UCLA Institute for Quantitative and Computational Biology

W5b: RNA-Seq II Analysis

Workshop W5b, 10th March 2020

Day 1

Karolina E. Kaczor-Urbanowicz, PhD

UCLA School of Dentistry, Department of Oral Biology & Medicine

(Area of interests: salivary RNA Sequencing, bioinformatics and biomarker development for many oral and systematic diseases)

&

Nicolas Rochette, PhD

UCLA Department of Ecology and Evolutionary Biology & the Institute for Society and Genetics

(Area of interests: population genomics and evolutionary physiology, the dynamics of adaptation)

W5b: RNA-Seq II Analysis

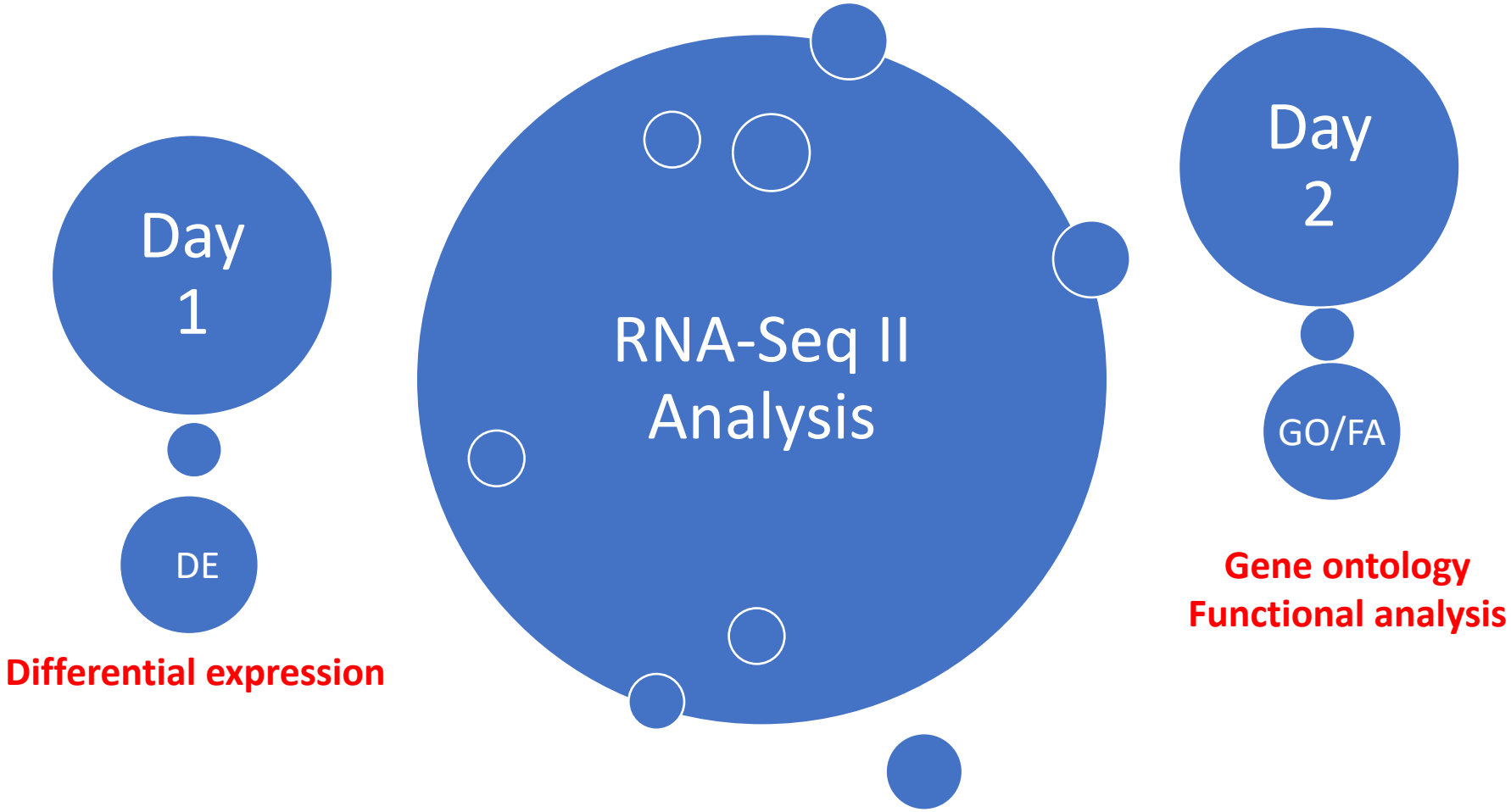
Workshop Description

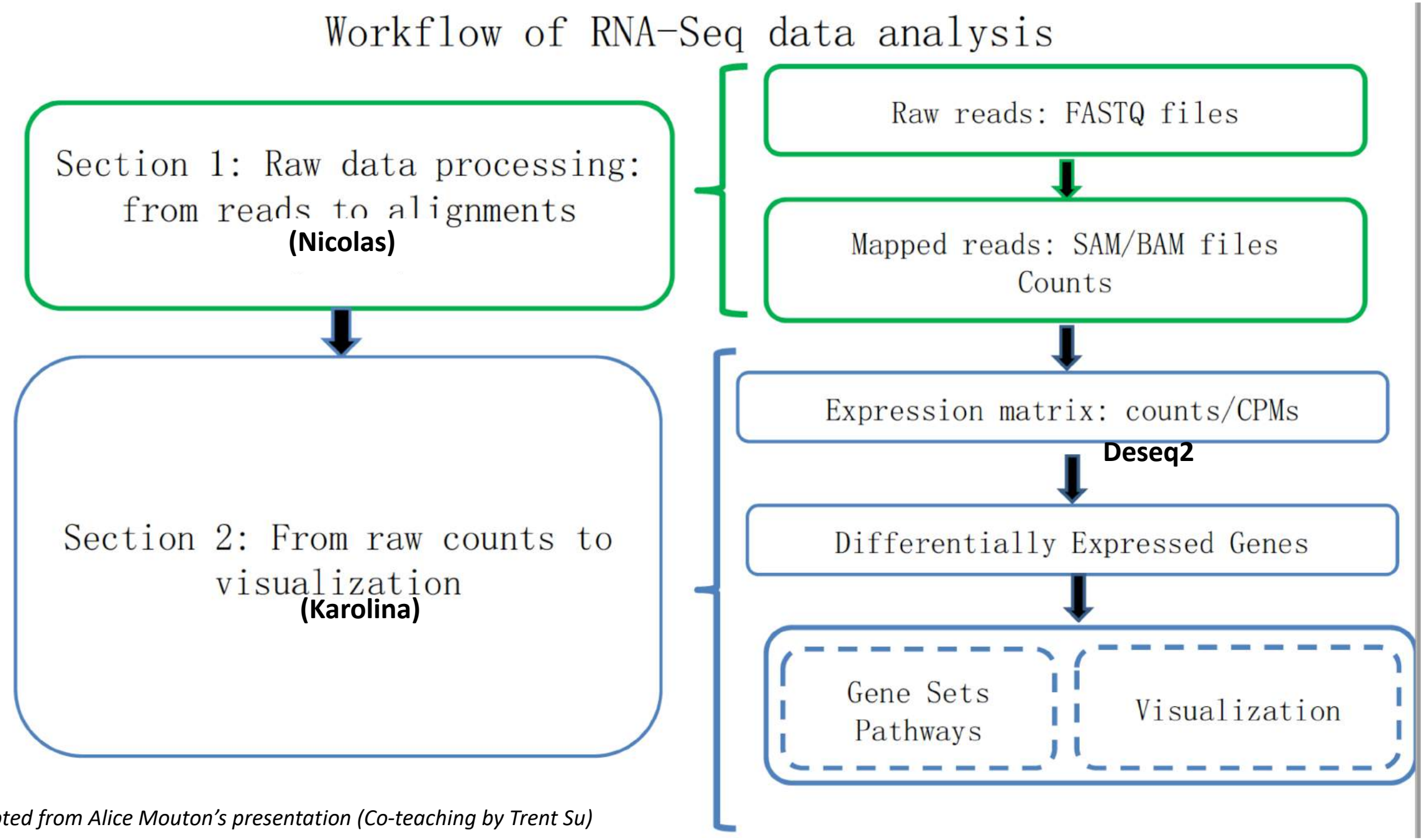
RNA-seq II aims to provide tools for analysis of gene expression data from read counts to biology. To facilitate learning, the workshop will use a real case study based approach appropriate for Illumina read data (same as RNA-seq I).

From **gene expression to biology** (gene expression analysis):

- Normalization of gene expression
- Exploratory analysis of RNA-Seq data
- Differential gene expression analysis
- Gene set and pathway enrichment analysis
- Visualization of gene expression

W5b: RNA-Seq II Analysis

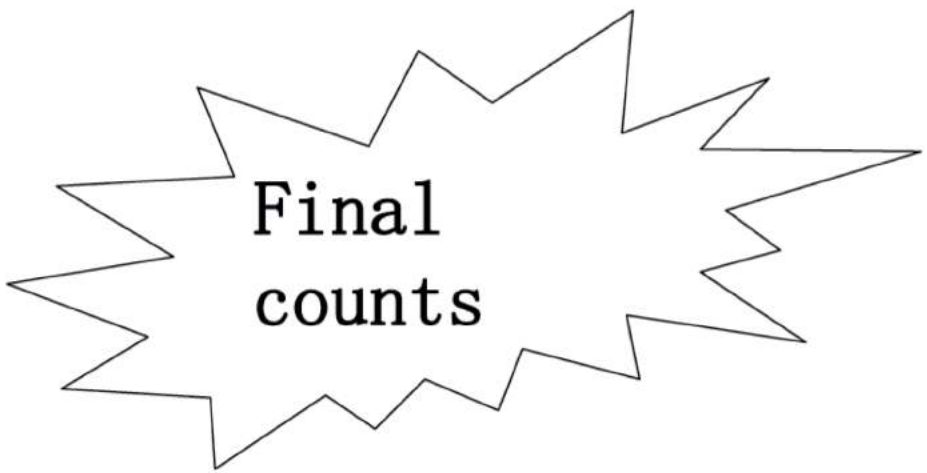




Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

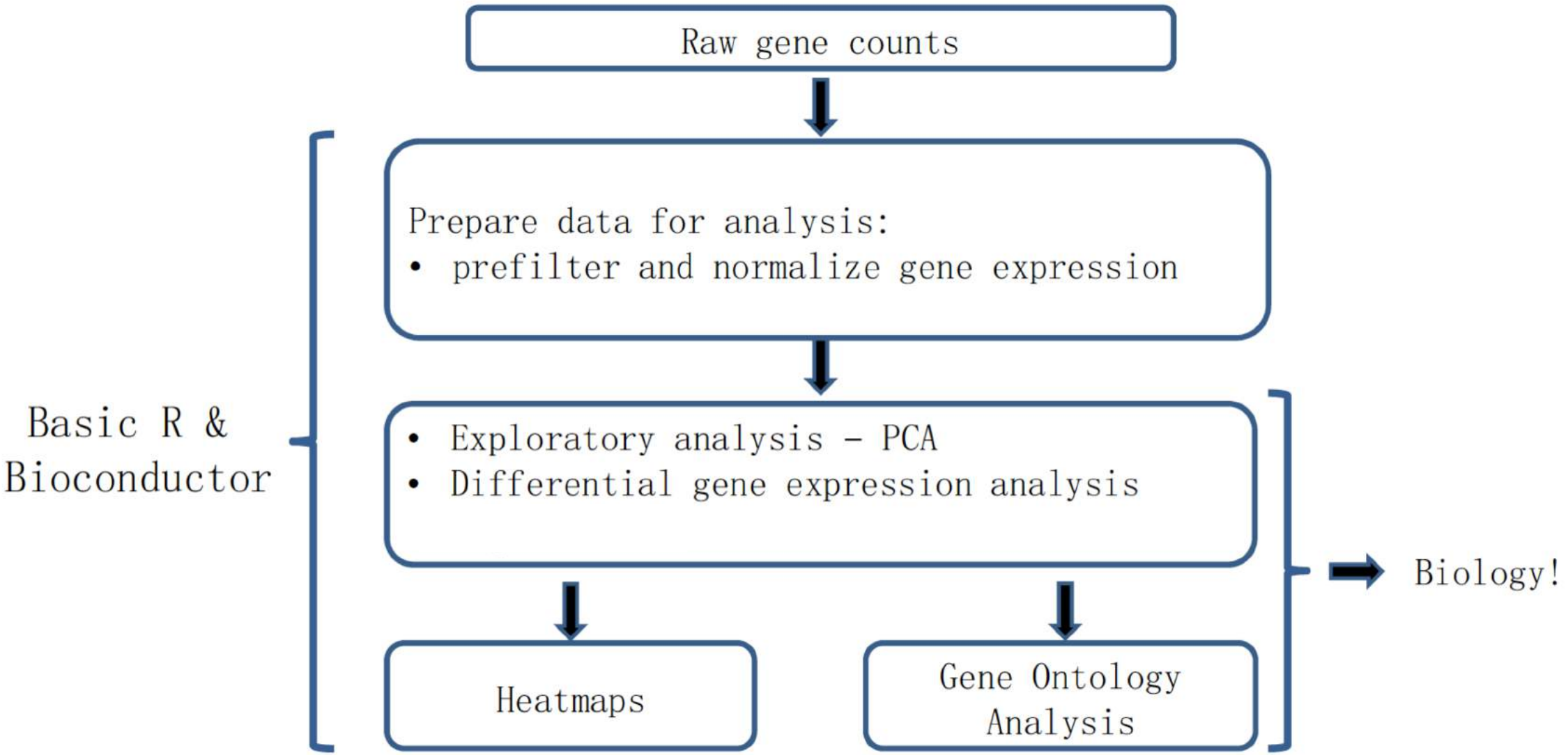
Background

	A	B	C	D	E	F	G	H
1	Ensembl_Gene_ID	661M_PEr1	828M_PEr1	759F_PEr1	870F_PEr1	871M_PEr1	825F_PEr1	829F_PEr1
2	ENSCAFG00000000001	37	26	46	41	40	13	34
3	ENSCAFG00000000002	0	0	3	3	1	1	6
4	ENSCAFG00000000005	0	0	4	7	1	0	1
5	ENSCAFG00000000007	271	728	325	244	318	382	334
6	ENSCAFG00000000008	72	131	98	76	30	100	132
7	ENSCAFG00000000009	128	364	136	163	138	313	150
8	ENSCAFG00000000010	360	885	442	325	368	488	297
9	ENSCAFG00000000011	68	243	96	59	105	111	86
10	ENSCAFG00000000012	626	1119	852	565	590	936	898
11	ENSCAFG00000000013	10	2	4	3	5	0	6



Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

From Counts to Biology



Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

1. Concept of normalization

What can you tell just by looking the counts?

