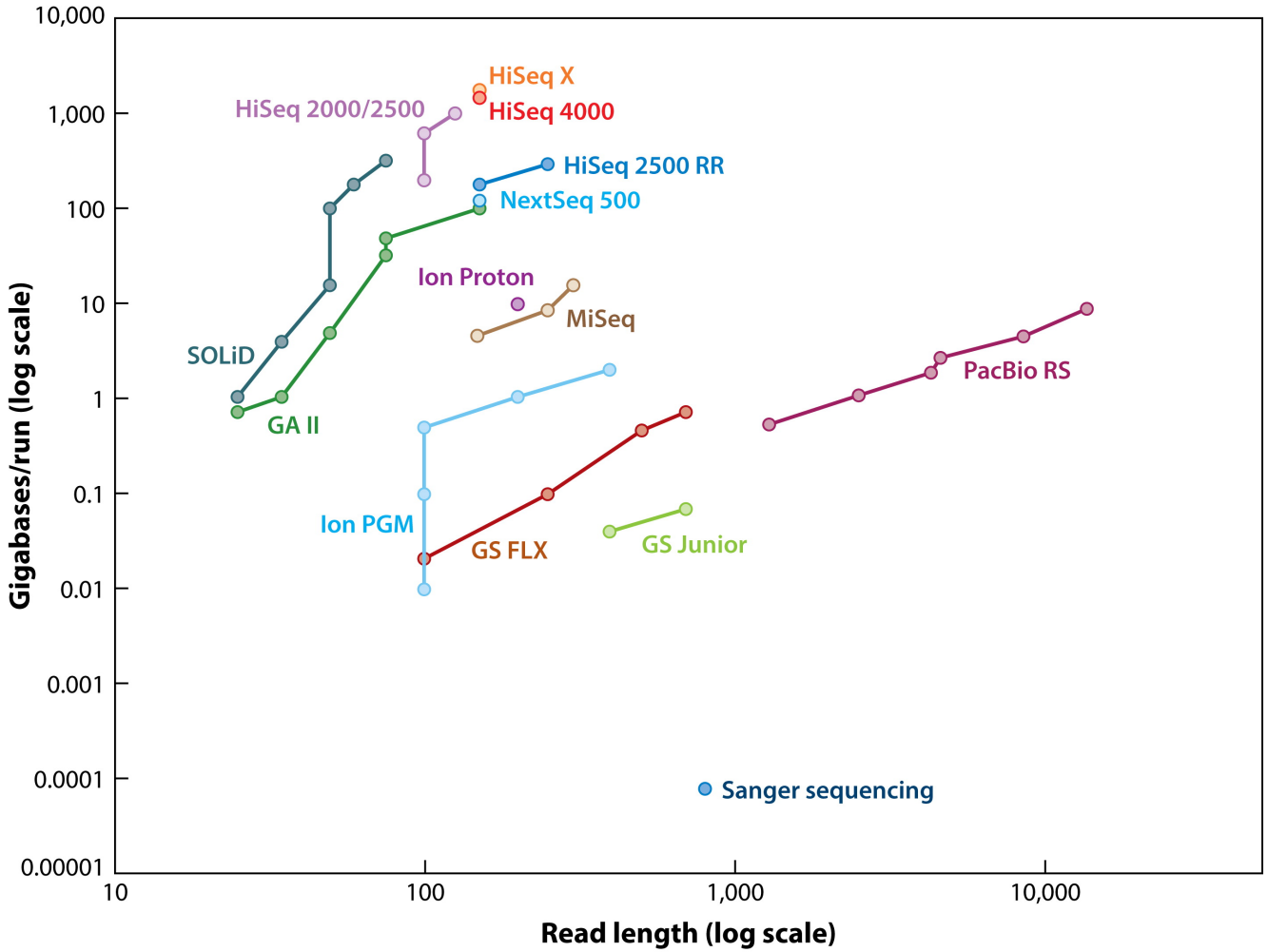
- to quantify the changing expression levels of each transcript under different biological conditions (**differential analysis**);
- to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs;
- to determine the transcriptional structure of genes: splicing patterns, post-transcriptional modifications;
- to discover allele-specific expression.

Estimate of RNA levels in a typical mammalian cell (Palazzo et al., 2015).



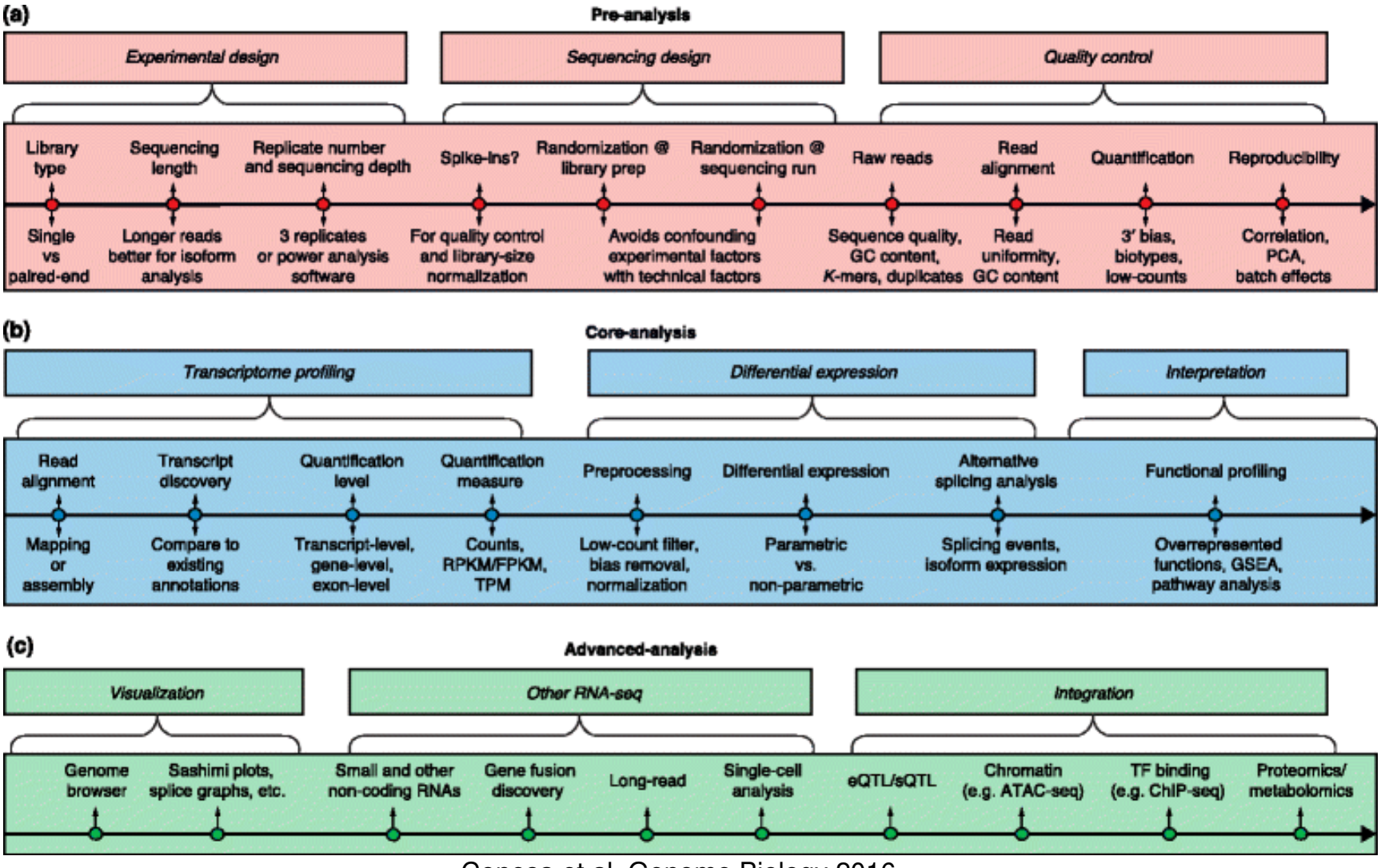
Which high-throughput sequencing technology to choose?

Illustrate the dynamic and changing nature of sequencing based on the number of reads and read length.



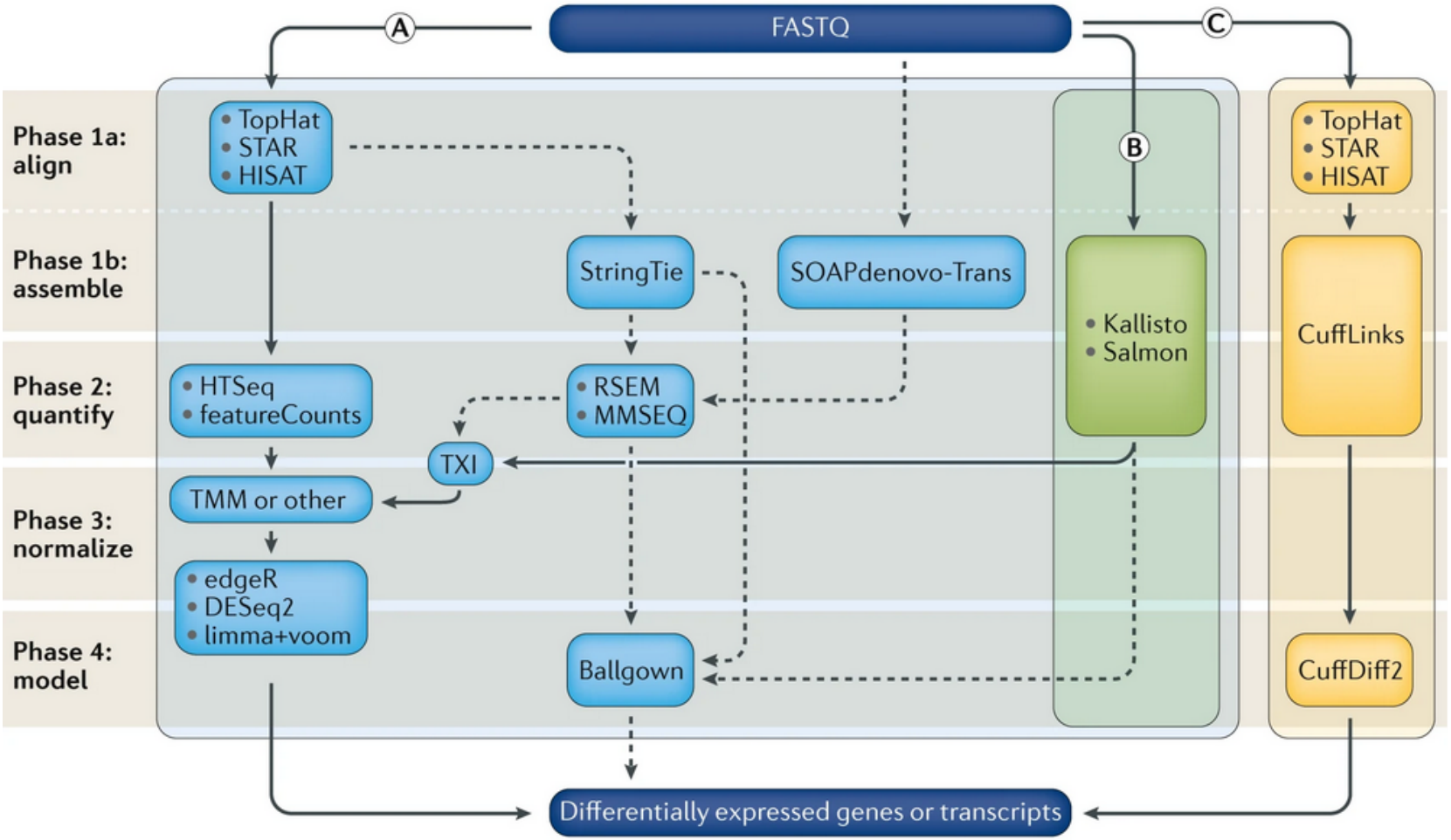
Levy SE, Myers RM. 2016.
Annu. Rev. Genom. Hum. Genet. 17:95–115

