Analysis of RNA-Seq data with Galaxy

Recherche des régions d'intérêt différentiellement exprimées

J. Aubert et C. Hennequet-Antier

Plateforme Migale

23-25 mai 2022
`https://migale.inra.fr/trainings`

Outline

Presentation

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

Introduction to statistical analysis of expression data with Galaxy

Objectifs

- Connaître le vocabulaire et les concepts statistiques utiles pour analyser des données type RNA-Seq
- Savoir enchaîner de façon pertinente un ensemble d'outils bioinformatiques et biostatistiques dans l'environnement Galaxy
- 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

Programme : alternance Cours / TP

- Se familiariser à l'environnement Galaxy
- Construire un plan d'expérience simple
- Explorer les données
- Identifier les transcrits différentiellement exprimés
- Se sensibiliser aux tests multiples

Migale Galaxy instance : `https://galaxy.migale.inra.fr`
RNA-seq tools

Reference

citation("SARTools")
Hugo Varet, Loraine Brillet-Guéguen, Jean-Yves Coppée and Marie-Agnès Dillies (2016): “SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data.” PLoS One, doi: `http://dx.doi.org/10.1371/journal.pone.0157022`

Details about this tool

`https://github.com/PF2-pasteur-fr/SARTools`

To share a common vocabulary

between Biology, Bioinformatics and Statistics.

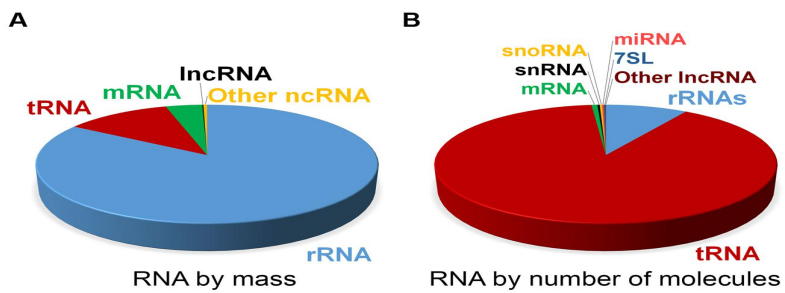
Transcriptomics

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:

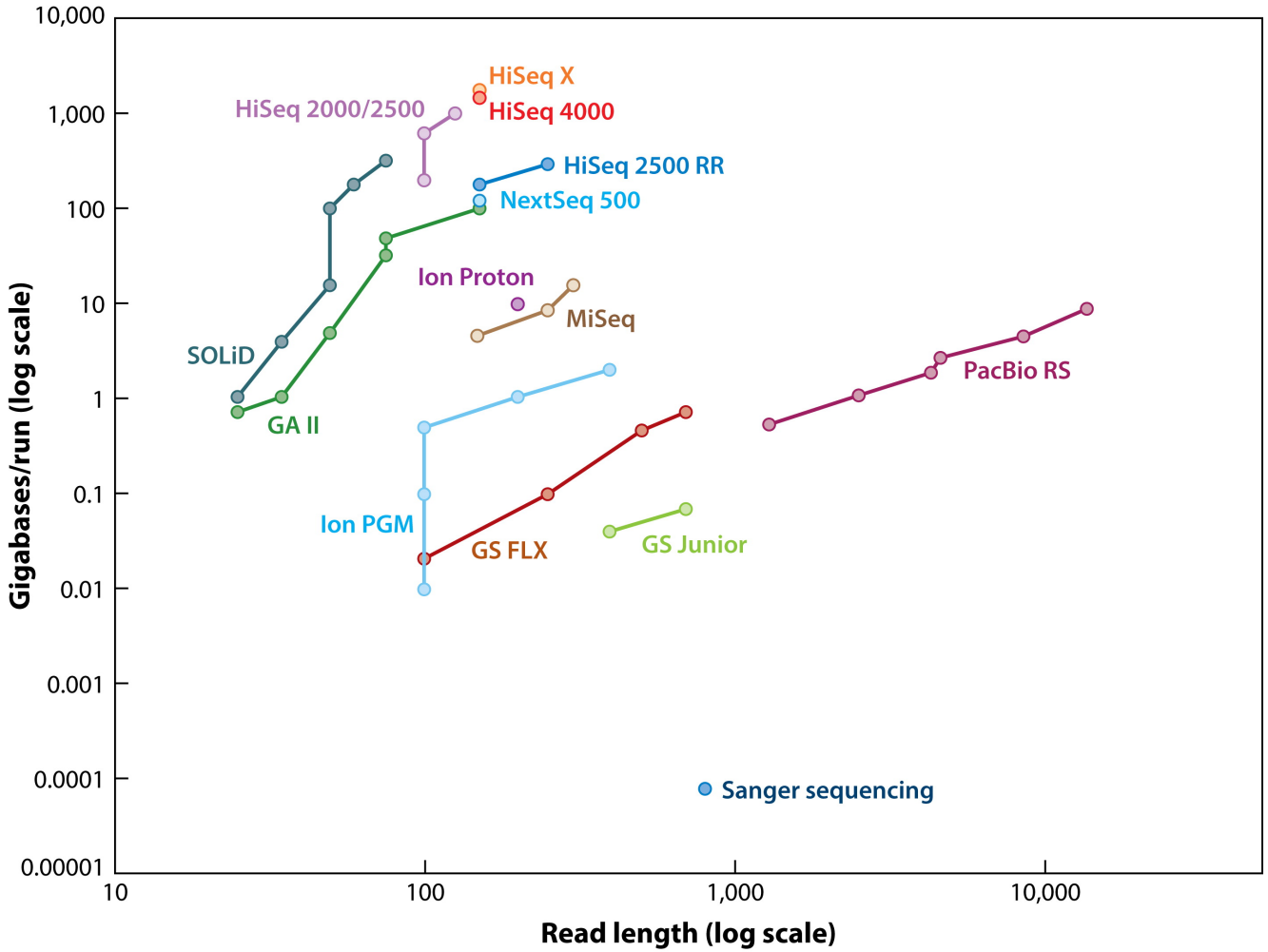
- to quantify the changing expression levels of each transcript under different biological conditions (**differential analysis**);
- to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs;
- to determine the transcriptional structure of genes: splicing patterns, post-transcriptional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

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



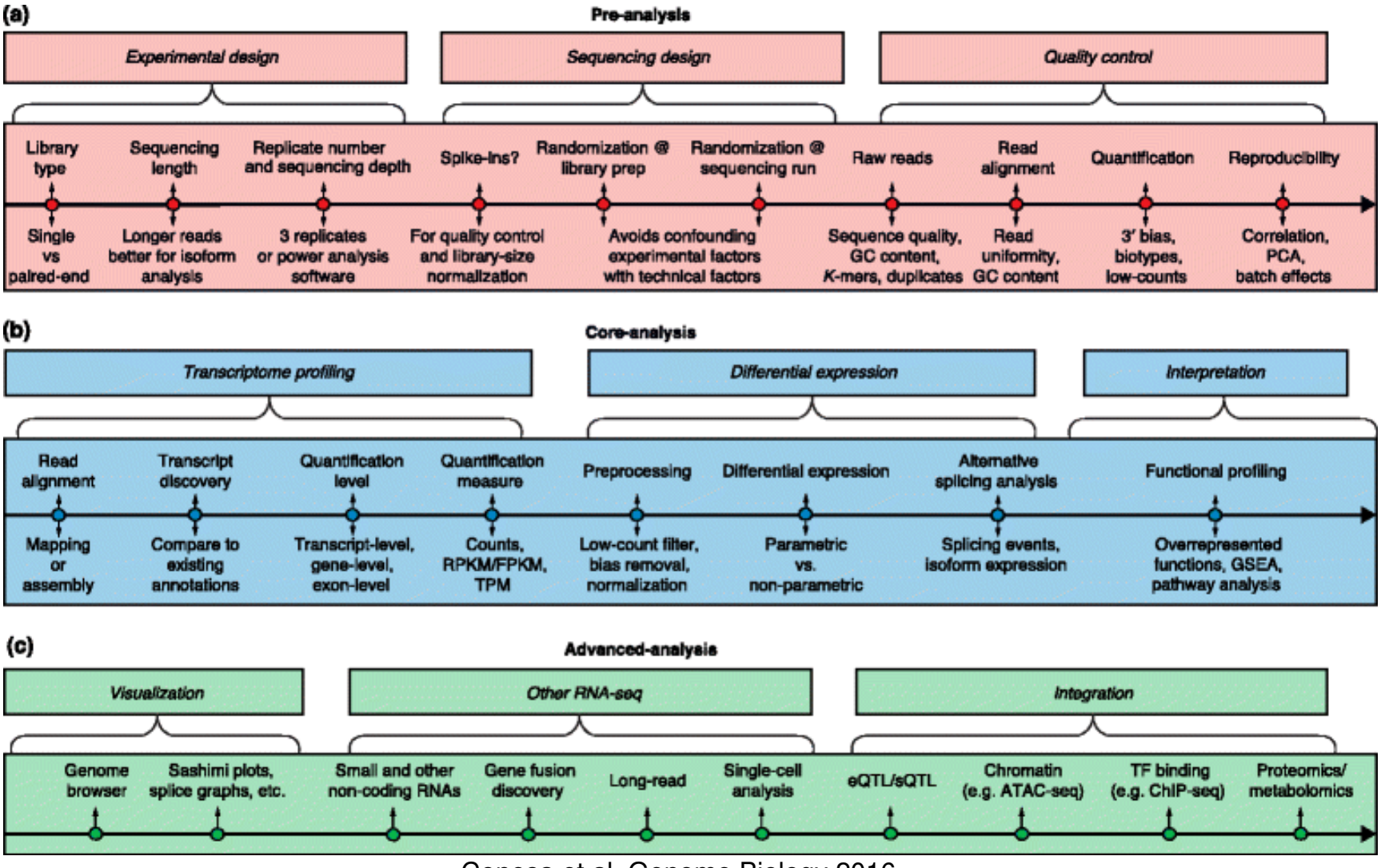
Which high-throughput sequencing technology to choose?

