Analysis of RNASeq data

Study of differential gene expression

Alexis Dereeper

8th - 9th of February, 2024

UNAL, Bogota, Colombia

Some definitions

- Sequencing: Determine the linear succession of DNA bases A,T,C,G, reading of this sequence allow to study the included biological information
- Next Generation Sequencing (NGS): High throughput sequencing, generation of a high number of sequences simultaneously
- RNA-seq: transcriptome sequencing. Informations about RNAs using the sequencing of complementary DNA (cDNA)
- Re-sequencing: sequencing of a genome that could be compared to a known reference sequence (the genome of the species has been sequenced already)
- *de-novo* sequencing: sequencing of a genome for which there is no reference genome, determination of a unknown sequence

Why using RNA-seq?

Access to sequences of RNA allows to:

- Annotate a genome
- Establish the catalog of expressed genes
- Identify new genes
- Identify alternative transcripts
- Quantify gene expression and compare between different experimental conditions
- Identify small RNAs (regulation of expression, silencing…)

...

Advantages RNA-seq / microarray

- More accurate and sensitive: allows to discover more
- RNA-seq allows detection of alternative splicing
- Possibility to study transcripts that are lowly expressed
- No need reference genome
  (for microarray, it is required to design probes)

Objectives of the trainings

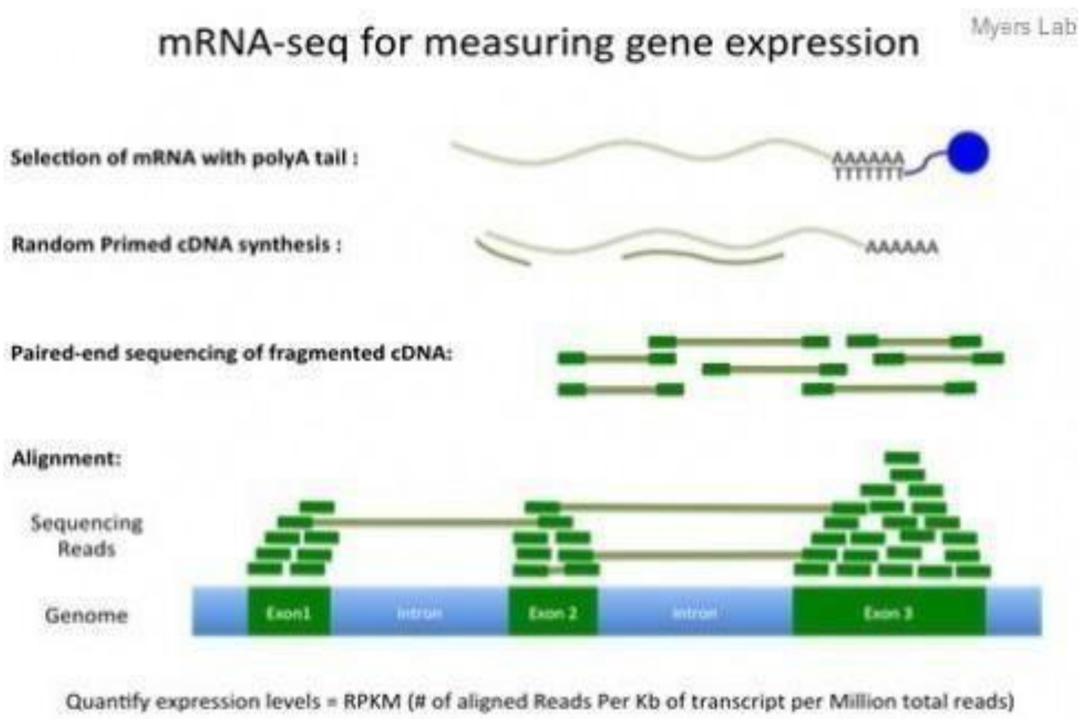
- Know and manipulate packages/tools available for the identification of differentially expressed genes

- Think about different techniques of normalization of data

- Detect genes that are differentially expressed between 2 conditions

- Compare results obtained with two different approaches/tools. Understand differences

General principle based on read counting



Myers Lab

mRNA-seq for measuring gene expression

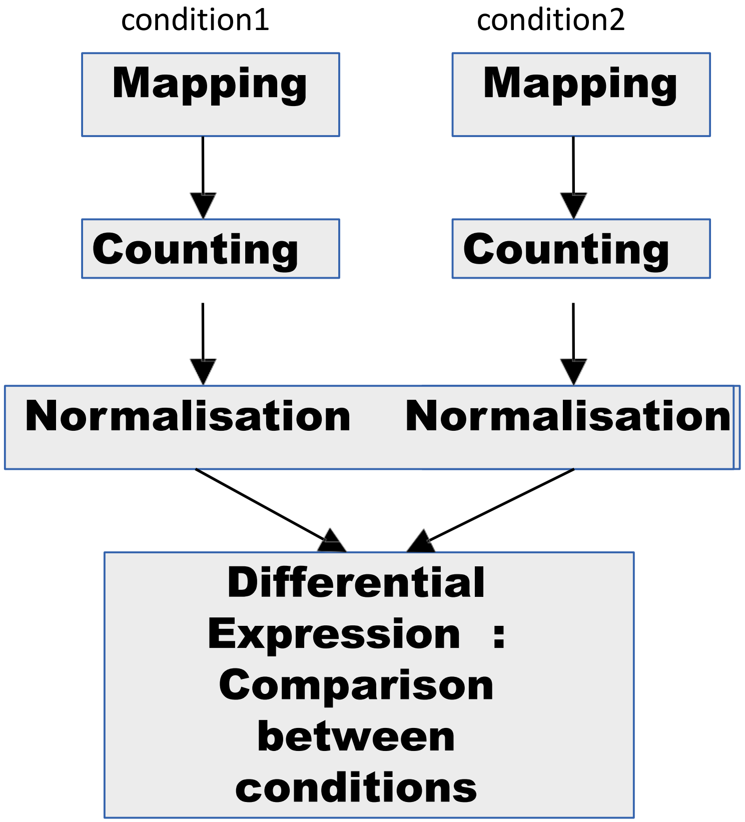
Selection of mRNA with polyA tail :

Random Primed cDNA synthesis :

Paired-end sequencing of fragmented cDNA:

Alignment:

Quantify expression levels = RPKM (# of aligned Reads Per Kb of transcript per Million total reads)