A generic roadmap for RNA-Seq data analyses



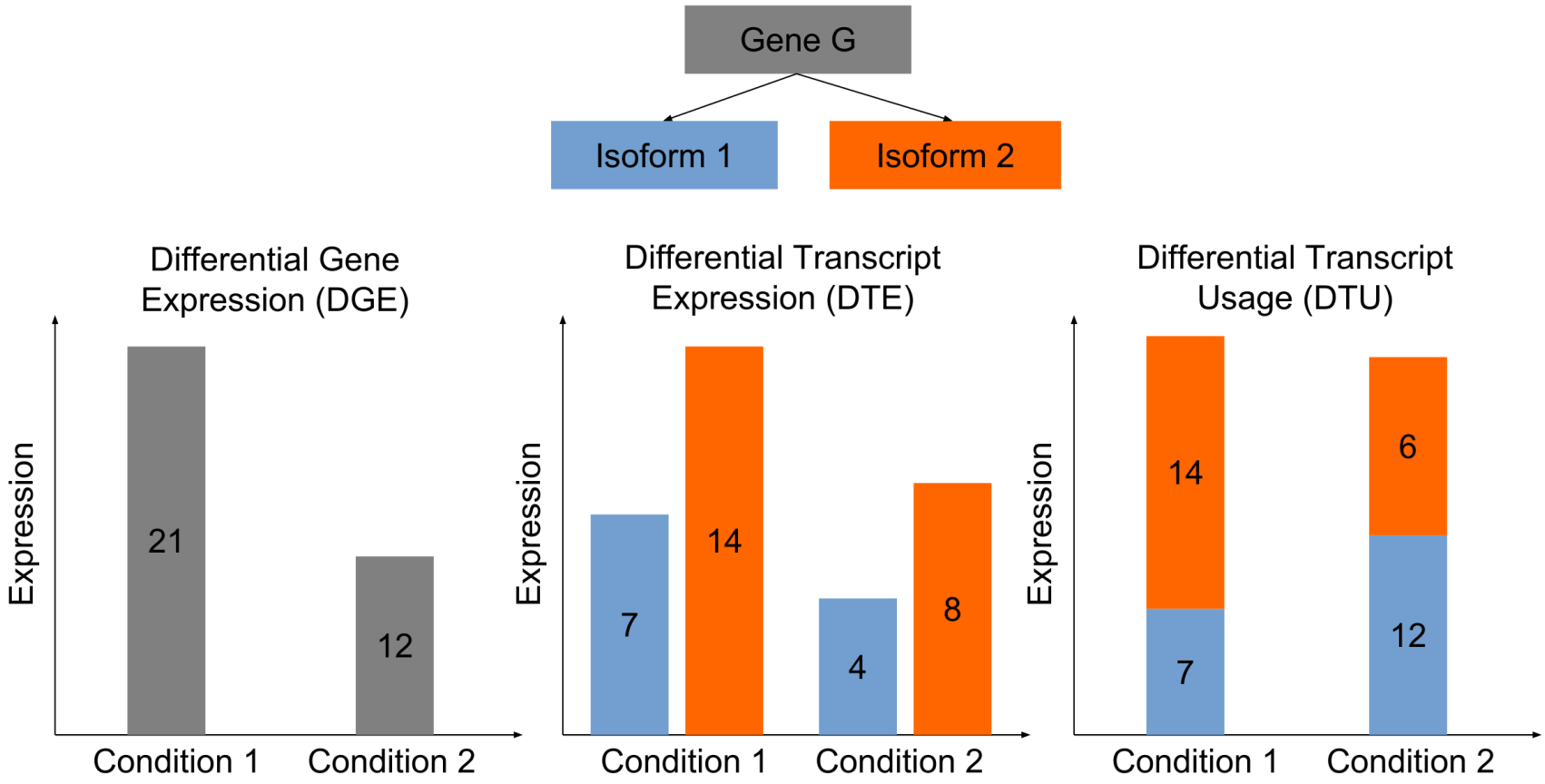
Conesa et al, Genome Biology 2016

RNA-seq data analysis workflow for differential gene expression



Stark et al., Nature Reviews Genetics 2019

Several strategies

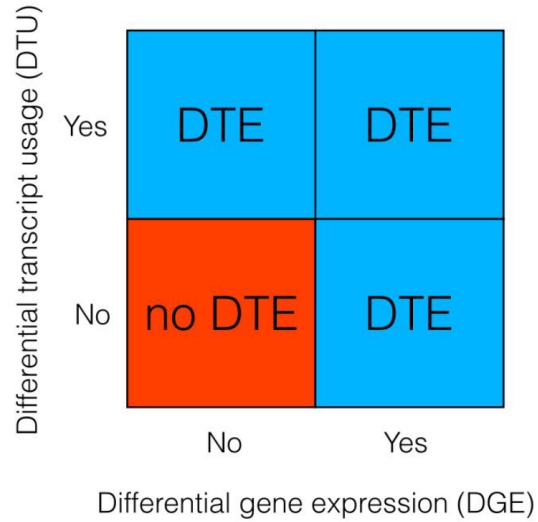


Source: H. Varet, Institut Pasteur

DTE, DGE & DTU

If a transcript is DE (DTE): what about the other transcripts of the gene?

- Because of DGE?
- Because of DTU?
- Because of both DGE and DTU?



Soneson et al. *Differential analyses for RNA-Seq: transcript-level estimates improve gene-level inferences*, F1000Research, 2016

Source: H. Varet, Institut Pasteur

A typical raw dataset

	S_1	S_2	...	S_j	...	S_n
Gene 1	16	9	...	y_{1j}	...	15
Gene 2	4448	3973	...	y_{2j}	...	3964
...	...	...	...	...	...	...
Gene i	y_{i1}	y_{i2}	...	y_{gj}	...	y_{in}
...	...	...	...	...	...	...
Gene G	59	164	...	y_{Gj}	...	143
Seq. depth	6865057	11127087	...	$n_j = \sum_{g=1}^{G} y_{gj}$	...	11320226

y_{gj} = number of sequences from sample j assigned to gene g.

Remark: one row = one region of interest (gene, exon, transcript, $\cdots$).

Example Data from Fu et al. (Nature Cell Biology, 2015)

RNA-Seq profiles of mouse mammary gland

Expression profiles of

- basal stem-cell enriched cells (B) and committed luminal cells (L)
- in the mammary gland of virgin, pregnant and lactating mice.

The dataset consists of a matrix $\mathbf{Y} = [y_{gj}]$ or data frame (gene $\times$ sample) of counts.

- Each row g = one gene
- Each column j = one experimental sample

Your turn ! Exercise 1 - Description of the biological experiment

Part1

- Import the data in R
- Data visualization

Part2

- Create a single grouping factor condition combining CellType and Status
- Check the number of replicates per condition

Your turn ! Exercise 2 - Formatting the count data

Part1

- Import the count data in R
- Create a data.frame that contains only the counts
- Visualize the first and last lines

Part2

- Have a look at the column names
- Rename rownames of your data.frame using EntrezGeneID.

Your turn ! Exercise 3 - Creating a DGEList

Part1

- Convert geneCount to y DGEList with DGEList()
- Examine the elements of y
- Give the number of genes

Part2

- Provide samples and group argument giving the experimental condition for each sample
- Set remove.zeros to TRUE. How many genes with all zero counts are removed?

Challenge

- Add annotation information for each gene into y DGEList object. We will use org.Mm.eg.db package which gives genome wide annotation for Mouse based on mapping using Entrez Gene identifiers.

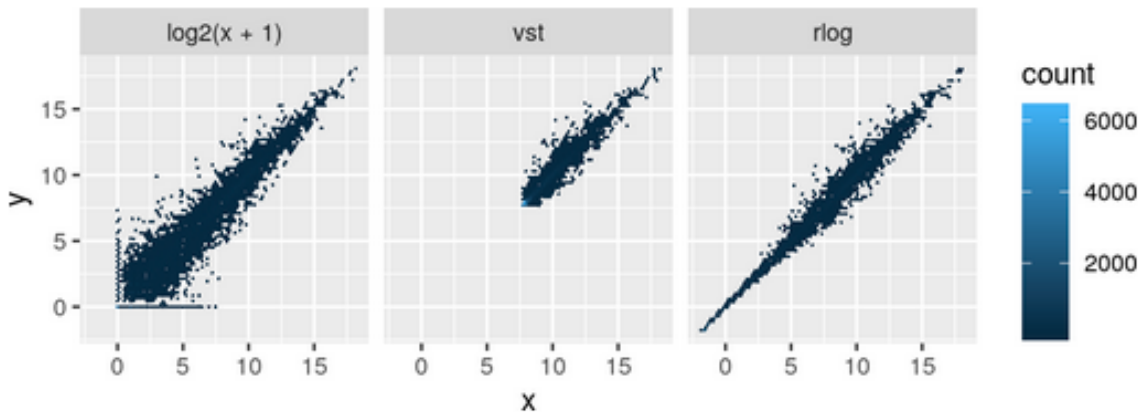
Data transformation (for visualisation or clustering)

- Pseudo-counts

$$\widetilde{y} = \log_2(y + k)$$

where y is the count values and k some chosen positive constant.

- Variance-stabilizing transformation DESeq2::varianceStabilizingTransformation
- Regularized logarithm transformation DESeq2::rlogTransformation



Data transformation (for visualisation or clustering)

- VST: much faster to compute and less sensitive to high count outliers than the rlog.
- rlog tends to work well on small datasets (n < 30), potentially outperforming the VST when there is a wide range of sequencing depth across samples (an order of magnitude difference).
- The authors recommend VST for medium-to-large datasets ($n > 30$).

Principal Component Analysis (PCA)

Aim

To reduce multidimensional datasets to lower dimensions analysis

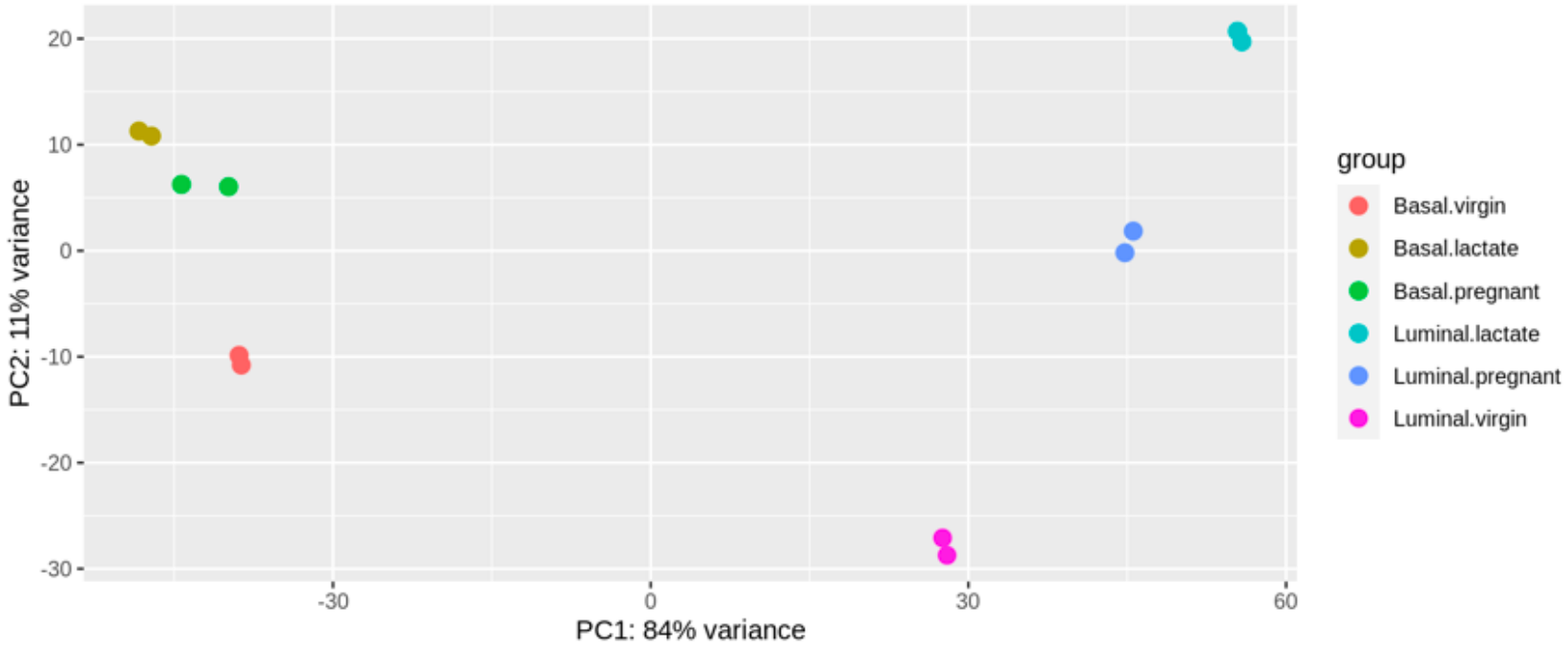
How ?

Transformation of a set of observations of possible correlated variables (genes) into a set of values of linearly uncorrelated variables (principal components)

- Property: the first principal component has the largest possible variance.
- PCA is sensitive to the scaling of the data.