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



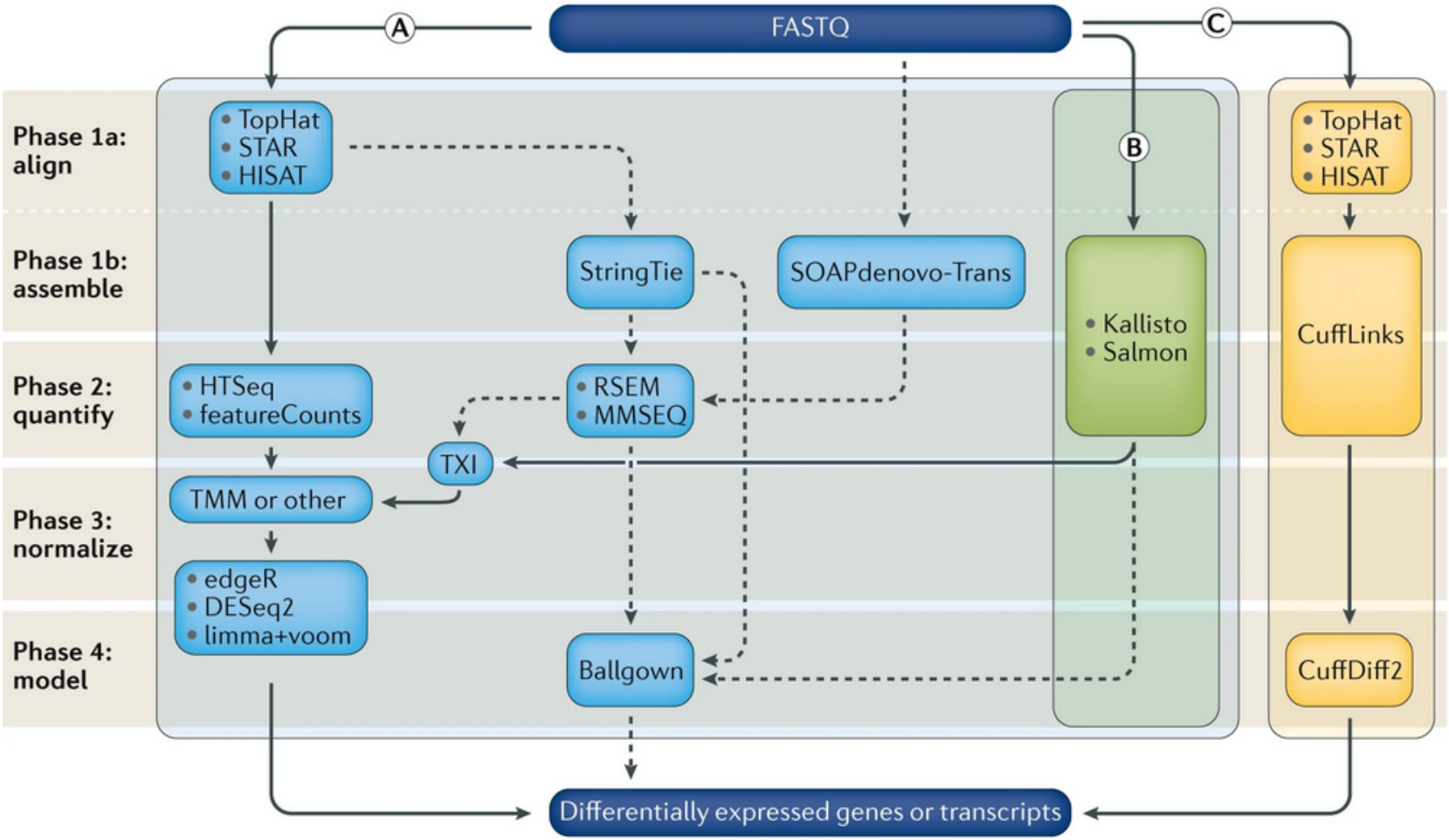
Levy SE, Myers RM. 2016.
Annu. Rev. Genom. Hum. Genet. 17:95–115

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

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.

Statistical Test

A test allows to choose given the observations between two hypotheses H_0 and H_1.

		Null hypothesis H_0 is	
		True	False
Judgement of H_0	Fail to reject	OK	β type II error (False Negative)
	Reject	α type I error (False Positive)	OK

Remark: a null hypothesis is a statement that one seeks to nullify with evidence to the contrary.

Example

H_0 = {The mean gene expression is the same in the two conditions }
H_1 = {The mean gene expression is different in the two conditions }

Test statistique

Risque de première espèce est la probabilité de rejeter H_0 alors qu'elle est vraie.

Niveau ou seuil noté α est la valeur la plus élevée du risque de première espèce.

Risque de deuxième espèce noté β est la probabilité de ne pas rejeter H_0 alors qu'elle est fausse.

Puissance du test notée 1 - β est la probabilité de rejeter H_0 alors qu'elle est fausse.

***p*-valeur** est le seuil limite auquel H_0 est rejetée compte tenu des observations (nombre compris entre 0 et 1). C'est la probabilité d'obtenir une statistique de test plus grande que la statistique observée (calculée) sous l'hypothèse nulle.

Statistical issues of gene expression analysis from RNA-Seq experiment

- A large number of genes and few replicates
- Discrete, positive and skewed data
- Large dynamic range with presence of 0 counts
- The total number of sequences 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_G	...	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 Lobel et Herskovits (2016)

- Study of CodY's regulatory repertoire in *Listeria monocytogenes*;
- 2 conditions (Wild Type and codY mutant) $\times$ 2 growth conditions (rich and minimal)

- 11 raw files, one per sequenced sample.
- each file contains the raw counts after bioinformatic steps.

SARTools in Galaxy environment

Provide Design/target file (tabular format) with one row per sample and is composed of at least three columns with headers:

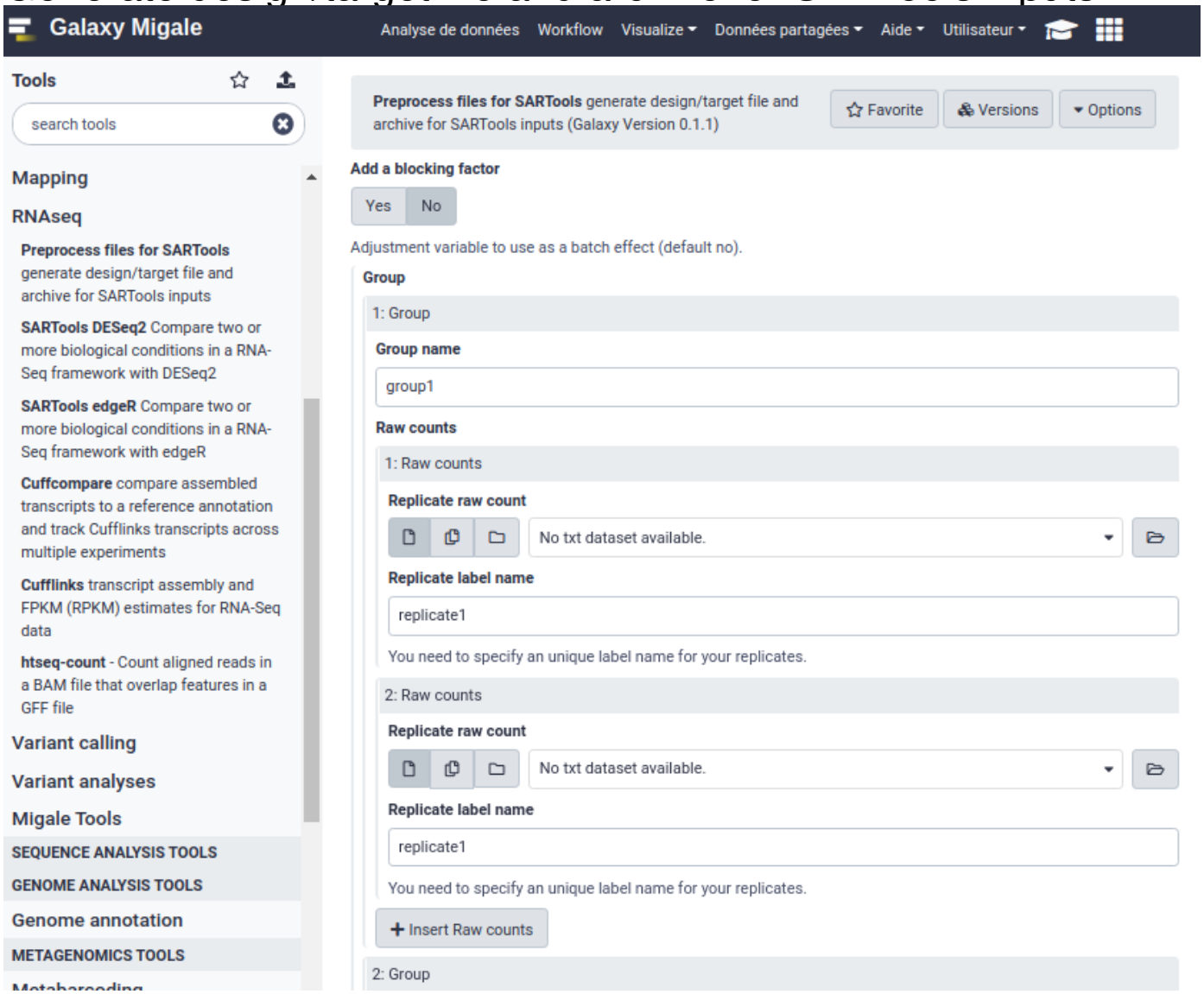
- column 1 : unique names of the samples (short but informative as they will be displayed on all the figures)
- column 2 : name of the count files;
- column 3 : biological conditions;
- optional columns : further information about the samples (day of library preparation for example).