condition1

condition2

Mapping

Mapping

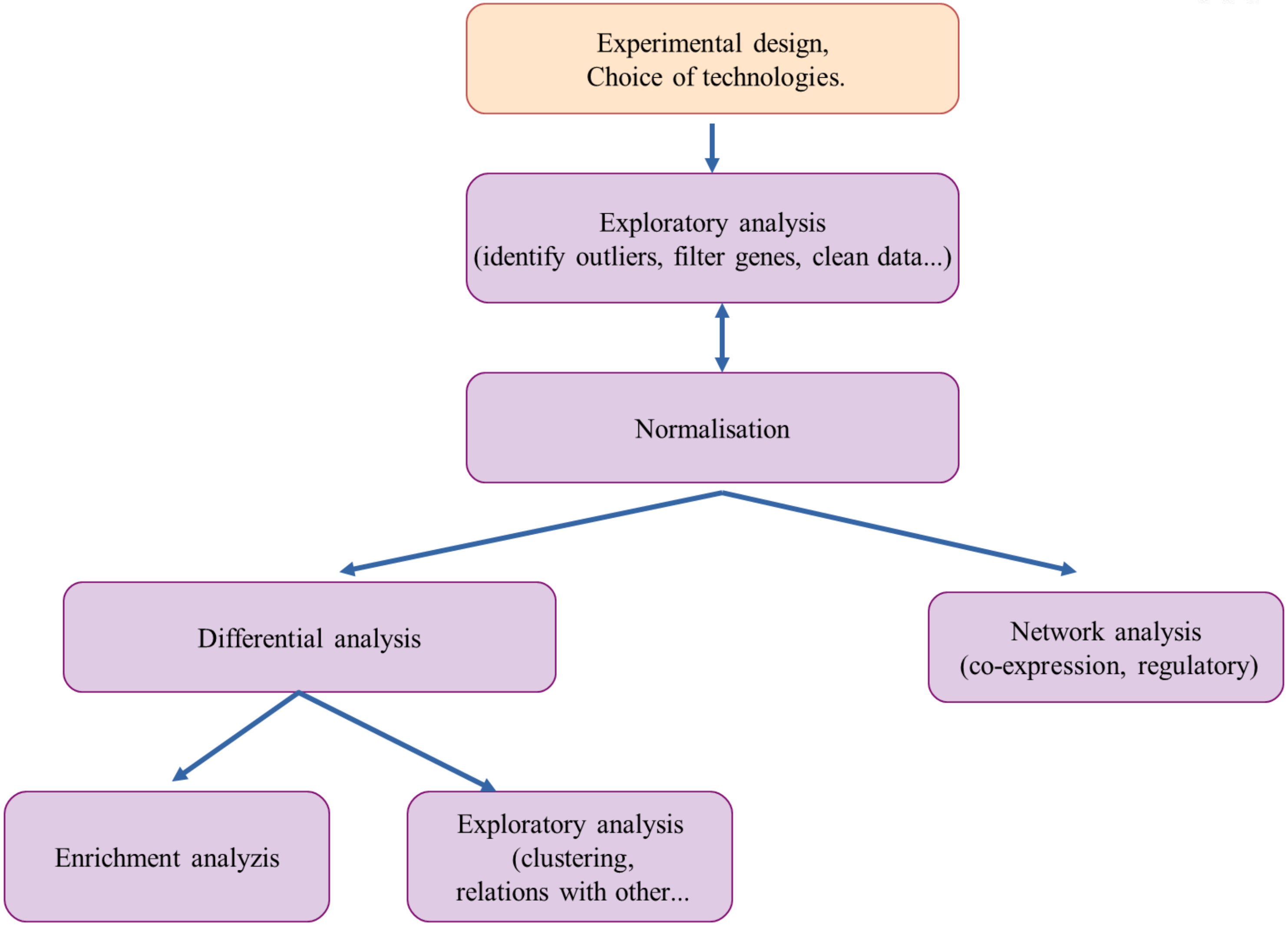
Counting

Counting

Normalisation

Normalisation

Differential Expression : Comparison between conditions



Experimental design,
Choice of technologies.

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

Normalisation

Differential analysis

Network analysis
(co-expression, regulatory)

Enrichment analyzis

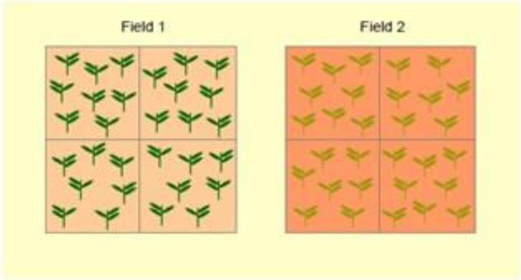
Exploratory analysis
(clustering,
relations with other...

1) Experimental design

Basic experiment : Find differences between conditions control/treated

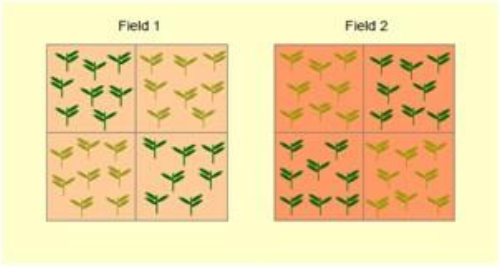
control group plant
treated group plant

Bad experimental design: treated plants and control plants are located in 2 different fields



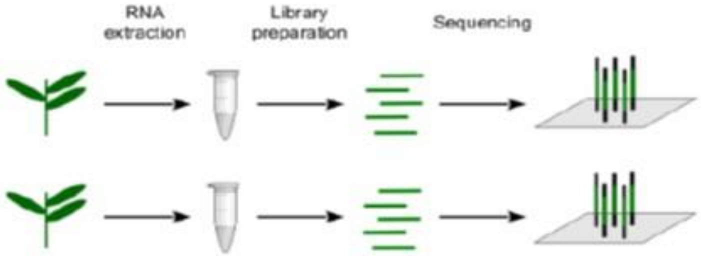
Not possible to differentiate between treatment effect and field effect

Good experimental design: treated plants and control plants are mixed in the 2 fields



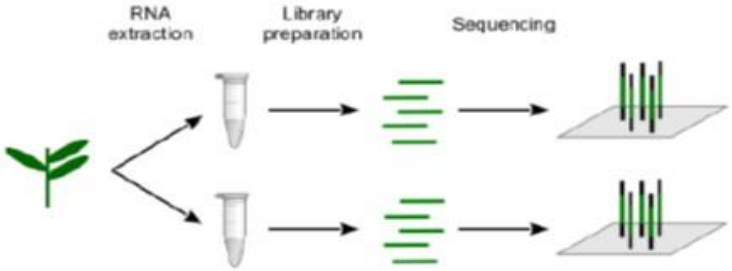
Possible to differentiate between treatment effect and field effect

Biological replicates: Different biological samples, repeated several times (at least 3 times)



Technical replicates: Same biological material, repeated several times

- Several extractions from the same sample
- Several sequencing from the same library



2) Mapping

Choice of mapping software

1) If we hold a reference genome

Use of « splice junction mapper »
(ex : TopHat2, CRAC, MapSplice)

1) If we have annotation
=> Optimize alignment by considering GFF annotation
=> Allow to search for new genes

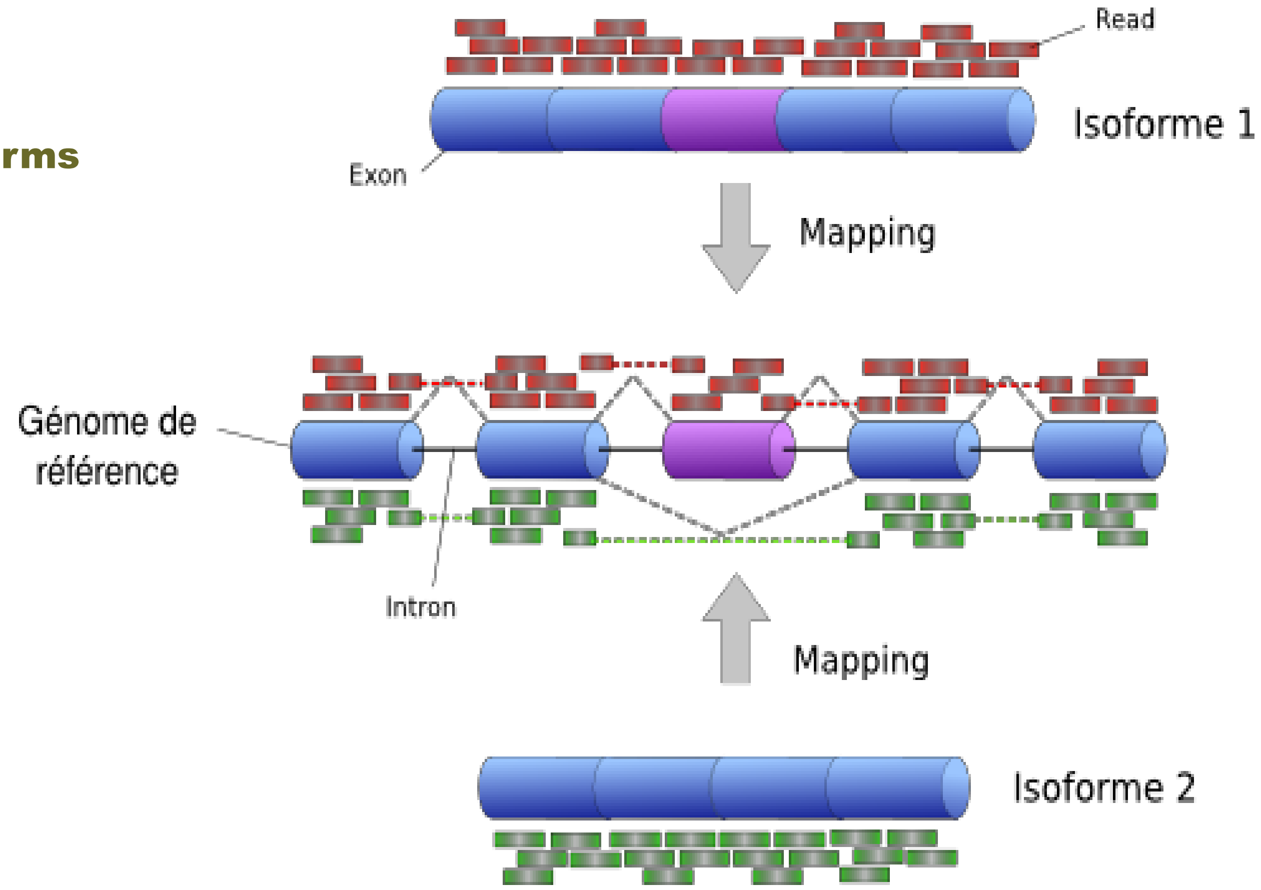
2) If we don't have annotation
=> Help for structural annotation (gene identification)