In DESeq2, the PCA is performed on the top genes selected by highest row variance (*ntop* argument) of the plotPCA function

DESeq2::plotPCA



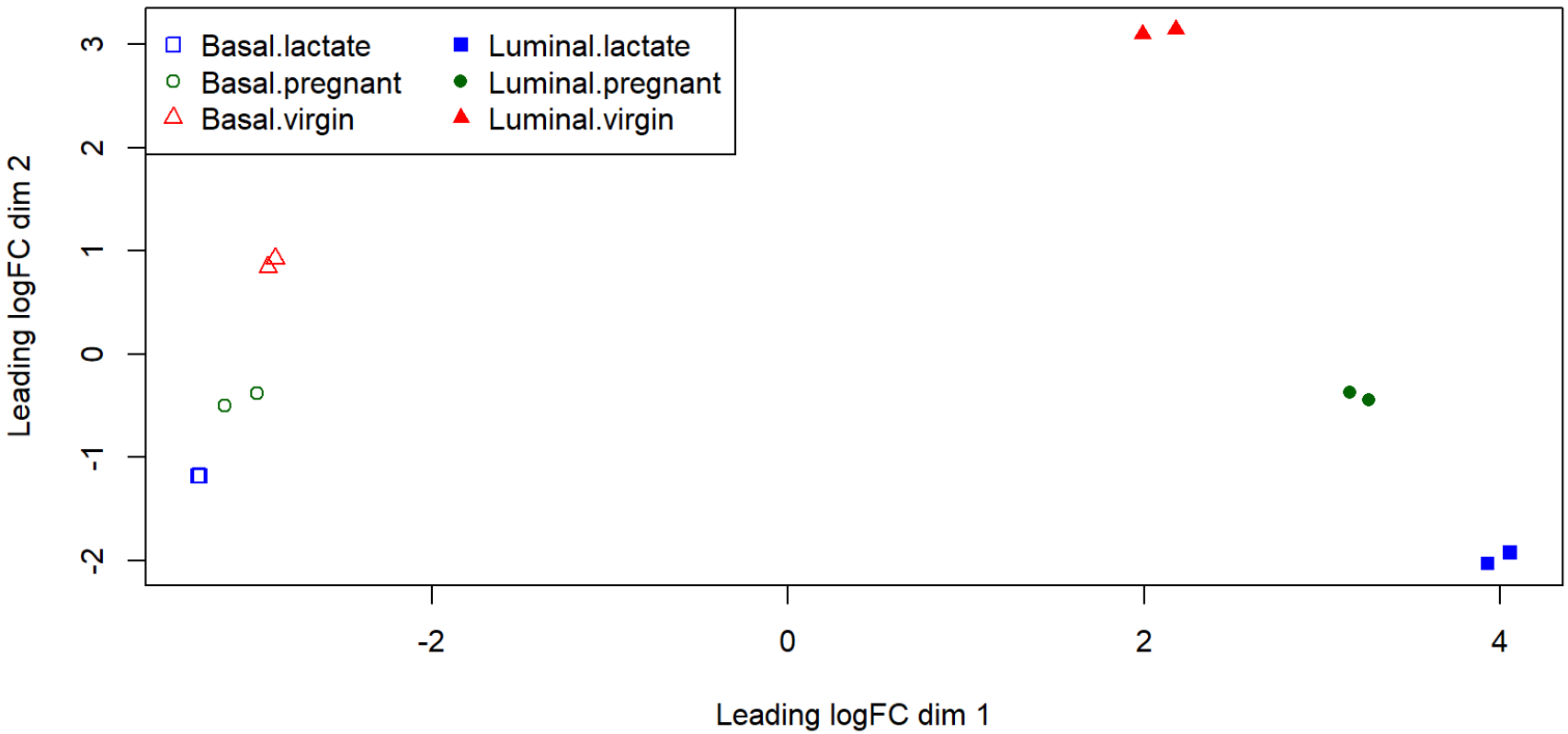
Mouse mammary gland dataset (Fu et al. 2015), PCA plot

edgeR::plotMDS

MDSPlot Multidimensional scaling plot
A means of visualizing the level of similarity of individual cases of a dataset. The distances between points on the plot reflects the level of similarity between them. The argument *gene.selection* of the plotMDS edgeR function corresponds to top genes chosen for the calculation of the MDS.

- *common* : top genes with the largest root-mean-square deviations between samples
- *pairwise* (default value) : a different set of top genes is selected for each pair of samples

edgeR::plotMDS



Mouse mammary gland dataset (Fu et al. 2015), MDS plot

Your turn ! Exercise 4 - Data filtering

Part1

- Calculate counts-per-million (cpm)

Part2

- Keep genes if they are expressed at a counts-per-million (cpm) above 0.5 in at least two samples corresponding to the number of replicates
- How many genes are kept?

Your turn ! Exercise 5 - Quality Control

Part1

Barplot of library sizes

- Explore graphically library size with bar chart representation
- Add color on bar for each group

Multimensional scaling (MDS) plot

- Produce a Multidimensional scaling plot (MDSplot)
- What is the greatest source of variation in the data (i.e. what does dimension 1 represent)?
- What is the second greatest source of variation in the data?

Part2

- Explore graphically library size with ggplot2 package to make beautiful and customizable plots

Modélisation statistique

Définir un modèle statistique permettant de décrire le phénomène d'intérêt.

modèle ?

- une formule mathématique,
- une représentation simplifiée de la réalité
- qui fait des hypothèses explicites, potentiellement fausses
- et permet de raisonner.

statistique ?

On va supposer que le processus qui a généré les observations est stochastique.

Traduire la ou les questions biologiques en terme statistique. Revient souvent à faire des tests sur les paramètres du modèle.

Comment s'y prendre ?

The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill. Albert Einstein

Planification expérimentale

Cette étape permet à partir d'un modèle et de premières données de définir le plan d'expérience optimal.

Dans la pratique, il est fréquent que les expériences soient produites en se contentant d'une formulation verbale de la démarche, en faisant varier quelques facteurs d'intéret et en ajustant le nombre de répétitions aux contraintes expérimentales ou budgétaires.

Démarche

Bien comprendre . . .

- le contexte
- les questions
- comment les données sont recueillies

Identifier . . .

- l'**unité statistique**, élément de base sur lequel des données sont observées ou mesurées.
- la ou les **variables d'intérêt** ou à expliquer $\rightarrow$ **Y**
- les sources de variabilité

Your turn ! Mouse mammary gland dataset

What are experimental units ? Factors ? Discuss biological questions.

File	Sample	CellType	Status
GSM1480297_MCL1-DG_BC2CTUACXX_ACTTGA_L002_R1.txt	GSM1480297	Basal	virgin
GSM1480298_MCL1-DH_BC2CTUACXX_CAGATC_L002_R1.txt	GSM1480298	Basal	virgin
GSM1480299_MCL1-DI_BC2CTUACXX_ACAGTG_L002_R1.txt	GSM1480299	Basal	pregnant
GSM1480300_MCL1-DJ_BC2CTUACXX_CGATGT_L002_R1.txt	GSM1480300	Basal	pregnant
GSM1480301_MCL1-DK_BC2CTUACXX_TTAGGC_L002_R1.txt	GSM1480301	Basal	lactate
GSM1480302_MCL1-DL_BC2CTUACXX_ATCACG_L002_R1.txt	GSM1480302	Basal	lactate
GSM1480291_MCL1-LA_BC2CTUACXX_GATCAG_L001_R1.txt	GSM1480291	Luminal	virgin
GSM1480292_MCL1-LB_BC2CTUACXX_TGACCA_L001_R1.txt	GSM1480292	Luminal	virgin
GSM1480293_MCL1-LC_BC2CTUACXX_GCCAAT_L001_R1.txt	GSM1480293	Luminal	pregnant
GSM1480294_MCL1-LD_BC2CTUACXX_GGCTAC_L001_R1.txt	GSM1480294	Luminal	pregnant

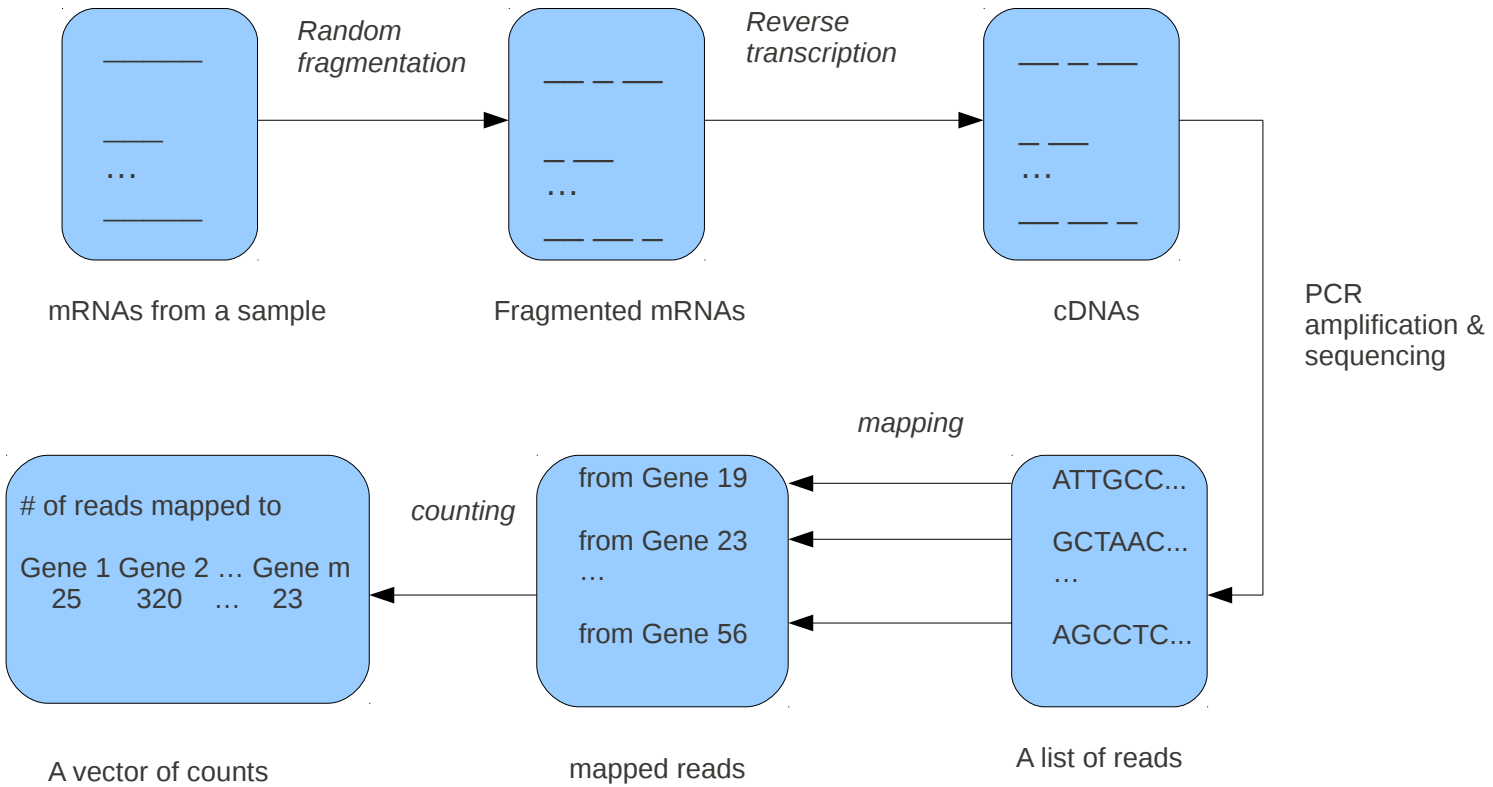
Differential analysis

Identification of differentially expressed (DE) genes

A gene is declared **differentially expressed** (DE) between two conditions if the observed difference is statistically significant, i.e. greater than a natural random variation.

- Need of statistical tools to make a decision.

RNA-sequencing



Adapted from Li et al. (2011)

Statistical issues of gene expression analysis from RNA-Seq experiment

- A large number of genes and few replicates
- Non-negative integers with asymmetric distribution
- From 0 up to millions with different variance within different parts of the dynamic range (**heteroskedasticity**)
- Systematic sampling biases, e.g. the total number of sequences (= **library size**) is not the same for all the samples

A typical raw dataset

	S_1	S_2	...	S_j	...	S_n
Gene 1	16	9	...	y_{1j}	...	15
Gene 2	4448	3973	...	y_{2j}	...	3964
...	...	...	...	...	...	...
Gene i	y_{i1}	y_{i2}	...	y_{gj}	...	y_{in}
...	...	...	...	...	...	...
Gene G	59	164	...	y_{Gj}	...	143
Seq. depth	6865057	11127087	...	$n_j = \sum_{g=1}^{G} y_{gj}$	...	11320226

y_{gj} = number of sequences from sample j assigned to gene g.

Remark: one row = one region of interest (gene, exon, transcript, $\cdots$).

Normalization or how to make measurements comparable ?

Definition

Normalization is a process designed to identify and correct **technical biases** removing the least possible biological signal. This step is technology and platform-dependant.

Technical biases

Some biases may be **controlled** by an adapted experimental design or a good experimental protocol.
Normalization aims to correct systematic **uncontrollable** biases such as those induced by sequencing process.

Within and between normalization

Within-sample normalization enabling comparisons of fragments (genes) from a **same** sample.
Between-sample normalization enabling comparisons of fragments (genes) from **different** samples.

Sources of variability

Read counts are proportional to expression level, gene length and sequencing depth (same RNAs in equal proportion).