Provide Zip file containing raw counts files:

- the unique IDs of the features in the first column;
- the raw counts associated with these features in the second column (null or positive integers).

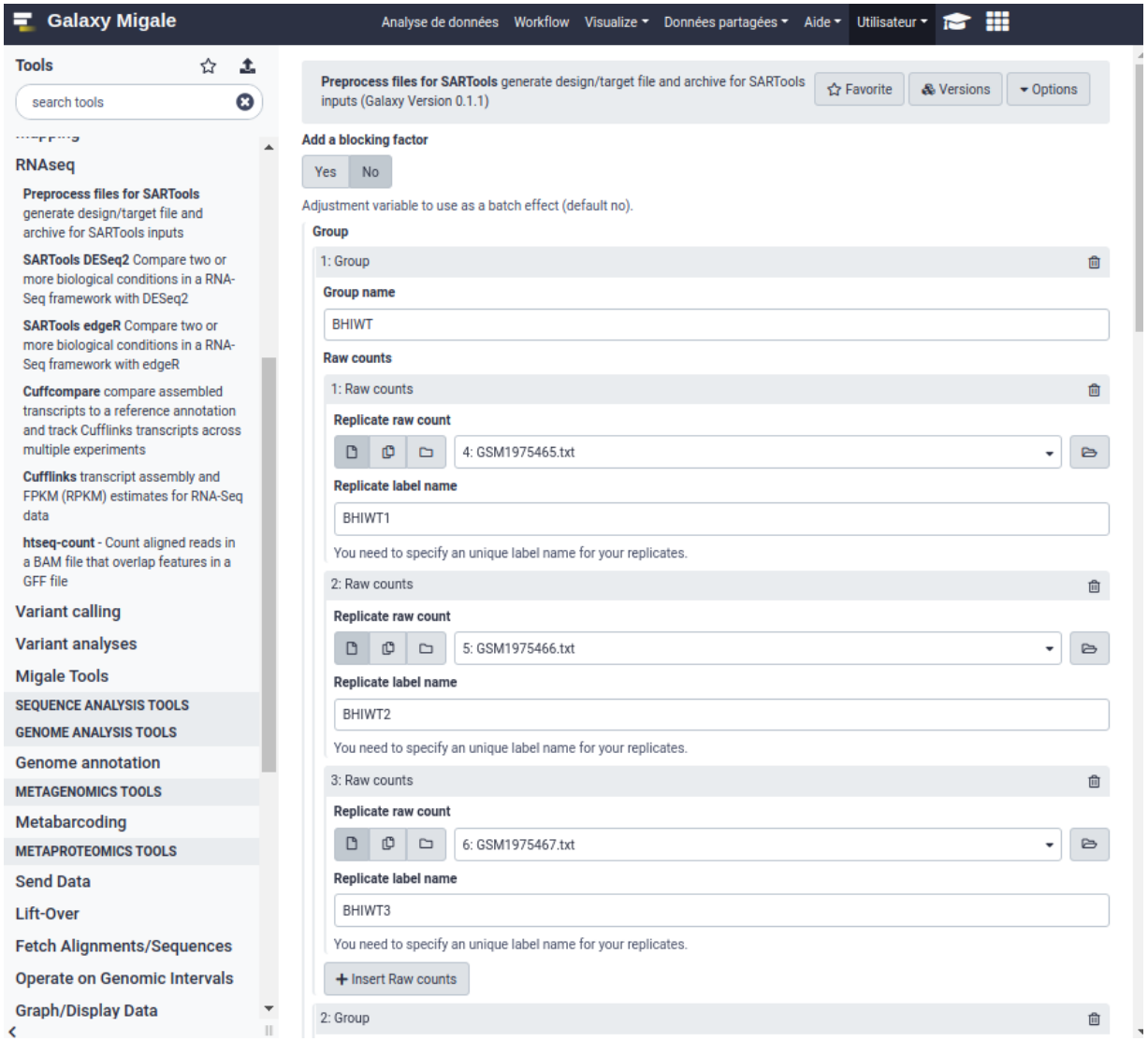
Your turn ! TP - Preprocess files for SARTools

Generate design/target file and archive for SARTools inputs.



Your turn ! TP - Preprocess files for SARTools

With data from Lobel et Herskovits (2016)



Your turn ! TP - Design/target file

Design Lobel et Herskovits (2016)

1	2	3
label	files	group
BHIWT1	dataset_286945.dat	BHIWT
BHIWT2	dataset_286946.dat	BHIWT
BHIWT3	dataset_286947.dat	BHIWT
BHIcodY1	dataset_286948.dat	BHIcodY
BHIcodY2	dataset_286949.dat	BHIcodY
BHIcodY3	dataset_286950.dat	BHIcodY
LBMMWT1	dataset_286969.dat	LBMMWT
LBMMWT2	dataset_286970.dat	LBMMWT
LBMMWT3	dataset_286971.dat	LBMMWT
LBMMcodY1	dataset_286973.dat	LBMMcodY
LBMMcodY2	dataset_286974.dat	LBMMcodY

Example Data from Lobel et Herskovits (2016)

The input dataset is a matrix $\mathbf{y} = [y_{ij}]$ or data frame (gene $\times$ sample) of counts.

- Each row i = one experimental unit (feature or gene)
- Each column j = one variable (experimental sample)

Statistical modelling : $\mathbf{y_i} = f(\mathbf{X}) + \epsilon$

- where $\mathbf{y_i}$ denotes the $(n \times 1)$ vector of expression intensities of the feature i,
- $\mathbf{X}$ denotes the $(n \times p)$ design matrix,
- and ϵ is a $(n \times 1)$ stochastic random error vector

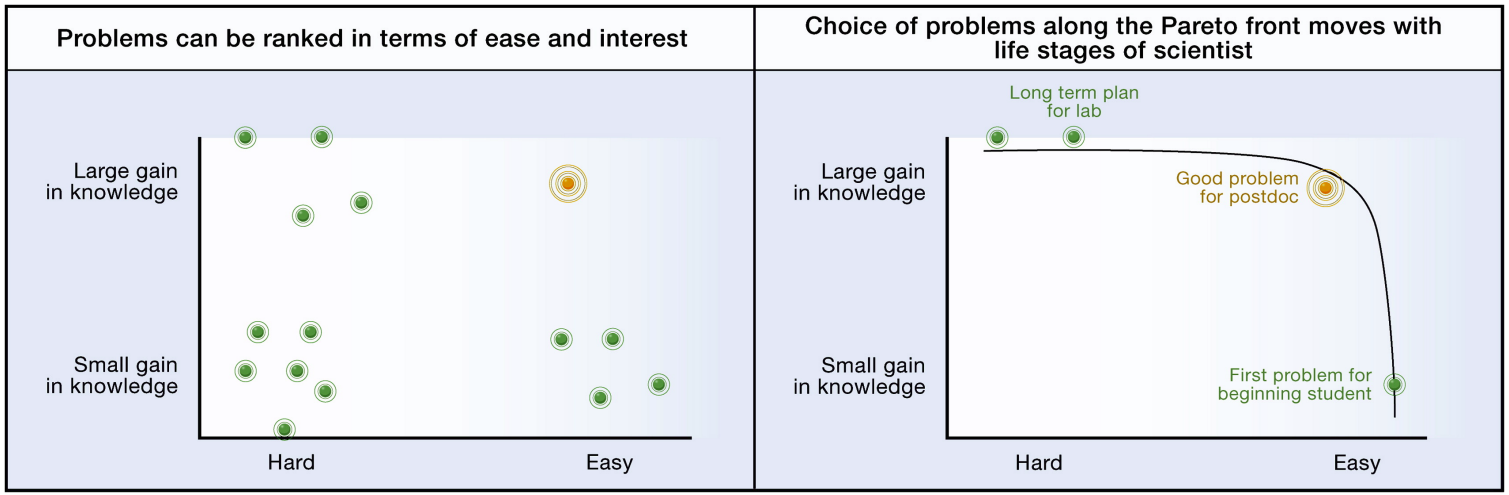
Experimental Design

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/browse/summary/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.