Gene 1	0	0	0
Gene 2	15	22	40
Gene 3	10	13	27
Gene 4	265	300	350
Gene 5	1500	1200	2400
Gene 6	700	500	1000
Gene 7	…..	…	…
……	…	….	….
	2490	2035	3817

➔ Differences between genes

➔ Differences between samples (sequencing depth)

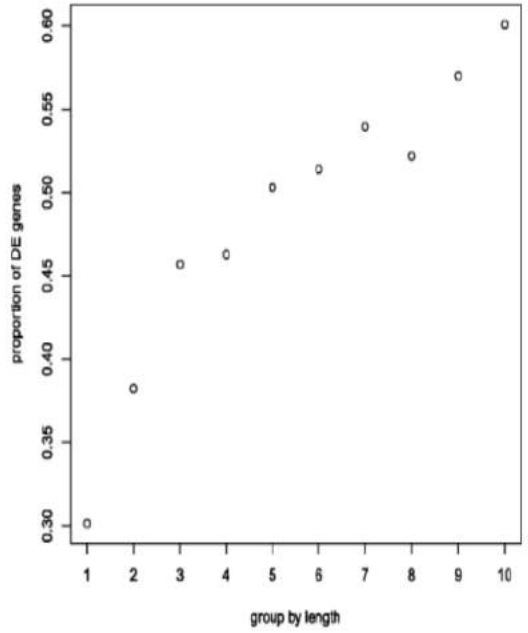
Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

1. Concept of normalization

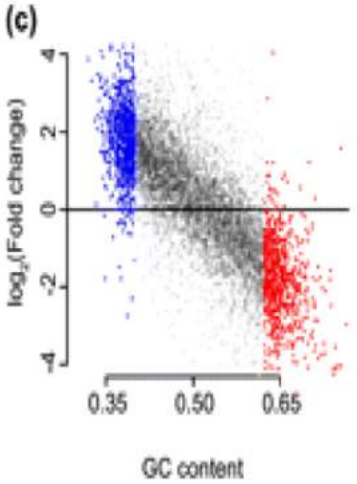
Normalization => raw count are adjusted to account for factors that prevent direct comparison of expression (identify and remove systematic technical differences between samples)

Source of Variability

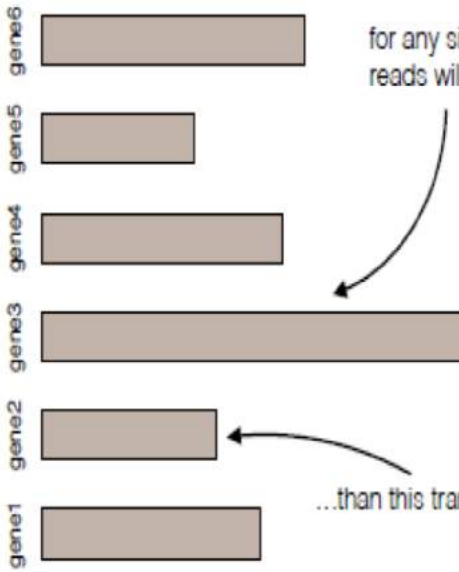
1. Within samples : Gene length
   GC content



Ren *et al* 201



Hansen *et al*. 2012



for any single expressed mRNA, more reads will be generated from this transcript

...than this transcript

Tung 2016 Pers. Comm

You don't have to worry about that in the context of DE between samples!

Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

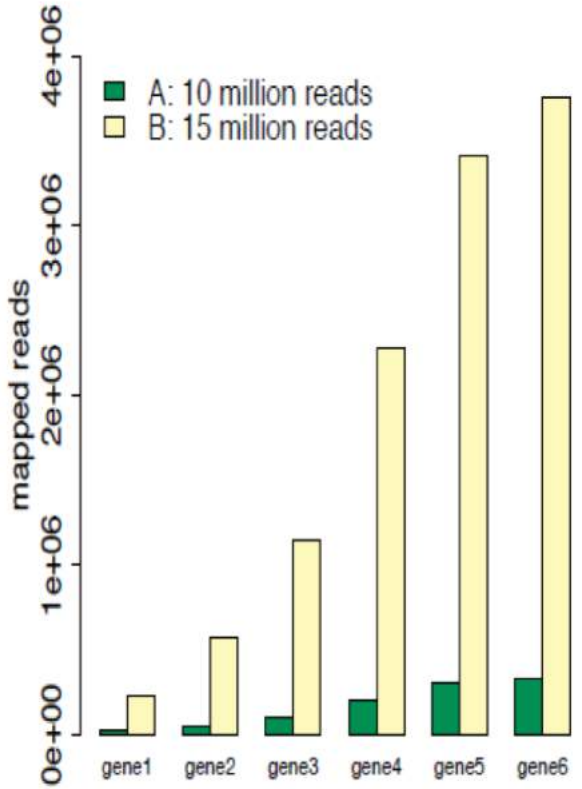
1. Concept of normalization

Normalization => raw count are adjusted to account for factors that prevent direct comparison of expression (identify and remove systematic technical differences between samples)

Source of Variability

2. Between samples

- Sequencing depth
- Batch effects
- …



Tung 2016 Pers. Comm

Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

1. Concept of normalization

Global procedure of normalization (without length => between samples)

adjustments of values or distributions in statistics by using correction multiplicative factors

	control				treated		
Gene 1	5	1	0	0	4	0	0
Gene 2	0	2	1	2	1	0	0
Gene 3	92	161	76	70	140	88	70
:		:				:	
:		:				:	
:		:				:	
Gene G	15	25	9	5	20	14	17

Correction multiplicative factor:

C_j	1.1	1.6	0.6	0.7	1.4	0.7	0.8

Column multiplication by factor C_j:

Gene 3	92	161	76	70	140	88	70
C_j	1.1	1.6	0.6	0.7	1.4	0.7	0.8
							x
Gene 3	101.2	257.6	45.6	49	196	61.6	56

From Gonzalez 2014 Statistical analysis of RNA-Seq data

Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

Comon Normalization methods

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

RPKM/FPKM is not recommended

(because the normalized count values output by the RPKM/FPKM method are not comparable between samples)

Using RPKM/FPKM normalization, the total number of RPKM/FPKM normalized counts for each sample will be different. Therefore, you cannot compare the normalized counts for each gene equally between samples.

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

DeSeq2-normalized counts: Median of ratios method

DESeq2-normalized counts: Median of ratios method

Since tools for differential expression analysis are comparing the counts between sample groups for the same gene, gene length does not need to be accounted for by the tool. However, **sequencing depth** and **RNA composition** do need to be taken into account.

To normalize for sequencing depth and RNA composition, DESeq2 uses the median of ratios method. On the user-end there is only one step, but on the back-end there are multiple steps involved, as described below.

> **NOTE:** The steps below describe in detail some of the steps performed by DESeq2 when you run a single function to get DE genes. Basically, for a typical RNA-seq analysis, **you would not run these steps individually**.

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

For each gene, a pseudo-reference sample is created that is equal to the geometric mean across all samples.

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**

Step 2: calculates ratio of each sample to the reference

For every gene in a sample, the ratios (sample/ref) are calculated (as shown below). This is performed for each sample in the dataset. Since the majority of genes are not differentially expressed, the majority of genes in each sample should have similar ratios within the sample.

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

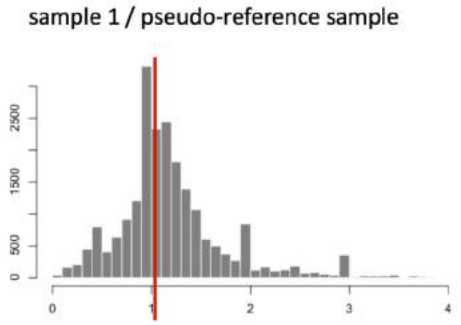
Step 3: calculate the normalization factor for each sample (size factor)