Within-sample

- Gene length
- Sequence composition (GC content)

Between-sample

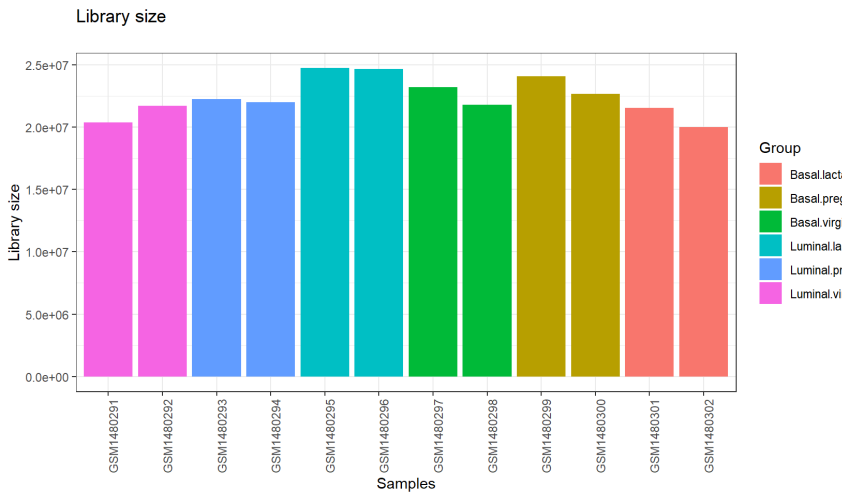
- Depth (total number of sequenced and mapped reads)
- RNA-composition or presence of majority fragments
- Sequence composition du to PCR-amplification step in library preparation (Pickrell et al. 2010, Risso et al. 2011)

Normalization and differential expression (DE) analysis

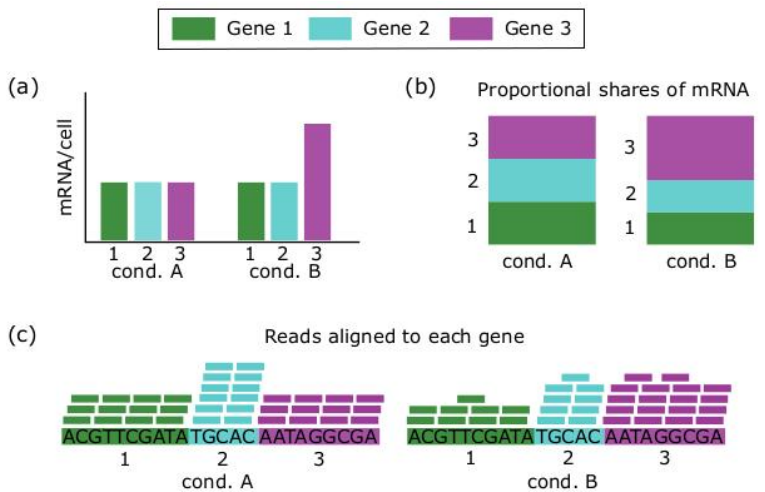
DE analysis concerned with **relative changes** in expression levels between conditions rather than estimating absolute expression levels.

Normalization: identify and correct technical effects related to the experimental conditions (sample-specific effects) without altering the biological signal.

Sequencing depth



RNA composition



from Evans et al. (2017)

Typology of normalization methods

according to the underlying assumptions (Evans et al. 2017).

Normalization by library size

Same total expression, same amount of mRNA/cell for each experimental condition.

Normalization by distribution or testing

- DE and non-DE genes have the same behaviour.
- Balanced expression (up/down).

Normalization by controls

- Existence of control (invariant set of genes).
- Control genes behave like non-control genes (same technical effects).

Relative library size

y_{gj}: raw read counts of gene g in sample j

$n_j = \sum_{g=1}^{G} y_{gj}$: relative library size of sample j after sequencing

Warning: n_j have only a technical, not a biological meaning.

Absolute counts and effective library size

a_{gj}: *unknown* absolute counts (average number of mRNAs from a given gene in the cells before seq.) We observed counts prop. to a_{gj} and L_g, the length of the gene g.

Effective library size: $\sum_{g=1}^{G} a_{gj}$.

Normalization by library size: effective Library Size concept

Motivation

Different biological conditions express different RNA repertoires, leading to different total amounts of RNA

Assumption

A majority of transcripts is not differentially expressed

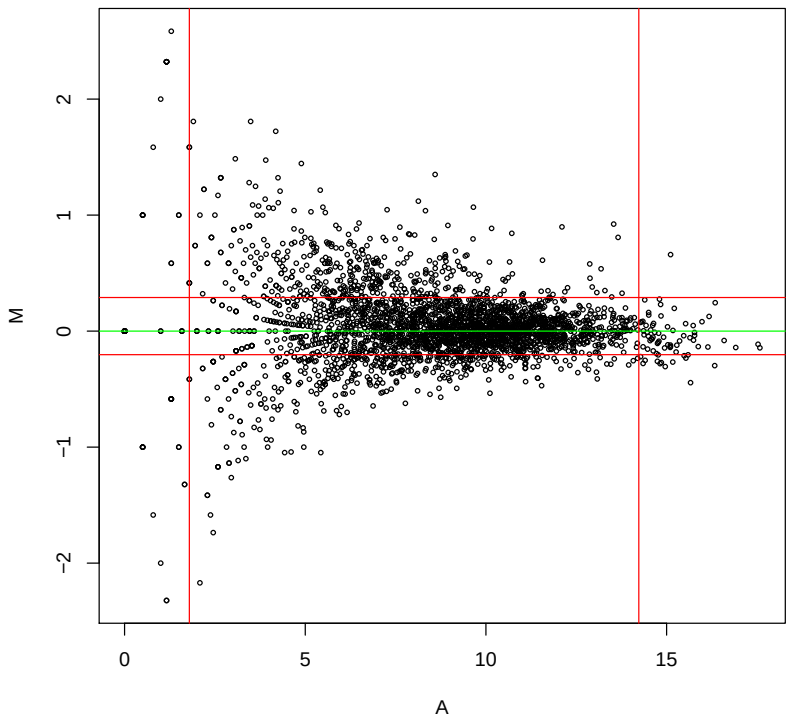
Aim

Minimizing effect of (very) majority sequences

- Trimmed Mean of M-values, Robinson and Oshlack 2010 (edgeR)
- Relative Log-Expression, Anders and Huber 2010 (DESeq2)

Normalization by library size: Trimmed Mean of M-values (TMM)

Idea: we may not estimate the total ARN production in one condition but we may estimate a global expression change between two conditions from non extreme M_g distribution.



Filter on:

- transcripts with nul counts,
- the 30% more extreme $M^r_{gj} = log_2(\frac{y_{gj}/N_j}{y_{gr}/N_r})$ values,
- the 5% more extreme $A^r_{gj} = 0.5 \times [log_2(\frac{y_{gj}}{N_j}) + log_2(\frac{y_{gr}}{N_r})]$ values.

Normalization by library size: Trimmed Mean of M-values

1. Select the reference sample r
2. Define a set of genes G^* for which neither the M_{gj}^r or the A_{gj}^r value was trimmed
3. Calculate the scaling factors $TMM_j^{(r)}$ such as

$$log_2(TMM_j^{(r)}) = \frac{\sum_{g \in G^*} w_{gj}^r M_{gj}^r}{\sum_{g \in G^*} w_{gj}^r}$$

with $w_{gj}^r = \frac{N_j - y_{gj}}{N_j y_{gj}} - \frac{N_r - y_{gr}}{N_r y_{gr}}$

4. Rescale the factors to avoid dependance on a specific reference sample

$$\hat{s}_j = \frac{TMM_j^{(r)}}{exp(\sum_\ell TMM_\ell^{(r)} / n)}$$

Normalization by library size: Relative Log-Expression method (RLE, DESeq)

1. Compute a pseudo-reference sample: geometric mean across samples (less sensitive to extreme value than standard mean)

$$y^r_{gj} = \left(\pi_{j=1}^{n} y_{gj}^{1/n}\right)$$

with y_{gj} number of reads in sample j assigned to gene j, n number of samples in the experiment.

2. Calculate scaling factors

$$\hat{s}_j = \underset{g:y^r_{gj}\neq 0}{median} \frac{y_{gj}}{y^r_{gj}}$$

Normalization by library size: Some remarks about TMM and RLE normalization

Interpretation of the scaling factors

- The normalization factors of all the libraries multiply to 1.
- $\hat{s}_j < 1$: a small number of high count genes are monopolizing the sequencing. $\Rightarrow$ Need of downscaling.

	condA.1	condA.2	condA.3	condB.1	condB.2	condB.3
RLE	1.05	1.05	0.87	1.06	1.06	0.93
TMM	1.02	1.00	0.97	1.01	1.05	0.95

Model-based normalization, not transformation

In edgeR and DESeq2, normalization factors $=$ correction factors that enter into the model.

Where conventional methods fail

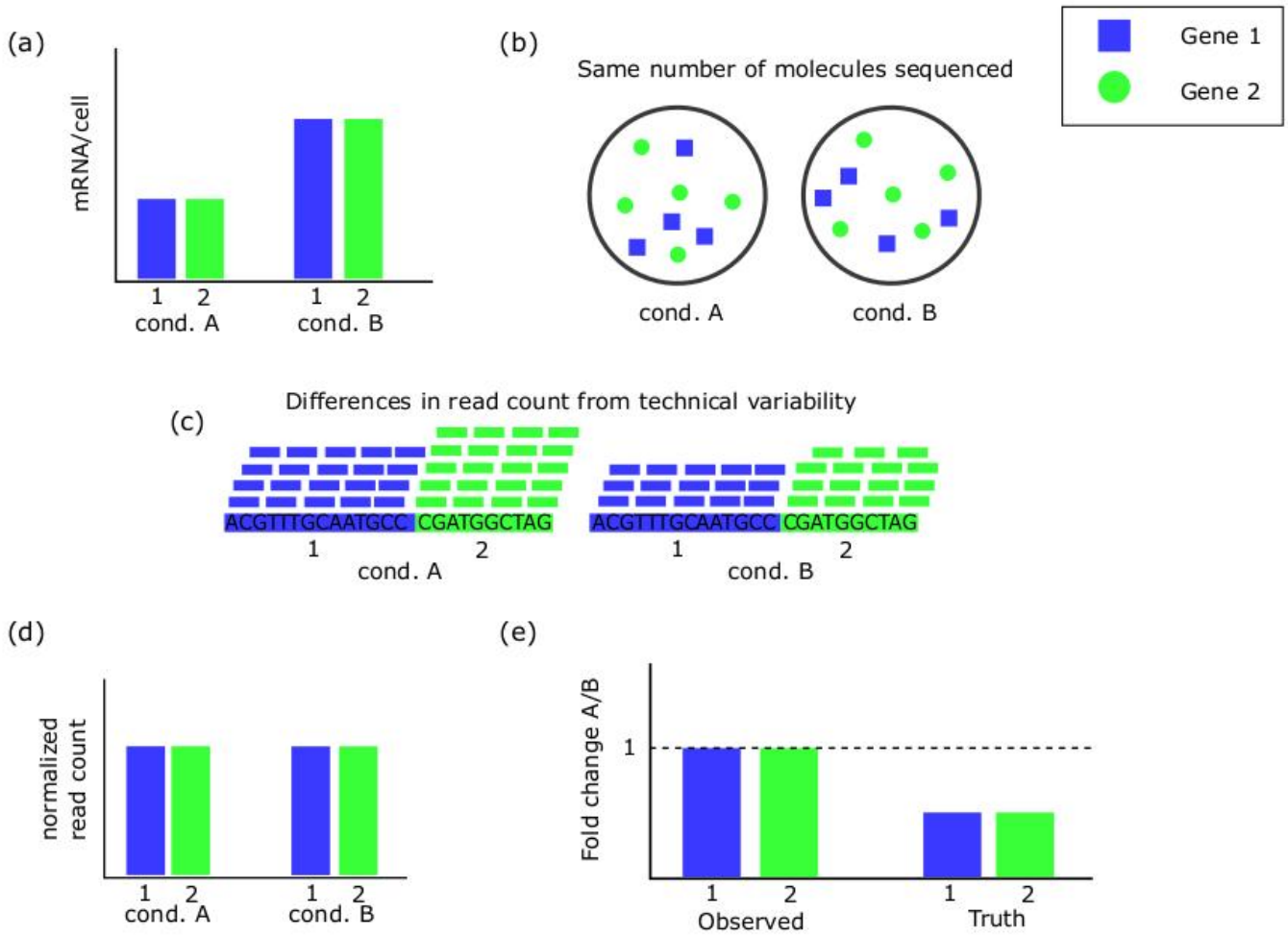


Figure: from Evans et al. (2017)

Normalization by controls

Assumptions

- Existence of controls and behavior as expected (negative controls = non-DE)
- Controls behave like non-control genes (affected by same technical effects)

Methods

- Housekeeping genes
- "Conventional normalization" with Spike-ins
- Factor analysis of controls: Remove Unwanted Variation (RUV) (Risso et al., 2014)

Normalization: key points (1/2)

Dillies et al. 2013, Evans et al. 2017

- A normalization is needed and has a great impact on the DE genes,
- RNA-seq data are affected by technical biaises (total number of mapped reads per lane, gene length, composition bias...),
- Do not normalize by gene length in a context of differential analysis,
- Performant and robust methods in a DE analysis context on the gene scale:
  - Trimmed Mean of M-values, (Robinson and Oshlack 2010, edgeR)
  - Relative Log-Expression, (Anders and Huber 2010, DESeq2)

Normalization: key points (2/2)

Dillies et al. 2013, Evans et al. 2017

- The correct normalization method to use depends on which assumptions are valid for the biological experiment:
  - same / different amount of mRNA / cell
  - majority of genes is invariant between conditions, low number of DE genes
  - symmetry of differential expression
  - absence of high count genes, similar library size
- Incorrect normalization leads to problem in downstream analysis, such as inflated FP.
- There are examples of global shifts in expression that violate assumptions of conventional normalization methods, requiring controls.