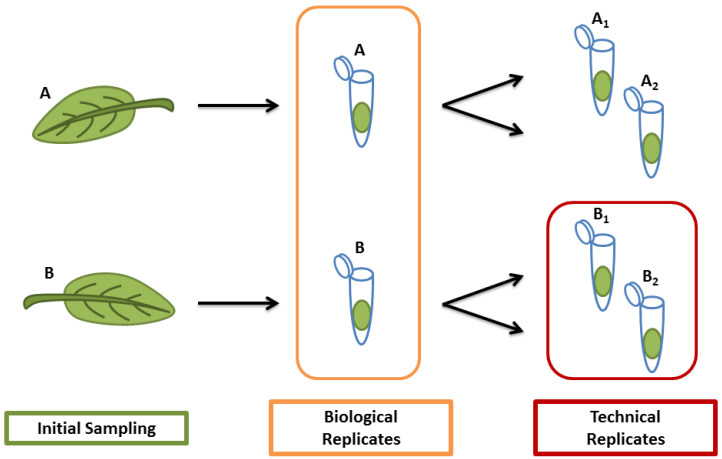
Define the biological question

- Q1: To identify differentially expressed genes between WT and codY mutant in minimal growth conditions
- Q2: To identify differentially expressed genes between WT and codY mutant in rich growth conditions

Rule 2: Factors of variation - Metadata (1)

Basic principles - Fisher (1935)

- 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).

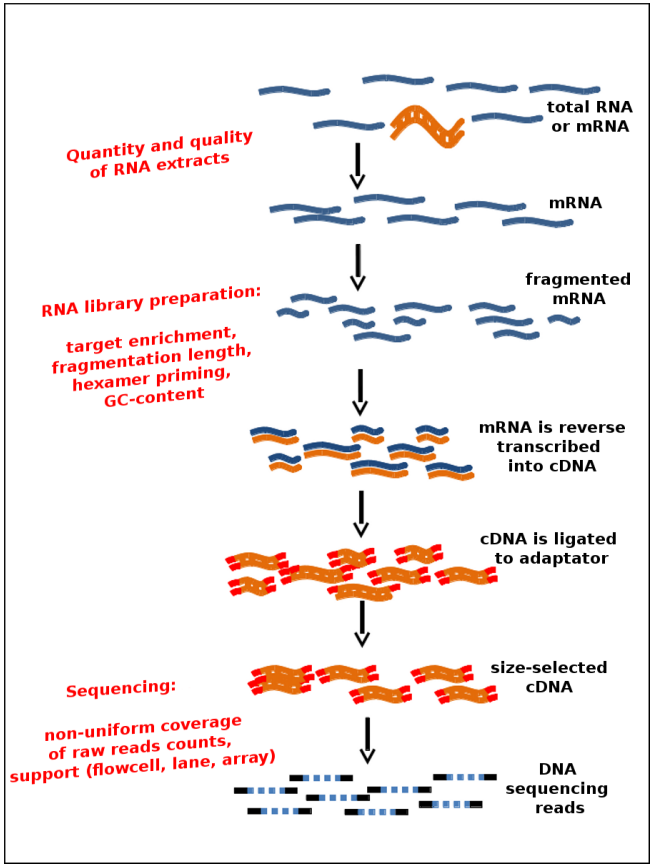
- Randomization

  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 (e.g. lane). Experimental units are grouped into homogeneous block. Random allocation within each block.

Rule 2: Factors of variation - Metadata (2)



(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.

lane effect $<$ run effect $<$ library prep effect $<<$ 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

SARTools DESeq2

Compare two or more biological conditions in a RNA-Seq framework with DESeq2.

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

SARTools edgeR

Compare two or more biological conditions in a RNA-Seq framework with edgeR.

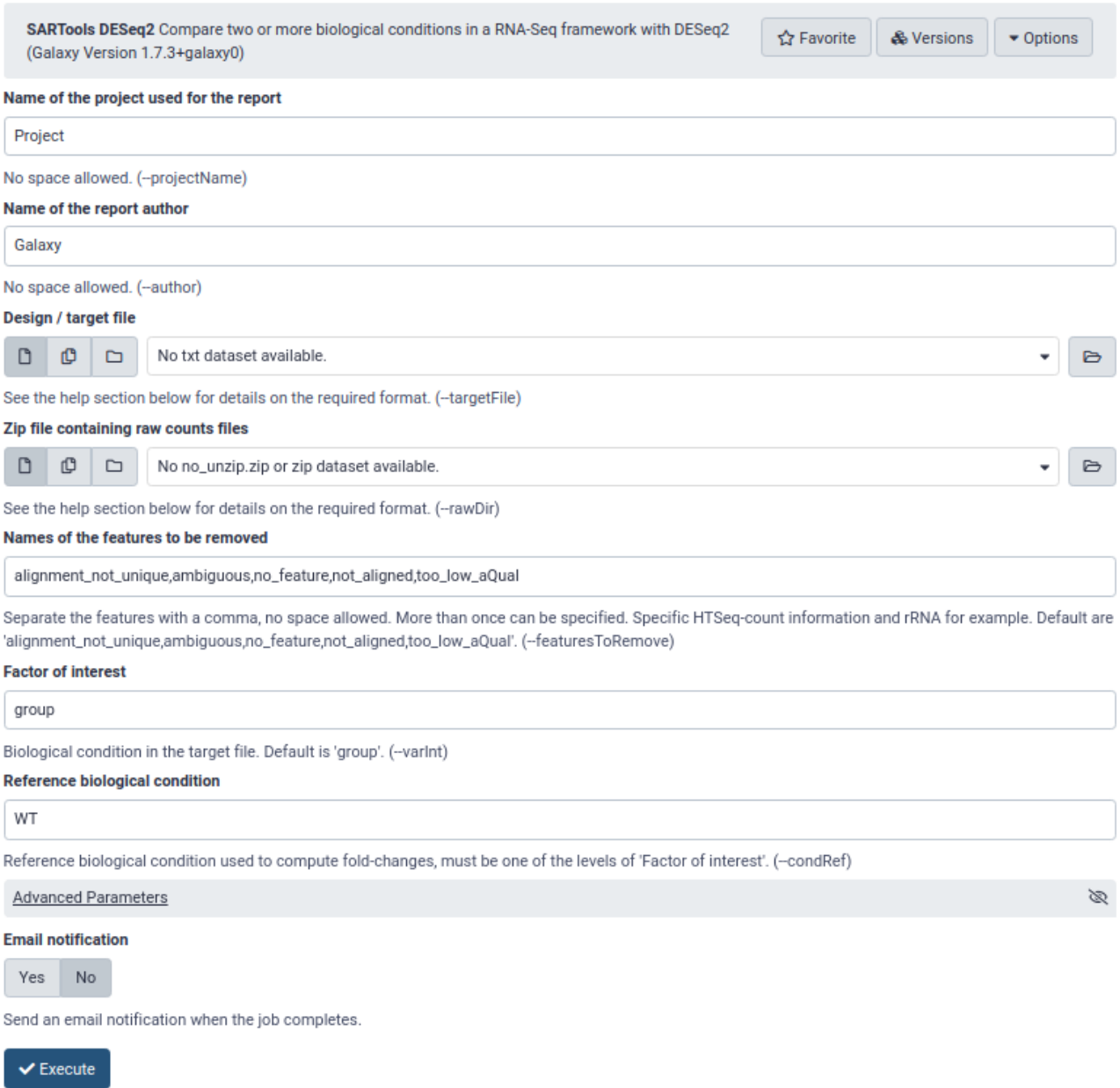
Reference

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.

Your turn ! TP - SARTools on Galaxy - DESeq2



Your turn ! TP - SARTools on Galaxy - Fill it for DESeq2

SARTools DESeq2 Compare two or more biological conditions in a RNA-Seq framework with DESeq2 (Galaxy Version 1.7.3+galaxy0)

☆ Favorite | Versions | ▾ Options

Name of the project used for the report

Formation_Lobel_DESeq2

No space allowed. (--projectName)

Name of the report author

CHA

No space allowed. (--author)

Design / target file

17: design file for SARTools (on data 16, data 15, and others)

See the help section below for details on the required format. (--targetFile)

Zip file containing raw counts files

18: counts files for SARTools (on data 16, data 15, and others)

See the help section below for details on the required format. (--rawDir)

Names of the features to be removed

alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual

Separate the features with a comma, no space allowed. More than once can be specified. Specific HTSeq-count information and rRNA for example. Default are 'alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual'. (--featuresToRemove)

Factor of interest

group

Biological condition in the target file. Default is 'group'. (--varInt)

Reference biological condition

BHIWT

Reference biological condition used to compute fold-changes, must be one of the levels of 'Factor of interest'. (--condRef)

Advanced Parameters

Email notification

Yes | No

Send an email notification when the job completes.

