

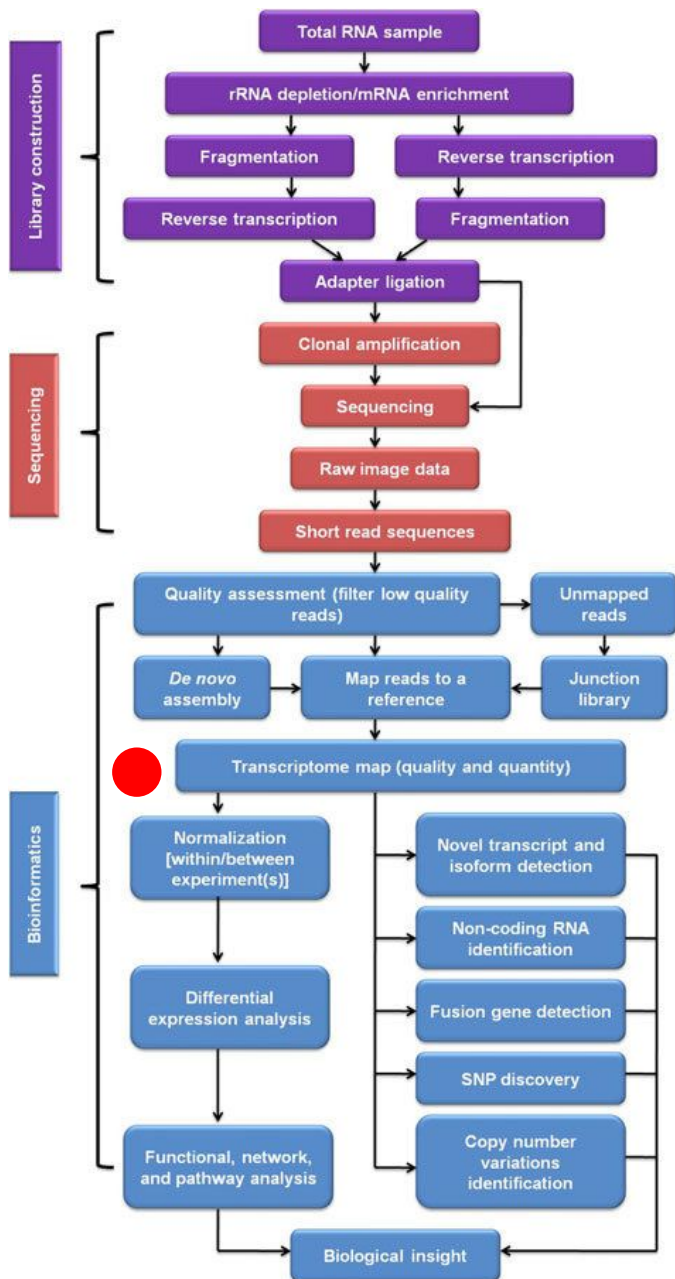
Introduction to RNA-seq count data analysis with R

Alessia Buratin

Why use R ?

- Script vs. Menu driven software
 - Can be re-rerun with new data
- R works with vectors and tables
 - $Z \leftarrow X + Y$ where X and Y are vectors
- Incredible graphics and plots
 - ggplot2 family
- Work environment
 - R studio
- Document your data processing
 - R markdown
- Share your data and workflow
 - GitHub





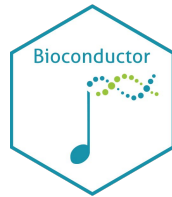
● You are here !

Possible Biological Questions:

- Exist a significant difference in gene expression between group comparison?
- ... or across sites and times of sampling?
- ... or after a treatment? (drugs, pollutants, etc.)

Graphical Abstract

Let's start !



Open Rstudio and packages

**Counts and clinical matrices
pre-processing**

```
dds <- DESeqDataSetFromMatrix(countData = raw_counts,  
                               colData = clindata,  
                               design = ~ positivity + batch)  
  
dds <- estimateSizeFactors(dds)  
  
norm_counts <- counts(dds, normalized=TRUE)
```

Quality Control

Counts normalization

Data transformation

Import data



Keyword or GEO Accession

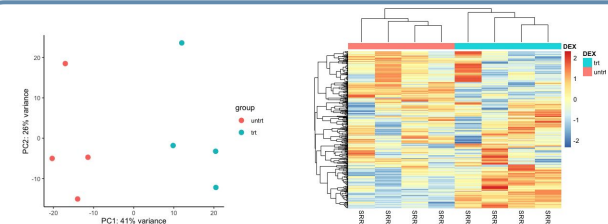
Browse Content

Repository Browser

DataSets: 4348

- Use your own RNA-seq experiment
- Browse public repositories

Data analysis



Differential gene expression analysis
Samples clusterization based on gene expression

Public data resources and Bioconductor

▣ Browsing Gene Expression Omnibus (GEO) repositories

- ▣ Download raw sequencing data (.fastq sequencing files)
 - ▣ process and align sequencing data before performing expression analyses using HISAT2, StringTie and Ballgown
- ▣ Download pre-processed files (tab-delimited.txt files containing matrices with sequence read counts **after** trimming and alignment to the reference genome)
 - ▣ The pre-processed files may contain a raw-counts matrix (non-normalized) or a normalized counts matrix
 - ▣ Sample information is also available to download
 - ▣ Finally, the platform used, and pre-processing algorithm are specified

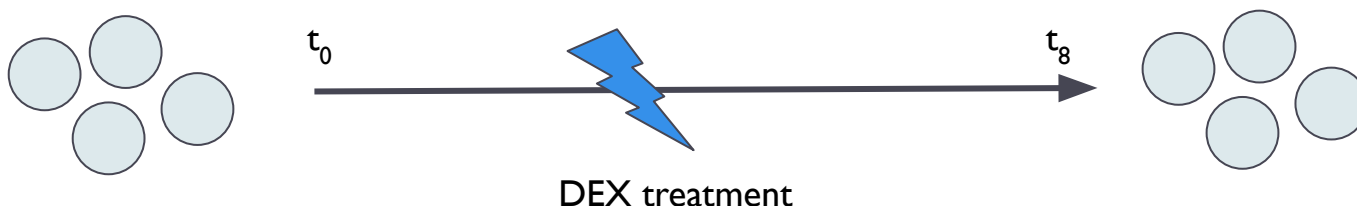
▣ R/Bioconductor packages used

- ▣ [GEOquery](#): Access to the NCBI Gene Expression Omnibus (GEO), a public repository of gene expression (primarily microarray) data
- ▣ [SRADBv2](#): A compilation of metadata from the NCBI Sequence Read Archive, the largest public repository of sequencing data from the next generation of sequencing platforms
- ▣ [curatedTCGAData](#): Curated data from The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects
- ▣ [curatedMetagenomicData](#): Curated metagenomic data of the human microbiome
- ▣ [HMPI6SData](#): Curated metagenomic data of the human microbiome
- ▣ [Bioconductor R package](#) with stored pre-processed count data



Download pre-processed data

- The data used in this workflow is stored in the [airway package](#) that summarizes an RNA-seq experiment wherein airway smooth muscle cells were treated with dexamethasone, a synthetic glucocorticoid steroid with anti-inflammatory effects (Himes et al. 2014)
- In the experiment, four primary human airway smooth muscle cell lines were treated with micromolar dexamethasone for 8 hours. For each of the four cell lines, we have a treated and an untreated sample

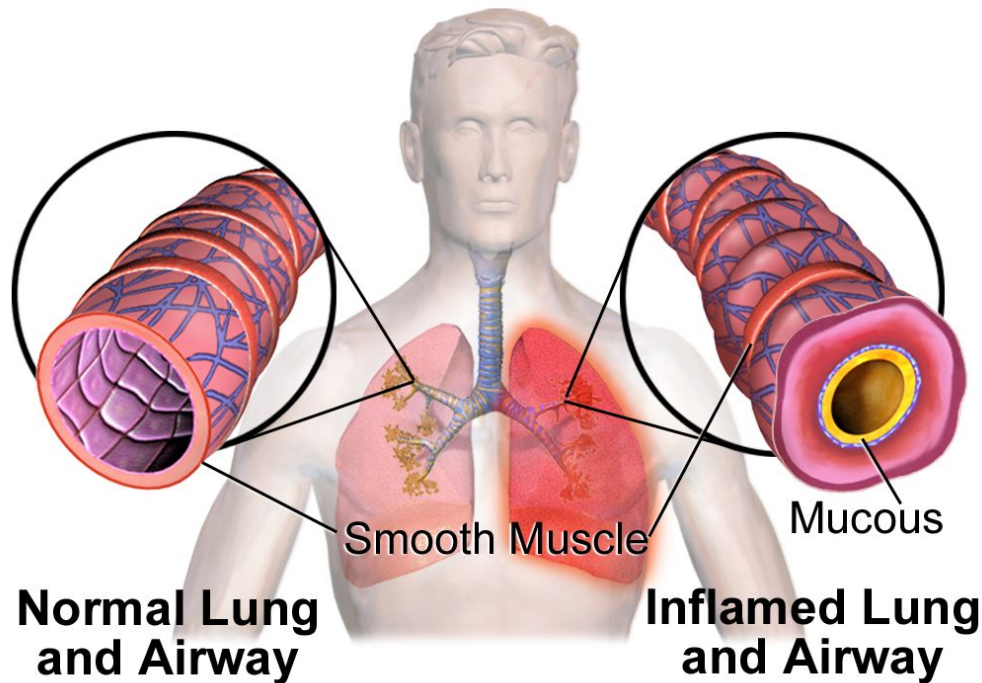


- For more description of the experiment see the PubMed entry [24926665](#) and for raw data see the GEO entry [GSE52778](#)