Your turn ! Exercise 6 - Normalization

Part1

- Calculate normalization factors

Part2

- Perform a bar chart to represent normalization factors
- Add a horizontal line fixed at 1
- Add color on bar corresponding to each group

Differential Analysis

Identification of differentially expressed genes (DE)

A gene is declared differentially expressed (DE) between two conditions if the observed difference is statisticially significant, ie more than only du to natural random variation.

- Statistical tools are necessary to take this decision.
- The main steps are : experimental design, normalisation and differential analysis, multiple testing.

Fold Change approach and ideal cut-off values

Cut-off values for gene expression fold change when performing RNA seq

I would like to know what the general consensus is regarding cut-off values for gene expression fold changes (is it mainly >2 up and down-regulated?). Also, is this cut-off applied together with the cut-off for p-value which is p<0.05?

I think the general consensus is > and < than 2-fold, however, we should all justify our rationale for using 2-fold. In our specific case, a difference

> For most gene expression change, people always use fold change 2 as a
> cutoff for microarray or qPCR. As for RNAseq, since the method is much
> more sensitive, I guess it must lose some specificity, so I think it may
> need a higher cutoff number than 2.

Fold Change approach and ideal cut-off values

$$FC_g = \frac{x_{g\cdot}}{y_{g\cdot}}$$

	Gene	CondA1	CondA2	CondB1	CondB2	FC	pvalue
1	Gene1	5.00	7.00	2.00	2.00	3.00	0.06
2	Gene2	800.00	1000.00	350.00	250.00	3.00	0.03
3	Gene3	700.00	1100.00	350.00	250.00	3.00	0.10
4	Gene4	500.00	1300.00	550.00	50.00	3.00	0.33

FC does not take the variance of the samples into account.

Problematic since variability in gene expression is partially gene-specific.

Differential analysis

Aim : To detect differentially expressed genes between two conditions

- Discrete quantitative data
- Few replicates
- Overdispersion problem

Challenge: method which takes into account overdispersion and few number of replicates

- Proposed methods : edgeR, DESeq(2) for the most used and known *Anders et al. 2013, Nature Protocols*
- An abundant litterature
- Comparison of methods : Pachter et al. (2011), Kvam and Liu (2012), Soneson and Delorenzi (2013), Rapaport et al. (2013), Schurch et al. (2016)

Hypothesis testing

Definition

A general method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample.

Four ingredients

1. Experimental data $x_1, x_2, \ldots, x_n$
2. Statistical model : assumptions about the independence or distributions of the observations with parameter θ
3. Hypothesis to test : assumption about one parameter of the distribution
4. Region of rejection (or critical region): the set of values of the test statistic T for which the null hypothesis H_0 is rejected. $T = f(X_1, X_2, \ldots, X_n)$ is a function which summarizes the data without any loss of information about θ. The distribution of T under H_0 is known.

Critical region and p-value

p-value $p(t)$

For a realisation t of the T test statistic $p(t)$ is the probability (calculating under H_0) of obtaining a test statistic at least as extreme as the one that was actually observed.

In bilateral case :

$$p(t) = \mathbb{P}_{H_0}\{|T| \geq |t|\}$$

The p-value measures the agreement between H_0 and obtained result.

Link with the critical region

$$\mathbb{P}_{H_0}\{T \in \mathcal{R}\} = \mathbb{P}\{p(t) \leq \alpha\}$$

with α the significance level.

Differential analysis gene-by-gene- with replicates

For each gene g

Is there a significant difference in expression between condition A and B?

- Statistical model (definition and parameter estimation) - Generalized linear framework Y_{gjk} follows $\mathbf{f}(\theta_{gjk})$
- Hypothesis to test : H_{0g} Equality of relative abundance of gene g in condition A and B vs H_{1g} non-equality
- Critical region - Wald Test or Likelihood Ratio Test

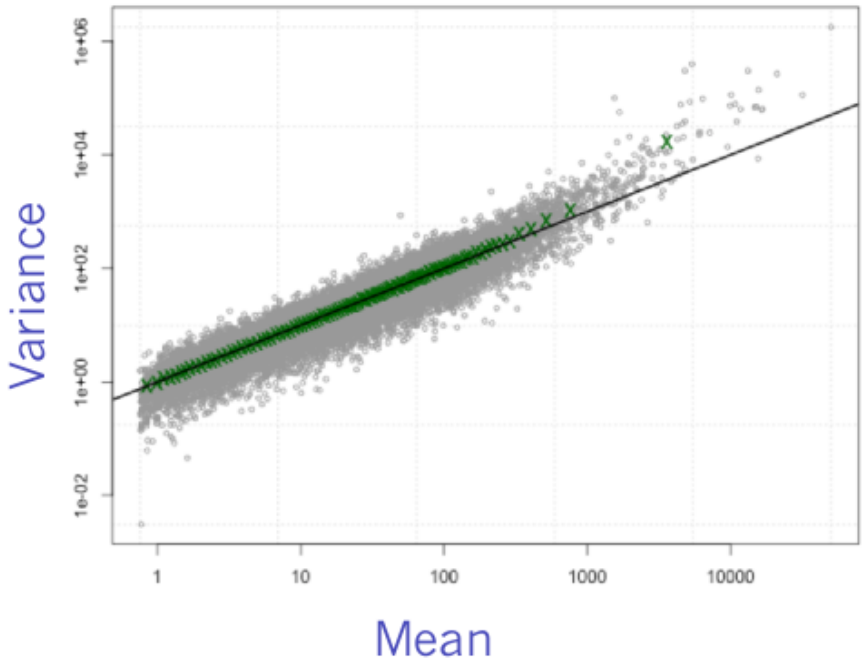
The Poisson Model

Let be Y_{gj} the read count for gene g in sample j

- Y_{gj} follows a **Poisson** distribution ($\mu_{gj} = s_{gj} * q_{gj}$), with s_{gj} library size and $\log q_{gj} = \sum_r x_{jr}\beta_{gr}$, $\mathbf{X} = [x_{jr}]$ is the design matrix and β is the vector of coefficients.
- Property : $\mathbb{V}(Y_{gj}) = \mathbb{E}(Y_{gj}) = \mu_{gj}$

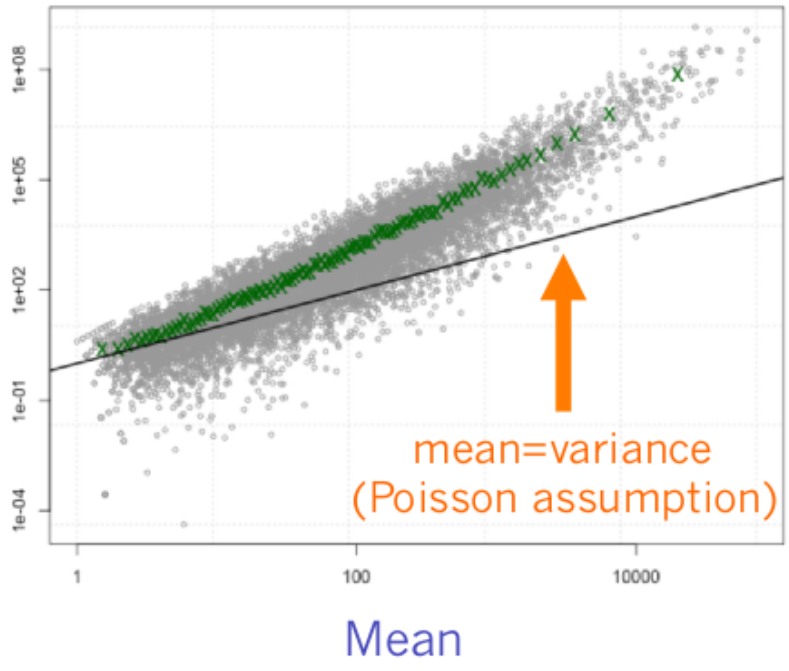
Mean-Variance Relationship

Technical replicates



data from Marioni et al. Gen Res 2008

Biological replicates



data from Parikh et al. *Genome Bio* 2010

From D. Robinson and D. McCarthy

Overdispersion in RNA-seq data

Counts from biological replicates tend to have variance exceeding the mean (= overdispersion relative to the Poisson distribution). Poisson describes only technical variation.

What causes this overdispersion?

- Correlated gene counts
- Clustering of subjects
- Within-group heterogeneity
- Within-group variation in transcription levels
- Different types of noise present...

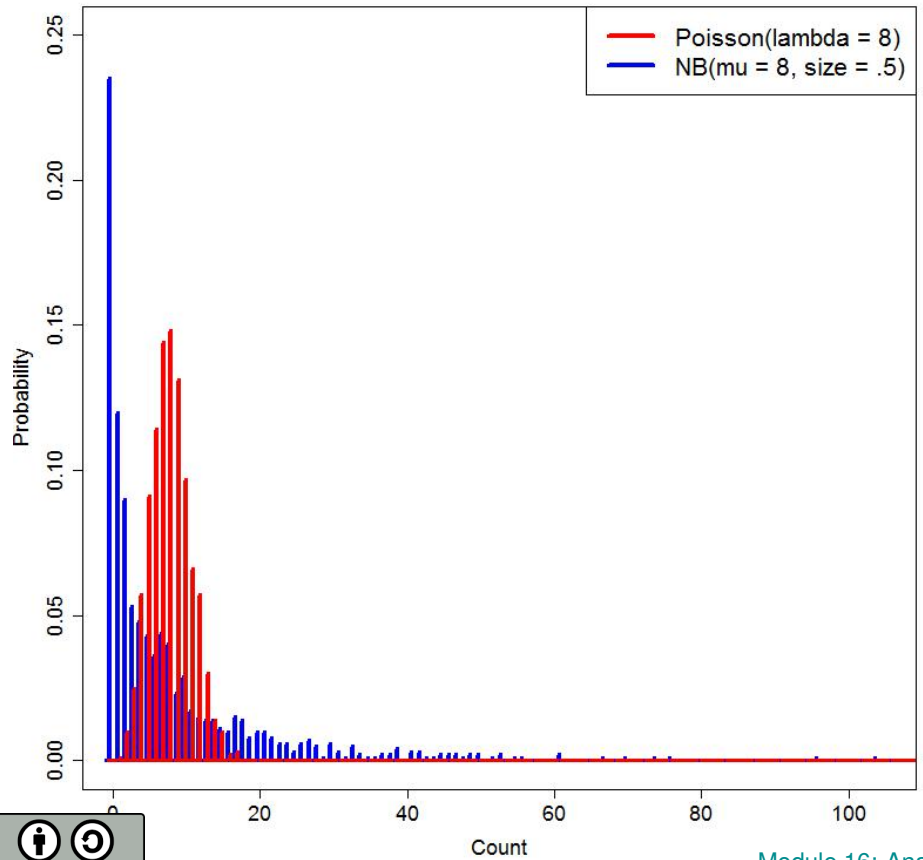
In case of overdispersion, $\uparrow$ of the type I error rate (prob. to declare incorrectly a gene DE).

Alternative : Negative Binomial Models

A supplementary dispersion parameter ϕ to model the variance

Y_{gj} follows a **Negative Binomial** distribution (mean = μ_{gj}, dispersion = ϕ_g)

Poisson vs Negative Binomial Models



Types of noise in data

1. Shot noise: unavoidable noise inherent in counting process (dominant for weakly expressed genes)
2. Technical noise: from sample preparation and sequencing, hopefully negligable
3. Biological noise: unaccounted for differences between samples (dominant for strongly expressed genes)

Estimation the dispersion (biological noise)

How to estimate a reliable dispersion from a very small number of replicates (sometimes less than 5) ?

- gene-specific tests: lack of sensitivity (proportion of true positives among positives) due to the lack of information
- common dispersion parameter for all tests $\rightarrow$ many false positives

Estimation the dispersion: the key question

One solution: compromise between gene-specific and common dispersion parameter estimation

- **edgeR**: borrow information across genes for stable estimates of ϕ
  3 ways to estimate ϕ (common, trended, tagwise)
- **DESeq**: data-driven relationship of variance and mean estimated using parametric or local regression for robust fit across genes

Method	Variance	Reference
DESeq	$\mu(1+\phi_\mu\mu)$	Anders et Huber (2010)
edgeR	$\mu(1+\phi\mu)$	Robinson et Smyth (2009)

the DESeq2 pipeline

Model

$Y_{gj} \sim$ NB(mean = μ_{gj}, dispersion = ϕ_g)

$\mu_{gj} = s_{gj} * q_{gj}$

$\log q_{gj} = \sum_r x_{jr}\beta_{gr}$, where $\mathbf{X} = [x_{jr}]$ is the design matrix and β is the vector of coefficients.