✓ Execute

Your turn ! TP - SARTools on Galaxy - DESeq2 parameters

Parameters

- **projectName:** name of the project;
- **author:** author of the analysis;
- **featuresToRemove:** character vector containing the IDs of the features to remove before running the analysis (default are "alignment not unique", "ambiguous", "no feature", "not aligned", "too low aQual" to remove HTSeq-count specific rows);
- **varInt:** variable of interest, i.e. biological condition, in the target file ("group" by default);
- **condRef:** reference biological condition used to compute fold-changes (no default, must be one of the levels of varInt);
- **batch:** adjustment variable to use as a batch effect, must be a column of the target file (NULL if no batch effect needs to be taken into account);
- **alpha:** significance threshold applied to the adjusted p-values to select the differentially expressed features (default is 0.05);
- **fitType:** type of model for the mean-dispersion relationship ("parametric" by default, or "local");
- **cooksCutoff:** TRUE (default) of FALSE to execute or not the detection of the outliers [4];
- **independentFiltering:** TRUE (default) of FALSE to execute or not the independent filtering [5];
- **pAdjustMethod:** p-value adjustment method for multiple testing [6, 7] ("BH" by default, "BY" or any value of p.adjust.methods);
- **typeTrans:** method of transformation of the counts for the clustering and the PCA (default is "VST" for Variance Stabilizing Transformation, or "rlog" for Regularized Log Transformation);
- **locfunc:** function used for the estimation of the size factors (default is "median", or "shorth" from the genefilter package);
- **colors:** colors used for the figures (one per biological condition), 8 are given by default.
- **forceCairoGraph:** TRUE or FALSE (default) to force the use of cairo with options(bitmapType="cairo").

Your turn ! TP - SARTools on Galaxy - DESeq2 output files

Output files

Report:

Give details about the methodology, the different steps and the results. It displays all the figures produced and the most important results of the differential analysis as the number of up- and down-regulated features.
The user should read the full HTML report and closely analyze each figure to check that the analysis ran smoothly.

Tables:

- **TestVsRef.complete.txt:** contains all the features studied;
- **TestVsRef.down.txt:** contains only significant down-regulated features, i.e. less expressed in Test than in Ref;
- **TestVsRef.up.txt:** contains only significant up-regulated features i.e. more expressed in Test than in Ref.

Figures:

- **MAplot.png:** MA-plot for each comparison (log ratio of the means vs intensity).
- **PCA.png:** first and second factorial planes of the PCA on the samples based on VST or rlog data;
- **barplotNull.png:** percentage of null counts per sample;
- **barplotTC.png:** total number of reads per sample;
- **cluster.png:** hierachical clustering of the samples (based on VST or rlog data);
- **countsBoxplot.png:** boxplots on raw and normalized counts;
- **densplot.png:** estimation of the density of the counts for each sample;
- **diagSizeFactorsHist.png:** diagnostic of the estimation of the size factors;
- **diagSizeFactorsTC.png:** plot of the size factors vs the total number of reads;
- **dispersionsPlot.png:** graph of the estimations of the dispersions and diagnostic of log-linearity of the dispersions;
- **majSeq.png:** percentage of reads caught by the feature having the highest count in each sample;
- **pairwiseScatter.png:** pairwise scatter plot between each pair of samples and SERE values;
- **rawpHist.png:** histogram of the raw p-values for each comparison;
- **volcanoPlot.png:** vulcano plot for each comparison (- log10 (adjusted P value) vs log ratio of the means).

R log file:

Give the R console outputs.

R objects (.RData file):

Give all the R objects created during the analysis is saved: it may be used to perform downstream analyses.

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.

The authors recommend to use the rlog transformation.
In DESeq2, the PCA is performed on the top genes selected by highest row variance (*ntop* argument) of the PCAplot function

Your turn ! TP - SARTools on Galaxy - DESeq2 - PCA

Visualize PCA from Lobel et Herskovits (2016)

Your turn ! TP - SARTools on Galaxy - edgeR

SARTools edgeR Compare two or more biological conditions in a RNA-Seq framework with edgeR (Galaxy Version 1.7.3+galaxy0)

☆ Favorite | Versions | ▾ Options

Name of the project used for the report

Project

No space allowed. (--projectName)

Name of the report author

Galaxy

No space allowed. (--author)

Design / target file

No txt dataset available.

See the help section below for details on the required format. (--targetFile)

Zip file containing raw counts files

No no_unzip.zip or zip dataset available.

See the help section below for details on the required format. (--rawDir)

Names of the features to be removed

alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual

Separate the features with a comma, no space allowed. More than once can be specified. Specific HTSeq-count information and rRNA for example. Default are 'alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual'. (--featuresToRemove)

Factor of interest

group

Biological condition in the target file. Default is 'group'. (--varInt)

Reference biological condition

WT

Reference biological condition used to compute fold-changes, must be one of the levels of 'Factor of interest'. (--condRef)

Advanced Parameters

Email notification

Yes | No

Send an email notification when the job completes.

✔ Execute

Your turn ! TP - SARTools on Galaxy - Fill it for edgeR

SARTools edgeR Compare two or more biological conditions in a RNA-Seq framework with edgeR (Galaxy Version 1.7.3+galaxy0)

Favorite | Versions | Options

Name of the project used for the report

Formation_Loble_edgeR

No space allowed. (--projectName)

Name of the report author

CHA

No space allowed. (--author)

Design / target file

17: design file for SARTools (on data 16, data 15, and others)

See the help section below for details on the required format. (--targetFile)

Zip file containing raw counts files

18: counts files for SARTools (on data 16, data 15, and others)

See the help section below for details on the required format. (--rawDir)

Names of the features to be removed

alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual

Separate the features with a comma, no space allowed. More than once can be specified. Specific HTSeq-count information and rRNA for example. Default are 'alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual'. (--featuresToRemove)

Factor of interest

group

Biological condition in the target file. Default is 'group'. (--varInt)

Reference biological condition

BHIWT

Reference biological condition used to compute fold-changes, must be one of the levels of 'Factor of interest'. (--condRef)

Advanced Parameters

Email notification

Yes | No

Send an email notification when the job completes.

Execute

INRAE

Your turn ! TP - SARTools on Galaxy - edgeR parameters

Parameters

- **projectName:** name of the project;
- **author:** author of the analysis;
- **featuresToRemove:** character vector containing the IDs of the features to remove before running the analysis (default are "alignment not unique", "ambiguous", "no feature", "not aligned", "too low aQual" to remove HTSeq-count specific rows);
- **varInt:** variable of interest, i.e. biological condition, in the target file ("group" by default);
- **condRef:** reference biological condition used to compute fold-changes (no default, must be one of the levels of varInt);
- **batch:** adjustment variable to use as a batch effect, must be a column of the target file (NULL if no batch effect needs to be taken into account);
- **alpha:** significance threshold applied to the adjusted p-values to select the differentially expressed features (default is 0.05);
- **pAdjustMethod:** p-value adjustment method for multiple testing [4, 5] ("BH" by default, "BY" or any value of p.adjust.methods);
- **cpmCutoff:** counts-per-million cut-off to filter low counts (default is 1, set to 0 to disable filtering);
- **gene.selection:** method of selection of the features for the MultiDimensional Scaling plot ("pairwise" by default or common);
- **normalizationMethod:** normalization method in calcNormFactors(): "TMM" (default), "RLE" (DESeq method) or "upperquartile";
- **colors:** colors used for the figures (one per biological condition), 8 are given by default.
- **forceCairoGraph:** TRUE or FALSE (default) to force the use of cairo with options(bitmapType="cairo").

Your turn ! TP - SARTools on Galaxy - edgeR output files

Output files

Report:

Give details about the methodology, the different steps and the results. It displays all the figures produced and the most important results of the differential analysis as the number of up- and down-regulated features.
The user should read the full HTML report and closely analyze each figure to check that the analysis ran smoothly.

Tables:

- **TestVsRef.complete.txt:** contains all the features studied;
- **TestVsRef.down.txt:** contains only significant down-regulated features, i.e. less expressed in Test than in Ref;
- **TestVsRef.up.txt:** contains only significant up-regulated features i.e. more expressed in Test than in Ref.

Figures:

- **MAplot.png:** MA-plot for each comparison (log ratio of the means vs intensity).
- **PCA.png:** first and second factorial planes of the PCA on the samples based on VST or rlog data;
- **barplotNull.png:** percentage of null counts per sample;
- **barplotTC.png:** total number of reads per sample;
- **cluster.png:** hierachical clustering of the samples (based on VST or rlog data);
- **countsBoxplot.png:** boxplots on raw and normalized counts;
- **densplot.png:** estimation of the density of the counts for each sample;
- **diagSizeFactorsHist.png:** diagnostic of the estimation of the size factors;
- **diagSizeFactorsTC.png:** plot of the size factors vs the total number of reads;
- **dispersionsPlot.png:** graph of the estimations of the dispersions and diagnostic of log-linearity of the dispersions;
- **majSeq.png:** percentage of reads caught by the feature having the highest count in each sample;
- **pairwiseScatter.png:** pairwise scatter plot between each pair of samples and SERE values;
- **rawpHist.png:** histogram of the raw p-values for each comparison;
- **volcanoPlot.png:** vulcano plot for each comparison (– log10 (adjusted P value) vs log ratio of the means).

R log file:

Give the R console outputs.

R objects (.RData file):

Give all the R objects created during the analysis is saved: it may be used to perform downstream analyses.