Biological Question

Identify the biological effect of a drug on gene expression in smooth muscle cells lines

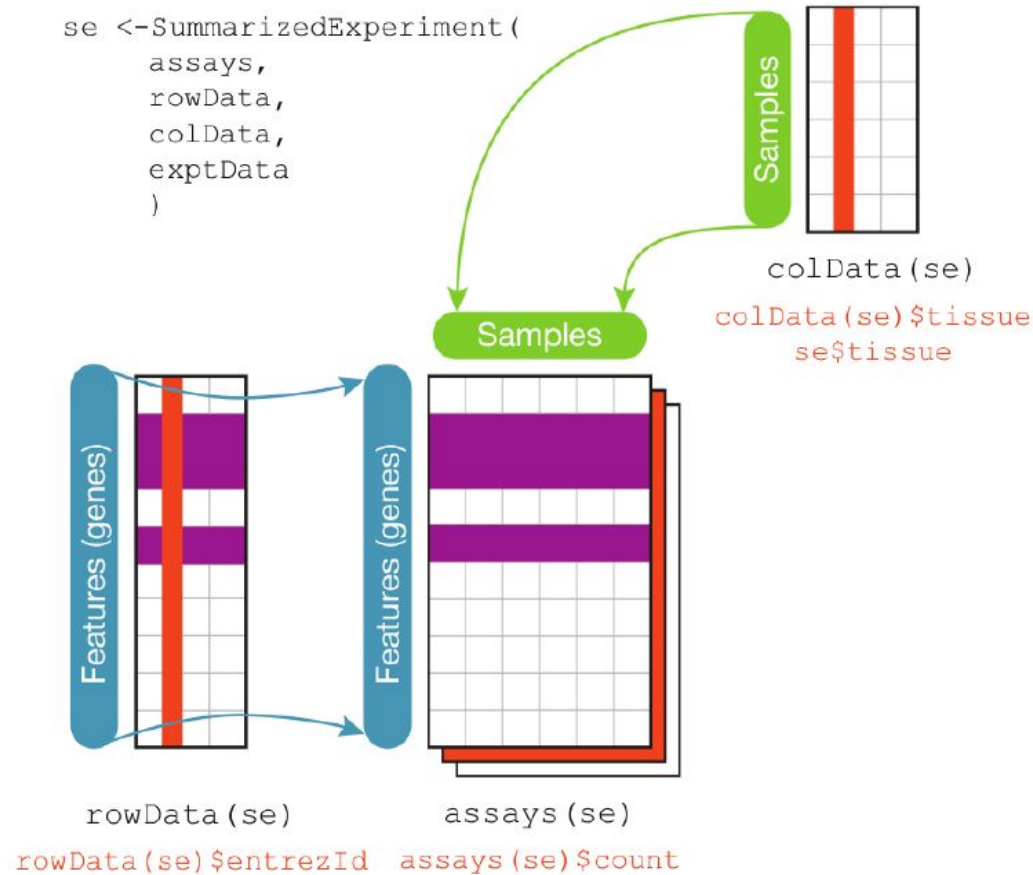


Data

□ Necessary

1. Count data
2. Sample variables

□ The *airway* experiment have already counts prepared in a *SummarizedExperiment* object

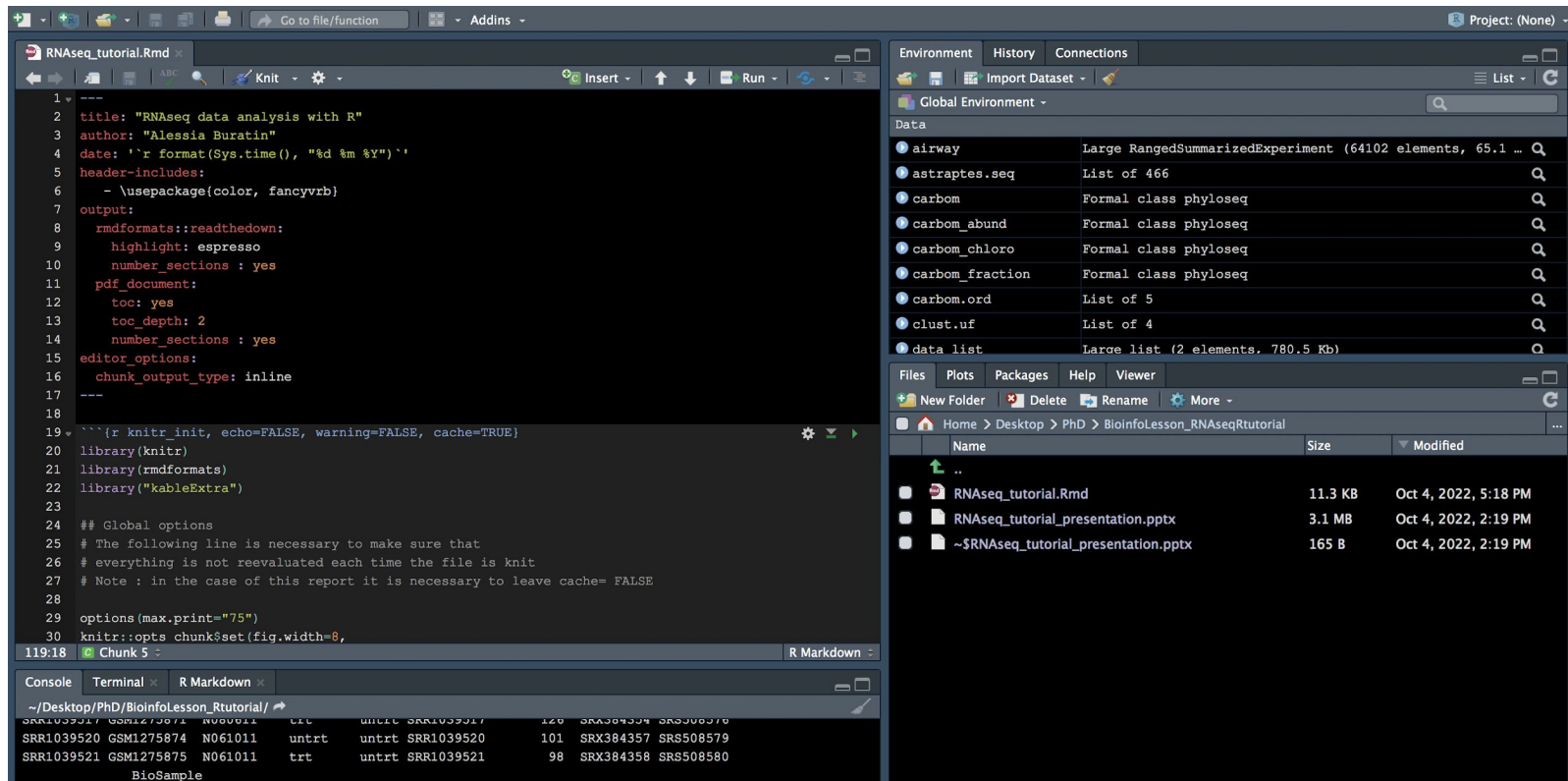


Steps

- Import data
- Quality Control
 - Visualizing Library sizes
 - Filtering non-expressed genes
 - Normalization
 - Multivariate analysis
- Data Analysis
 - Differential Expression analysis
 - Cluster Analysis



- Open R studio
- Open «esercitazioneR_botta.Rmd» with R studio



Library paths setup

```
.libPaths("/home/botta01/R/x86_64-pc-linux-gnu-library/4.2/")
```