3) If we hold a reference transcriptome
Use of traditional mapper (ex : BWA, bowtie)

Mapping onto a genomic reference

=> Allow to highlight isoforms

=> Help for the structural annotation of the genome



Suite TopHat / Cufflinks / CummeRbund

(update: HiSAT/StringTie)

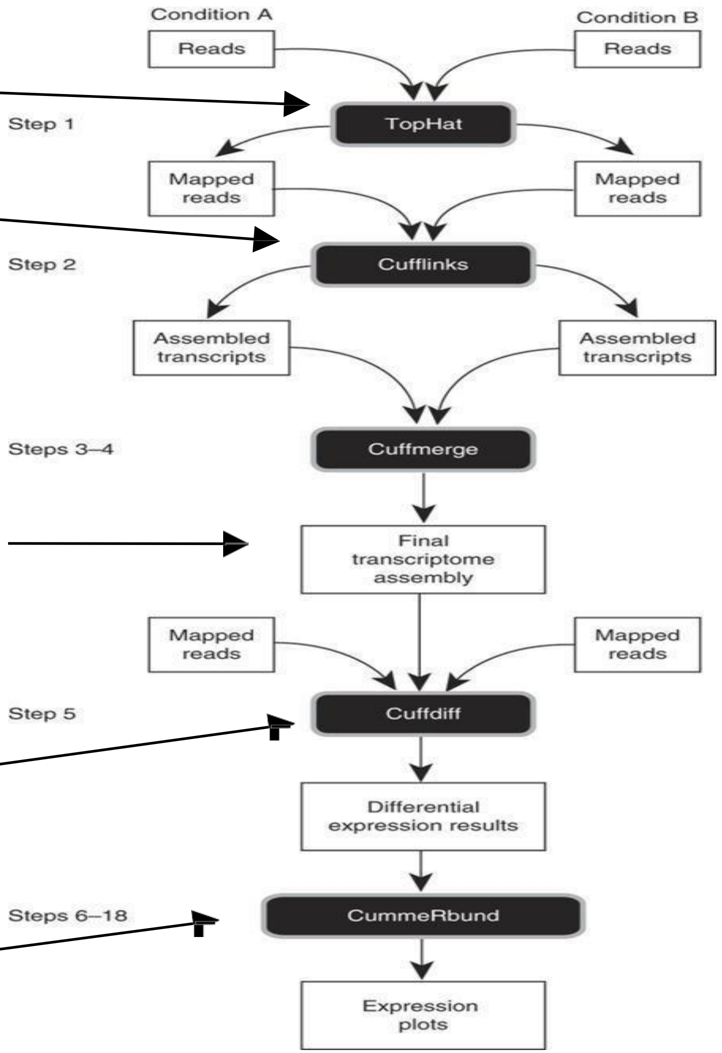
alignment

Assembly
(for each condition)

Gathering assemblies = meta-assembly on which we can calculate the level of expression