The median value (column-wise for the above table) of all ratios for a given sample is taken as the normalization factor (size factor) for that sample, as calculated below. Notice that the differentially expressed genes should not affect the median value:

```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))

normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```

The figure below illustrates the median value for the distribution of all gene ratios for a single sample (frequency is on the y-axis).



The median of ratios method makes the assumption that not ALL genes are differentially expressed; therefore, the normalization factors should account for sequencing depth and RNA composition of the sample (large outlier genes will not represent the median ratio values). **This method is robust to imbalance in up-/down-regulation and large numbers of differentially expressed genes.**

> Usually these size factors are around 1, if you see large variations between samples it is important to take note since it might indicate the presence of extreme outliers.

Step 4: calculate the normalized count values using the normalization factor

This is performed by dividing each raw count value in a given sample by that sample's normalization factor to generate normalized count values. This is performed for all count values (every gene in every sample). For example, if the median ratio for SampleA was 1.3 and the median ratio for SampleB was 0.77, you could calculate normalized counts as follows:

SampleA median ratio = 1.3

SampleB median ratio = 0.77

Raw Counts

gene	sampleA	sampleB
EF2A	1489	906
ABCD1	22	13
...	...	...

Normalized Counts

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...

> Please note that normalized count values are not whole numbers.

Differential Expression Analysis

How do the expression levels differ across several conditions?

Challenges:

1. Count data is discrete – no normal distribution. Cannot perform t-test.
2. Small number of replicates – cannot use permutation methods.
3. Account for variability in measurements across biological replicates of an experiment.

*Adopted from Soumya Luthra's presentation ("*RNA-Seq analysis in R (Bioconductor)*")*

Poisson Distribution?

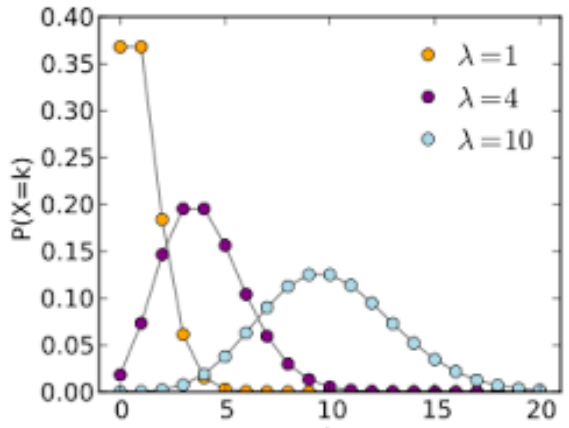
In probability theory and statistics, the **Poisson distribution** is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume, e.g. the number of phone calls received by a call center per hour.

- Mean = Variance

❖ **Mean** is the average of the numbers

❖ **Variance** (σ^2) in statistics is a measurement of the spread between numbers in a data set. That is, *it measures how far each number in the set is from the mean and therefore from every other number in the set.*

- Is read count data Poisson Distributed?
- **Over-dispersion** - variance in RNA-Seq measurements of gene expression are larger than the theoretical values



❖ In statistics, **overdispersion** is the presence of greater variability in a data set than would be expected based on a given statistical model.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

*Adopted from Soumya Luthra's presentation ("*RNA-Seq analysis in R (Bioconductor)"*)*

Negative Binomial Distribution

In probability theory and statistics, the **negative binomial distribution** is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures (denoted *r*) occurs. For example, if we define a 1 as failure, all non-1s as successes, and we throw a dice repeatedly until 1 appears the third time (*r* = three failures), then the probability distribution of the number of non-1s that appeared will be a negative binomial distribution.

- NB has been shown to be a good fit to RNA-Seq data
- It is flexible enough to account for biological variability

Model:

- Makes the assumption that an observation say $Y_{gj\ (observed\ number)}$ of reads for gene g sample j, has a mean μ_{gj} and a variance of $\mu_{gj} + \Phi_g \mu^2$, where Φ_g represents over-dispersion relative to poisson distribution.
- The mean parameter depends on the sequencing depth as well as on the amount of RNA from gene in the sample
- Obtaining good estimates of each gene's dispersion is critical for statistical testing.

Tools:

- EdgeR and DESeq count data using a Negative Binomial Distribution and perform statistical tests for differential expression.

*Adopted from Soumya Luthra's presentation ("*RNA-Seq analysis in R (Bioconductor)*")*

edgeR

EdgeR treats the Poisson variance as simple sampling variance, and refers to the dispersion estimate as the "biological coefficient of variation."

Estimating dispersion:

- EdgeR shares information across genes to determine a common dispersion. It then calculates a dispersion estimate per gene and shrinks it towards the common dispersion. The gene-specific (referred to in edgeR as tagwise) dispersion estimates are used in the test for differential expression.

Statistical Test:

- **Simple design** - Fischer's exact test

(statistical significance test that is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (e.g., P-value) can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as with many statistical tests).