Count data

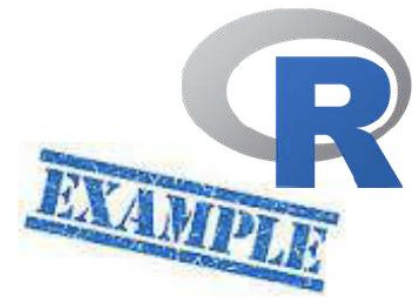


EXAMPLE

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520
ENSG00000000003	679	448	873	408	1138	1047	7
ENSG00000000005	0	0	0	0	0	0	
ENSG000000000419	467	515	621	365	587	799	4
ENSG000000000457	260	211	263	164	245	331	2
ENSG000000000460	60	55	40	35	78	63	
ENSG000000000938	0	0	2	0	1	0	
ENSG000000000971	3251	3679	6177	4252	6721	11027	51
ENSG000000001036	1433	1062	1733	881	1424	1439	13
ENSG000000001084	519	380	595	493	820	714	6
ENSG000000001167	394	236	464	175	658	584	3
ENSG000000001460	172	168	264	118	241	210	1
ENSG000000001461	2112	1867	5137	2657	2735	2751	24
ENSG000000001497	524	488	638	357	676	806	4
ENSG000000001561	71	51	211	156	23	38	1
ENSG000000001617	555	394	905	415	727	697	6
ENSG000000001626	10	2	9	2	10	6	
ENSG000000001629	1660	1251	2259	1079	2462	2514	18
ENSG000000001630	59	54	66	23	84	87	
ENSG000000001631	729	692	943	475	1034	1163	7
ENSG000000002016	201	161	256	99	268	257	1
ENSG000000002070	3	0	3	1	4	0	

Showing 1 to 21 of 1,000 entries

Sample variables

The image shows a screenshot of the RStudio interface. The top menu bar includes 'RStudio', 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', 'Window', and 'Help'. Below the menu bar, there are several tabs for different data objects: 'RNAseq_tutorial.kind', 'colData(airway)', 'assay(airway)', and 'rncanonical'. The 'colData(airway)' tab is selected, displaying a table with 10 columns: 'SampleName', 'cell', 'dex', 'albut', 'Run', 'avgLength', 'Experiment', 'Sample', and 'BioSample'. The table contains 9 rows of data, each representing a different sample. The first row is highlighted in blue.

	SampleName	cell	dex	albut	Run	avgLength	Experiment	Sample	BioSample
SRR1039508	GSM1275862	N61311	untrt	untrt	SRR1039508	126	SRX384345	SRS508568	SAMN02422669
SRR1039509	GSM1275863	N61311	trt	untrt	SRR1039509	126	SRX384346	SRS508567	SAMN02422675
SRR1039512	GSM1275866	N052611	untrt	untrt	SRR1039512	126	SRX384349	SRS508571	SAMN02422678
SRR1039513	GSM1275867	N052611	trt	untrt	SRR1039513	87	SRX384350	SRS508572	SAMN02422670
SRR1039516	GSM1275870	N080611	untrt	untrt	SRR1039516	120	SRX384353	SRS508575	SAMN02422682
SRR1039517	GSM1275871	N080611	trt	untrt	SRR1039517	126	SRX384354	SRS508576	SAMN02422673
SRR1039520	GSM1275874	N061011	untrt	untrt	SRR1039520	101	SRX384357	SRS508579	SAMN02422683
SRR1039521	GSM1275875	N061011	trt	untrt	SRR1039521	98	SRX384358	SRS508580	SAMN02422677

Quality Control



Data quality control (i.e. the removal of insufficiently good data) are essential steps of any data analysis



These steps should typically be performed very early in the analysis of a new data set, preceding or in parallel to the differential expression testing



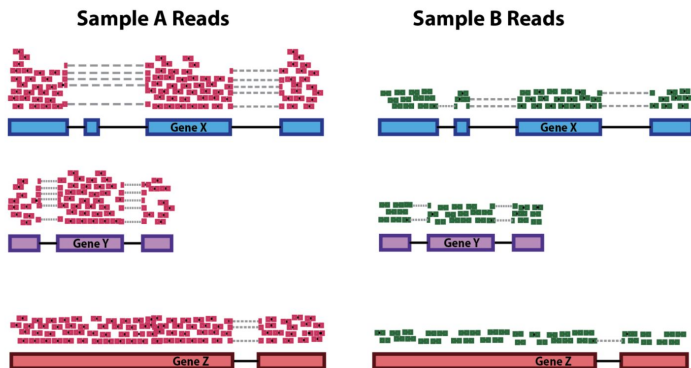
Our purpose is the detection of differentially expressed genes, and we are looking in particular for samples whose experimental treatment suffered from an **abnormality** that renders the data points obtained from these particular samples detrimental to our purpose!!



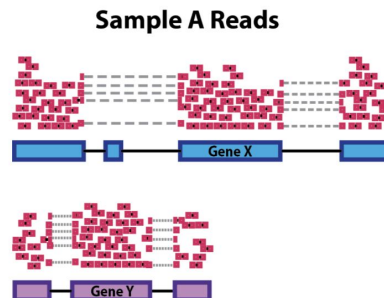
Normalization

- The counts of mapped reads for each gene is proportional to the expression of RNA (“interesting”) in addition to many other factors (“uninteresting”). Normalization is the process of scaling raw count values to account for the “uninteresting” factors. In this way the expression levels are more comparable between and/or within samples
- The main factors often considered during normalization are:

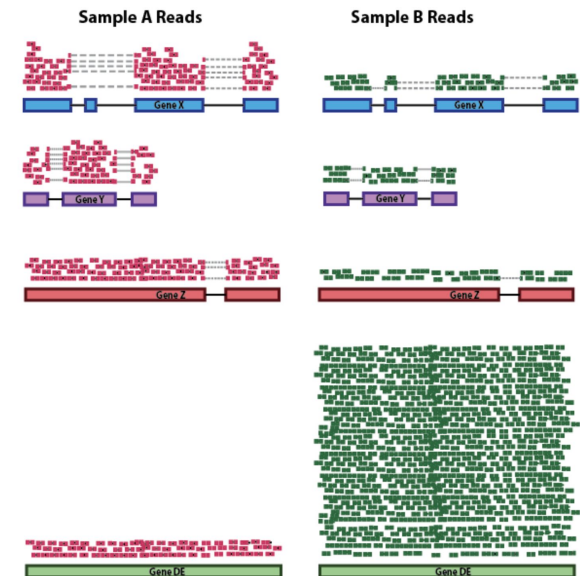
Sequencing depth



Gene length



RNA composition



Normalization and data transformation

Normalization Method	Description	Accounted factors	Recommendations
CPM (Count per Million)	Counts scaled by total number of reads	sequencing depth	gene count comparisons between samples; NOT for DE analysis
TMM (Trimmed Mean of M values)	Uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition	gene count comparisons between samples and for DE analysis
VST (Variance Stabilizing Transformation)	Transforms the count data yielding a matrix of values which are now approximately homoskedastic (having constant variance along the range of mean values).	Mean and variance of gene expression	Input for machine learning techniques (clustering, discriminant analysis)