Statistics DE

Plots



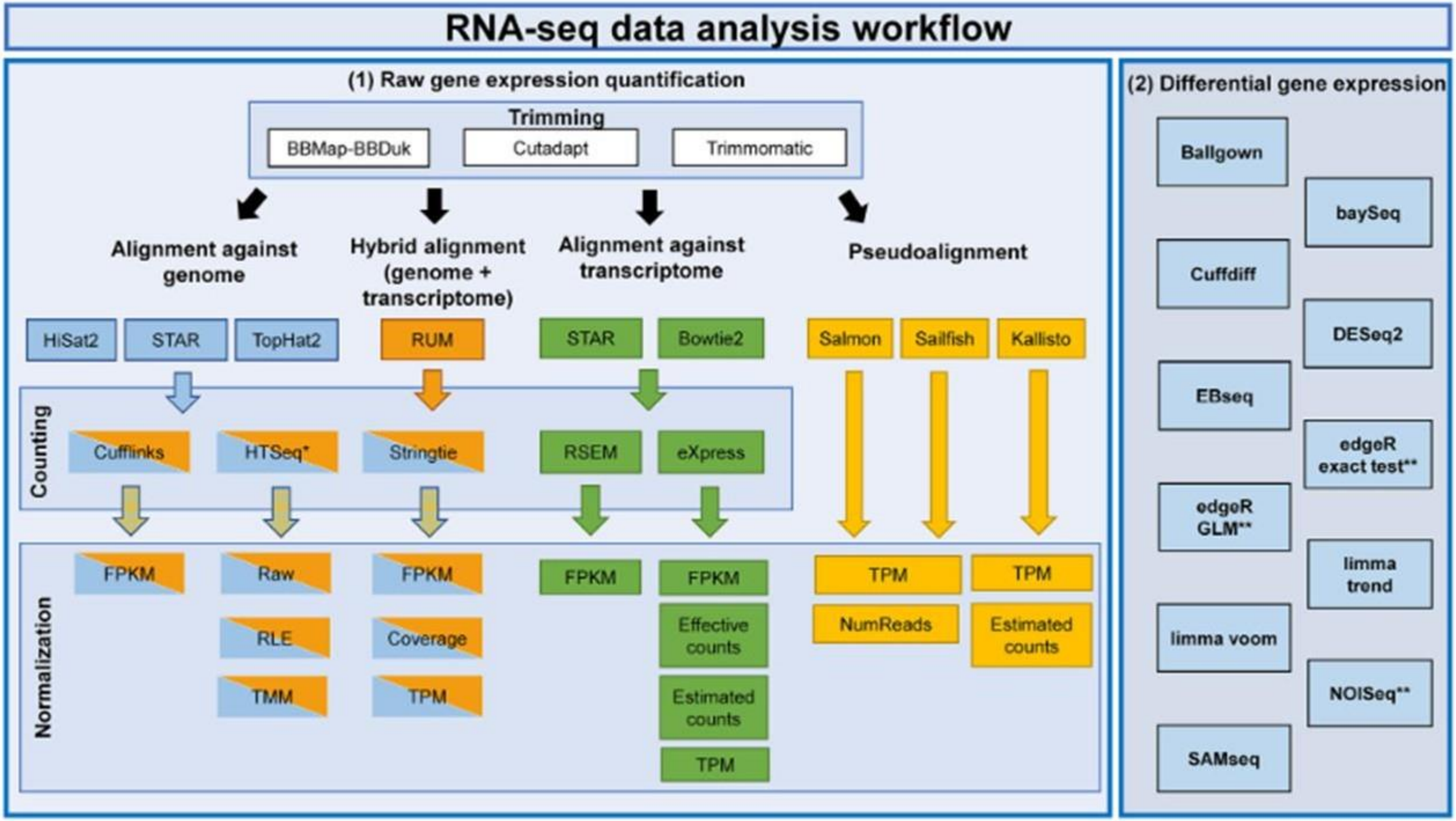


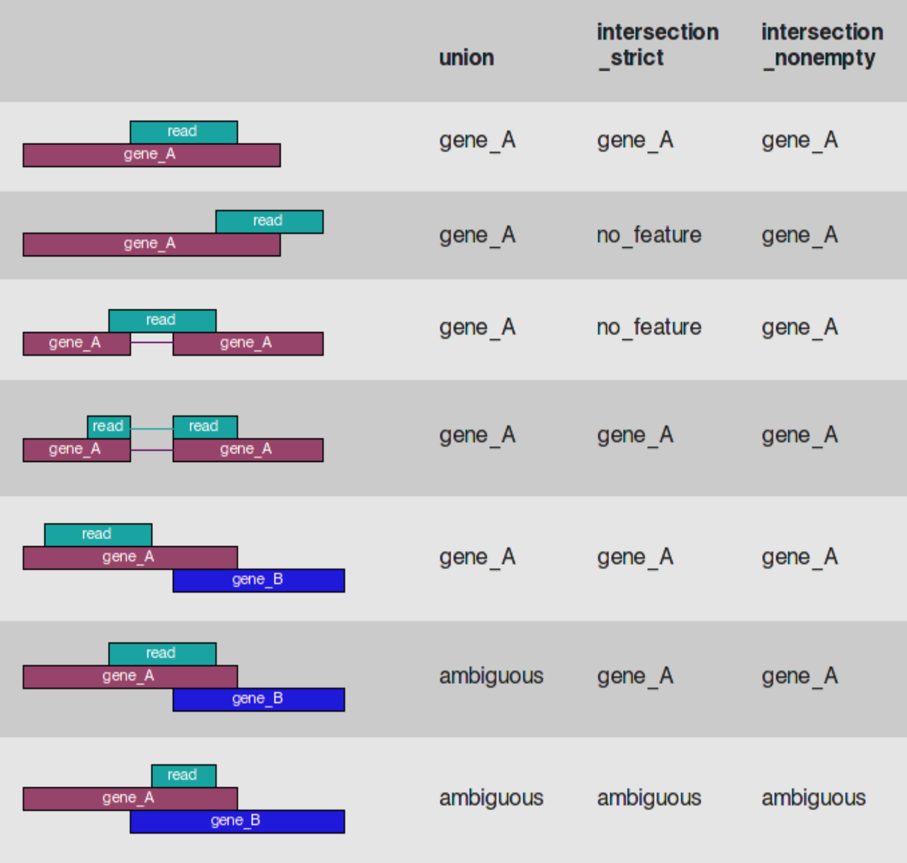
Figure 1. RNA-seq analysis workflow. Left panel (1) represents the raw gene expression quantification workflow. Every box contains the algorithms and methods used for the RNA-seq analysis at trimming, alignment, counting, normalization and pseudoalignment levels. The right panel (2) represents the algorithms used for the differential gene expression quantification. **HTSeq* was performed in two modes: union and intersection-strict. ***EdgeR exact* test, *edgeR GLM* and *NOISeq* have internally three normalization techniques that were evaluated separately.

3) Counting

Choice of the counting software

1) If mapping has been performed against an annotated reference genome

=> Use of HTSeq-count (takes as input GFF annotation)



	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

2) If mapping has been performed against reference transcriptome

=> samtools idxstats

=> Kallisto (pseudo-alignment)

4) Data Normalization

Objectives : allows to compare obtained values between different samples

Mistake to avoid: believe that RNA-seq data are more stable that those of DNA microarray and that normalization is not required

« One particularly powerful advantage of RNA-seq is that it can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets » (Wang et al.,Nat. Rev. Genet., 2009)

In reality, biases exist but are different
=> Need to realize specific normalization methods

Main biases currently identified :
- Size of the bank (= depth of coverage)
- Gene length
- GC content of genes

Effect of the size of the bank:

For two samples having the same RNA content, we product one bank
for each sample
We obtained 2 781 315 reads for bank A and 2 254 901 reads for
bank B
=> We have « artificially » 1.2334 times more RNA in bank A although « real » quantity
are identical

Effect of gene length:

For the same level of expression, a long transcript will have more chances to be sequenced (and thus more reads) than a shorter transcript

=> More relevant for highlighting DE
=> Need to correct this bias

Methods of normalization :

1) Methods of normalization inter-bank :

Objectives : calculate a scaling factor to be applied to each bank

-Total Count (TC) : we divide every number of reads by the total number of reads (i.e. size of the bank) and we multiply by the average total number of reads across banks

-Upper Quartile (UQ) : same as TC but we replace the total number of reads by the 3rd quartile of counts different to 0
=> normalization less sensitive to extreme values
normalization more robuste, notably in the case where several genes abundant are differentially expressed

- RLE (Relative log expression)
- TMM (Trimmed Means of M-Values)

http://biorxiv.org/content/biorxiv/early/2015/09/03/026062.full.pdf

Methods of normalization :

2) Reads Per Kilobase per Million (RPKM) :

Objectives : perform a normalization taking into account both size of the bank (using the method Total Count) AND gene length

=> Mix of normalization inter and intra-bank
=> Allows to compare genes between them but not necessarily usefull to compare 2 conditions on a same gene

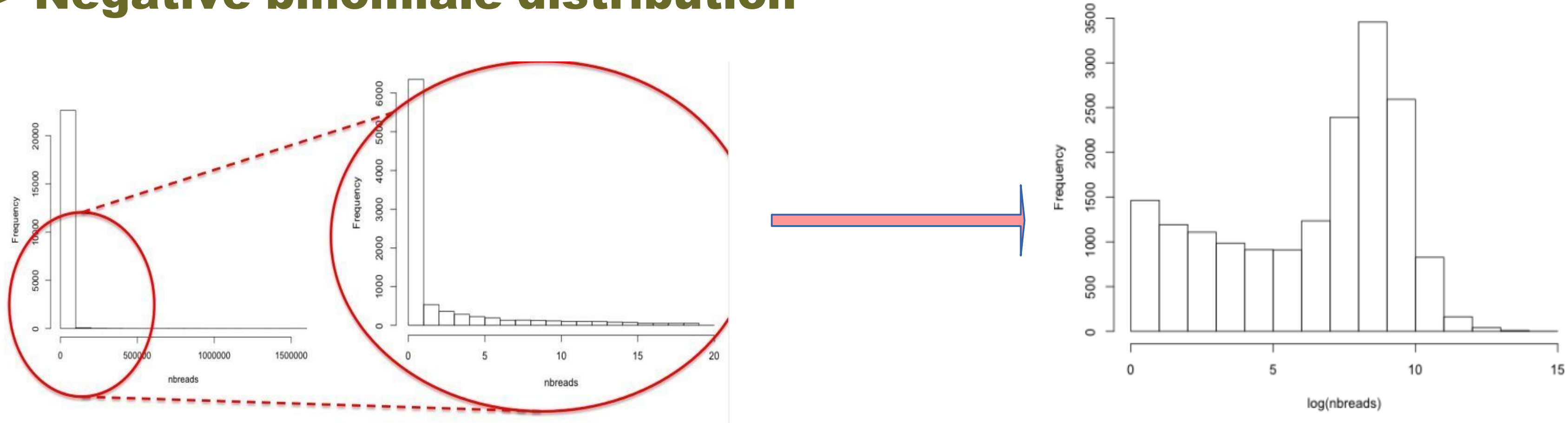
3) Normalization taking into account the bias associated to GC content

-Total Count method not really efficient (doesn’t take into account possible differences in RNA composition between conditions)
- RPKM method not efficient and sccessfull, is criticized (even for cases where there is bias related to gene length, the use of RPKM doesn’t allow to correct it completely)
- More successfull methods to prefer: Upper-Quartile, RLE, TMM

5) Search for differentially expressed genes

Modeling data

- In order to follow a statistics law, use of the log(number reads) instead of the number of reads
+ need to transform « 0 »
=> Negative binomiale distribution