- **Complex design** - Generalized linear model (GLM) framework

(In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.)

*Adopted from Soumya Luthra's presentation ("*RNA-Seq analysis in R (Bioconductor)*")*

DESeq

- Differential gene expression from count data based on negative binomial distribution.
- Offers two transformations for stabilizing the variance of count data:
  - **VST** – Variance stabilizing transformation
  - **Regularized log transformation** (**rlog**)

http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Variance stabilizing transformation

Above, we used a parametric fit for the dispersion. In this case, the closed-form expression for the variance stabilizing transformation is used by the *vst* function. If a local fit is used (option `fitType="locfit"` to *estimateDispersions*) a numerical integration is used instead. The transforme data should be approximated variance stabilized and also includes correction for size factors or normalization factors. The transformed data is o the log2 scale for large counts.

Regularized log transformation

The function *rlog*, stands for *regularized log*, transforming the original count data to the log2 scale by fitting a model with a term for each sample and a prior distribution on the coefficients which is estimated from the data. This is the same kind of shrinkage (sometimes referred to as regularization, or moderation) of log fold changes used by the *DESeq* and *nbinomWaldTest*. The resulting data contains elements defined as:

$$\log_2(q_{ij}) = \beta_{i0} + \beta_{ij}$$

where q_{ij} is a parameter proportional to the expected true concentration of fragments for gene i and sample j (see formula below), β_{i0} is an intercept which does not undergo shrinkage, and β_{ij} is the sample-specific effect which is shrunk toward zero based on the dispersion-mean trend over the entire dataset. The trend typically captures high dispersions for low counts, and therefore these genes exhibit higher shrinkage from the *rlog*.

Note that, as q_{ij} represents the part of the mean value μ_{ij} after the size factor s_j has been divided out, it is clear that the rlog transformation inherently accounts for differences in sequencing depth. Without priors, this design matrix would lead to a non-unique solution, however the addition of a prior on non-intercept betas allows for a unique solution to be found.

*Adopted from Soumya Luthra's presentation ("*RNA-Seq analysis in R (Bioconductor)*")*

How do I use VST or rlog data for differential testing?

The variance stabilizing and rlog transformations are provided for applications other than differential testing, for example clustering of samples or other machine learning applications. For differential testing we recommend the *DESeq* function applied to raw counts.

Anders and Huber *Genome Biology* 2010, **11**:R106
http://genomebiology.com/2010/11/10/R106

Genome Biology

METHOD Open Access

Differential expression analysis for sequence count data

Simon Anders*, Wolfgang Huber

Abstract

High-throughput sequencing assays such as RNA-Seq, ChIP-Seq or barcode counting provide quantitative readouts in the form of count data. To infer differential signal in such data correctly and with good statistical power, estimation of data variability throughout the dynamic range and a suitable error model are required. We propose a method based on the negative binomial distribution, with variance and mean linked by local regression and present an implementation, *DESeq*, as an R/Bioconductor package.

Background

High-throughput sequencing of DNA fragments is used in a range of quantitative assays. A common feature between these assays is that they sequence large amounts of DNA fragments that reflect, for example, a biological system's repertoire of RNA molecules (RNA-Seq [1,2]) or the DNA or RNA interaction regions of nucleotide binding molecules (ChIP-Seq [3], HITS-CLIP [4]). Typically, these reads are assigned to a class based on their mapping to a common region of the target genome, where each class represents a target transcript, in the case of RNA-Seq, or a binding region, in the case of ChIP-Seq. An important summary statistic is the number of reads in a class; for RNA-Seq, this *read count* has been found to be (to good approximation) linearly related to the abundance of the target transcript [2]. Interest lies in comparing read counts between different biological conditions. In the simplest case, the comparison is done separately, class by class. We will use the term *gene* synonymously to class, even though a class may also refer to, for example, a transcription factor binding site, or even a barcode [5].

We would like to use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

If reads were independently sampled from a population with given, fixed fractions of genes, the read counts would follow a multinomial distribution, which can be approximated by the Poisson distribution.