Main steps performed by the DESeq function:

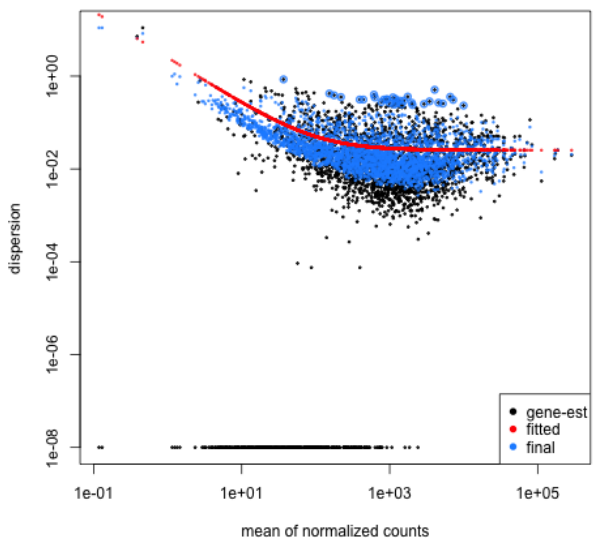
1. estimation of size factors $s_{gj} = s_j$ by estimateSizeFactors
2. estimation of dispersion by estimateDispersions
3. negative binomial GLM fitting for β_g and Wald statistics by nbinomWaldTest

Remark: the method implemented in the DESeq2 package is quite different than the method proposed in the DESeq paper (Anders and Huber 2010)

Estimating dispersion parameters

estimateDispersions

1. calculation of a preliminary gene-wise dispersion estimates by maximum likelihood
   few samples $\rightarrow$ *strong fluctuation aroung the true values*;
2. fitting of a trend curve to capture the dependence of these estimates on average expression strength;
3. the final estimates of dispersion results in a shrinkage of the noisy gene-wise estimates towards a consensus.



Shrinkage estimation of logFCs

Observation

Variance of logFCs depends on mean count (heteroskedasticity)
logFC estimates for genes with low read count have a strong variance

$\rightarrow$ effect sizes difficult to compare across the dynamic range of the data

Shrinkage estimation

DESeq2 propose to shrink logFCs estimates toward zero in a manner such that shrinkage is stronger when the available information for a gene is low (because low counts, high dispersion or few degrees of freedom)

More than two conditions - GLM framework

$Y_{gj} \sim$ NB(mean = μ_{gj}, dispersion = ϕ_g) with $\log(\mu_{gj}) = \log(s_j) + \log(q_{gj})$ in which:

- s_{gj} is the (gene-specific g) library size for sample j,
- $\log q_{gj} = \sum_r x_{jr}\beta_{gr}$ where $\mathbf{X} = [x_{jr}]$ is the design matrix and β is the vector of coefficients.

A Generalized Linear Model (GLM) allows to decompose the effects on the mean of

- different factors,
- their interactions.

More than two conditions - GLM framework

Alternative approach: linear model for count data (Law et al., 2014)

1. Data are transformed so that they are approximately normally distributed (voom)
2. A linear Gaussian model is fitted (limma)

Comparaison of 11 differential analysis methods

Soneson and Delorenzi, Rapaport et al. (2013), Schurch et al. (2016)

- The number of replicates matters!
  - Small number of replicates (2-3) or low expression $\rightarrow$ be careful!!
  - Large number of replicates (10 or so) or very high expression $\rightarrow$ method choice does not matter much.

Comparaison of 11 differential analysis methods

Soneson and Delorenzi, Rapaport et al. (2013), Schurch et al. (2016)

- Results are more accurate and less variable between methods if DE genes are regulated in both directions.
- Outlier counts affect different methods in different ways
  Removing genes with outlier counts or using non-parametric methods reduce the sensitivity to outliers
- The dispersion estimation method matters! Allow tagwise dispersion values is better.
- Normalization methods have problems when all DE genes are regulated in one direction. Iterative approaches like TCC improve performance

Interpretation - Statistical significance and practical importance

- Practical importance and statistical significance (detectability) have little to do with each other.
- An effect can be important, but undetectable (statistically insignificant) because the data are few, irrelevant, or of poor quality.
- An effect can be statistically significant (detectable) even if it is small and unimportant, if the data are many and of high quality.

Your turn ! Exercises 7.1 and 7.2

Exercise 7.1 - Experimental design matrix

- Construct a design matrix with ‘model.matrix()‘

Exercise 7.2 - Estimating dispersion

Part1

- Compute the estimate of the common, trended and tagwise dispersions across genes.
- What is the common dispersion value?

Part2

- Plot the genewise biological coefficient of variation (BCV) against gene abundance (in log2 counts per million)

Challenge

- Plot variance against mean of counts per gene in log2 scale.

Why is robustness needed?

Transcriptome genetics using second generation sequencing in a Caucasian population

Stephen B. Montgomery[1,2], Micha Sammeth[3], Maria Gutierrez-Arcelus[1], Radoslaw P. Lach[2], Catherine Ingle[2], James Nisbett[2], Roderic Guigo[3] & Emmanouil T. Dermitzakis[1,2]

Nature, 2010

Random split of dataset: n_1=5; n_2=5 → Very little true differential expression

Results driven by outliers

CPMs (counts per million)

	NA19222	NA12287	NA19172	NA11881	NA18871	NA12872	NA18916	NA18856	NA19193	NA19140
4004	0.0	1.9	178.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0
2538	2.0	0.6	235.5	6.8	60.2	1.0	0.0	0.0	2.5	1.3
4962	3.5	0.6	429.5	1.0	35.9	0.0	0.4	0.0	0.0	4.7
7921	1.0	5.1	78.9	2.9	0.0	0.0	0.8	0.0	0.0	0.4
6115	0.0	1.3	0.0	1.9	0.0	0.5	46.1	0.0	100.1	1.3
5156	13.8	1.3	30.7	0.0	7.1	0.0	0.0	1.0	0.0	1.3
2527	23.7	111.0	228.8	77.0	129.5	10.0	45.3	27.4	26.3	19.1
1115	2.0	15.2	1074.8	19.5	13.2	10.0	29.6	0.0	1.3	5.5
3175	3.0	6.3	181.0	7.8	7.6	0.0	5.5	3.0	3.1	2.5
7951	1.0	12.1	35.9	0.0	1.0	1.0	0.0	1.0	0.0	0.0
7631	0.0	1.9	0.4	1.0	0.0	0.5	29.6	0.0	24.4	5.5
3437	24.6	31.1	167.0	4.9	21.2	4.5	8.3	10.1	8.1	0.4

	logFC	logCPM	LR	PValue	FDR
4004	-10.413038	4.186203	30.07924	4.147469e-08	0.0002239513
2538	-5.942865	4.963086	29.60406	5.299369e-08	0.0002239513
4962	-6.387829	5.576979	26.06085	3.308237e-07	0.0009320406
7921	-5.808379	3.183079	22.51927	2.080466e-06	0.0043960241
6115	5.746084	3.921353	21.37010	3.786299e-06	0.0064003595
5156	-4.573655	2.512035	20.13483	7.217026e-06	0.0101663841
2527	-2.154480	6.128702	18.44343	1.750229e-05	0.0211327628
1115	-4.575934	6.873996	18.14127	2.051076e-05	0.0211672325
3175	-3.843458	4.473754	17.71318	2.568407e-05	0.0211672325
7951	-4.786326	2.416892	17.66324	2.636730e-05	0.0211672325
7631	4.311717	2.683367	17.57990	2.754846e-05	0.0211672325
3437	-3.014484	4.821100	17.05690	3.627624e-05	0.0255505626

Dealing with outliers

NB framework

DESeq2, edgeR rely on the NB distribution which is versatile in having a mean and dispersion parameter. Extreme counts in individual samples might not fit well to the NB.

DESeq2 strategy

1. calculate Cook's distance (measure of how much the fitted coefficients would change if an individual sample were remove)
2. filter genes with outliers

Can inadvertently filter interesting genes

Dealing with outliers

edgeR strategy : robust estimation (Zhou et al. 2014, Chen et al. 2017)

- edgeR::estimateDisp(y, design, robust = TRUE)
  This option protect the empirical Bayes estimates against the possibility of outliers genes with exceptionaly large or small individual dispersions.
- edgeR::glmQLFit(y, design, robust = TRUE)
  This allows gene-specific priori df estimates, with lower values for outlier genes and higher values for the main body of genes. Reduces the chance of getting FP from genes with extreme dispersions and increases power to detect the others as DE.

Your turn ! Exercise 7.3 - Testing for DE genes

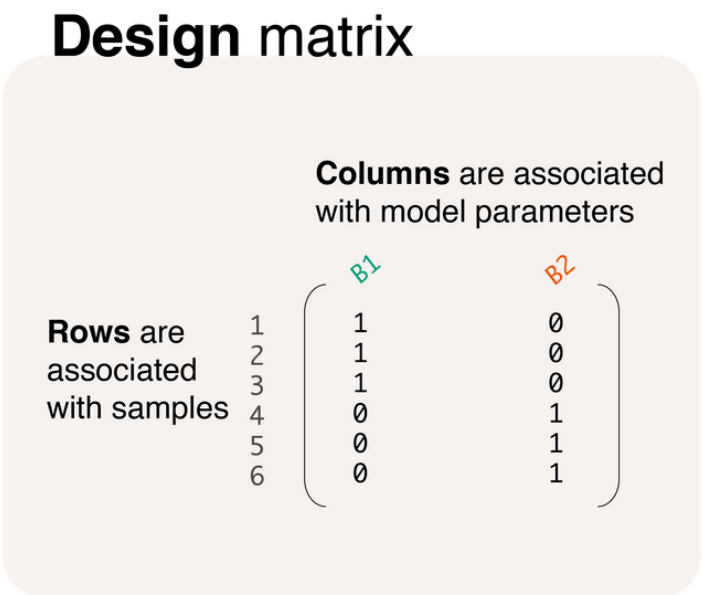
Part1

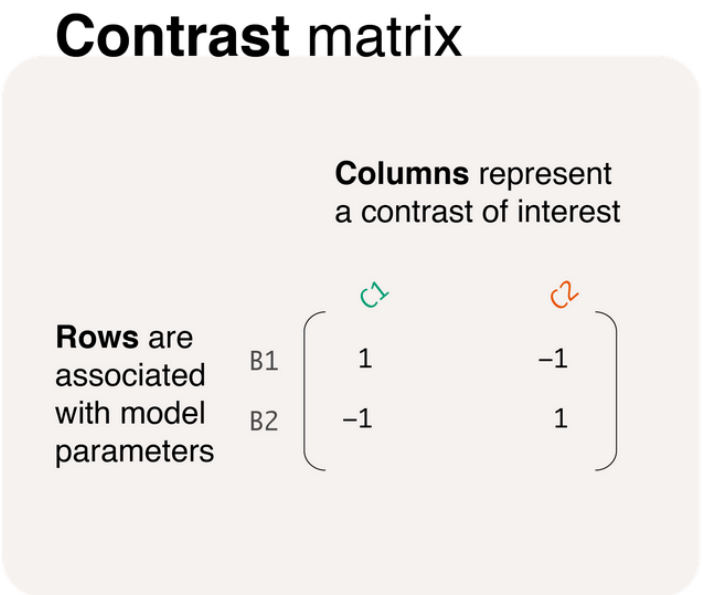
- Use glmFit to fit generalized linear model
- Use glmLRT to conduct likelihood ratio tests for one coefficient in the linear model

Creating a design matrix and contrasts

```
design <- model.matrix(~0+group)

contr.matrix <- makeContrasts(
   C1 = B1-B2,
   C2 = B2-B1)
```





Source: Law et al. 2018, package Bioconductor RNAseq123

Your turn ! Exercise 7.3 - Testing for DE genes

Part2

- Construct a specific contrast defined by difference between lactate and pregnant status within Basal CellType
- Compute likelihood ratio test for this contrast.
- Print the most differentially expressed genes.
- How many genes are up and down regulated between lactate and pregnant status within Basal CellType?

Differential Analysis : key points

- Methods dedicated to microarrays are not applicable to RNA-seq
- Small number of replicates (2-3) or low expression $\rightarrow$ be careful!!
- Large number of replicates (10 or so) or very high expression $\rightarrow$ method choice does not matter much.
- Filtering the data(genes with outliers or low counts) may be interesting
- Don't forget to correct for multiple testing !

Adapt the method to your data (nb of rep.)

Specific methods developped for few replicates.
The need for 'sophisticated' methods decreases when the number of replicates increases.

Multiple Testing

False positive (FP) (**type I error** : α) : A non differentially expressed (DE) gene which is declared DE.

For all 'genes', we test H_0 (gene i is not DE) vs H_1 (the gene is DE) using a statistical test (calcul of a score)

Pb :

Let assume all the G genes are not DE. Each test is realized at α level
Ex: $G = 10000$ genes and $\alpha = 0.05 \rightarrow \mathbb{E}(FP) = 500$ genes.

Simultaneous test of G null hypotheses

Reality	Declared non diff. exp.	Declared diff. exp.
G_0 non DE genes	**True Negatives** (*TN*)	**False Positives** (*FP*)
G_1 DE genes	**False Negatives** (*FN*)	**True Positives** (*TP*)
G Genes	N Negatives	P Positives

Aim : minimize *FP* and *FN*.