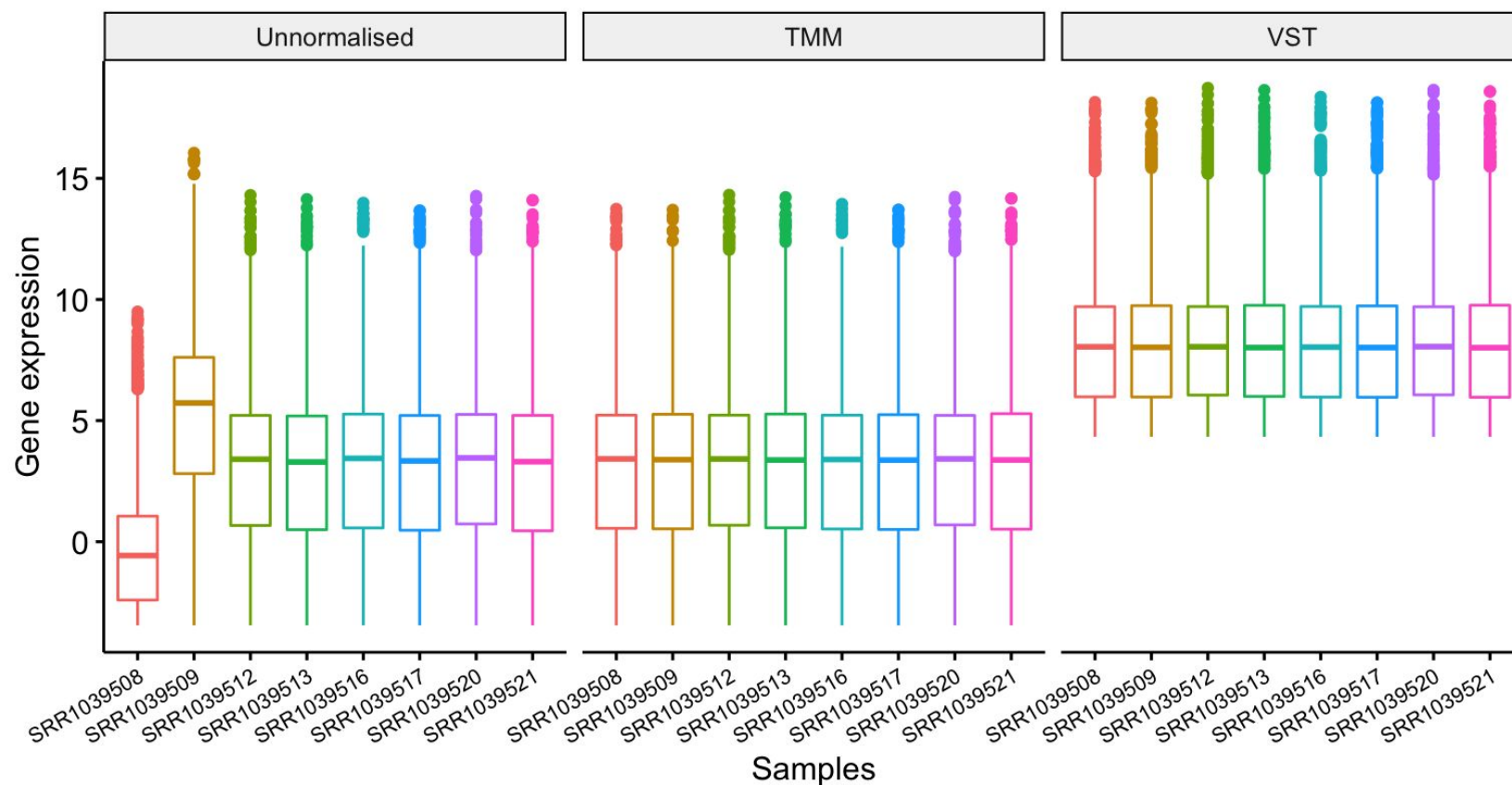
Normalization comparisons



EXAMPLE

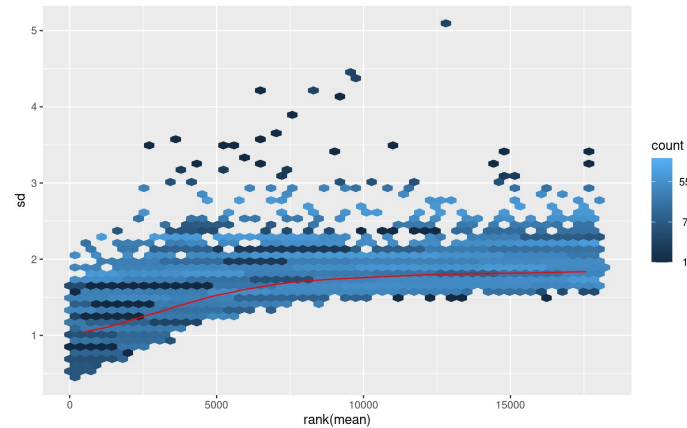
Var2

	SRR1039508		SRR1039512		SRR1039516		SRR1039520
	SRR1039509		SRR1039513		SRR1039517		SRR1039521



Effect of data transformation

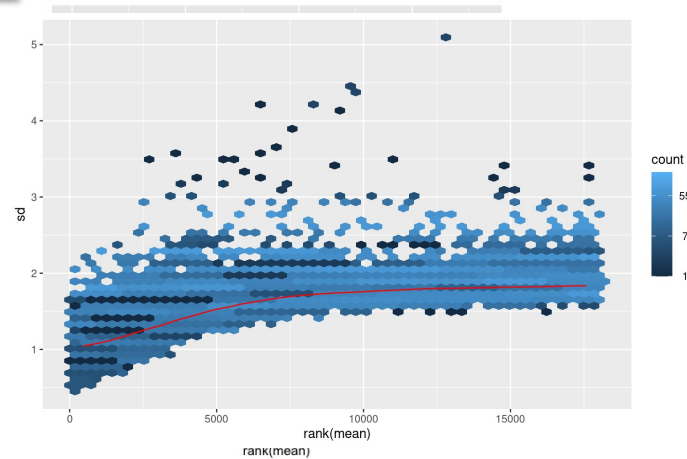
The point of the transformations is to remove the dependence of the variance on the mean particularly the high variance of the logarithm of count data when the mean is low.



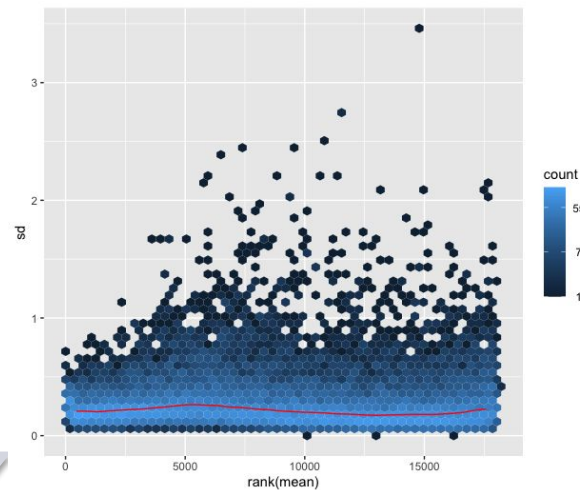
We do not require or desire that all the genes have *exactly* the same variance after transformation, but that the experiment-wide trend has flattened.

Effect of data transformation

The point of the transformations is to remove the dependence of the variance on the mean particularly the high variance of the logarithm of count data when the mean is low.



We do not require or desire that all the genes have *exactly* the same variance after transformation, but that the experiment-wide trend has flattened.

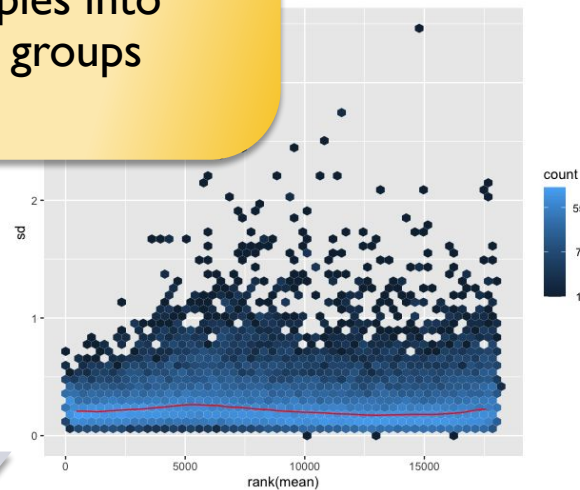
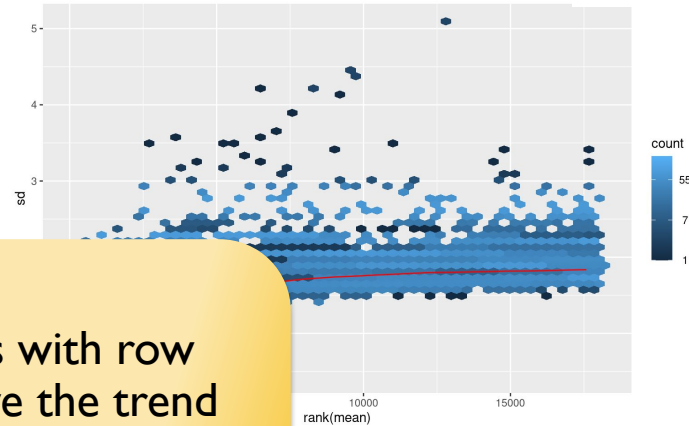


Effect of data transformation

The point of the transformations is to remove the dependence of the variance on the mean particularly the high variance of the logarithmic data when the mean is high.

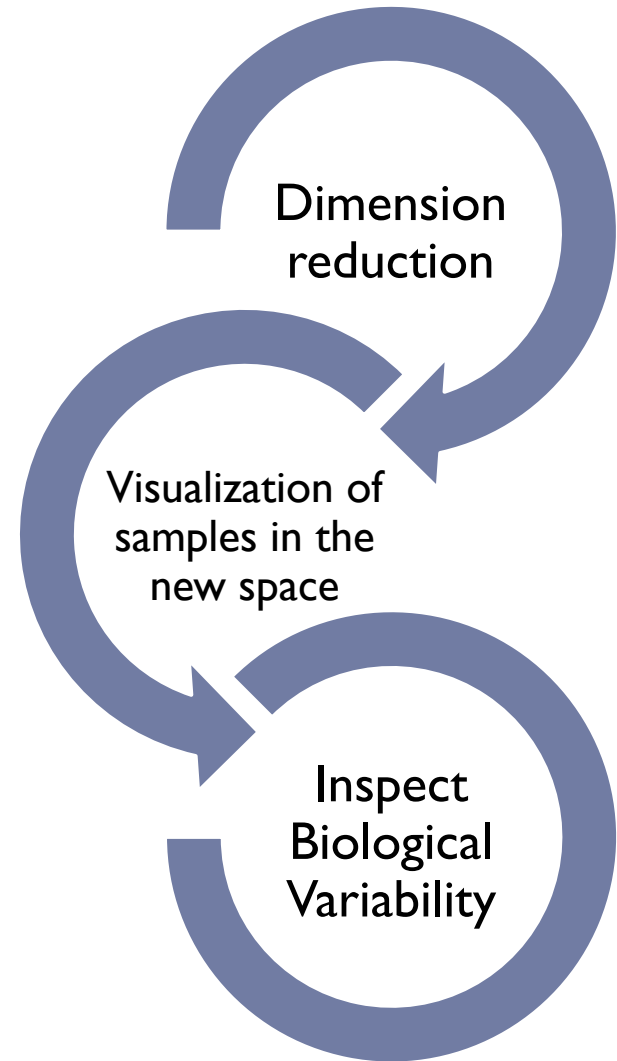
Those genes with row variance above the trend which will allow us to cluster samples into interesting groups

We do not require that all the genes have *exactly* the same variance after transformation, but that the experiment-wide trend has flattened.