Consequently, the Poisson distribution has been used to test for differential expression [6,7]. The Poisson distribution has a single parameter, which is uniquely determined by its mean; its variance and all other properties follow from it; in particular, the variance is equal to the mean. However, it has been noted [1,8] that the assumption of Poisson distribution is too restrictive: it predicts smaller variations than what is seen in the data. Therefore, the resulting statistical test does not control type-I error (the probability of false discoveries) as advertised. We show instances for this later, in the Discussion.

To address this so-called overdispersion problem, it has been proposed to model count data with negative binomial (NB) distributions [9], and this approach is used in the *edgeR* package for analysis of SAGE and RNA-Seq [8,10]. The NB distribution has parameters, which are uniquely determined by mean μ and variance σ^2. However, the number of replicates in data sets of interest is often too small to estimate both parameters, mean and variance, reliably for each gene. For *edgeR*, Robinson and Smyth assumed [11] that mean and variance are related by $\sigma^2 = \mu + \alpha\mu^2$, with a single proportionality constant α that is the same throughout the experiment and that can be estimated from the data. Hence, only one parameter needs to be estimated for each gene, allowing application to experiments with small numbers of replicates.

In this paper, we extend this model by allowing more general, data-driven relationships of variance and mean, provide an effective algorithm for fitting the model to

* Correspondence: sanders@fs.tum.de
European Molecular Biology Laboratory, Mayerhofstraße 1, 69117 Heidelberg, Germany

Genome Biol. 2010;11(10):R106. doi: 10.1186/gb-2010-11-10-r106. Epub 2010 Oct 27.

Differential expression analysis for sequence count data.

Anders S[1], Huber W.

Author information

1 European Molecular Biology Laboratory, Mayerhofstraße 1, 69117 Heidelberg, Germany. sanders@fs.tum.de

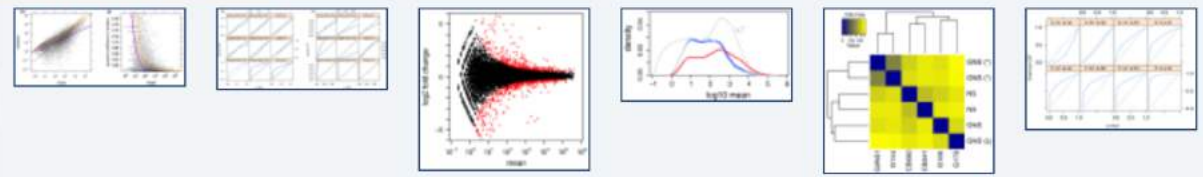
Abstract

High-throughput sequencing assays such as RNA-Seq, ChIP-Seq or barcode counting provide quantitative readouts in the form of count data. To infer differential signal in such data correctly and with good statistical power, estimation of data variability throughout the dynamic range and a suitable error model are required. We propose a method based on the negative binomial distribution, with variance and mean linked by local regression and present an implementation, DESeq, as an R/Bioconductor package.

PMID: 20979621 PMCID: PMC3218662 DOI: 10.1186/gb-2010-11-10-r106

[Indexed for MEDLINE] Free PMC Article

Images from this publication. See all images (7) Free text



Example of DE using DESeq2

mRNAs

adenine (3a) vs. control (1c)

mRNAs

DE mRNA expression, p <0.05 (Top 1000 mRNAs)

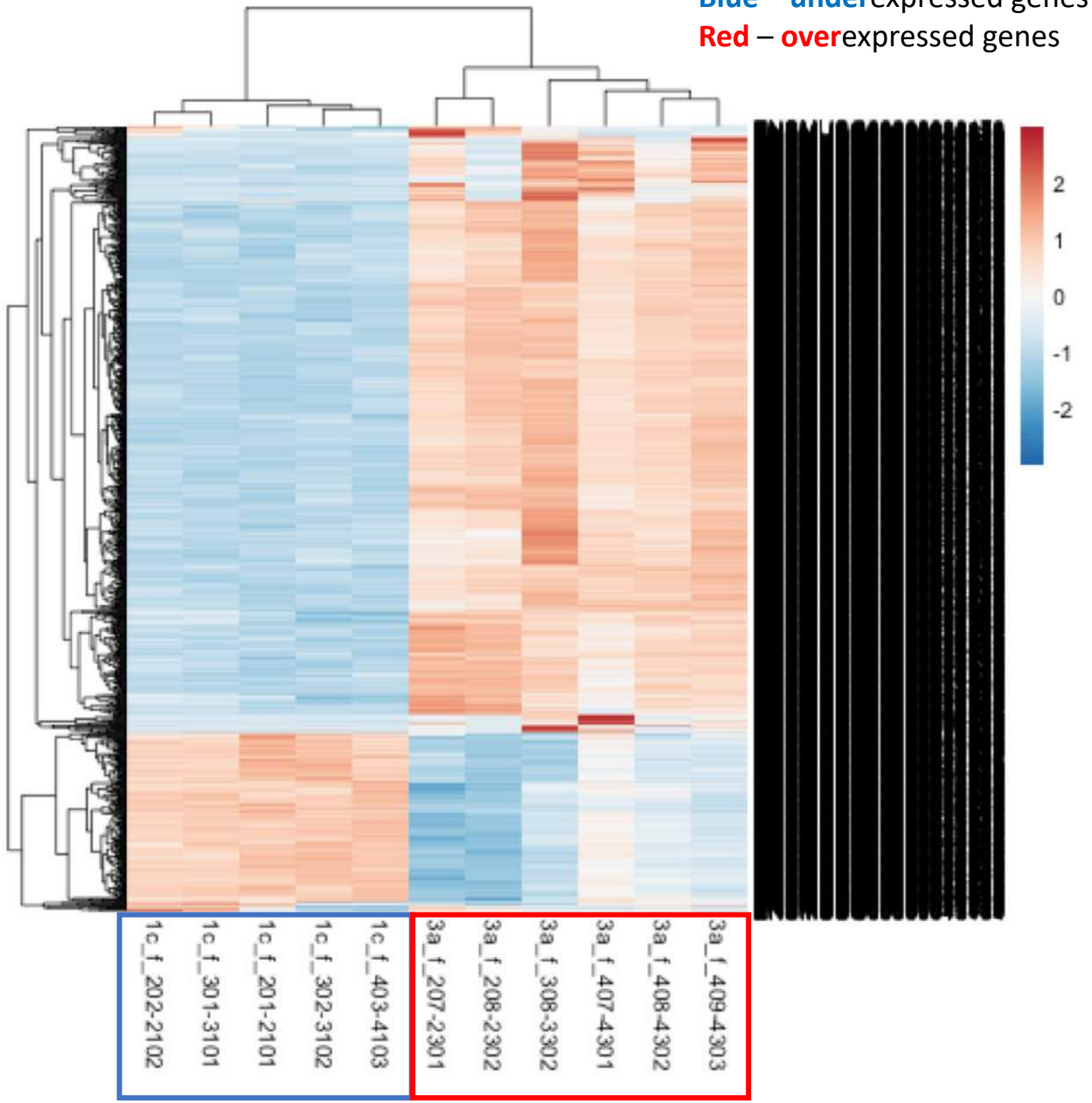
adenine (3a) vs. control (1c)

Blue – underexpressed genes
Red – overexpressed genes

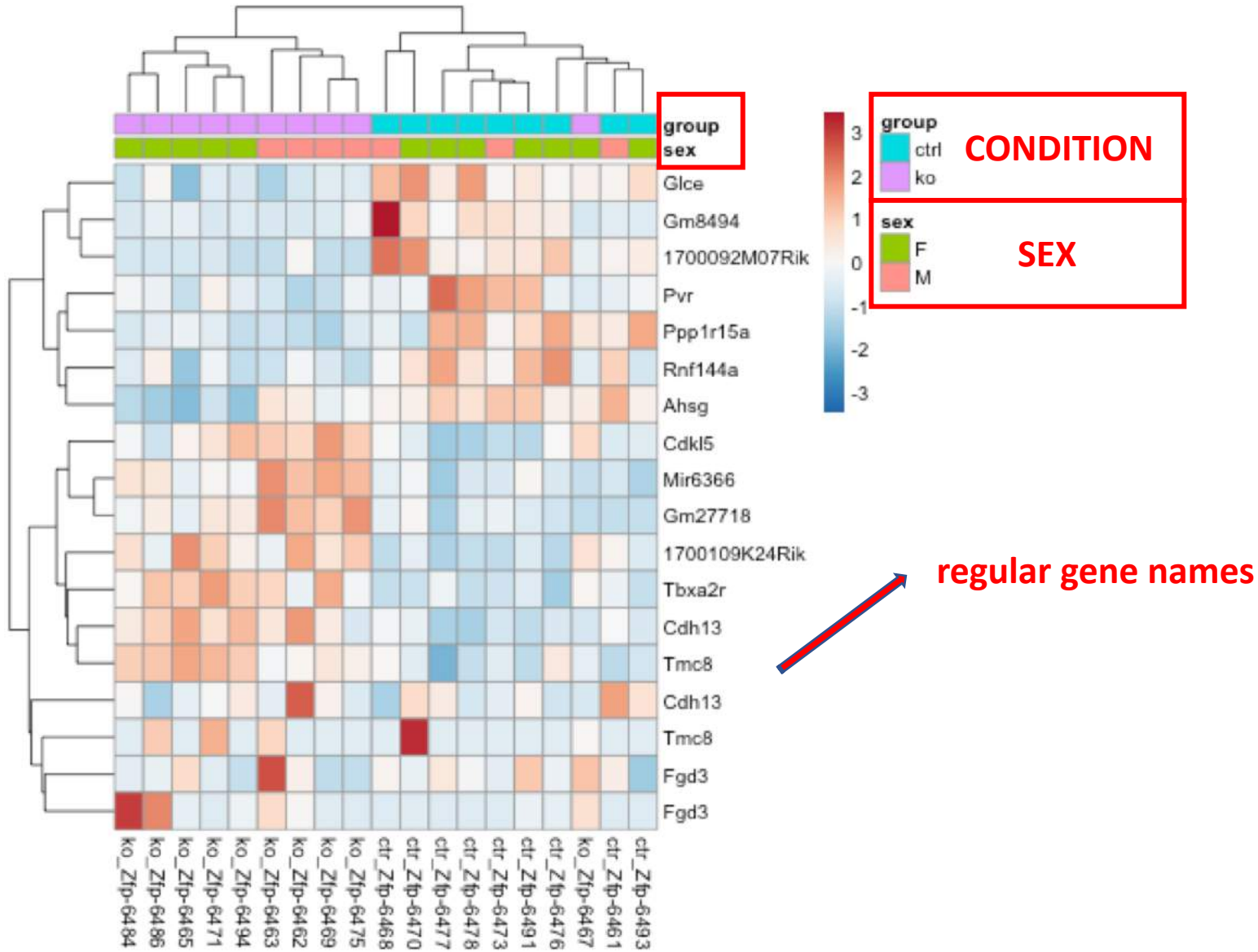
Dendrogram at the side shows us a hierarchical clustering for the genes.
Since the clustering is only relevant for genes that actually carry signal, one usually carries it out only for a subset of most highly variable genes (genes with the highest variance across samples)

The heatmap becomes more interesting if we do not look at absolute expression strength but rather at **the amount by which each gene deviates in a specific sample from the gene's average across all samples.** Hence, we center and scale each genes' values across samples, and plot a heatmap.

Heatmap is a graphical representation of data where individual values contained in a matrix are represented as colors. It allows to visualize expression of many genes in many samples.



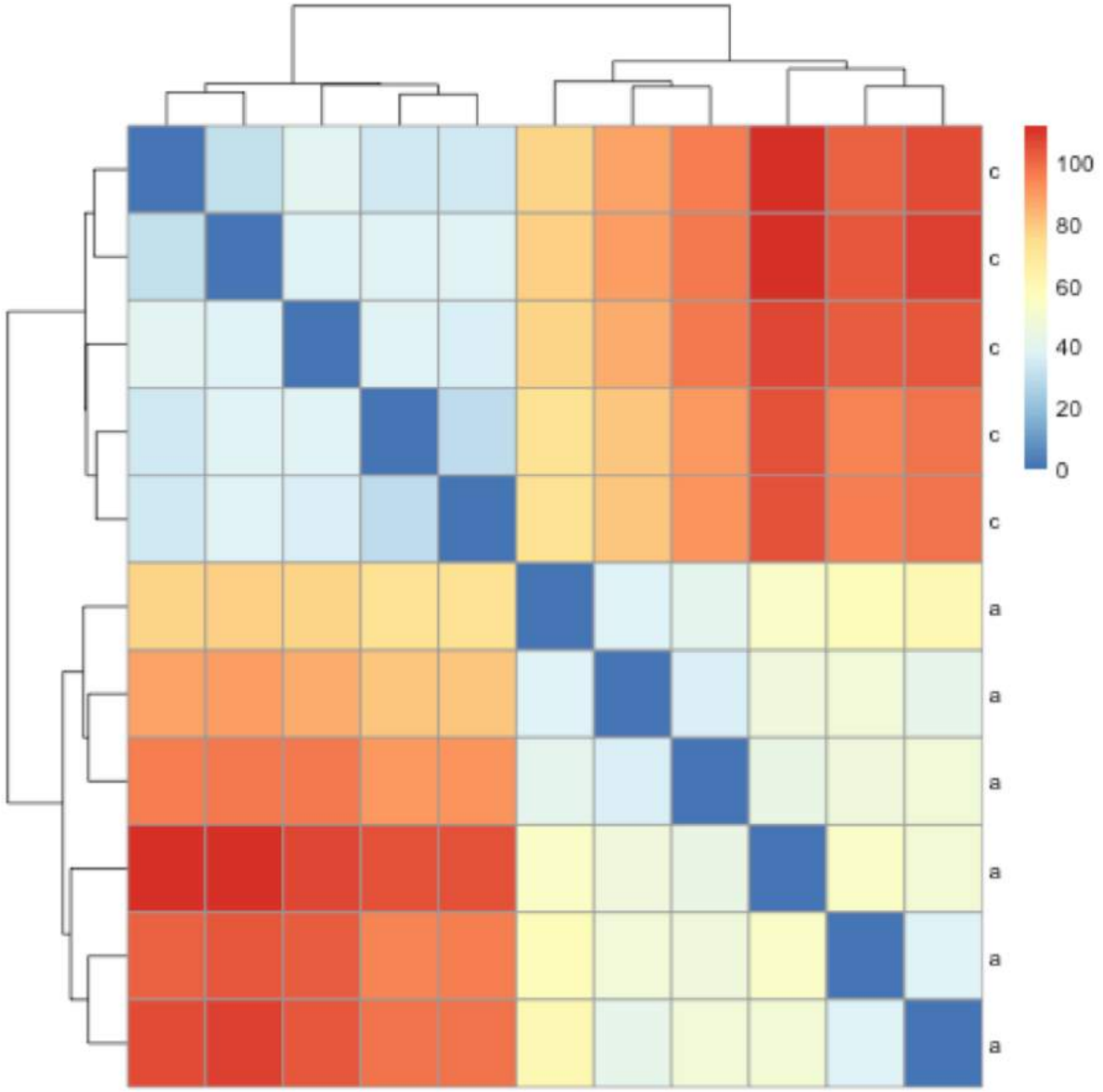
Adding other parameters for the heatmaps….



mRNAs

Sample-To-Sample distance (Euclidian)

adenine (a) vs. control (c)



Goal:
to assess overall similarity between samples

A heatmap of this distance matrix gives us an **overview over similarities and dissimilarities between samples.**

We have to provide a hierarchical clustering (hc) to the heatmap function based on the sample distances, or else the heatmap function would calculate a clustering based on the distances between the rows/columns of the distance matrix.

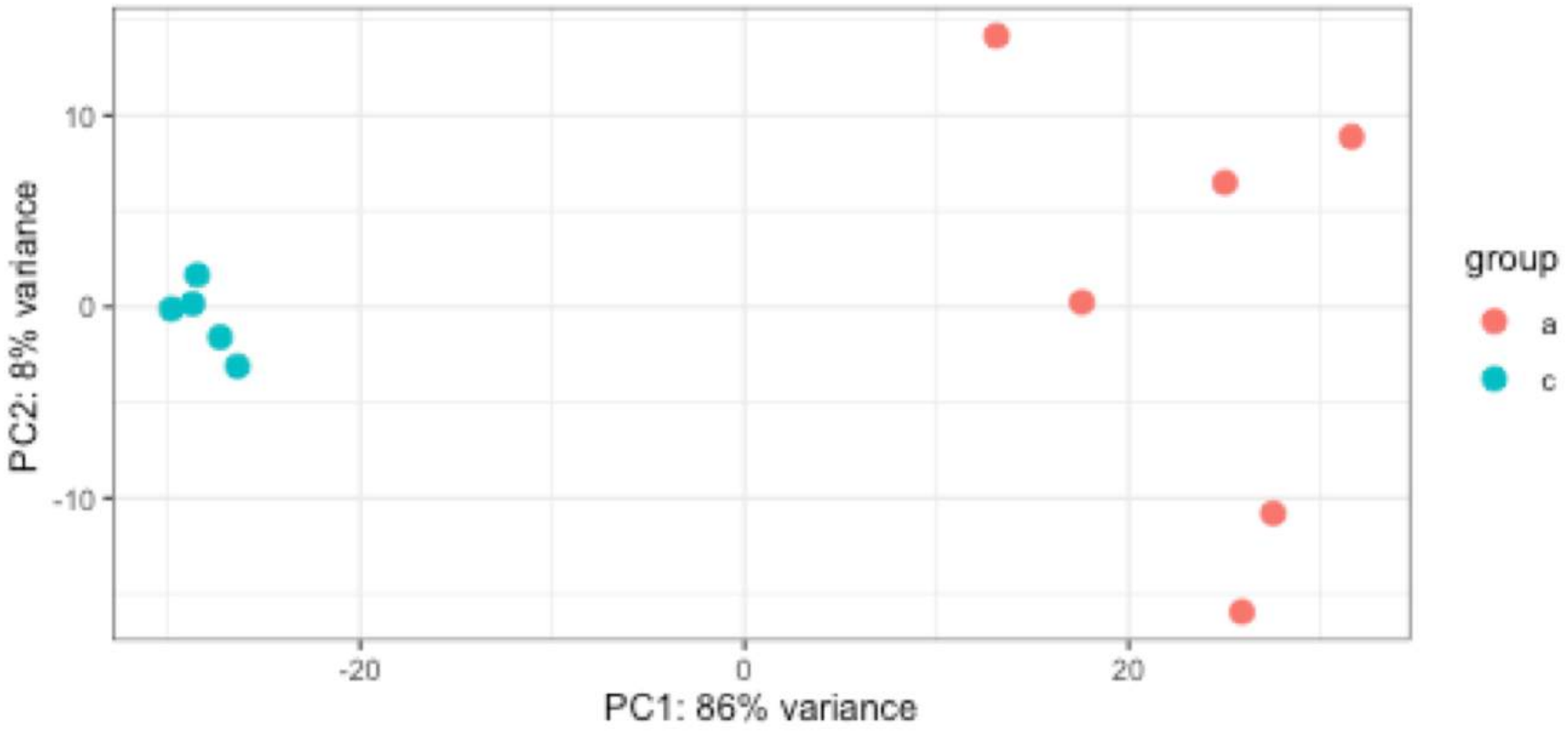
mRNAs

adenine (a) vs. control (c)

Principal component plot of the samples

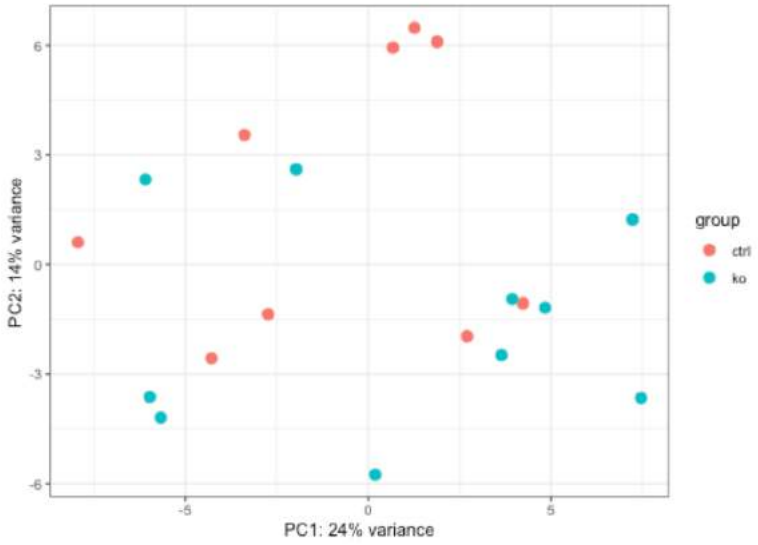
Related to the distance matrix is the PCA plot, which shows the samples in the 2D plane spanned by their first two principal components. This type of plot is useful for **visualizing the overall effect** of **experimental covariates** and **batch effects**.

PCA plot



PCA plot

- Principal Component analysis (PCA) is the most commonly used dimensionality reduction method;
- PCA projects multidimensional data onto lower uncorrelated dimensions (principal components – PCs), while retaining most of the information;
- PC1 is a projection that accounts for the most of the variation,
- PC2 is a projection that accounts for the most of the remaining information,
- PC3 is a next…, etc.



It informs us about:

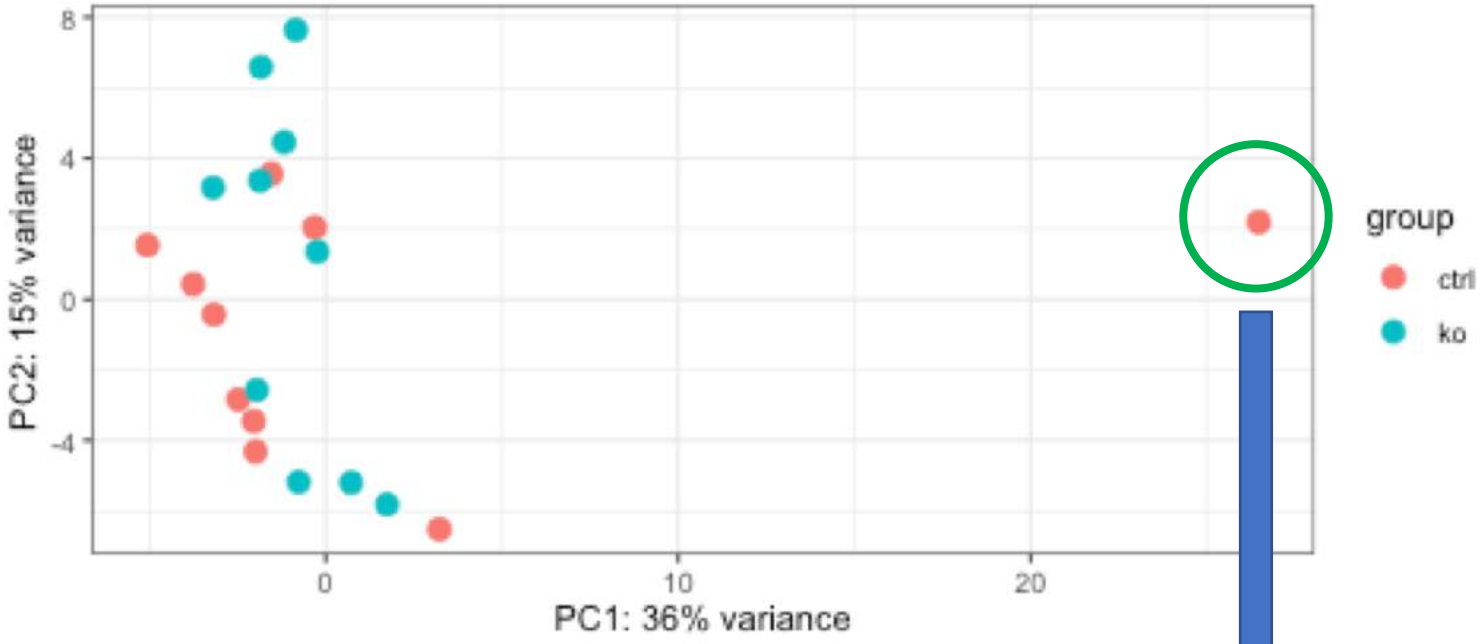
- PCA allows to characterize overall structure of the dataset;
- PCA allows to assess the quality of the dataset:
  - Concordance between replicates;