MDSPlot

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

Transform count data as moderated log-counts-per-million before performing MDSPlot.

Counts-per-million

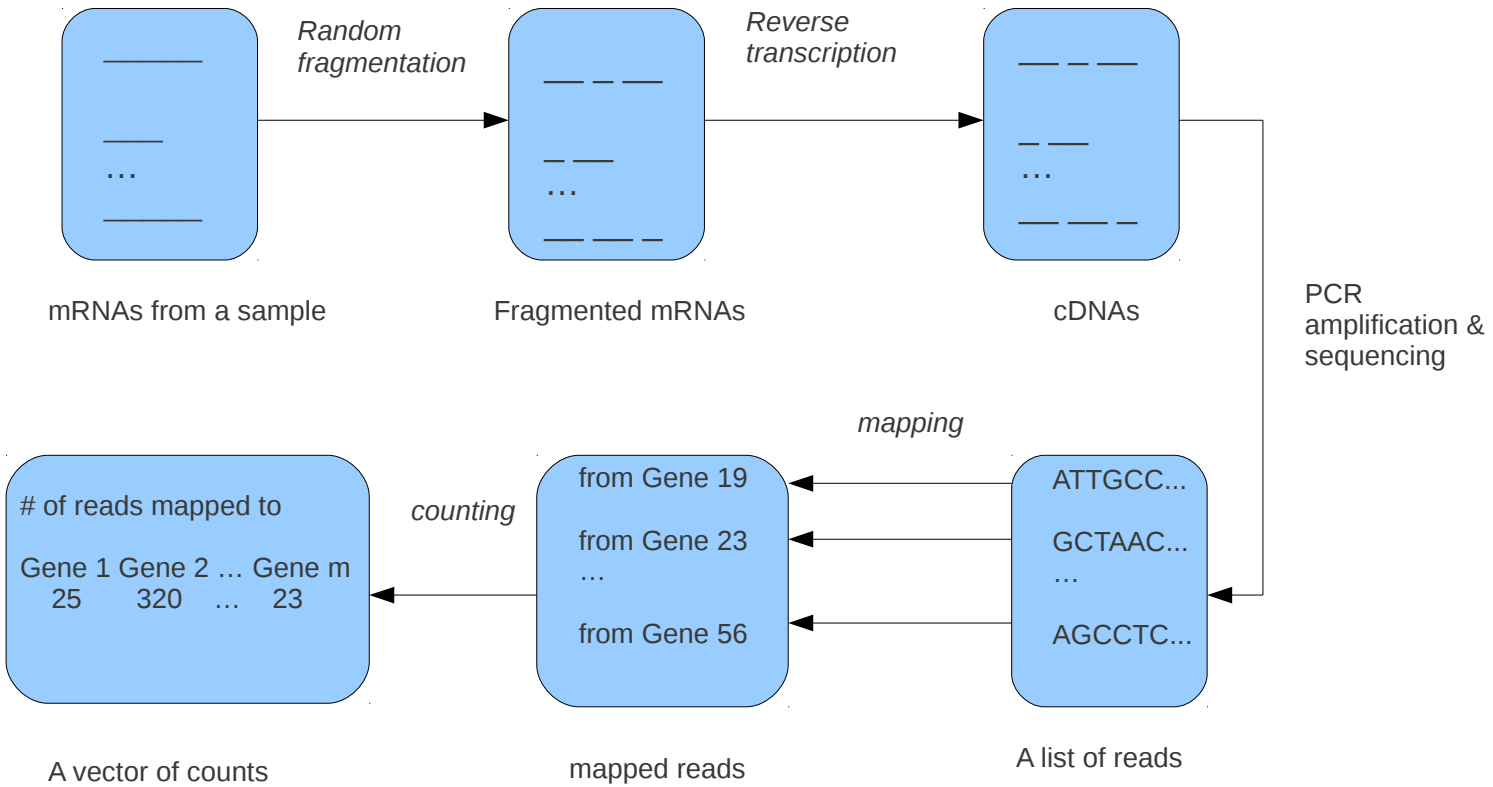
$$\text{CPM} = \frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$$

Your turn ! TP - SARTools on Galaxy - edgeR - MDSplot

Visualize MDSplot from Lobel et Herskovits (2016)

Introduction to statistical analysis of expression data with R

RNA-sequencing



Adapted from Li et al. (2011)

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_{ij}	...	y_{in}
...	...	...	...	...	...	...
Gene G	59	164	...	y_G	...	143
Seq. depth	6865057	11127087	...	$N_j = \sum_{i=1}^{G} y_{ij}$	...	11320226

y_{ij} = number of sequences from sample j assigned to gene i.

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

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

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

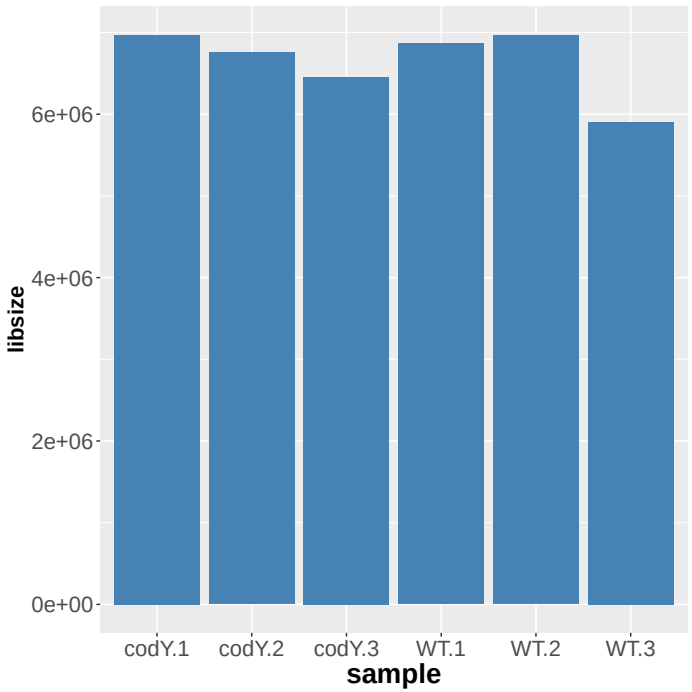
- Depth (total number of sequenced and mapped reads)
- RNA-composition or presence of majority fragments
- Sequence composition du to PCR-amplification step in library preparation (Pickrell et al. 2010, Risso et al. 2011)

Normalization and differential expression (DE) analysis

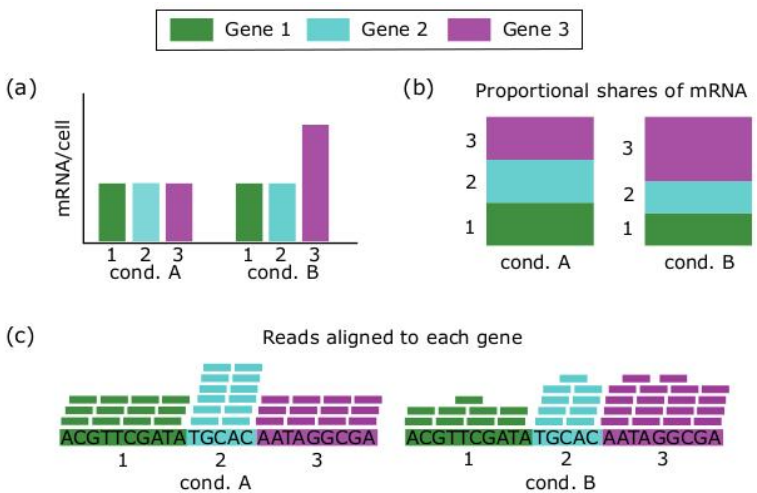
DE analysis concerned with **relative changes** in expression levels between conditions rather than estimating absolute expression levels.

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).

Understanding sequencing data

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

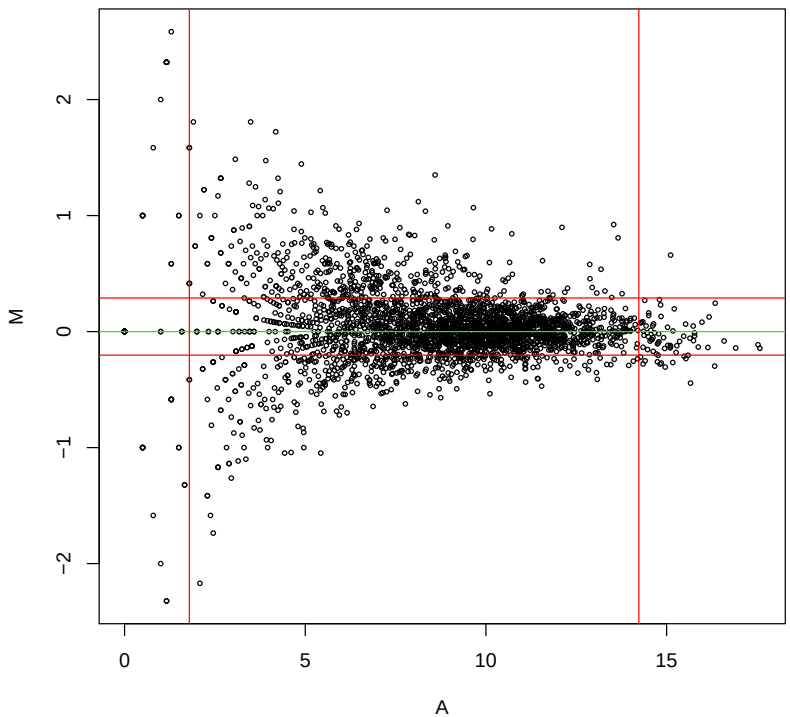
Aim

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_i distribution.