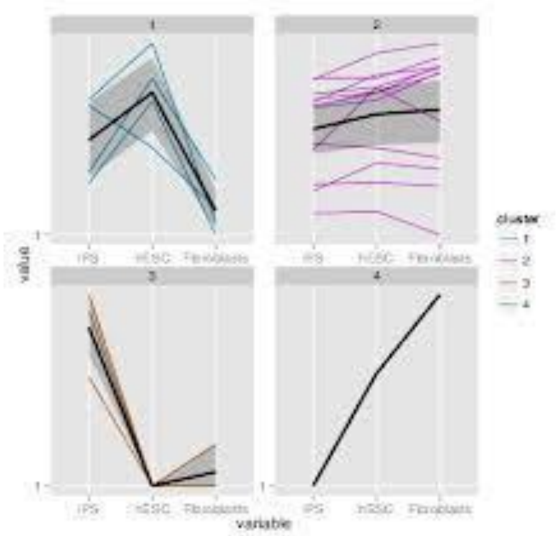


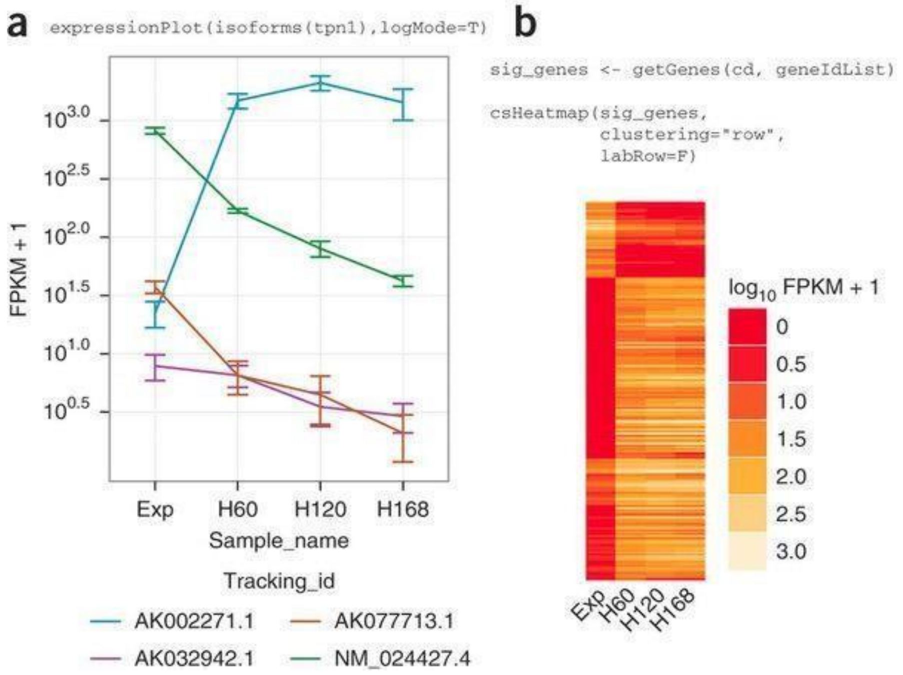
- Use of log(FoldChange)
Fold Change = ratio between 2 expression levels
= ratio final value / initial value

Methods based on RPKM

(Cuffdiff)

Cuffdiff - CummeRbund




a

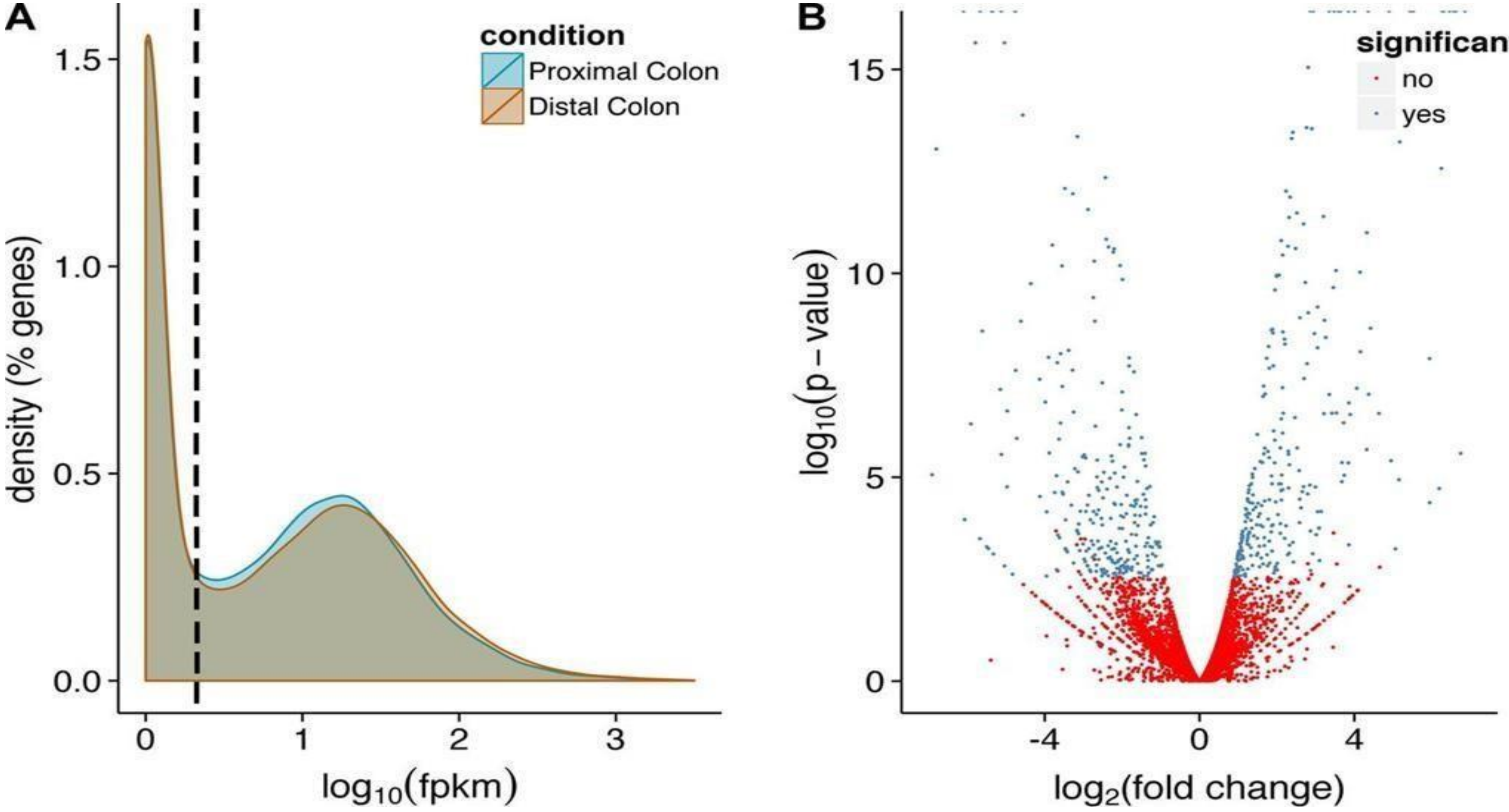
```
expressionPlot(isoforms(tpn1),logMode=T)
```

b

```
sig_genes <- getGenes(cd, geneIdList)

csHeatmap(sig_genes,
          clustering="row",
          labRow=F)
```

Cuffdiff - CummeRbund



Institut de Recherche pour le Développement
FRANCE

South Green bioinformatics platform

UNIVERSIDAD NACIONAL DE COLOMBIA

Methods based on inter-bank normalization

(RLE, TMM, Upper-Quartile)

(EdgeR et DESeq)

Comparison of softwares DESeq/EdgeR

DESeq uses an estimate of variance that makes it less permissive for high variability between conditions. If at least one of the conditions show a deviation, DESEq doen’t trust the gene et will not consider it as differentially expressed, even if there is a grande difference between conditions (logFC).