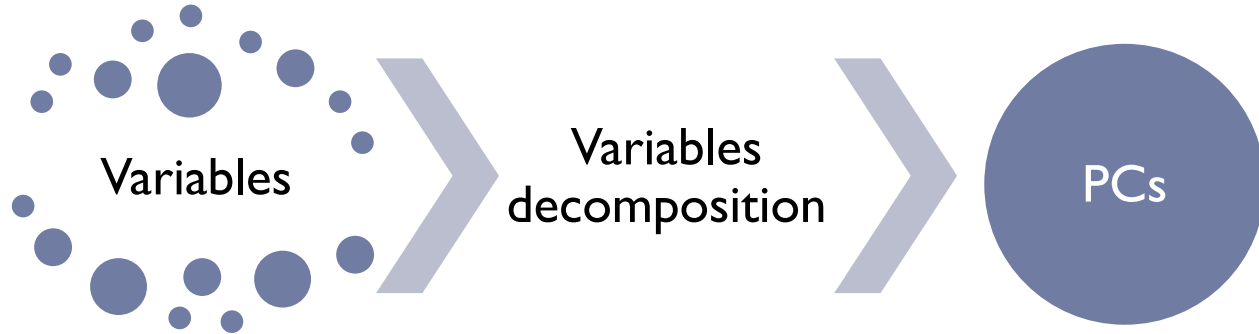


Multivariate analysis

- Multivariate Analysis is helpful for quality control showing distances, in terms of biological coefficient of variation, between samples
- What do you think of the quality of the data? Can we anticipate if exist an important biological variation?

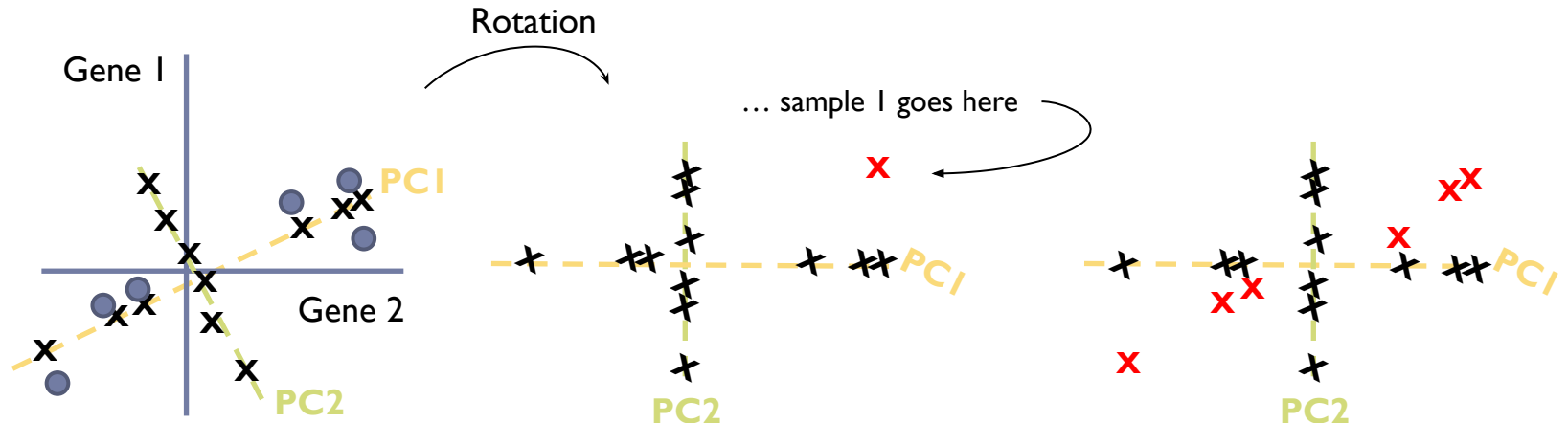


Principal Component Analysis (PCA)



Original
Dimensions

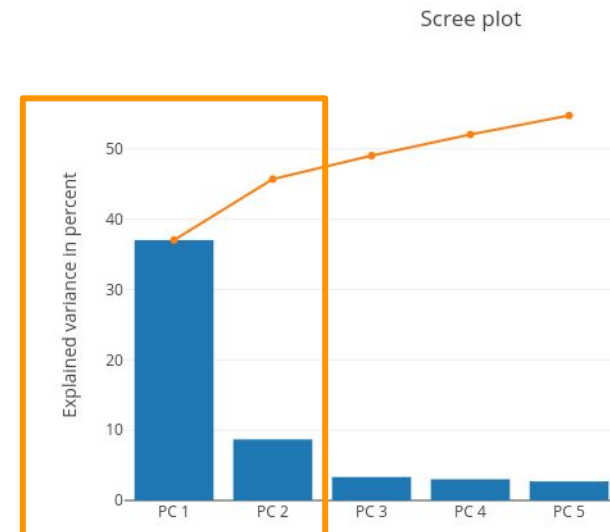
New
Dimensions



PCA

- Each component (PC) is a linear combination of **all** the original variables;
- They reproduce, in descending hierarchical order, the maximum variance reproducible in that turn
- The advantage of PCA is the ability to condense most of the variances and covariances present in the initial set of variables into the first components
- Thus, considering only the first principal components we obtain the best possible synthesis for the thousands of initial genes

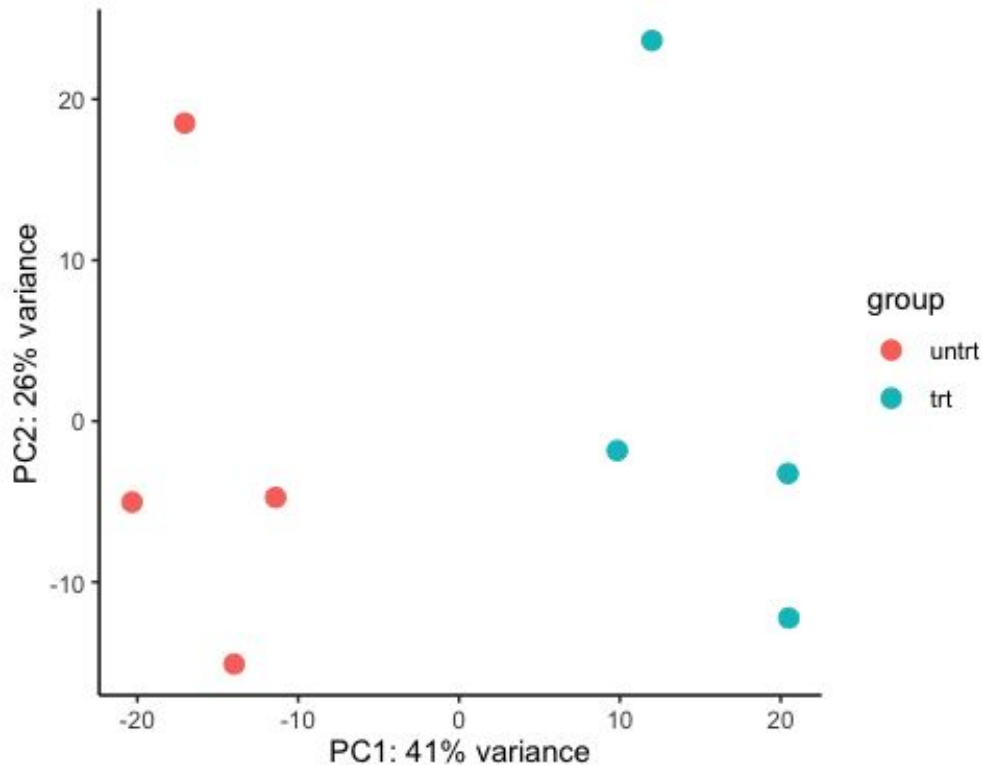
■ explained variance
— cumulative explained variance



PCA



EXAMPLE



- The first two components explain about 67% of the variability
- The first component is clearly associated to cells treatment
- Points that are close together correspond to samples that have similar values on the variables (genes)



Differential Expression Analysis

ASSUMPTIONS:

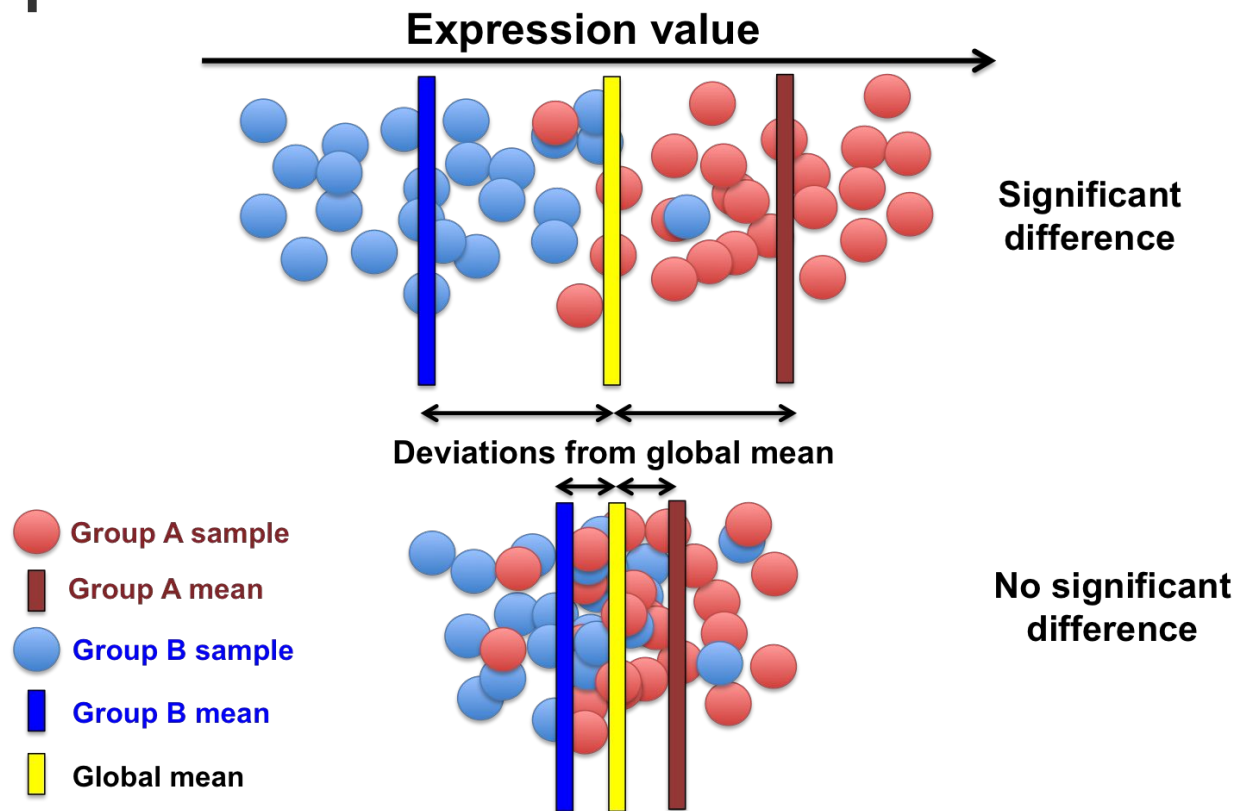
- We attend low number of genes differentially expressed (lower than 10%)
- Equal distribution of genes up and down deregulated
- The differential expression of genes is independent from the mean of expression

The goal of differential expression analysis is to perform statistical analysis to discover **changes in expression levels** of genes between experimental groups with **replicated samples**



Statistical Test

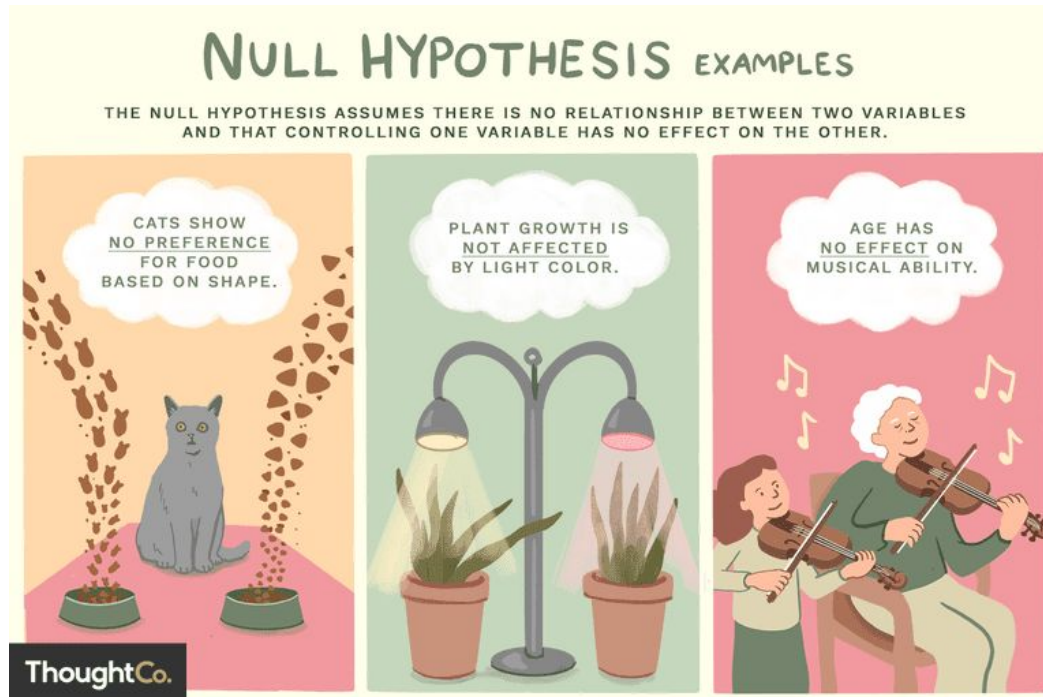
- Essentially, it aims at comparing the **average expression of a gene in group A** with the **average expression of this gene in group B**



Hypothesis Testing

Two hypotheses should be described upon designing an experiment:

- The NULL hypothesis H_0 : that there is NO difference of means
- The alternative hypothesis H_1 : there is difference



Statistically significant

- The **significance level** is the probability to reject the null hypothesis when it is assumed to be true

The significance level, *alpha*, is just the **threshold** of how big at most this error can be. Or, to put it another way, it is the maximum probability of false positive that can be accepted!

1/20=0.05 is the classical threshold

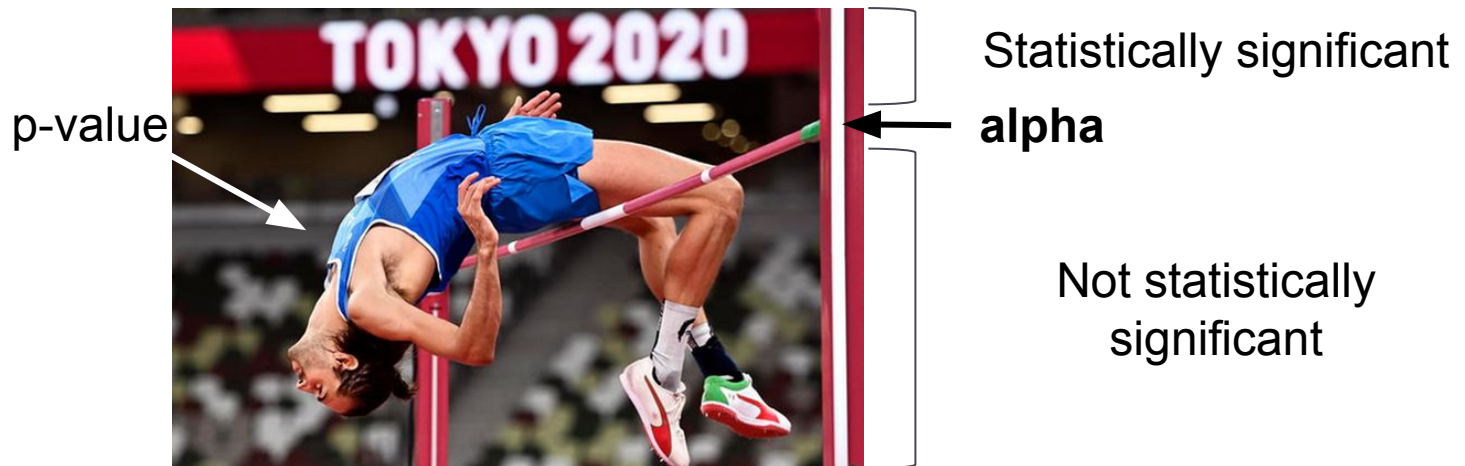
We accept that 1 in 20 times the observed difference could be due to chance



How to interpret p-value

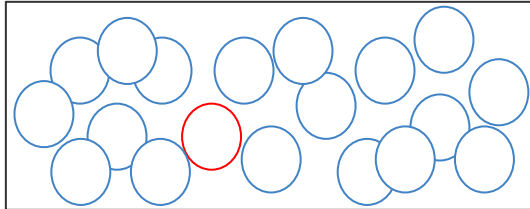
- Once the hypothesis are drawn and the significance level set, we perform a statistical test...
- The agreement between your calculated test statistic and the predicted values is described by the **p value**:

*If the p -value is **below** the significance level we can reject the null hypothesis in favor of the alternative hypothesis with statistical significance!*



Multiple testing correction

The incidence of false positives is proportional to the number of tests performed and the critical significance level (p-value cutoff):



19 blue balls

1 red ball

- What are the odds of randomly sampling the red ball by chance? It is 1 out of 20.
- Consider a case where you have 20 hypotheses to test, and a significance level of 0.05. What's the probability of observing at least one significant result just due to chance?
$$P(\text{at least one significant result}) = 1 - P(\text{no significant results}) = 1 - (1 - 0.05)^{20} \approx 0.64$$
- So, with 20 tests being considered, we have a 64% chance of observing at least one significant result, even if all of the tests are actually not significant.