  - Overall structure should correspond to design;
  - Identification of outlier samples;
  - Identification of confounding factors;
  - Identification of unwanted/unexpected patterns

Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

mRNAs

Ctrl vs. KO

PCA plot



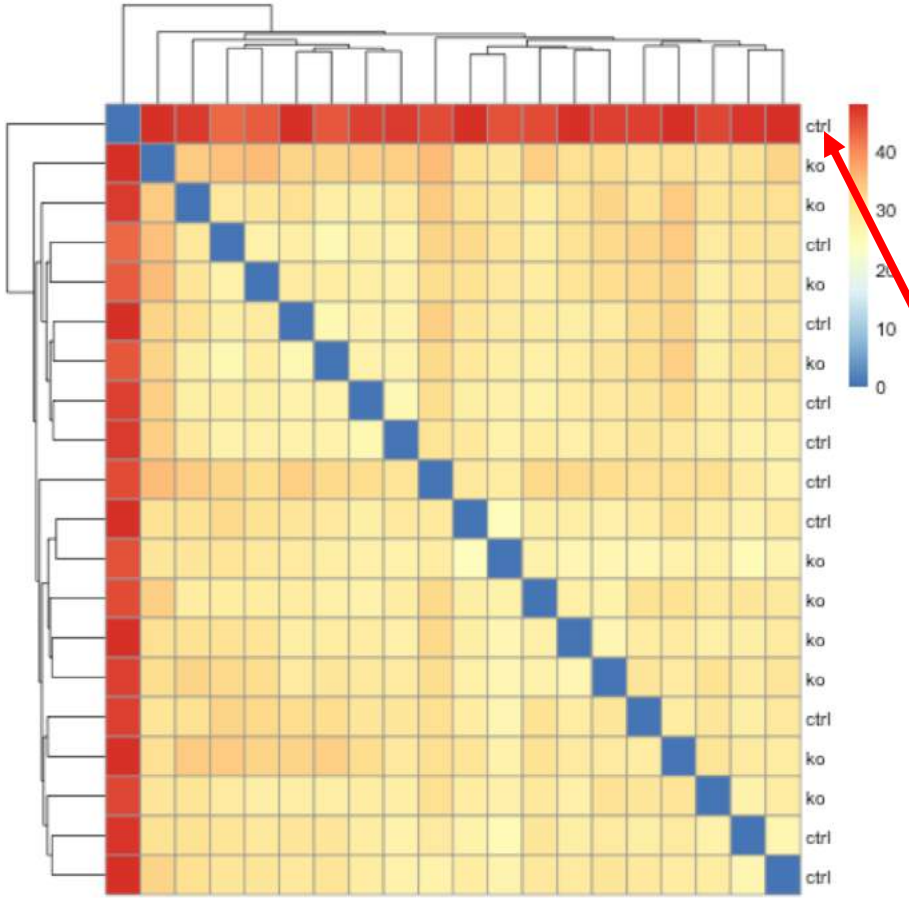
Outlier

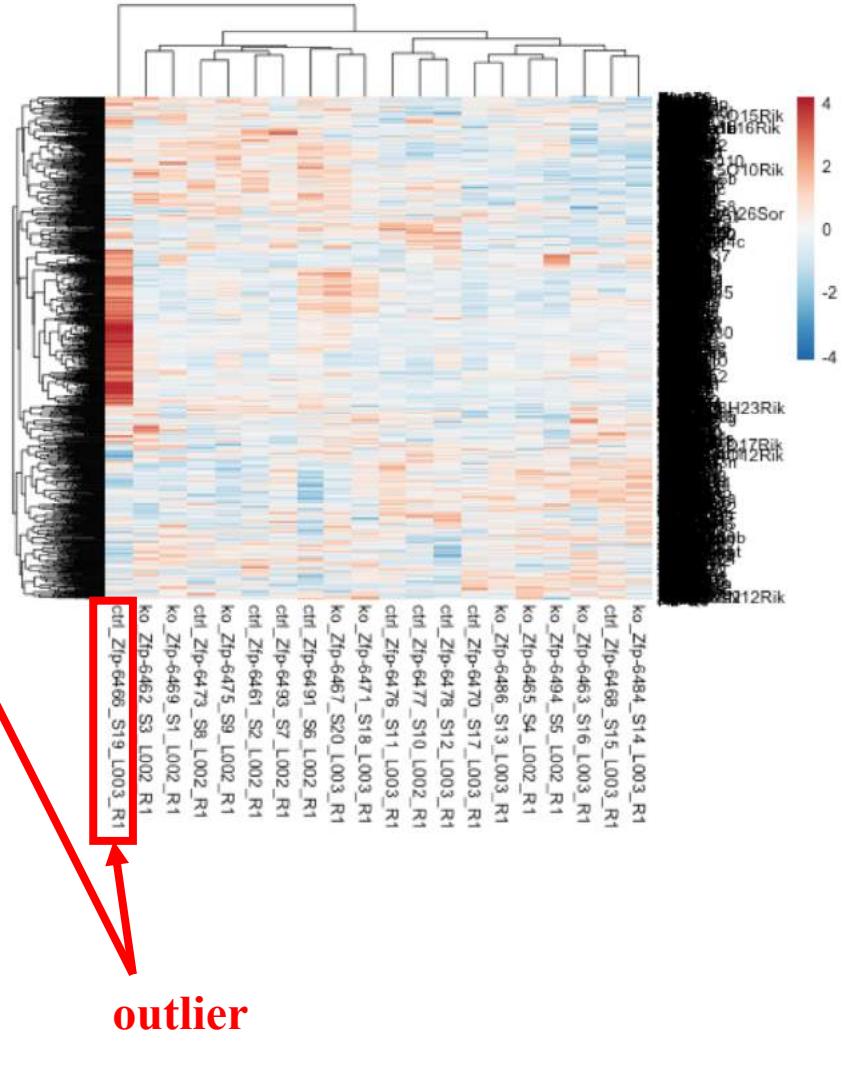
I had to exclude one Ctrl sample

mRNAs

Sample-To-Sample distance (Euclidian)

Ctrl vs. KO

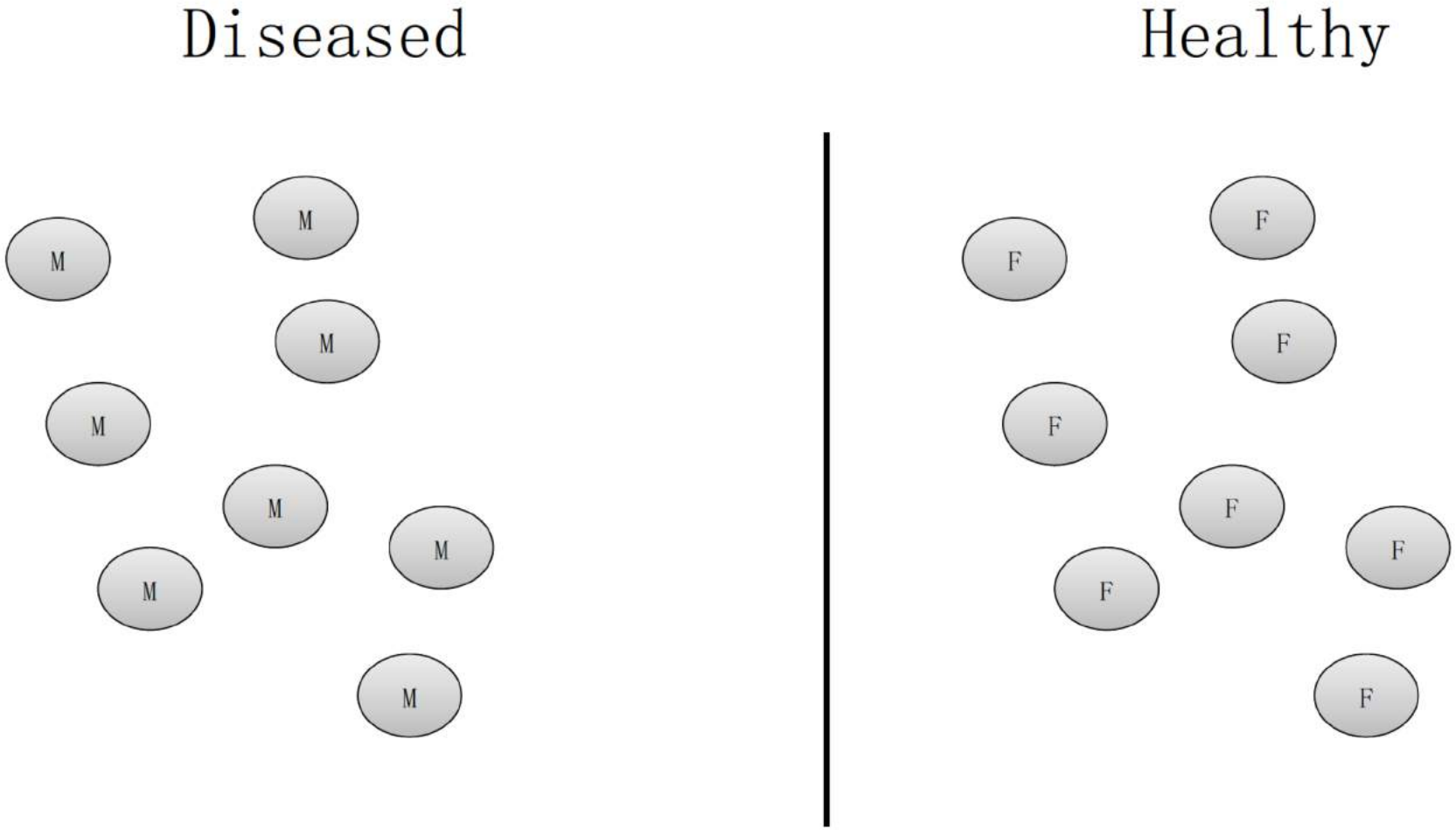




outlier

Quick reminder of study design

M = Male, F = Female



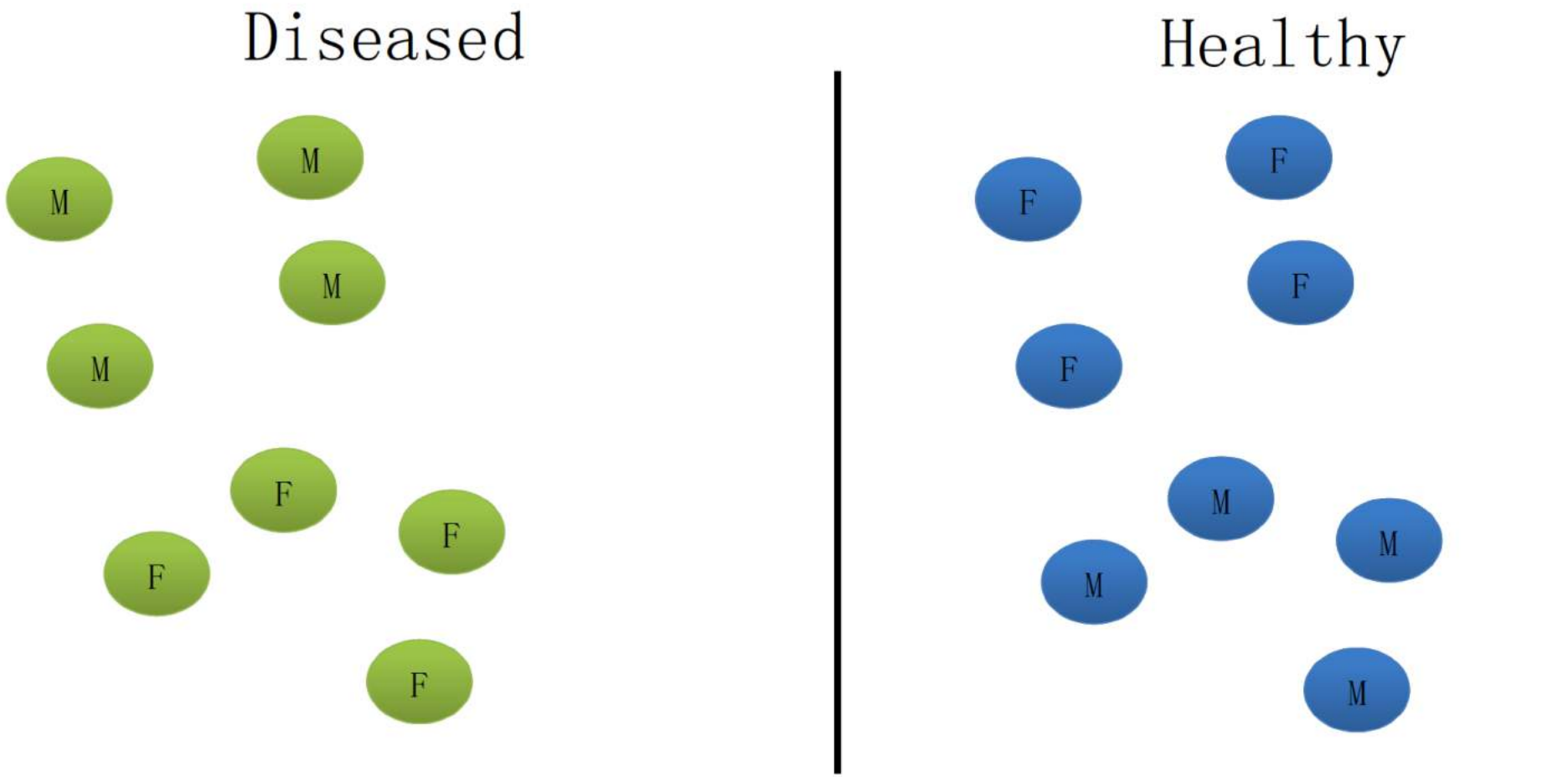
What's wrong with this design?

Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

Quick reminder of study design

M = Male, F = Female

Blue = Collected in Winter, Green = Collected in Summer



What's wrong with this design?

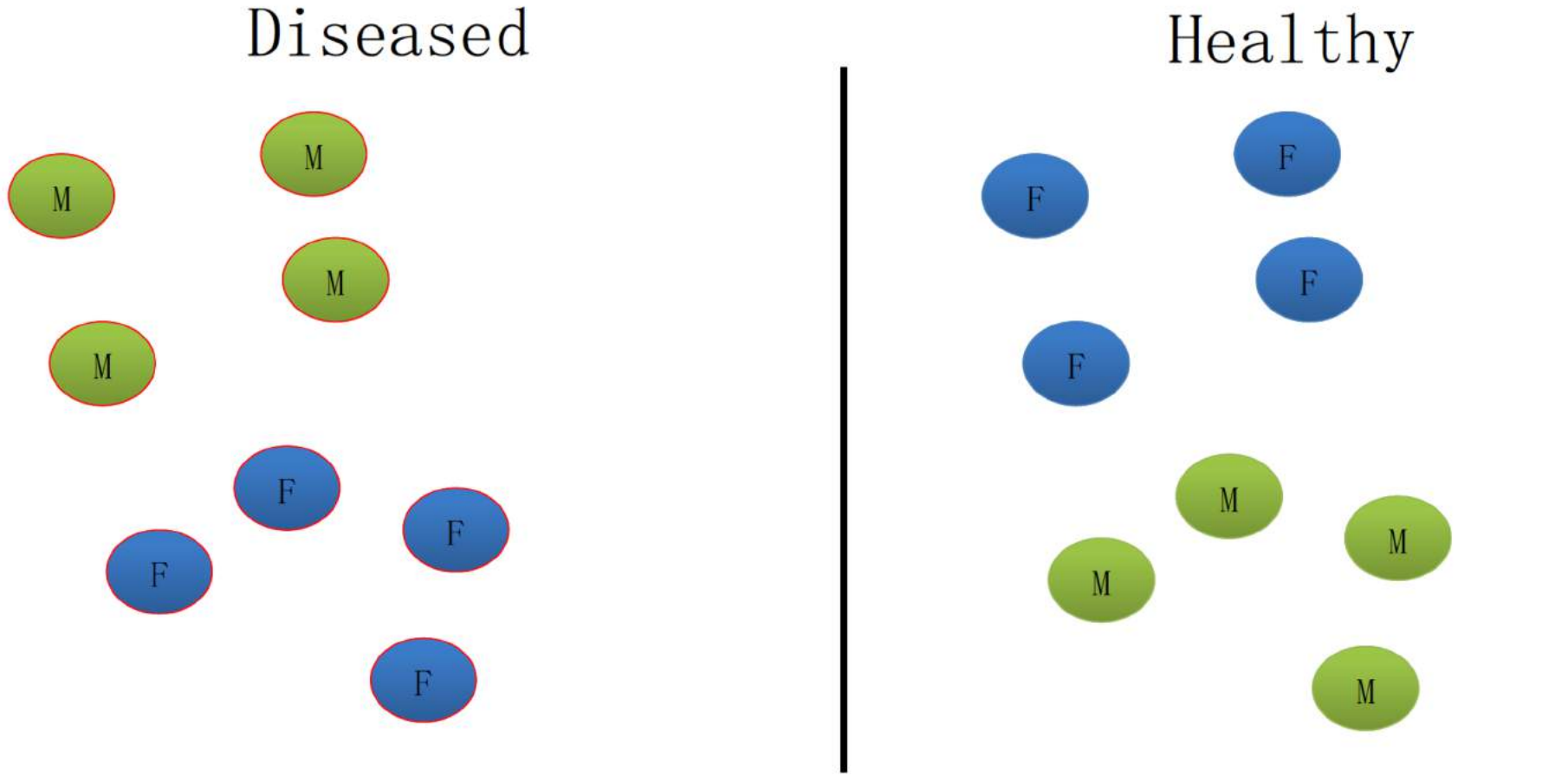
Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

Quick reminder of study design

M = Male, F = Female

Blue = Collected in Winter, Green = Collected in Summer

Red outline = all sequenced together on same lane



What's wrong with this design?

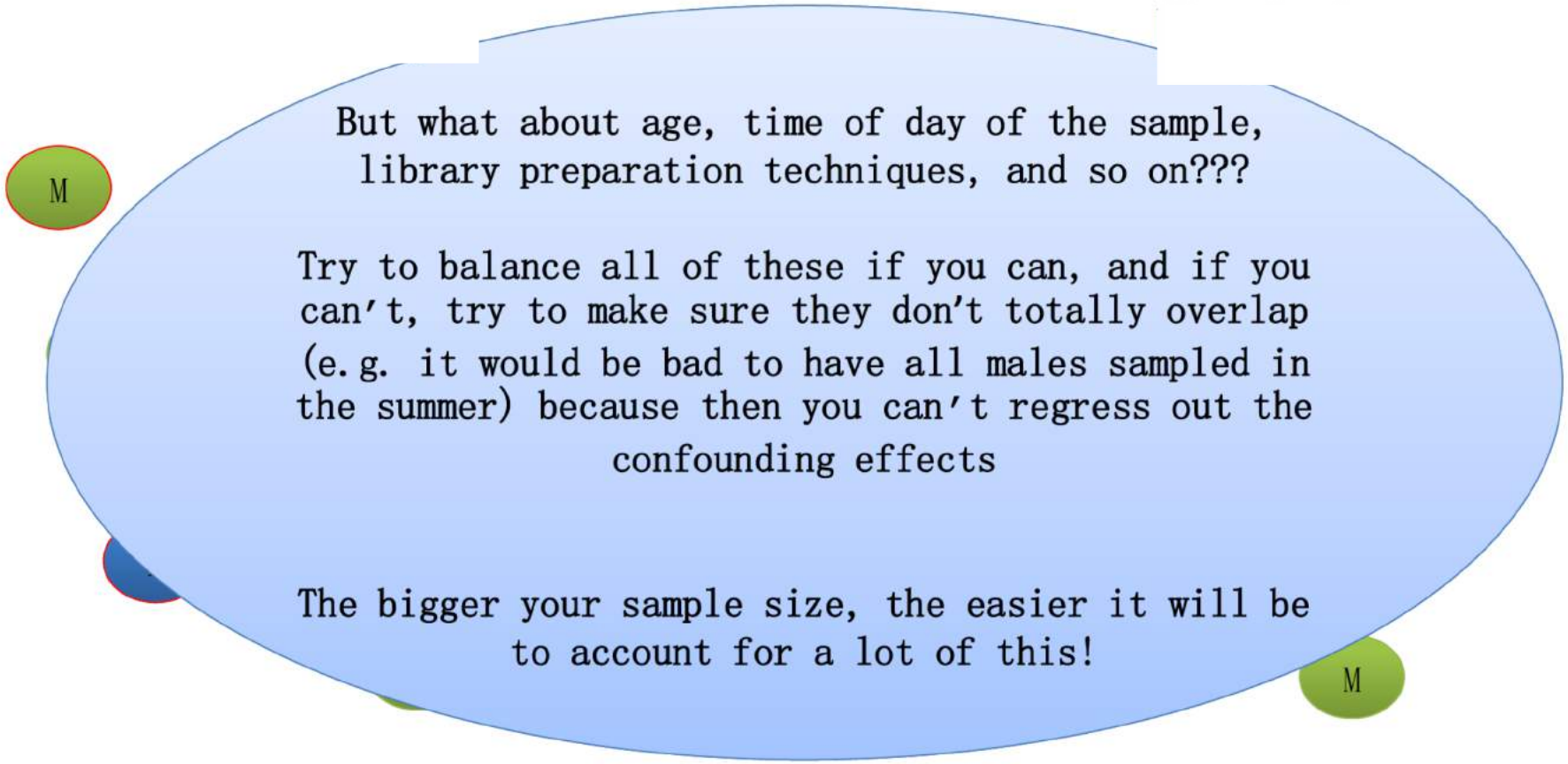
Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

Quick reminder of study design

M = Male, F = Female

Blue = Collected in Winter, Green = Collected in Summer

Red outline = all sequenced together on same lane

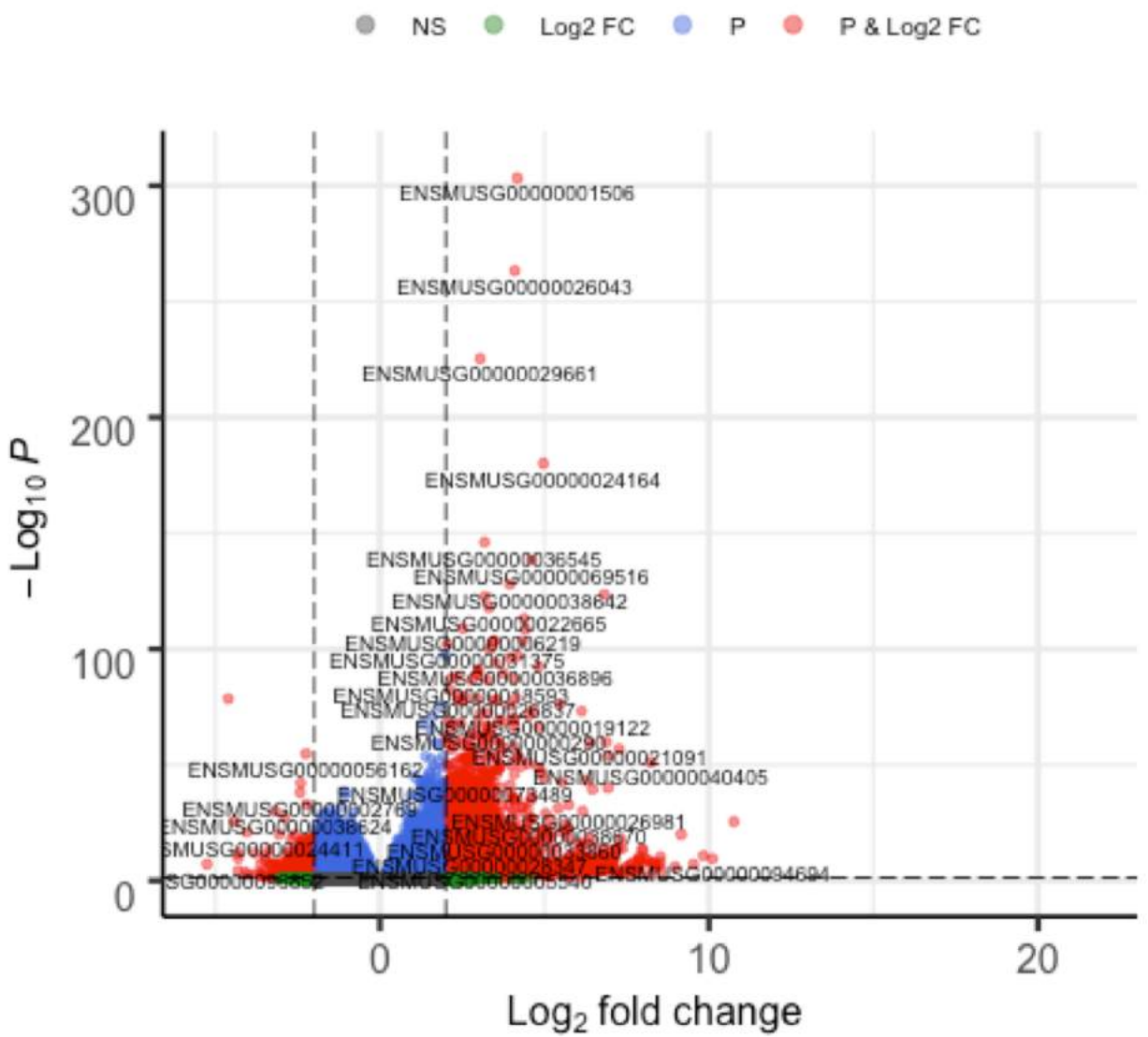


Balanced design

Adopted from Alice Mouton's presentation (Co-teaching by Trent Su)

mRNAs

Volcano plot - adenine (3a) vs. control (1c)



mRNAs

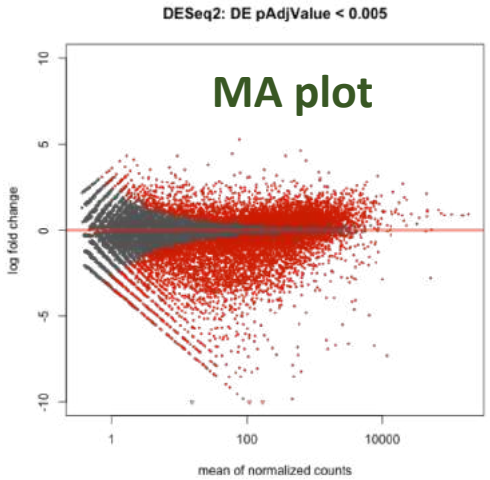
adenine (3a) vs. control (1c) vs. – first top 30 mRNAs

Exporting results to **CSV** files

sorted by padj

(from the smallest to the largest & expand selection)

No.	Gene name	Ensemble ID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
1	Col1a1	ENSMUSG00000001506	3568.664273	4.173044024	0.1119629	37.271677	4.72E-304	9.53E-300
2	Col3a1	ENSMUSG00000026043	3981.115995	4.096625696	0.1180196	34.711412	5.30E-264	5.35E-260
3	Col1a2	ENSMUSG00000029661	3278.481421	3.046653811	0.0949071	32.10144	4.21E-226	2.83E-222
4	C3	ENSMUSG00000024164	6370.458168	4.965836181	0.1731614	28.677494	7.28E-181	3.67E-177
5	Adamts2	ENSMUSG00000036545	282.842667	3.174079506	0.1229816	25.809376	6.96E-147	2.81E-143
6	Lyz2	ENSMUSG00000069516	3832.940699	4.607881751	0.1833924	25.12581	2.60E-139	8.74E-136
7	Ctss	ENSMUSG00000038642	1942.984658	3.945495455	0.1634801	24.134411	1.09E-128	3.14E-125
8	Ltbp2	ENSMUSG00000002020	555.4845735	6.817605323	0.2875797	23.706838	3.07E-124	7.73E-121
9	Mmp14	ENSMUSG00000000957	802.6007325	3.190026073	0.1350697	23.617637	2.54E-123	5.69E-120
10	Ccdc80	ENSMUSG00000022665	526.500283	3.30840364	0.1428569	23.158869	1.18E-118	2.39E-115
11	Thy1	ENSMUSG00000032011	302.3139572	4.378722698	0.1929659	22.691688	5.41E-114	9.93E-111
12	Fblim1	ENSMUSG00000006219	422.8370972	2.525093978	0.1134461	22.258099	9.42E-110	1.58E-106
13	Cd44	ENSMUSG00000005087	499.6884963	4.419823127	0.1989163	22.219515	2.22E-109	3.45E-106
14	C1qa	ENSMUSG00000036887	1190.967138	3.485798461	0.1607421	21.685666	2.80E-104	4.04E-101
15	C4b	ENSMUSG00000073418	322.2700895	4.342341384	0.2003426	21.674584	3.56E-104	4.80E-101
16	Mmp2	ENSMUSG00000031740	323.8005229	3.448601945	0.1592072	21.661095	4.78E-104	6.03E-101
17	Bgn	ENSMUSG00000031375	6234.405188	2.043288181	0.0949252	21.525254	9.03E-103	1.07E-99
18	C1qb	ENSMUSG00000036905	1102.615883	3.343560524	0.1568949	21.310826	9.01E-101	1.01E-97
19	Axl	ENSMUSG00000002602	1174.057444	1.97962705	0.0941189	21.033254	3.26E-98	3.46E-95
20	Siglec1	ENSMUSG00000027322	211.5220632	4.178186433	0.1993228	20.961913	1.46E-97	1.47E-94
21	Vcam1	ENSMUSG00000027962	1477.894386	3.945857289	0.1889041	20.888147	6.86E-97	6.60E-94
22	C1qc	ENSMUSG00000036896	1026.016346	3.511749191	0.1699725	20.660692	7.82E-95	7.18E-92
23	Aoc1	ENSMUSG00000029811	746.4090098	4.815845493	0.2350584	20.487864	2.76E-93	2.42E-90
24	Mpeg1	ENSMUSG00000046805	1715.418785	2.992047135	0.1469941	20.354874	4.20E-92	3.53E-89
25	Laptm5	ENSMUSG00000028581	975.9864598	2.963124346	0.1469332	20.166479	1.93E-90	1.56E-87
26	Runx1	ENSMUSG00000022952	208.4711273	3.750183207	0.1861044	20.15096	2.64E-90	2.05E-87
27	Tnfrsf1b	ENSMUSG00000028599	359.1322718	2.972165405	0.1476976	20.123313	4.61E-90	3.45E-87
28	Sh3pxd2b	ENSMUSG00000040711	345.0522222	2.342656605	0.117242	19.981376	8.00E-89	5.76E-86
29	Sparc	ENSMUSG00000018593	3983.544768	2.161385143	0.1085952	19.903143	3.82E-88	2.66E-85
30	Ccl6	ENSMUSG00000018927	287.4907186	4.074499712	0.2049543	19.880041	6.06E-88	4.08E-85



- The function *plotMA* shows **the log2 fold changes** attributable to a given **variable** over the mean of normalized counts for all the samples in the *DESeqDataSet*.
- Points will be colored **red** if the **adjusted *p* value is < 0.1**.

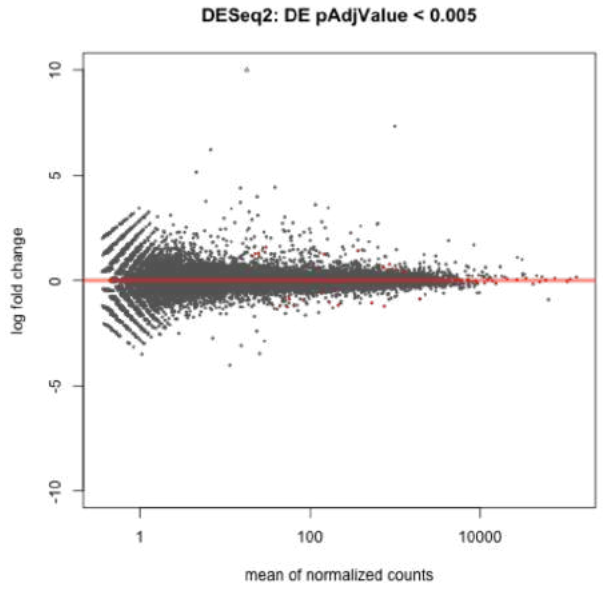
- `baseMean` : mean of normalized counts for all samples
- `log2FoldChange` : log2 fold change
- `lfcSE` : standard error
- `stat` : Wald statistic
- `pvalue` : Wald test p-value
- `padj` : BH adjusted p-values