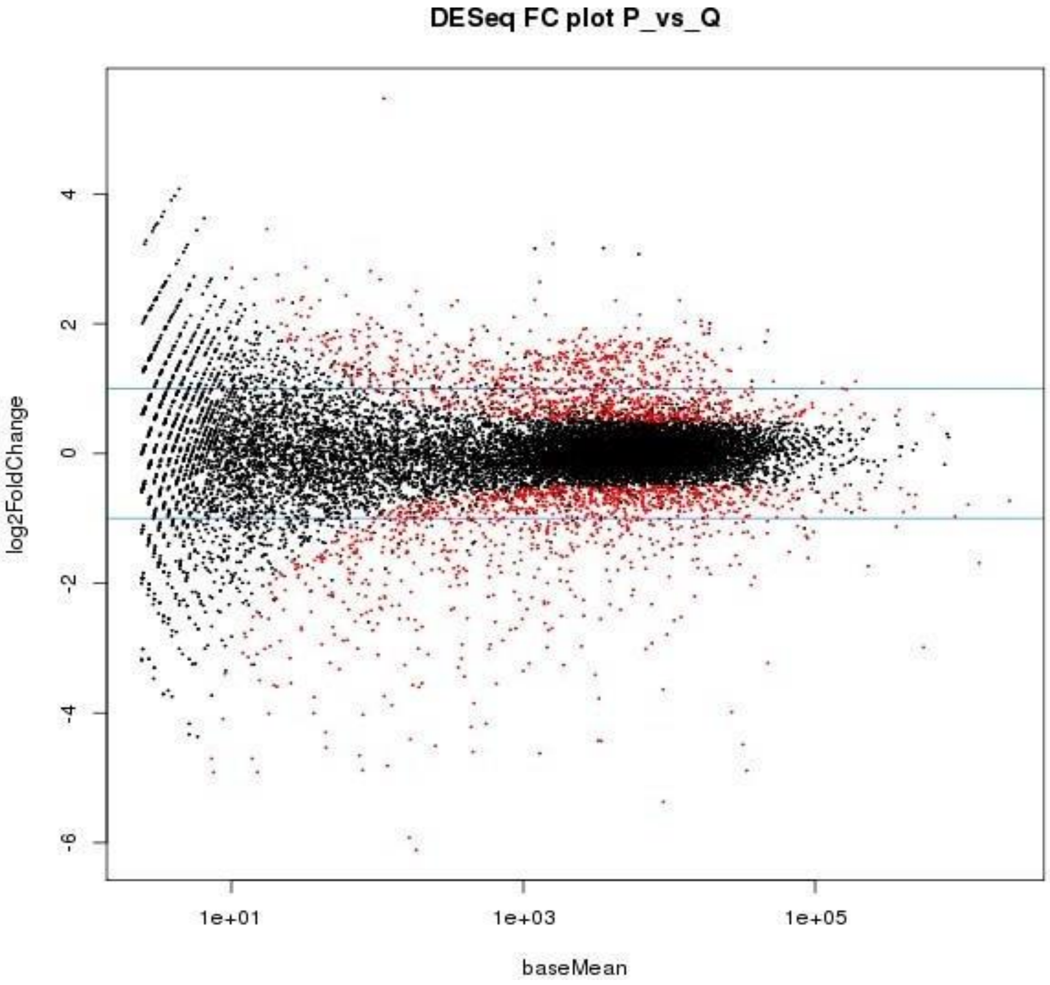
At the opposite, when the variability intra-condition is low, DESeq trust more and may select genes for which fold-Change is low even those discarded by EdgeR.

=> DESeq is to prefer for experimentations very repeatable

DESeq2 is more flexible than DESeq plus souple, will be less stringent and detect more DE genes