Standard approach to the multiple testing problem

Dudoit et al. (2003)

1. Computing a test statistic for each gene *g*
2. Applying a multiple testing procedure to determine which hypotheses to reject while controlling a suitable defined type I error rate

Multiple testing prodecure

It controls a particular type I error rate at level α if the error rate is $\leq \alpha$ when the procedure is applied to produce a list of P rejected hypotheses (DE genes).

The Family Wise Error Rate (FWER)

Definition

Probability of having at least one Type I error (false positive), of declaring DE at least one non DE gene.

$$FWER = \mathbb{P}(FP \geq 1)$$

The Bonferroni procedure

- Either each test is realized at $\alpha = \alpha^* / G$ level
- or use of adjusted pvalue $pBonf_g = min(1, p_g * G)$ and FWER $\leq \alpha^*$.

For $G = 2000$, $\leq \alpha^* = 0.05$, $\alpha = 2.510^{-5}$.

Easy but conservative and not powerful.

When the number of tests increases, the FWER $\rightarrow$ 1 with constant FP.

The False Discovery Rate (FDR)

Idea: Do not control the error rate but the proportion of error
$\Rightarrow$ less conservative than control of the FWER.

Definition

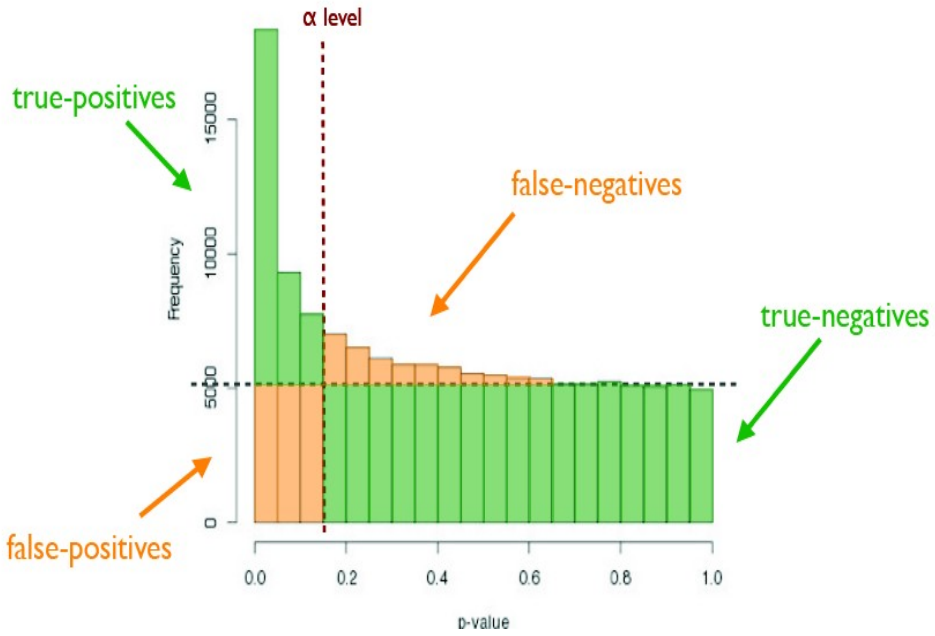
The false discovery rate of Benjamini and Hochberg (1995) is the expected proportion of Type I errors among the rejected hypotheses

$$\text{FDR} = \mathbb{E}(FP/P) \text{ if } P > 0 \text{ and } 0 \text{ if } P = 0$$

Prop

$$\text{FDR} \leq \text{FWER}$$

Standard assumption for p-value distribution



Source : M. Guedj, Pharnext

The False Discovery Rate - Benjamini et Hochberg (95)

Principle: The number of declared positive elements P is given by the greater g $p_{(g)} \leq g\alpha^*/G$.

Prop

In case of independant tests, $FDR \leq (G_0/G)\alpha^* \leq \alpha^*$

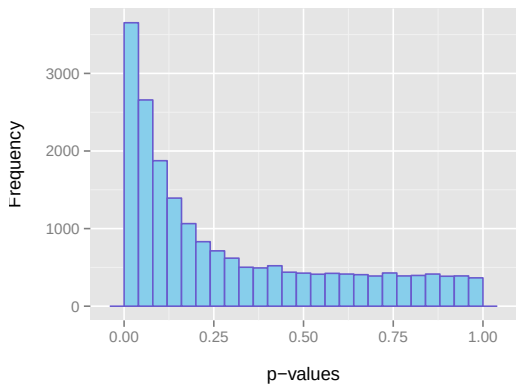
Prop

FDR Benjamini-Hochberg : $\pi_0 = \frac{G_0}{G}$=1

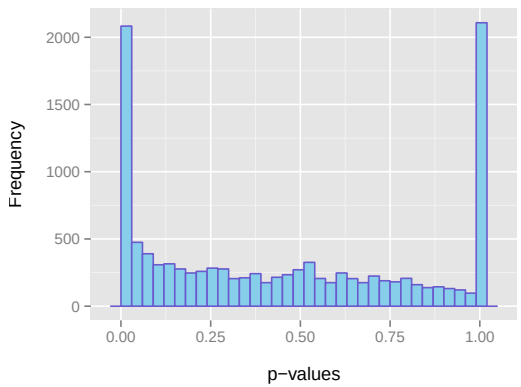
p-values histograms for diagnosis

Examples of expected overall distribution

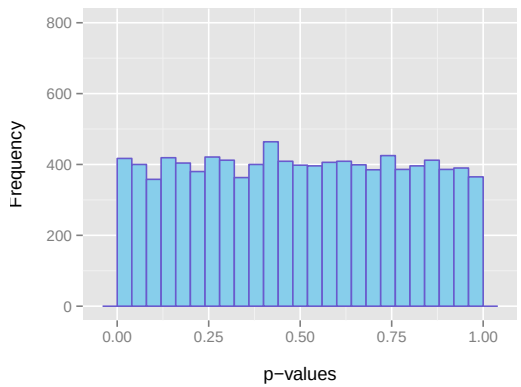
(a)



(b)



(c)



(a): the most desirable shape

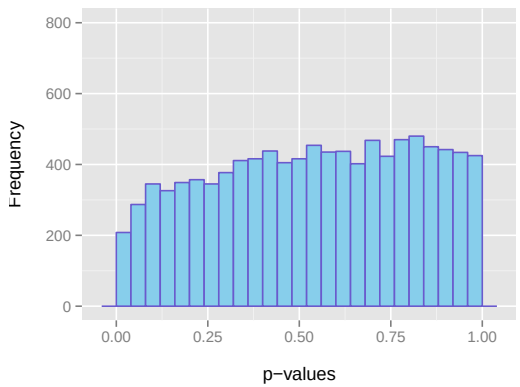
(b): very low counts genes usually have large p-values

(c): do not expect positive tests after correction

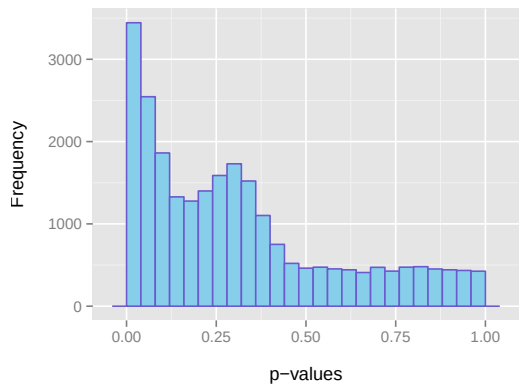
p-values histograms for diagnosis

Examples of not expected overall distribution

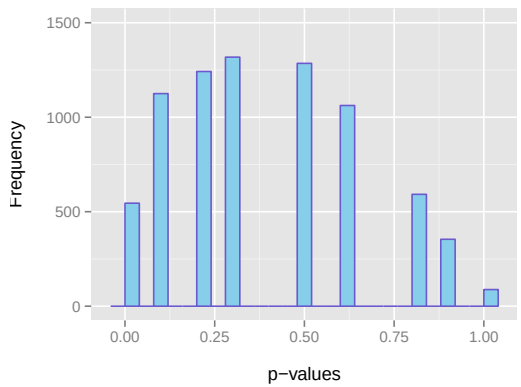
(a)



(b)



(c)



(a): indicates a batch effect (confounding hidden variables)

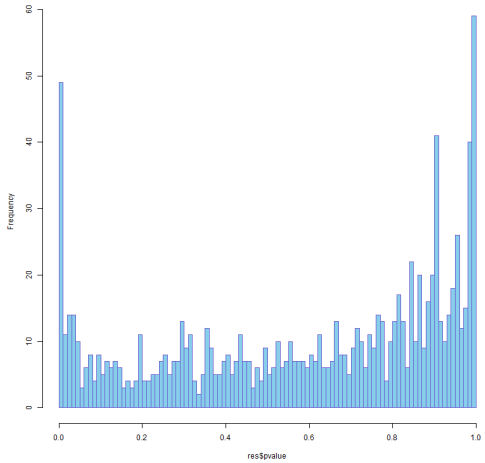
(b): the test statistics may be inappropriate (due to strong correlation structure for instance)

(c): discrete distribution of p-values: unexpected

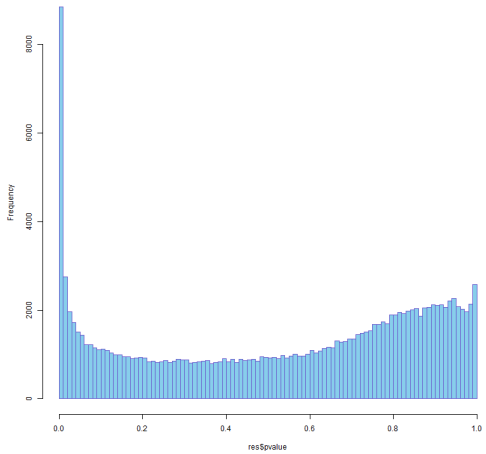
p-values histograms for diagnosis

Examples of expected or not overall distribution ?

(a)



(b)



Your turn ! Exercise 7.4 - Explore and save the results 1/2

Part1

- Plot the histogram of the raw pvalues. How do you interpret this plot?
- Produce a Volcano plot which displays log fold changes on the x-axis versus a measure of statistical significance on the y-axis
- Plot a MA-plot. Interpret the plot
- Save the results of the differential analysis in a csv file

Part2

- Order results by raw pvalues
- Save all the differentially expressed genes up and down in separate files

Your turn ! Exercise 7.4 - Explore and save the results 2/2

Challenge

- Create another contrast: difference between lactate and pregnant status within Luminal CellType.
- Compare results with Venn diagram.
- How many genes are commonly differentially expressed?

Venn diagramm and upset plot

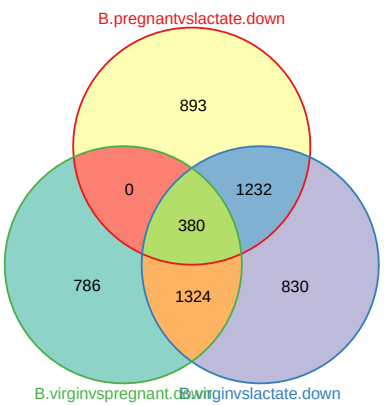
jvenn (Bardou et al. 2014)

- Venn diagramm with the venn function
  `http://bioinfo.genotoul.fr/jvenn/`
- How to export results ?

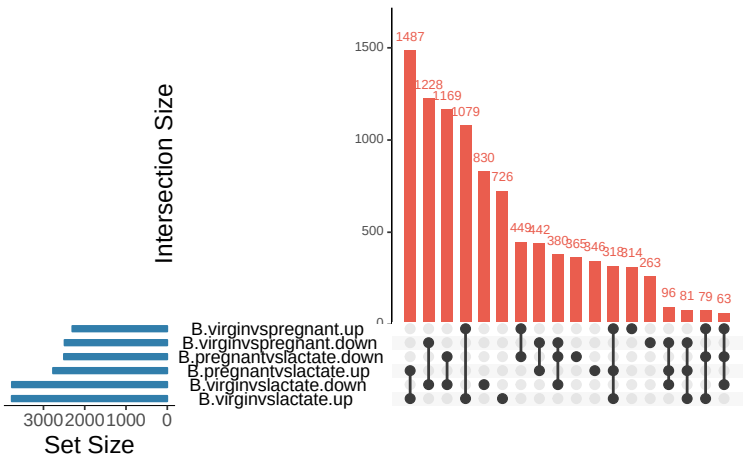
intervene Shiny app (Khan et Mathelier 2017)

`https://asntech.shinyapps.io/intervene/`

Venn diagramm



Upset plot



Multiple testing: key points

- Important to control for multiple tests
- FDR or FWER depends on the cost associated to FN and FP

Controlling the FWER

Having a great confidence on the DE elements (strong control). Accepting to not detect some elements (lack of power $\Leftrightarrow$ a few DE elements)

Controlling the FDR

Accepting a proportion of FP among DE elements. Very interesting in exploratory study.