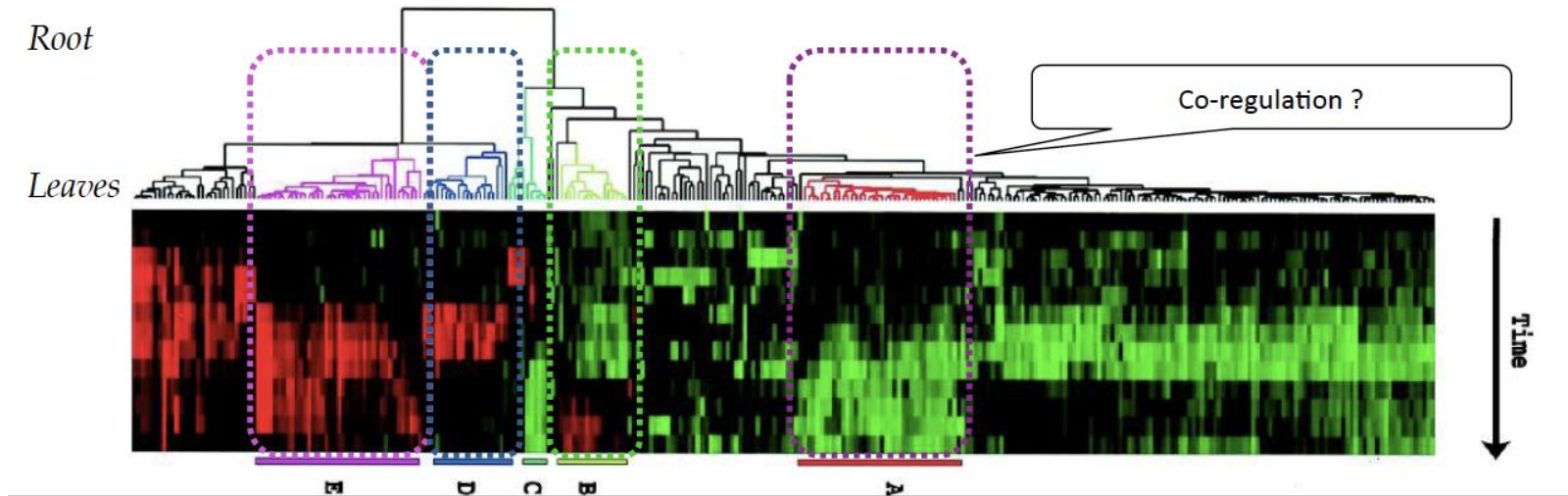
In genomics and other biology-related fields the probability of getting a significant result simply due to chance keeps going up:

If 10,000 genes are tested, 5% or 500 genes might be called significant by chance alone !

*Multiple testing correction adjusts the individual p-value for each gene to keep the overall error rate (or false positive rate) to less than or equal to the **user-specified** p-value cutoff or error rate*

Cluster analysis

- Purpose : cluster samples according to gene expression
- Output : dendrogram/tree



Cluster analysis: distance matrix

- The Hierarchical Clustering technique builds clusters based on the similarity between different objects in the set:

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix} \quad \triangleright \quad \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

$$dist(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \quad \text{Euclidean}$$

Which measures ?

$$dist(x,y) = \sum_i |x_i - y_i| \quad \text{Manhattan}$$

$$sim(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \cdot \sqrt{\sum_i (y_i - \bar{y})^2}} \quad \text{Pearson Correlation}$$



Cluster analysis: steps

□ Starting point: Distance matrix

1. Take the minimum distance and connect the two elements
2. Re-calculate the distance matrix ($n-1 \times n-1$) using the **linkage**
3. Take the minimum distance of the new matrix and connect the two elements
4. Re-calculate the distance matrix ($n-2 \times n-2$) using the **linkage**

...

□ Stop: when the distance matrix had dimension (2×2)

1

	A	B	C	D	E
A	0				
B	5	0			
C	10	3	0		
D	15	6	7	0	
E	20	8	2	11	0

2

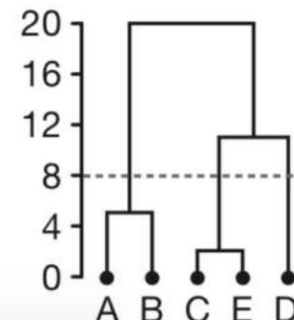
	A	B	CE	D
A	0			
B	5	0		
CE	20	8	0	
D	15	6	11	0

3

	AB	CE	D
AB	0		
CE	20	0	
D	15	11	0

4

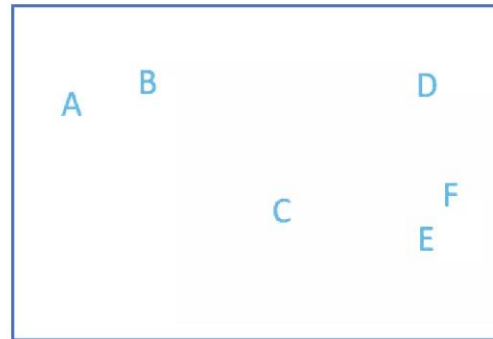
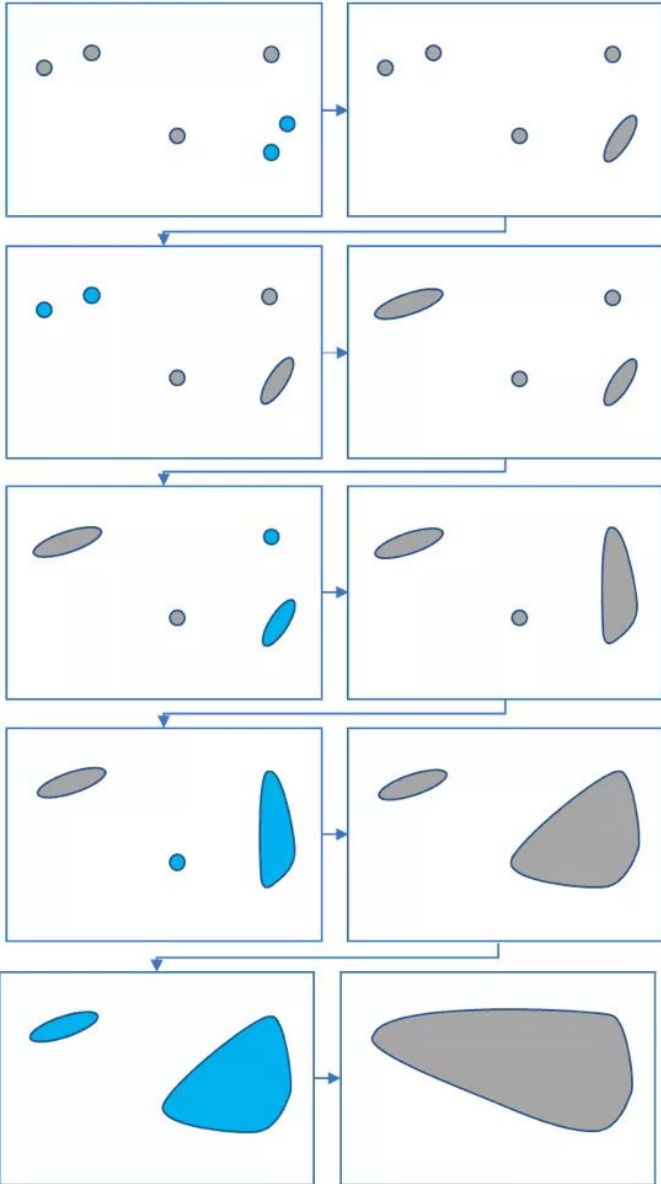
	AB	CED
AB	0	
CED	20	0