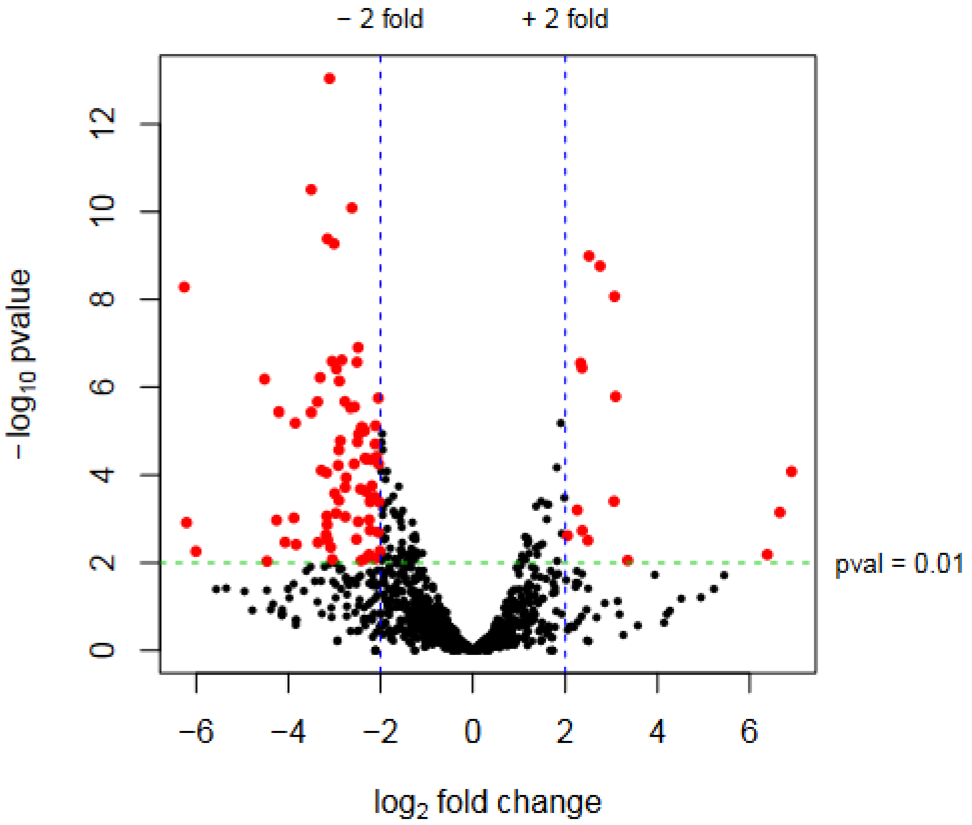
Smear plot / MA plot

Pvalue adj < 0.05



Volcano plot

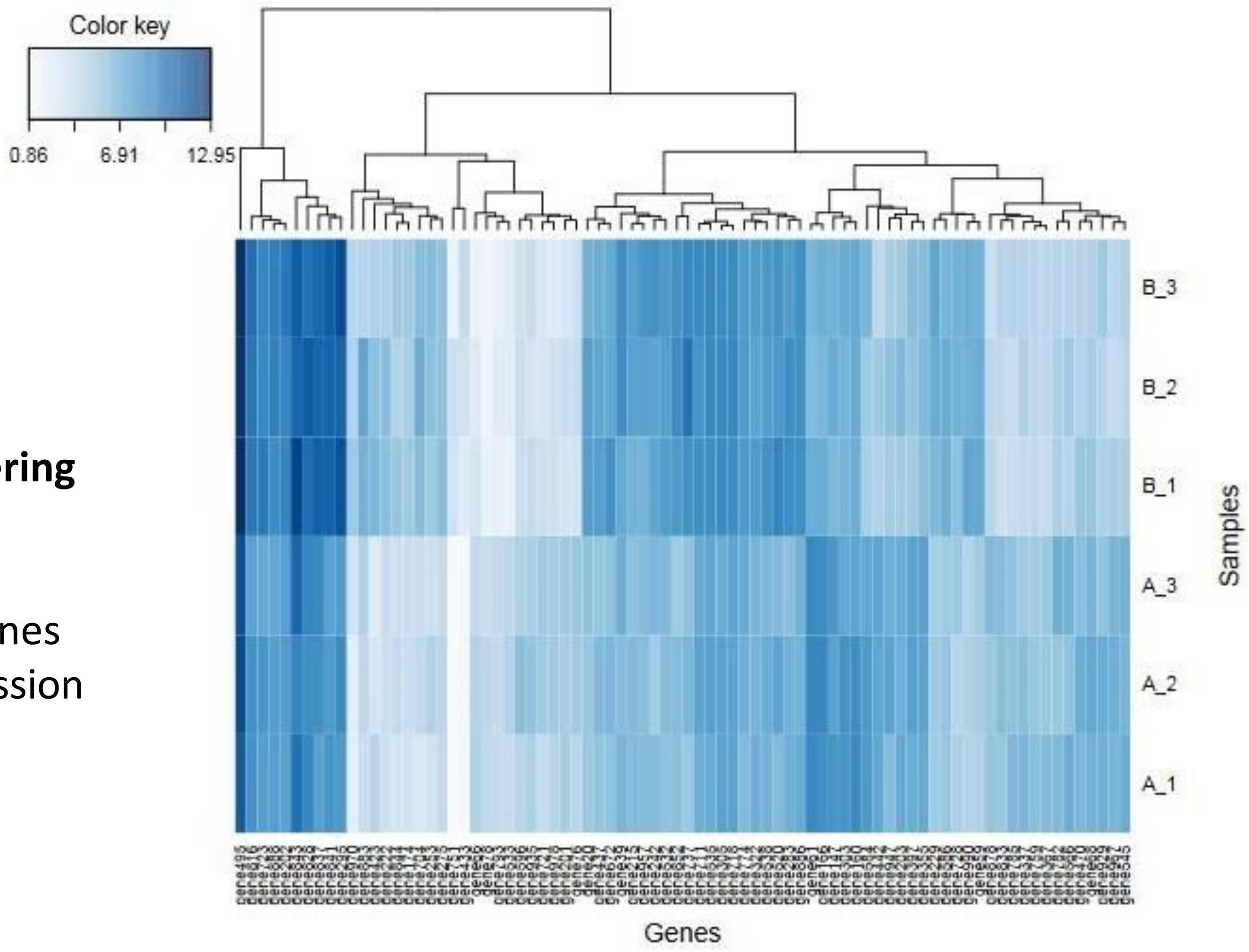
Pvalue adj < 0.01



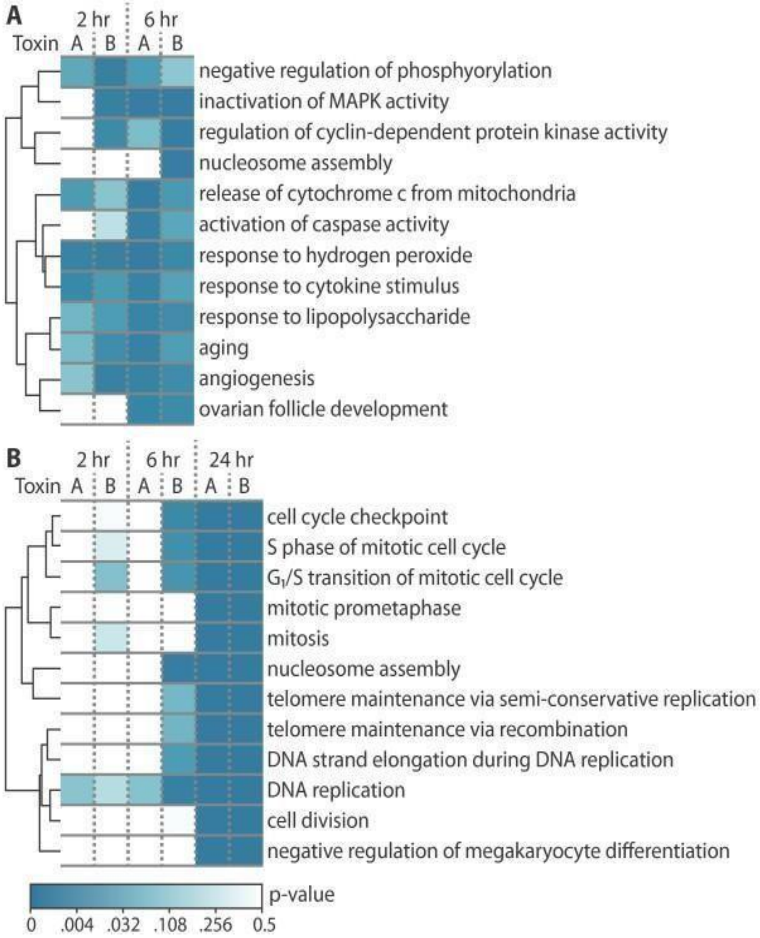
Tutorial: http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

Hierarchical Clustering et Heatmap

=> Clustering of genes according to expression patterns

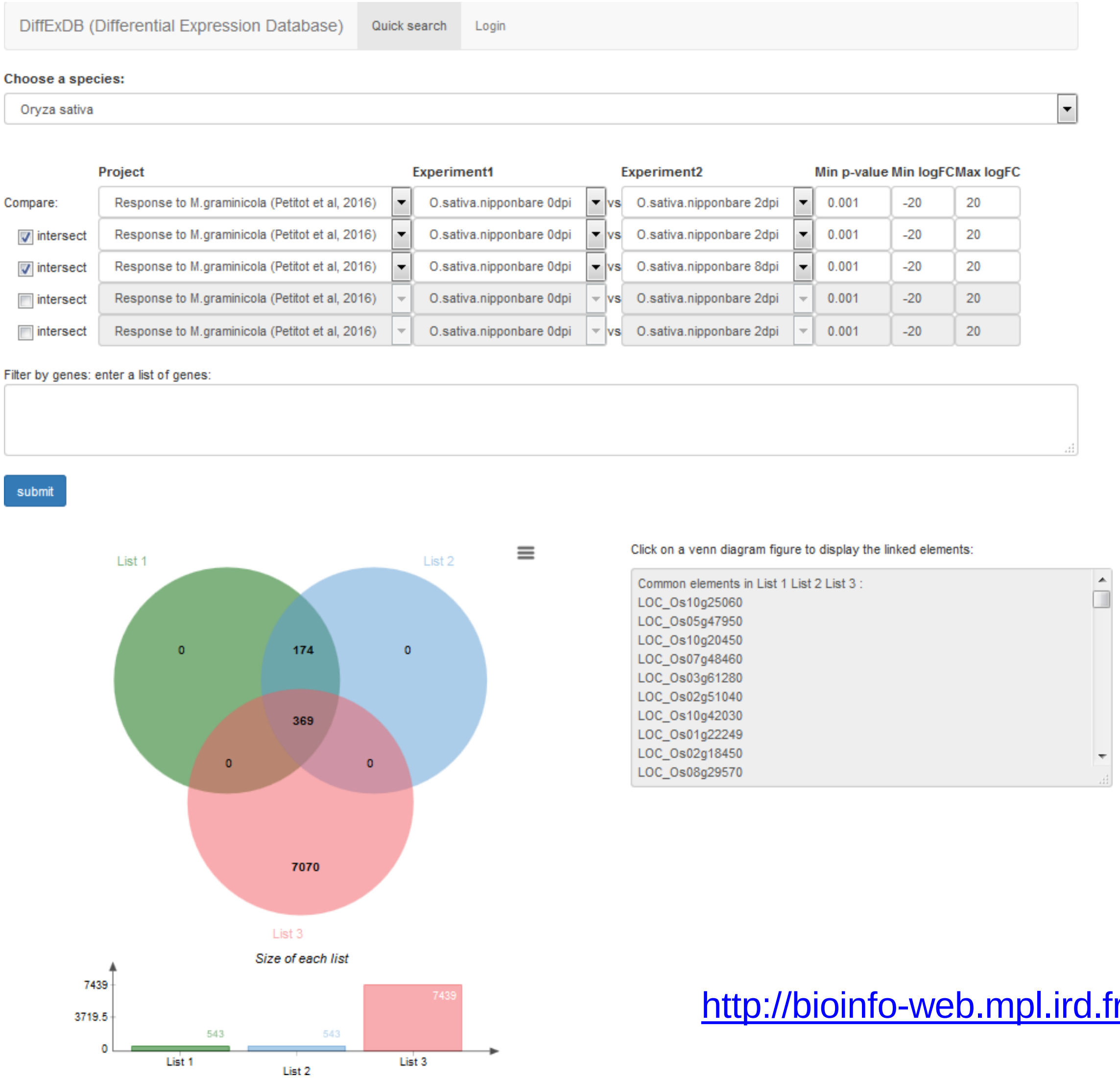


TopGO : Study of Gene Ontology terms enrichment



Need to have a GO functional annotation of transcripts

=> Test if it exist significant enrichments of GO functions between DE genes and non-DE genes (between 2 conditions)



Choose a species:

Oryza sativa

	Project	Experiment1		Experiment2	Min p-value	Min logFC	Max logFC
Compare:	Response to M.graminicola (Petitot et al, 2016)	O.sativa.nipponbare 0dpi	vs	O.sativa.nipponbare 2dpi	0.001	-20	20
☑ intersect	Response to M.graminicola (Petitot et al, 2016)	O.sativa.nipponbare 0dpi	vs	O.sativa.nipponbare 2dpi	0.001	-20	20
☑ intersect	Response to M.graminicola (Petitot et al, 2016)	O.sativa.nipponbare 0dpi	vs	O.sativa.nipponbare 8dpi	0.001	-20	20
□ intersect	Response to M.graminicola (Petitot et al, 2016)	O.sativa.nipponbare 0dpi	vs	O.sativa.nipponbare 2dpi	0.001	-20	20
□ intersect	Response to M.graminicola (Petitot et al, 2016)	O.sativa.nipponbare 0dpi	vs	O.sativa.nipponbare 2dpi	0.001	-20	20

Filter by genes: enter a list of genes:

Click on a venn diagram figure to display the linked elements:

Common elements in List 1 List 2 List 3 :
LOC_Os10g25060
LOC_Os05g47950
LOC_Os10g20450
LOC_Os07g48460
LOC_Os03g61280
LOC_Os02g51040
LOC_Os10g42030
LOC_Os01g22249
LOC_Os02g18450
LOC_Os08g29570

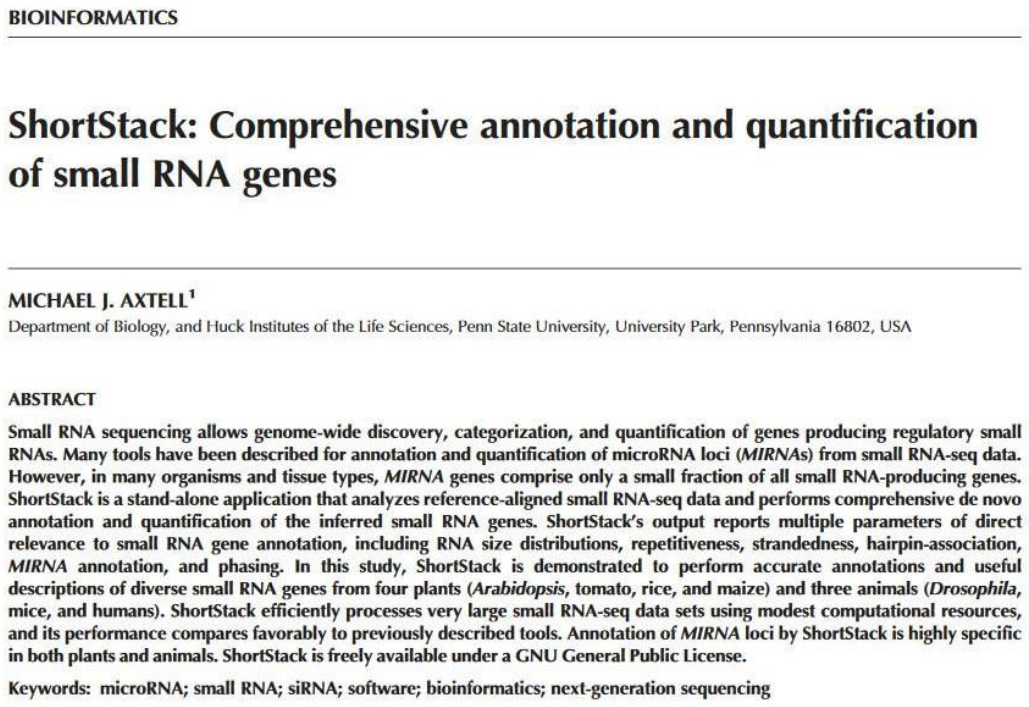
DiffExDB

Web application to explore data from differential expression analysis:

- Overlap between comparisons
- Heatmap of expression

http://bioinfo-web.mpl.ird.fr/cgi-bin2/microarray/public/diffexdb.cgi

ShortStack: Management of small RNA data



BIOINFORMATICS

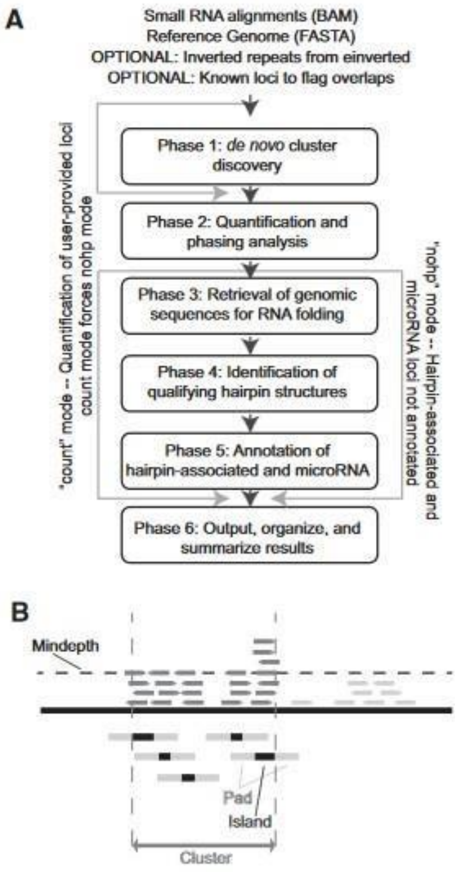
ShortStack: Comprehensive annotation and quantification of small RNA genes

MICHAEL J. AXTELL[1]
Department of Biology, and Huck Institutes of the Life Sciences, Penn State University, University Park, Pennsylvania 16802, USA

ABSTRACT

Small RNA sequencing allows genome-wide discovery, categorization, and quantification of genes producing regulatory small RNAs. Many tools have been described for annotation and quantification of microRNA loci (*MIRNAs*) from small RNA-seq data. However, in many organisms and tissue types, *MIRNA* genes comprise only a small fraction of all small RNA-producing genes. ShortStack is a stand-alone application that analyzes reference-aligned small RNA-seq data and performs comprehensive de novo annotation and quantification of the inferred small RNA genes. ShortStack's output reports multiple parameters of direct relevance to small RNA gene annotation, including RNA size distributions, repetitiveness, strandedness, hairpin-association, *MIRNA* annotation, and phasing. In this study, ShortStack is demonstrated to perform accurate annotations and useful descriptions of diverse small RNA genes from four plants (*Arabidopsis*, tomato, rice, and maize) and three animals (*Drosophila*, mice, and humans). ShortStack efficiently processes very large small RNA-seq data sets using modest computational resources, and its performance compares favorably to previously described tools. Annotation of *MIRNA* loci by ShortStack is highly specific in both plants and animals. ShortStack is freely available under a GNU General Public License.

Keywords: microRNA; small RNA; siRNA; software; bioinformatics; next-generation sequencing



A

Small RNA alignments (BAM)
Reference Genome (FASTA)
OPTIONAL: Inverted repeats from einverted
OPTIONAL: Known loci to flag overlaps

Phase 1: *de novo* cluster discovery

Phase 2: Quantification and phasing analysis

Phase 3: Retrieval of genomic sequences for RNA folding

Phase 4: Identification of qualifying hairpin structures

Phase 5: Annotation of hairpin-associated and microRNA

Phase 6: Output, organize, and summarize results

"count" mode -- Quantification of user-provided loci
count mode forces nohp mode

"nohp" mode -- Hairpin-associated and microRNA loci not annotated

B

Exercise:

1) Perform a counting per gene from a BAM file using the software samtools idxstats.

1) In Galaxy, import a complete dataset that will be used for differential expression analysis

Shared data => Data libraries => Formation 2015 => RNASeq

1) Pre-filter sequences in order to keep only those that have at least 10 reads across the whole conditions. How many genes have been filtered?

It is not possible to perform reliable tests using low values of counting. This is to limit the number of statistics tests and thus decrease the effect of corrections for multiple tests

1) Perform a differentially expressed genes study using the EdgeR software. Observe the graphical outputs. Setting a p-value threshold to 0.01, how many genes are found to be DE?