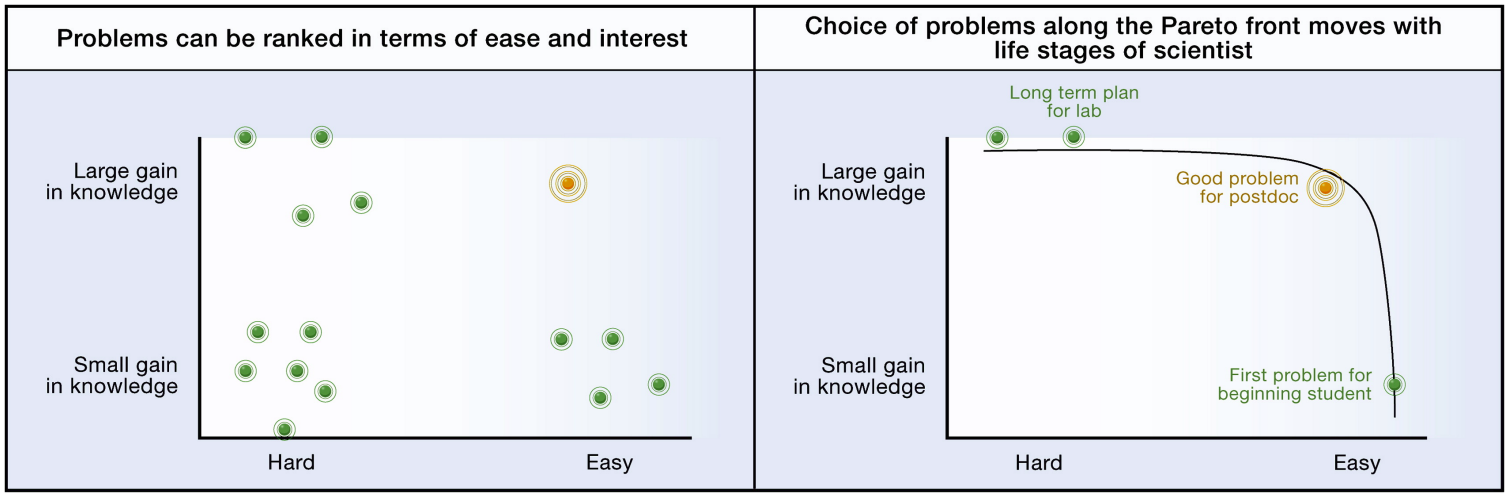
Experimental Design

A good design is a list of experiments to conduct in order to answer to the asked question which maximize collected information and minimize experiments cost with respect to constraints.

- Rule 1: Well define the biological question, get together and collect a priori knowledge (e.g. reference genome, splicing . . .),
- Rule 2: Anticipate, Identify all factors of variation and adapt Fisher's principles (1935), collect metadata from experiment and sequencing,
- Rule 3: Choose a priori tools/methods for bioinformatics and statistical analyses,
- Rule 4: Draw conclusions on results.

And do not forget: budget also includes cost of biological data acquisition, sequencing data backup, bioinformatics and statistical analysis.
`http://f1000.com/posters/1096840`

Rule 1: Well define the biological question



Choosing scientific problems on feasibility and interest [Alon 2009]

Make a choice

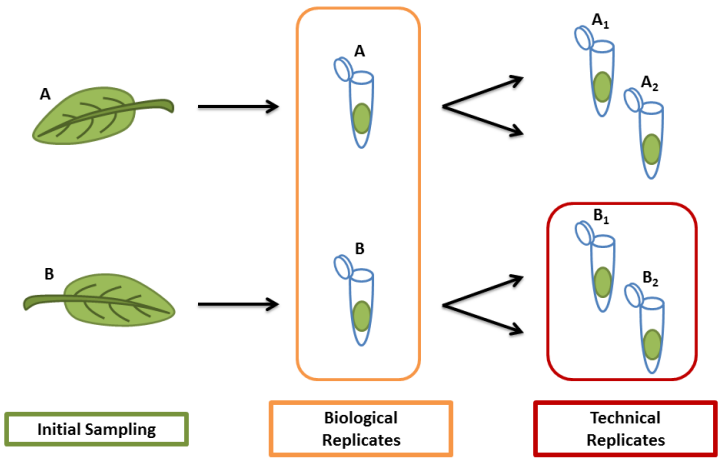
- Identify differentially expressed genes (between which conditions),
- Detect and estimate isoforms,
- Construct a de novo transcriptome.

Rule 2: Factors of variation - Metadata (1)

Basic principles - Fisher (1935), George Box (1978)

To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died. (Ronald A. Fisher, 1938).

- Technical or/and biological replications



Biological replicate:

Repetition of the same experimental protocol but independent data acquisition (several samples).

Technical replicate:

Same biological material but independent replications of the technical steps (several extracts from the same sample).

Rule 2: Factors of variation - Metadata (2)

Basic principles - Fisher (1935), George Box (1978)

Block what you can, randomize what you cannot. (George Box, 1978)

- **Randomization**

  Process of random assignment of individuals to group, block. Reduces bias caused by factors that have not been accounted for in the experimental design.

- **Blocking**

  Isolating variation attributable to a nuisance variable which has an effect on the response, but is of no interest to the experimenter (e.g. run, day, sex...). Experimental units are grouped into homogeneous block. Random allocation within each block.

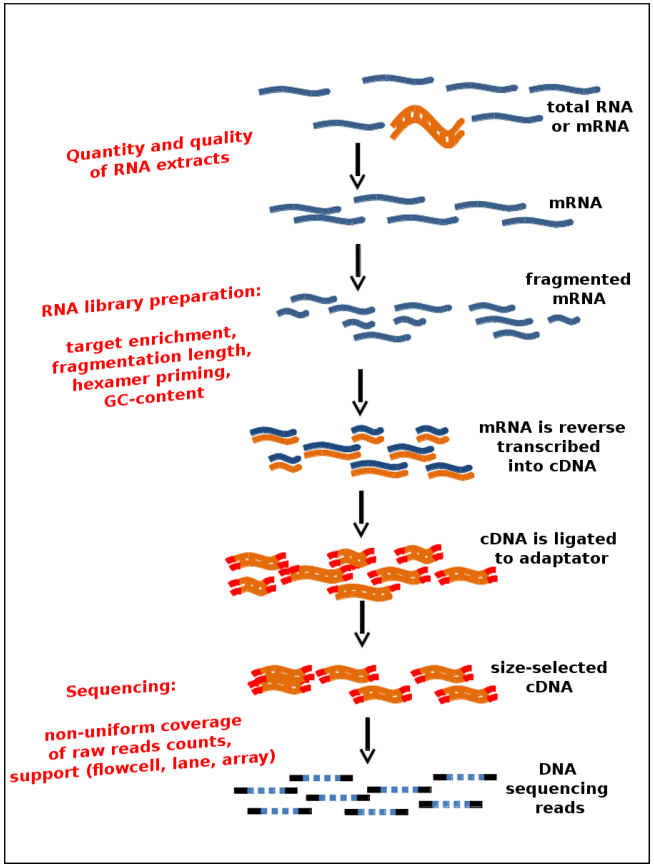
Rule 2: Factors of variation - Metadata (3)

Basic principles - Fisher (1935), George Box (1978)

Questions:

- Give the levels of replication: different operator or machines from the same technology, different variants of the protocol, different strains, different animals...
- Do you rather do five replicates on the same leaf, or one replicate each on three different leaves? How does the effect generalize to different leaves?
- Each treatment level is represented in each Block, but only once. This is a Randomized Complete Block Design. Why is this design useful? Why is each treatment level is represented only once within blocks?
- Two sequencing lanes are available and two treatments are tested. Samples fron one treatment level are assigned on the same lane. Why is this design unsuitable?

Rule 2: Factors of variation - Metadata (3)



(Source PEPI IBIS)

"Sequencing technology does not eliminate biological variability."

(Nature Biotechnology Correspondence, 2011)

Anticipate

- Identify factors of variation: controllable bias and technical specificity,
- Collect metadata from experiment and sequencing.

$$\text{lane effect} < \text{run effect} < \text{library prep effect} << \text{biological effect}$$

(Marioni, 2008), (Bullard, 2010)

Rule 3: Choose bioinformatics and statistics models (1)

- Related to technical choices

  Choice of sequencing technology, type of reads (paired-end ?), type of sequencing (directional ?), library preparation protocol

- Related to biological question
  - How many reads, which sequencing depth? which number of biological replicates ?

Why increasing the number of biological replicates?

- To generalize to the population level
- To estimate with a higher degree of accuracy variation in individual transcript (Hart, 2013)
- To improve detection of DE transcripts and control of false positive rate: TRUE with at least 3 (Sonenson 2013, Robles 2012)
- To focus on detection of low mRNAs, inconsistent detection of exons at low levels (<5 reads) of coverage (McIntyre, 2011)

Rule 3: Choose bioinformatics and statistics models (2)

More biological replicates or increasing sequencing depth?

It depends! (Haas, 2012), (Liu, 2014)

- DE transcript detection: (+) biological replicates
- Construction and annotation of transcriptome: (+) depth and (+) sampling conditions
- Transcriptomic variants search: (+) biological replicates and (+) depth

Support

- An experimental design using multiplexing,
- Tools for experimental design decisions: Scotty (Busby, 2013), RNAseqPower (Hart, 2013), PROPER (H. Wu, 2014), RNAseqPS (Guo, 2014)

Multiplexing:

Tag or bar coded with specific sequences added during library construction and that allow multiple samples to be included in the same sequencing reaction (lane).

Experimental Design: Conclusion

A good design is a list of experiments to conduct in order to answer to the asked question which maximize collected information and minimize experiments cost with respect to constraints.

- Well define the biological question, get together and collect a priori knowledge (e.g. reference genome, splicing . . .),
- Anticipate, Identify all factors of variation and adapt Fisher's principles (1935), collect metadata from experiment and sequencing,
- Include independent biological replicates to ensure reproducibility and accuracy of results

Conclusion

Exploring and interpreting results - a lot of to do

Exploration with basic plots

- the histogram of raw p-values
- the M-A plot
- an ordination plot
- a heatmap

Comparaisons of DE genes

- the venn diagramm
- the upset plot

General conclusions and perspectives

Practical conclusions

- Need to collaborate between biologists, bioinformaticians et statisticians and in a ideal world since the project construction
- Collect knowledge on the project and metadata from experiment and sequencing
- Choose and adapt the methods and tools to the asked question (no pipeline)
- Checks all the steps of the data analysis (quality, alignment, quantification, normalization, differential analysis . . .)

And after ?

- Interpretation
- Functional analysis
- Gene network

References

Normalization

- The French StatOmique Consortium (2012); Dillies, M.A.; 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.; Jaffrezic, F.;. **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.**, *Briefings in Bioinformatics* Vol. 17 Sept, 13 p; open access : doi : 10.1093./bib/bbs046.
- Robinson MD, Oshlack A. (2010) **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biology*, 11 :R25.
- Evans C., Hardin J., Stoebel D. (2016) **Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions**. *arXiv:1609.00959*
- Quinn TP., Erb I., Richardson MF., Crowley TM. (2018) **Understanding sequencing data as compositions: an outlook and review.**. *Bioinformatics*, 34(16):2870-2878. doi: 10.1093/bioinformatics/bty175.

Review

- Rory Stark R., Grzelak M., Hadfield J. (2019) **RNA sequencing: the teenage years**. *Nature Reviews Genetics*, vol. 20, pp631?656.

Differential analysis

DESeq2

- Anders, S, Huber, W. (2010) **Differential expression analysis for sequence count data**, *Genome Biology*,11:R106.
- Love, Michael and Huber, Wolfgang and Anders, Simon. (2014) **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**, *Genome Biology*.

edgeR

- Robinson MD, McCarthy DJ, Smyth, GK. (2009) **edgeR : a Bioconductor package for differential expression analysis of digital gene expression data**, *Bioinformatics*.
- McCarthy, DJ, Chen, Y, Smyth, GK (2012) **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation**, *Nucleic acids research*.

Other references for differential analysis

- Varet, H, Brillet-Guéguen, L, Coppée, J-Y and Dillies, M-A (2016) **SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data**, *Plos One*.
- Anders, S, McCarthy, DJ, Chen, Y, Okoniewski, M, Smyth GK, Huber, W and Robinson, MD (2013) **Count-based differential expression analysis of RNA sequencing data using R and Bioconductor**, *Nature Protocols*, doi:10.1038.
- Chen, Y, Lun, ATL and Smyth, GK. (2016) **From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline** *[version 2; peer review: 5 approved]. F1000Research*, `https://doi.org/10.12688/f1000research.8987.2`.

Multiple Hypothesis Testing

- Benjamini and Hochberg (1995), **Controlling the false discovery rate : a practical and powerful approach to multiple testing**, *JRSS B*, 57(1),289-300.
- Dudoit, S., Popper Shaffer, J and Boldrick, JC (2003), **Multiple Hypothesis Testing in Microarray Experiments**,*Statistical Science*, 28(1), 71-103.
- Storey and Tibshirani (2003), **Statistical significance for genome-wide studies**, *PNAS*, 100(16), 9440-9445.

Venn diagram

- Bardou, P. and Mariette, J. and Escudie, F. and Djemiel, C. and Klopp, C. (2014), **jvenn: an interactive Venn diagram viewer**. *BMC Bioinformatics*, 15:293.
- Khan A, Mathelier A. (2017), **Intervene: a tool for intersection and visualization of multiple gene or genomic region sets**. *BMC Bioinformatics*, 18:287.