The **Wald statistic** is the logfoldchange (LFC) divided by its standard error (lfcSE) . This Wald statistic is used to calculate p-values (it is compared to a standard normal distribution) . So it's the ratio of LFC and SE which determines significance.

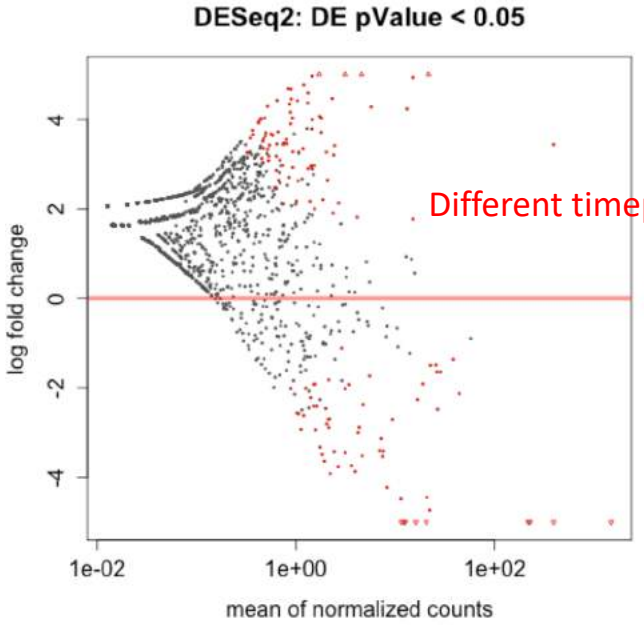
The Benjamini-Hochberg (BH) procedure is a powerful tool that decreases the false discovery rate. Adjusting the rate helps to control for the fact that sometimes small p-values (less than 5%) happen by chance, which could lead you to incorrectly reject the true null hypotheses. In other words, the BH Procedure helps you to avoid Type I errors (false positives).

MA plot

Normal

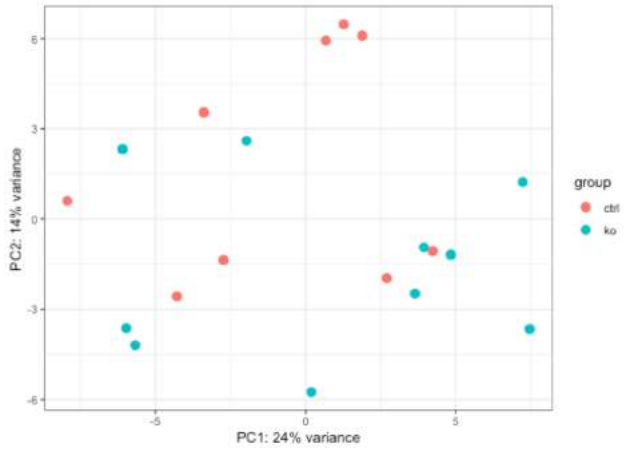


Batch effect

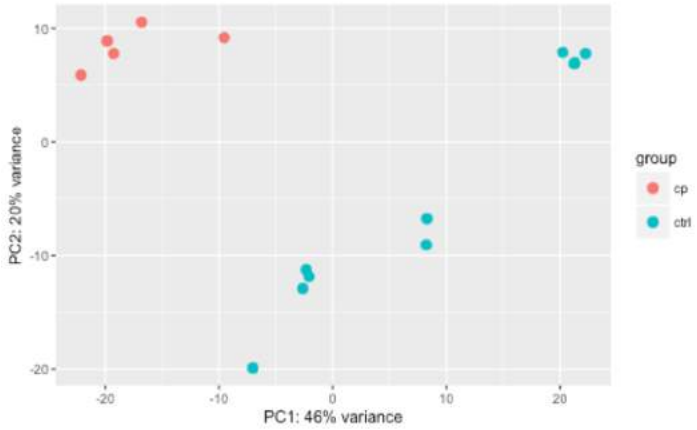


Different timepoints of sample collection

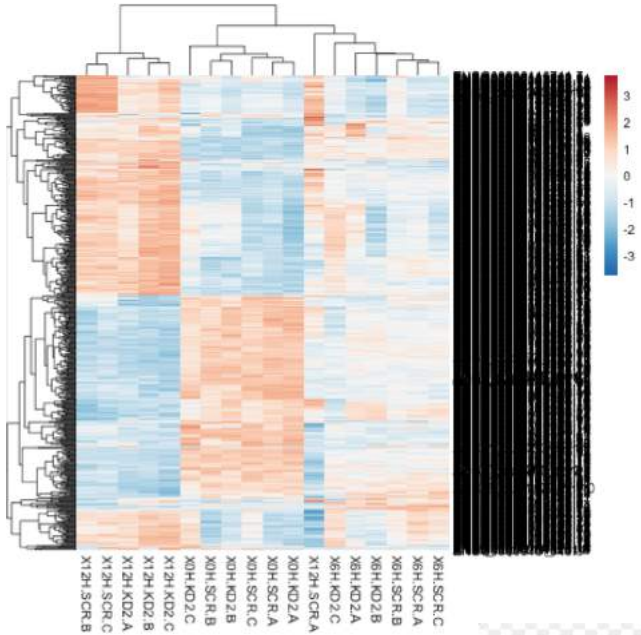
Ctrl vs. KO



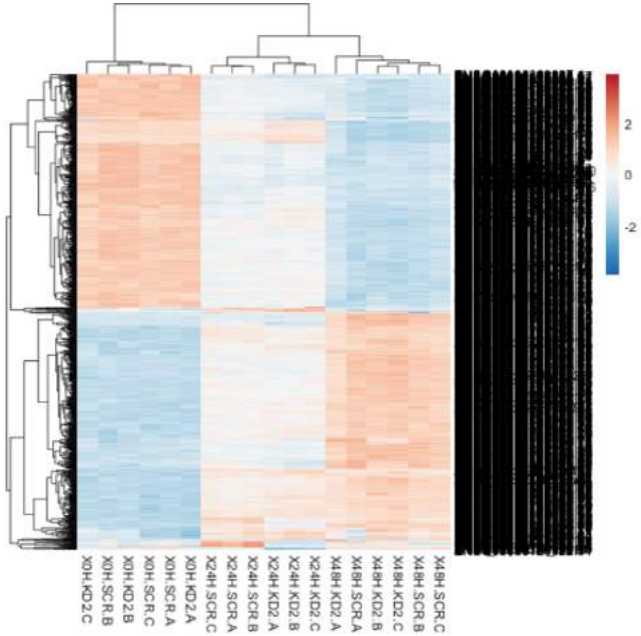
Disease vs. Ctrl



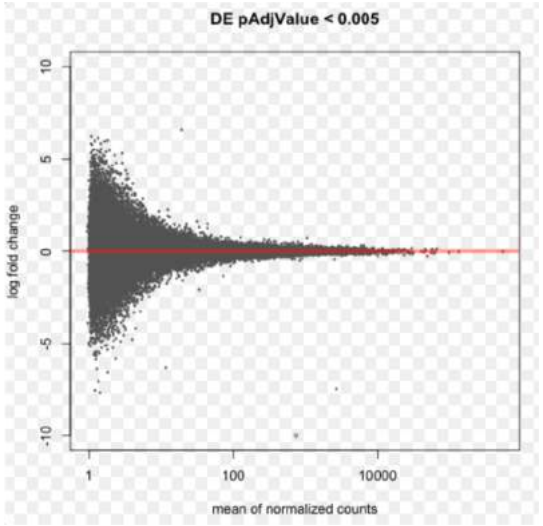
DESeq2 – difference over time



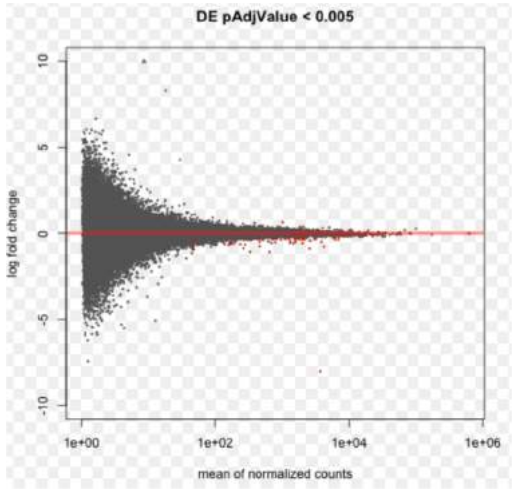
Ctrl vs. **KO**



Early timepoints
[0h, 6h, 12h)



Late timepoints
[0h, 24h, 48h)



DESeq2 – difference over time

condition : kdr2 or scr
time: 0, 6, 12 (early) & 0, 24, 46 (late)

Full model: design formula that models **the condition difference at time 0, the difference over time, and any condition-specific differences over time** (the interaction term condition:time).

Reduce model: from full model removed the condition-specific differences over time.

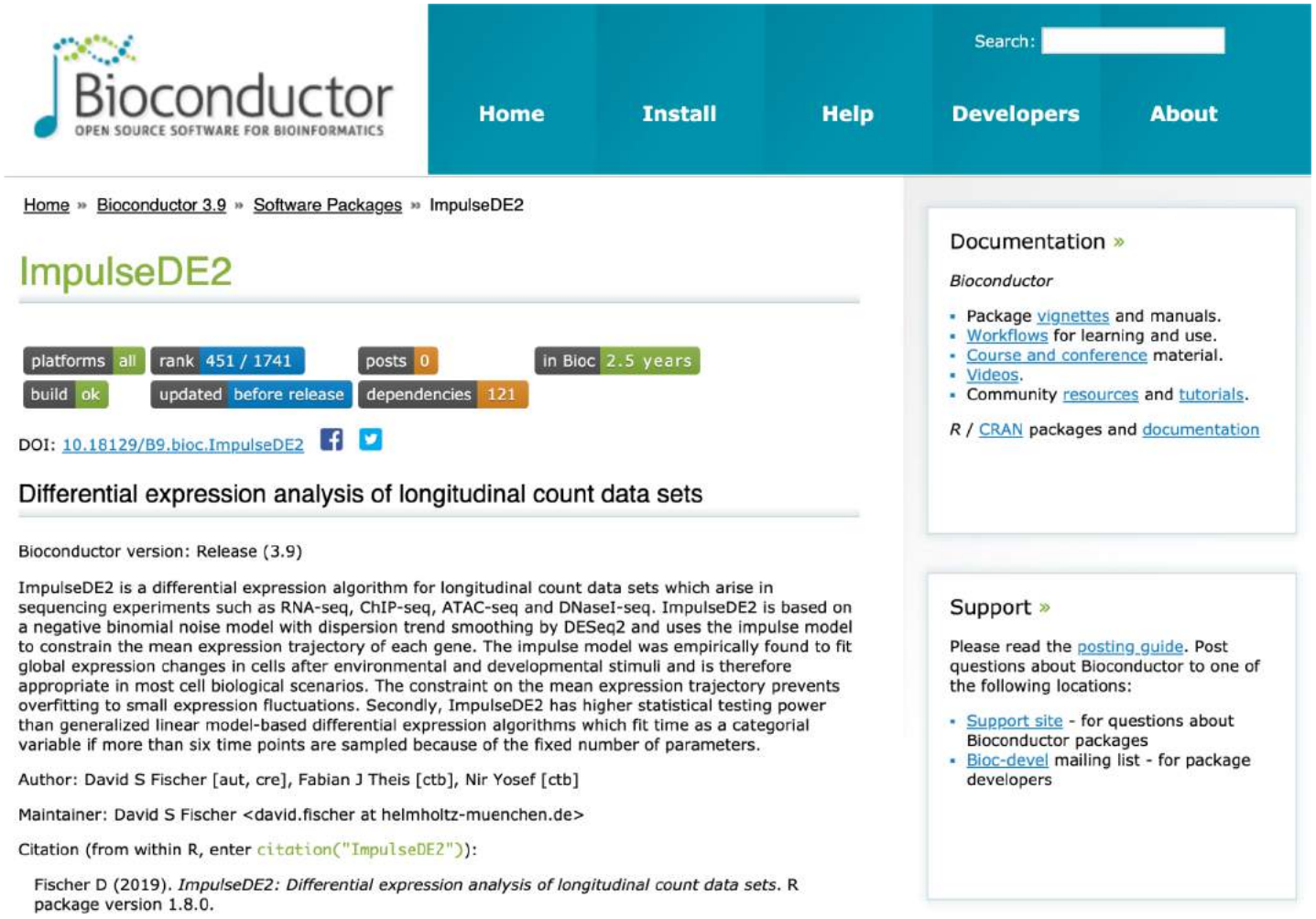
The **LRT** examines two models for the counts, a *full* model with a certain number of terms and a *reduced* model, in which some of the terms of the *full* model are removed. The test determines if the increased likelihood of the data using the extra terms in the *full* model is more than expected if those extra terms are truly zero

Genes with small *p* values from this test are those which at one or more time points after time 0 showed a condition-specific effect. Note therefore that this will not give small *p* values to genes that moved up or down over time in the same way in both conditions.

Impulse DE2

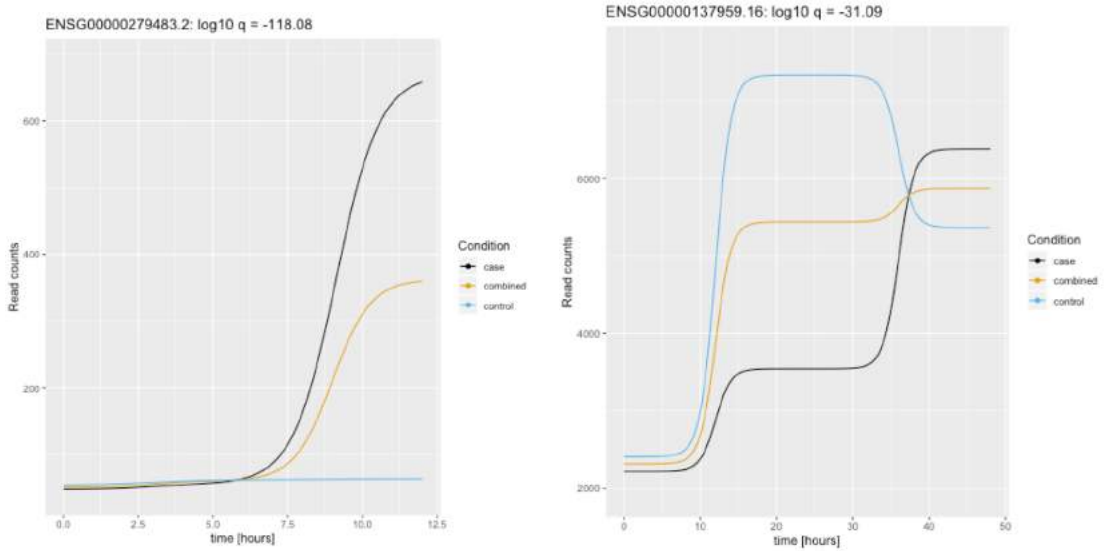
Differential expression analysis of longitudinal count data sets

https://bioconductor.org/packages/release/bioc/html/ImpulseDE2.html

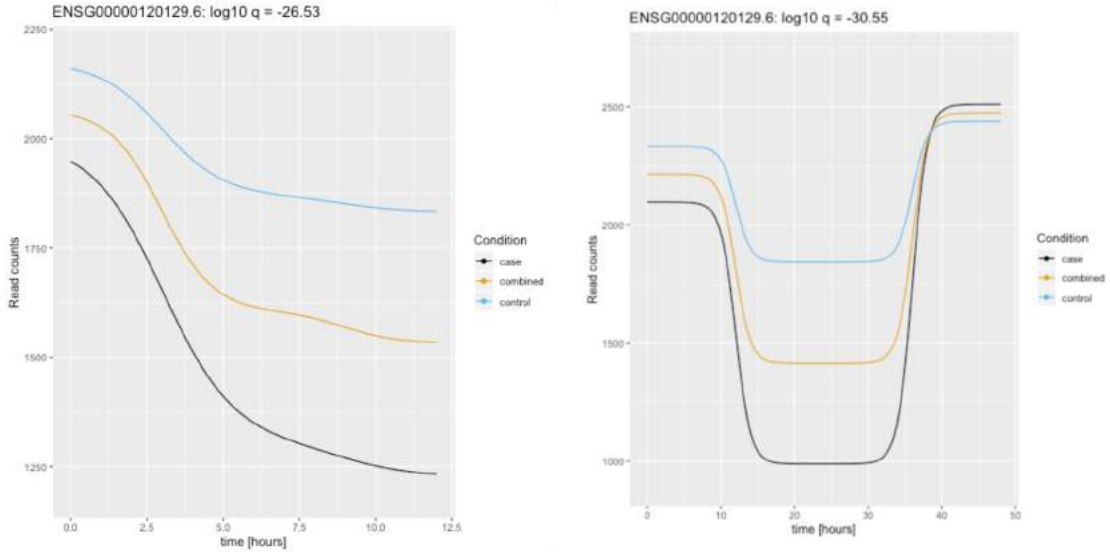


Fischer D (2019). ImpulseDE2: Differential expression analysis of longitudinal count data sets. R package version 1.8.0.

Impulse DE2



Early **Late**



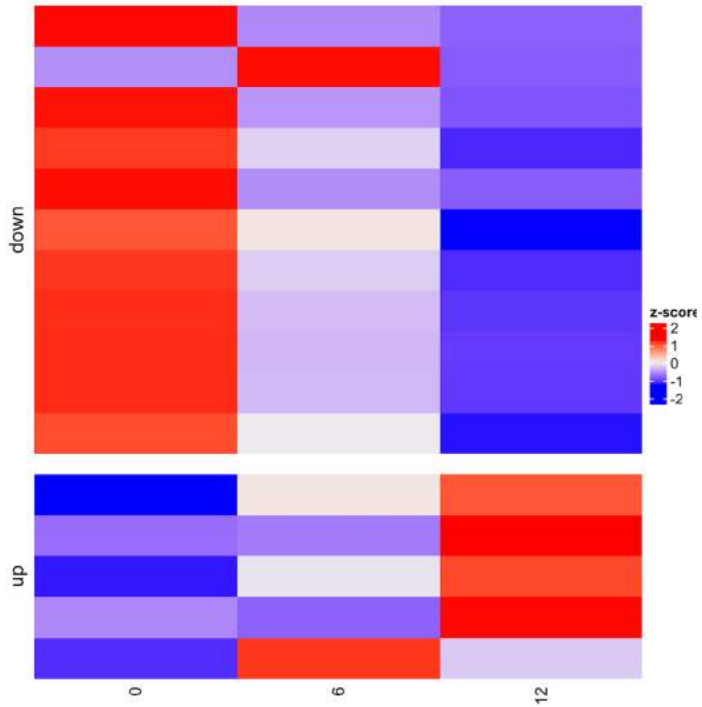
IMPULSEDE2 Heatmaps
scaQThres < 0.01 (scalar)
FDR-corrected p-value threshold for calling differentially expressed genes

case- KD2,
ctrl - SCR,
combined (both)

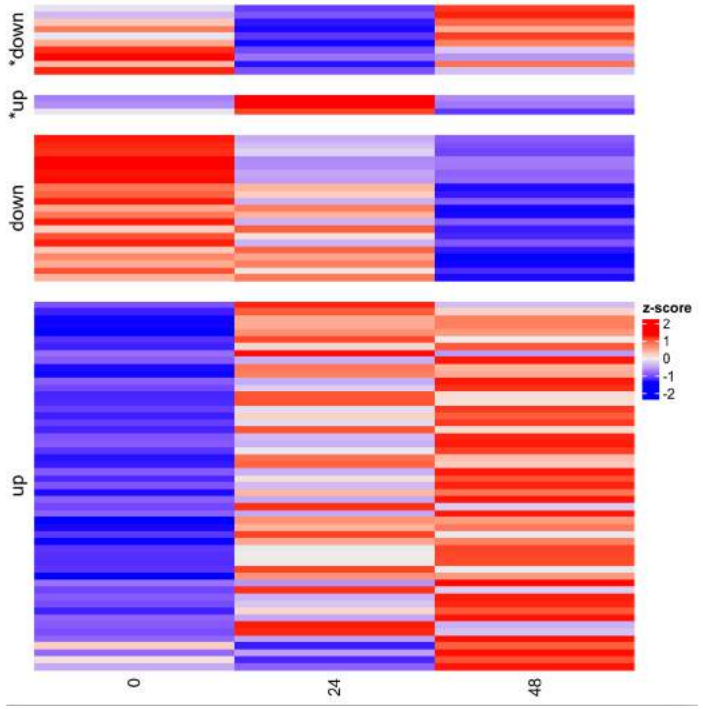
The Standard normalization from Impulse De2

The normalisation constant is the median of the ratio of gene counts versus the geometric gene count mean. There is one normalisation constant per replicate. An intuitive alternative would be the sequencing depth, the median ratio is however less sensitive to highly differentially expressed genes with high counts (ref. DESeq). The normalisation constants are used to scale the mean of the **negative binomial model** inferred during fitting to the sequencing depth of the given sample. The normalisation constants therefore replace normalisation at the count data level, which is not supposed to be done in the framework of ImpulseDE2. There is the option to supply size factors to this function to override its size factor choice.

Impulse DE2



Early



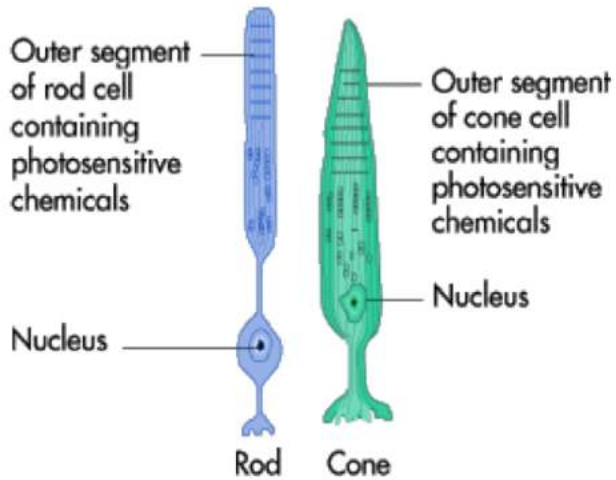
Late

Heatmap genes

Filter by genes where scaQThres < 0.01 (scalar)
FDR-corrected p-value threshold for calling differentially expressed genes: Only genes below this threshold are included in the heatmap.

FDR - the false discovery rate

Case Study: RNA-Seq profiling of developing photoreceptors



- There are two major types of photoreceptor cells in vertebrate retina, rods and cones;
- Rods and cones differ in many aspects of cell morphology and physiology, including light sensitivity;
- Transcription factor NRL is a key regulator of rod cell fate. In the absence of NRL rods convert into cone-like photoreceptors ("cods").

Cell Reports

Volume 17, Issue 9, 22 November 2016, Pages 2460–2473

Open Access

Resource

NRL-Regulated Transcriptome Dynamics of Developing Rod Photoreceptors

Jung-Woong Kim[1,2,7], Hyun-Jin Yang[1,7], Matthew John Brooks[1,7], Lina Zelinger[1], Gökhan Karakülah[1,8], Norimoto Gotoh[1,3], Alexis Boleda[1], Linn Gieser[1], Felipe Giuste[1], Dustin Thad Whitaker[1,4], Ashley Walton[1], Rafael Villasmil[5], Jennifer Joanna Barb[6], Peter Jonathan Munson[6], Koray Dogan Kaya[1], Vijender Chaitankar[1], Tiziana Cogliati[1], Anand Swaroop[1,9]

- Transcriptome profiling of differentiating rods at multiple time points;
- Transcriptome profiling was performed for rod cells from wild-type (WT) and NRL-knockout mutant (KO) mice;
- RNA-Seq data was generated from purified (FACS sorted) cell populations.

Workshop’s google drive folder

Slides & supporting documents:

➢R scripts and files with htseq counts:
https://drive.google.com/open?id=16akcy3Mrb8Jd5VLv57coSi2KGw1aLE72

EXERCISE

STEP 1: Download the htseq data from the google drive folder (30 files):