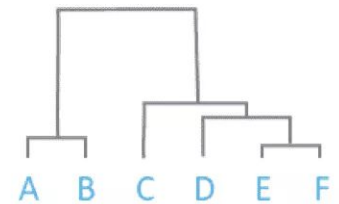
Cluster analysis: visual interpretation

Identify the two clusters that
are **closest** together

Merge the two most similar
clusters

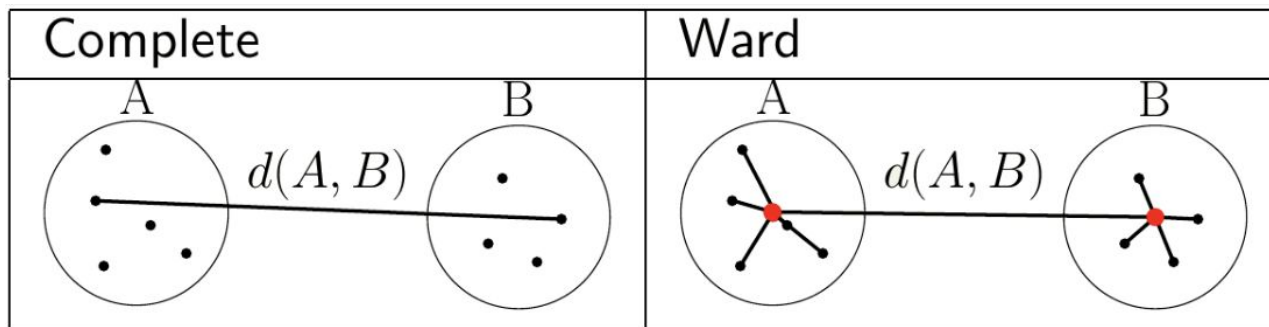


Dendrogram

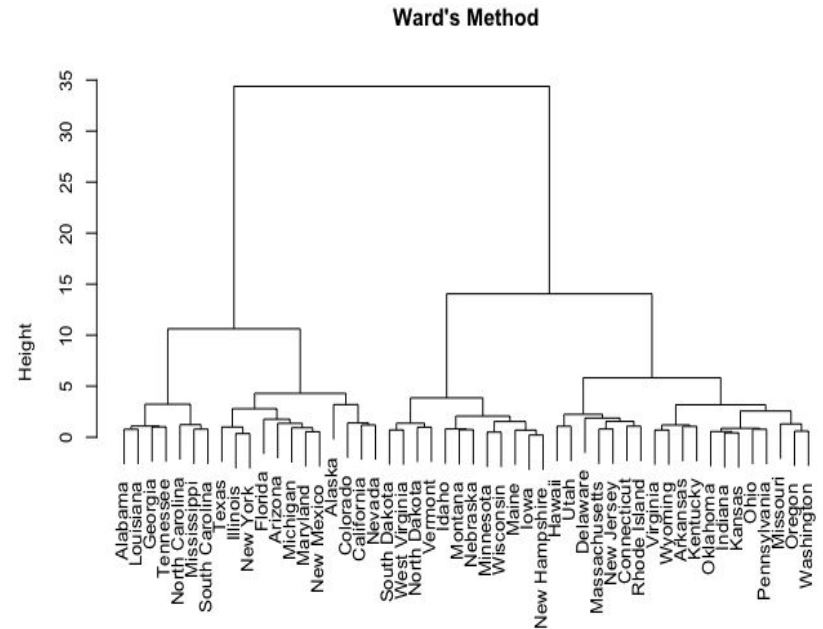
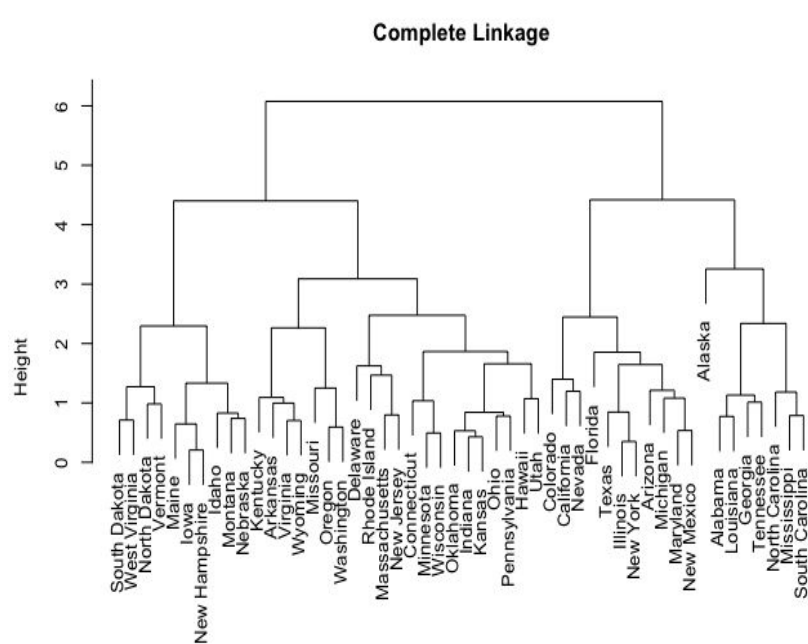


Cluster Analysis: Linkage methods

- A bigger question is: **How do we measure the dissimilarity between two clusters of observations?** A number of different cluster agglomeration methods (i.e, linkage methods) have been developed to answer to this question. The most common types methods are:
 - Complete linkage : tends to produce compact clusters maximising distance between less similar points in two clusters
 - Ward linkage: tends to also produces compact clusters but it minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged

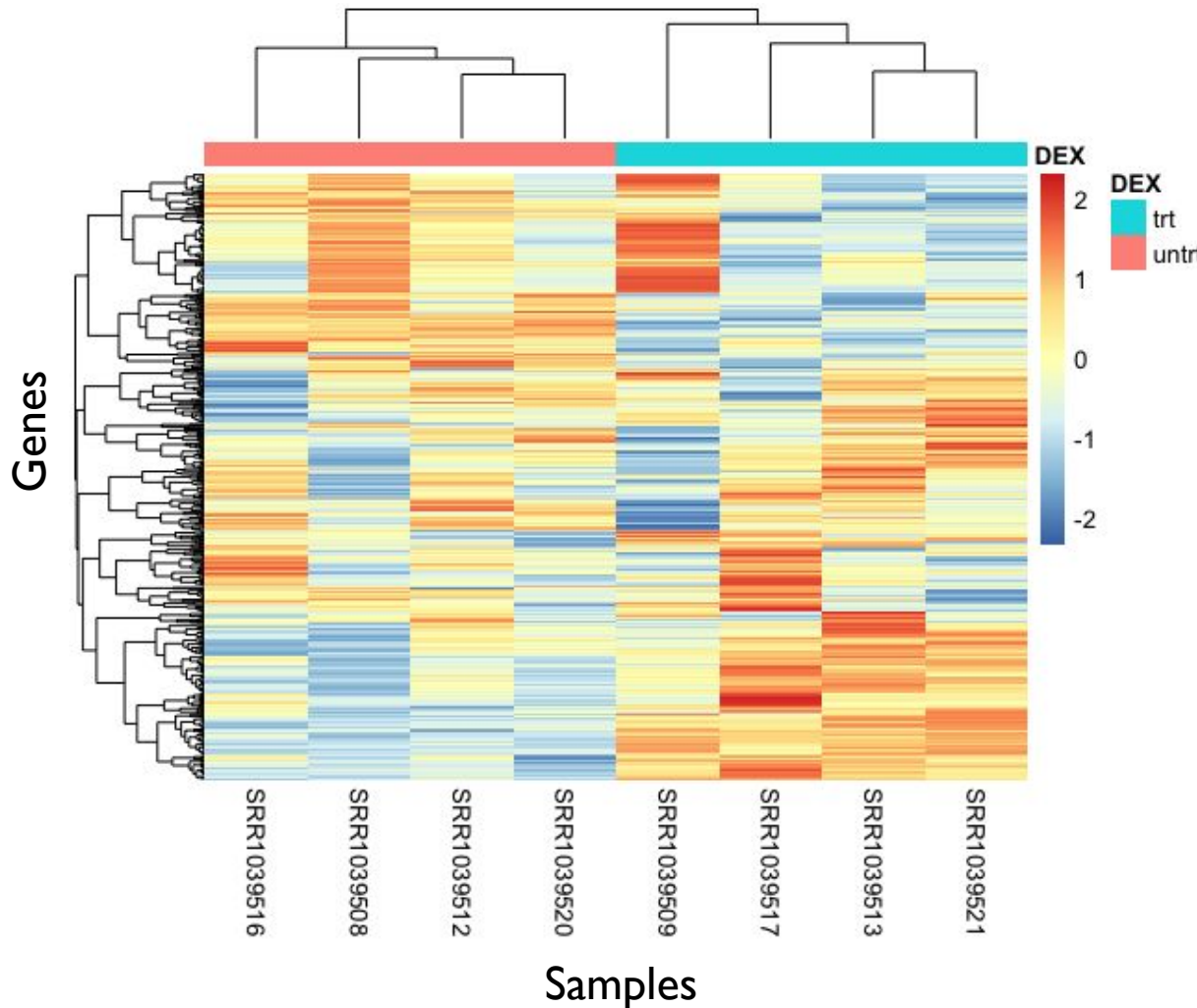


Cluster Analysis: Linkage methods



Heatmap with most varied genes

EXAMPLE



- Hierarchical clustering
- Clustering distance of rows: pearson correlation
- Clustering distance of cols: Euclidean
- Clustering method: Complete

Beyond gene expression ...

- Differentially expressed gene were selected for follow-up based on each gene's potential to be a **novel** steroid responsiveness gene
- Molecular mechanism can be developed and further investigated by functional experiments...

RNA-Seq transcriptome profiling identifies
CRISPLD2 as a glucocorticoid responsive gene that
modulates cytokine function in airway smooth
muscle cells