Filter on:

- transcripts with nul counts,
- the 30% more extreme $M_{ij}^r = log_2(\frac{y_{ij}/N_j}{y_{ir}/N_r})$ values,
- the 5% more extreme $A_{ij}^r = 0.5 \times [log_2(\frac{y_{ij}}{N_j}) + log_2(\frac{y_{ir}}{N_r})]$ 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_{ij}^r or the A_{ij}^r value was trimmed
3. Calculate the scaling factors $TMM_j^{(r)}$ such as

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

with $w_{ij}^r = \frac{N_j - y_{ij}}{N_j y_{ij}} - \frac{N_r - y_{ir}}{N_r y_{ir}}$

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)

$$y_{ij}^{r} = \left(\pi_{j=1}^{n} y_{ij}^{1/n}\right)$$

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

2. Calculate scaling factors

$$\hat{s}_j = \underset{i:y_{ij}^{r}\neq 0}{median} \frac{y_{ij}}{y_{ij}^{r}}$$

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.

	WT.1	WT.2	WT.3	codY.1	codY.2	codY.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.

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 ! TP - Exploration and normalisation

- Estimation of size factors
- Data normalisation
- Boxplot of raw and normalized data
- MA-plot of raw and normalized data

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

$$FC_i = \frac{x_i.}{y_i.}$$

	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)

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 i

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

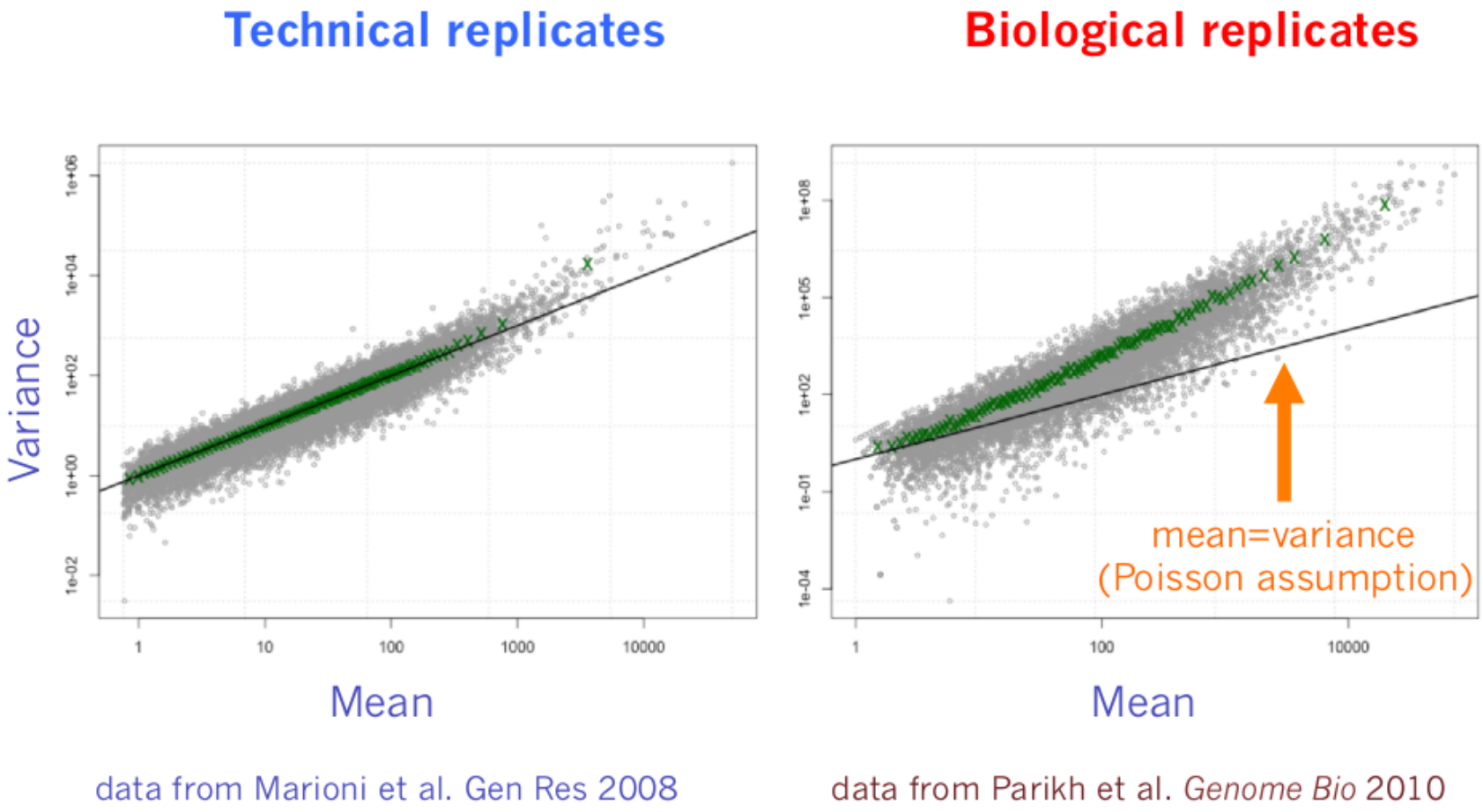
- Statistical model (definition and parameter estimation) - Generalized linear framework Y_{ijk} follows $\mathbf{f}(\theta_{ijk})$
- Hypothesis to test : H_{0i} Equality of relative abundance of gene i in condition A and B vs H_{1i} non-equality
- Critical region - Wald Test or Likelihood Ratio Test

The Poisson Model

Let be Y_{ij} the read count for gene i in sample j

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

Mean-Variance Relationship



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...

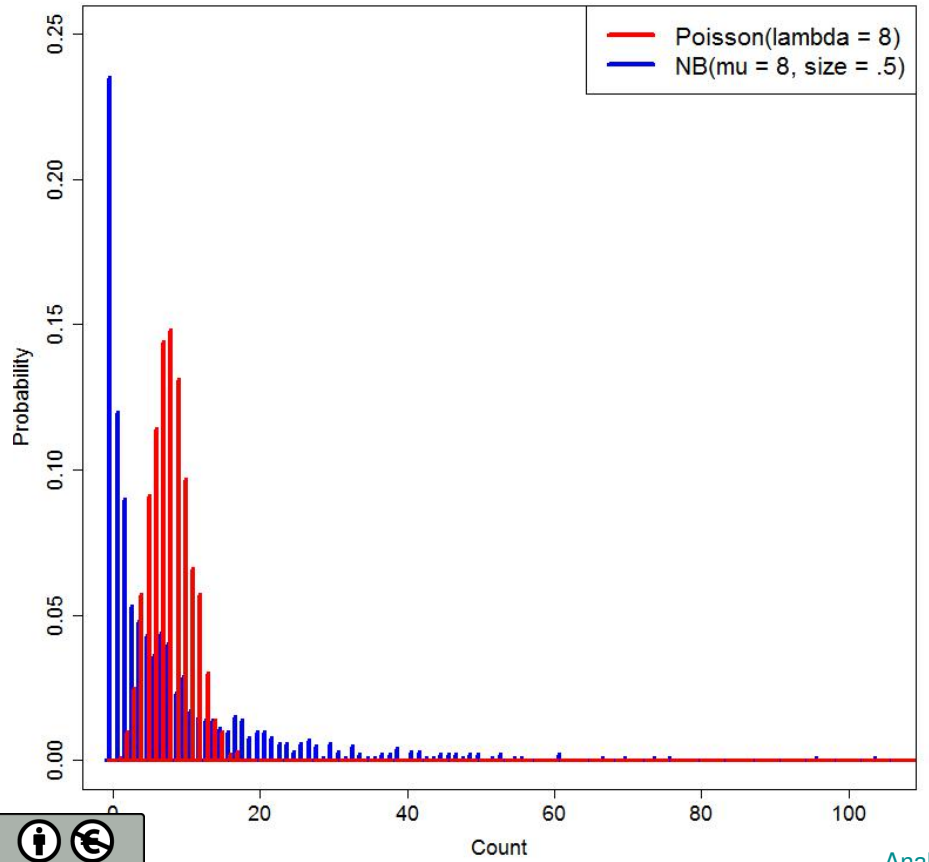
In case of overdispersion, ↑ of the type I error rate (prob. to declare incorrectly a gene DE).

Alternative : Negative Binomial Models

A supplementary dispersion parameter ϕ to model the variance

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

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)

Coefficient of Variation

Normalized measure of dispersion, ratio of the standard deviation to the mean

In the negative binomial model,

$$\begin{aligned} CV^2 &= CV^2_{technique} + CV^2_{biologique} \\ &= \frac{1}{\mu_{ij}} + \phi_i \end{aligned}$$

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

$Y_{ij} \sim$ NB(mean = μ_{ij}, dispersion = ϕ_i)

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

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

Main steps performed by the DESeq function:

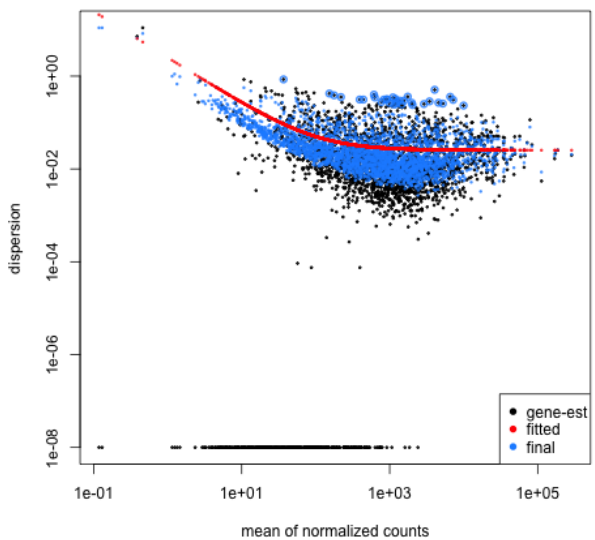
1. estimation of size factors $s_{ij} = s_j$ by estimateSizeFactors
2. estimation of dispersion by estimateDispersions
3. negative binomial GLM fitting for β_i and Wald statistics by nbinomWaldTest

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.

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

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

DESeq2 robustness policy

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

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 ! TP - Differential analysis

DE genes between WT and CodY mutant in rich growth conditions

- differential analysis with DESeq2 and edgeR
- MA-plot
- Volcano-plot

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.

Introduction to statistical analysis of expression data with R

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 *i*
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 $pBonf_i = min(1, p_i * G)$ 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

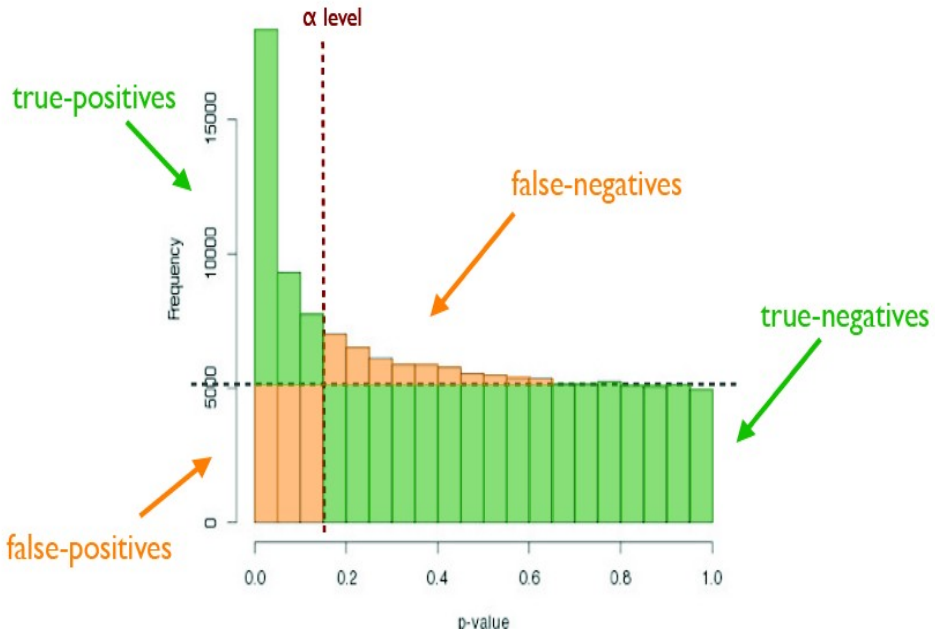
The false discovery rate of Benjamini and Hochberg (1995) is the expected proportion of Type I errors among the rejected hypotheses

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

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 i $p_{(i)} \leq i\alpha^*/G$.

Prop

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

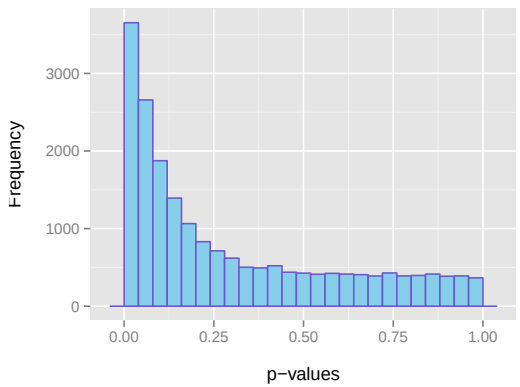
Prop

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

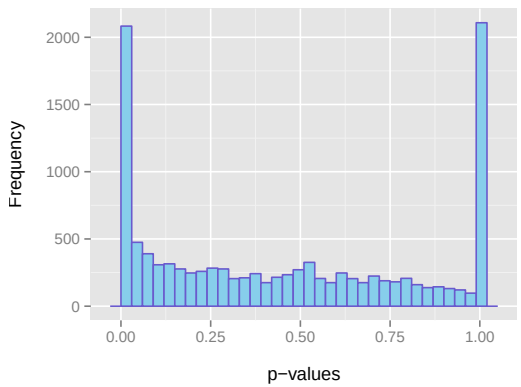
p-values histograms for diagnosis

Examples of expected overall distribution

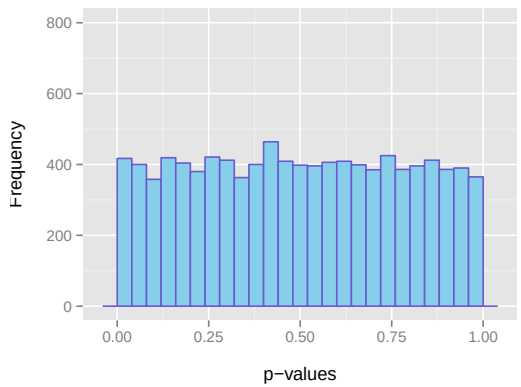
(a)



(b)



(c)



(a): the most desirable shape

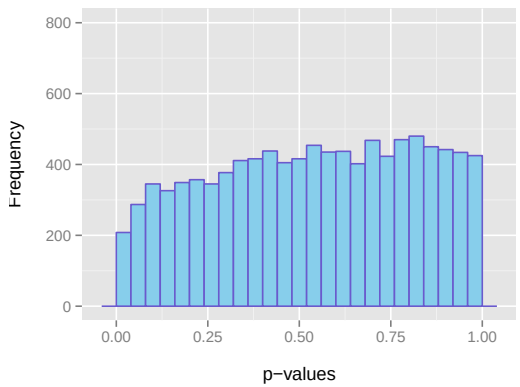
(b): very low counts genes usually have large p-values

(c): do not expect positive tests after correction

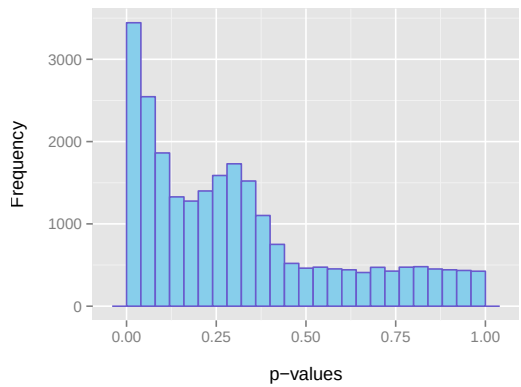
p-values histograms for diagnosis

Examples of not expected overall distribution

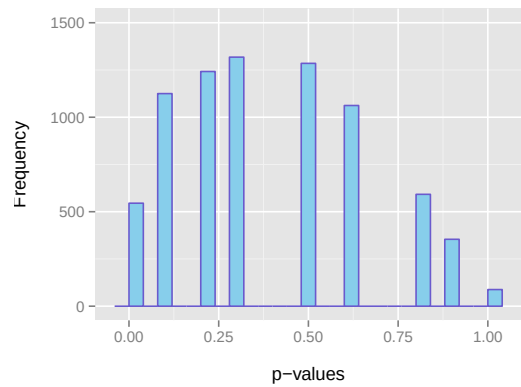
(a)



(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

Your turn ! TP - Multiple tests

Calculate adjusted pvalues with the Bonferroni and BH procedures for the difference CodY vs WT in minimal growth conditions

padjust

- Histogram of raw Pvalues
- How many DE genes with $\alpha = 0.01$ for each procedure ?

Your turn ! TP - Venn diagramm

Comparaison between minimal and rich growth conditions

- Venn diagramm
  the venn function
  `http://bioinfo.genotoul.fr/jvenn/` (Bardou et al. 2014)
- How to export results ?

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 confidendence 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.

Conclusion

General conclusions and perspectives

Pratical 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

More ressources for Galaxy trainings

RNA-seq counts to genes

https://www.usegalaxy.fr/training-material/topics/transcriptomics/tutorials/rna-seq-counts-to-genes/tutorial.html

Visualization of RNA-Seq results with heatmap2

https://www.usegalaxy.fr/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-heatmap2/tutorial.html

Visualization of RNA-Seq results with Volcano Plot

https://www.usegalaxy.fr/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-volcanoplot/tutorial.html

DESeq2

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

Other references for differential analysis

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*

Differential analysis

- 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*.
- 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*.

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.