https://drive.google.com/open?id=16akcy3Mrb8Jd5VLv57coSi2KGw1aLE72

STEP 2: Perform DE WT and KO ($NRL^{-/-}$) samples ("case-control design")

SraRunTable.csv

Run	Assay Type	AssemblyName	AvgSpotLen	BioProject	BioSample	cell_type	Center Name	Consent	DATASTORE filetyp	DATASTORE provid	DATASTORE region	Experiment	Genotype	GEO_Accession	Instrument	LibraryLayout	LibrarySelection	LibraryS
SRR2936836	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235754	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411331	Nrlp-GFP	GSM1924968	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936837	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235755	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411332	Nrlp-GFP	GSM1924969	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936838	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235756	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411333	Nrlp-GFP	GSM1924970	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936839	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235757	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411334	Nrlp-GFP	GSM1924971	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936840	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235758	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411335	Nrlp-GFP	GSM1924972	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936841	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235759	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411336	Nrlp-GFP	GSM1924973	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936842	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235760	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411337	Nrlp-GFP	GSM1924974	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936843	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235761	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411338	Nrlp-GFP	GSM1924975	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936844	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235762	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411339	Nrlp-GFP	GSM1924976	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936845	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235763	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411340	Nrlp-GFP	GSM1924977	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936846	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235764	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411341	Nrlp-GFP	GSM1924978	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936847	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235765	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411342	Nrlp-GFP	GSM1924979	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936848	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235766	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411343	Nrlp-GFP	GSM1924980	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936849	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235767	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411344	Nrlp-GFP	GSM1924981	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936850	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235768	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411345	Nrlp-GFP	GSM1924982	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936851	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235769	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411346	Nrlp-GFP	GSM1924983	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936852	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235770	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411347	Nrlp-GFP;Nrl-/-	GSM1924984	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936853	RNA-Seq	GCF_000001635.20	85	PRJNA301098	SAMN04235771	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411348	Nrlp-GFP;Nrl-/-	GSM1924985	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936854	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235772	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411349	Nrlp-GFP;Nrl-/-	GSM1924986	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936855	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235773	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411350	Nrlp-GFP;Nrl-/-	GSM1924987	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936856	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235774	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411351	Nrlp-GFP;Nrl-/-	GSM1924988	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936857	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235775	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411352	Nrlp-GFP;Nrl-/-	GSM1924989	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936858	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235776	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411353	Nrlp-GFP;Nrl-/-	GSM1924990	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936859	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235777	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411354	Nrlp-GFP;Nrl-/-	GSM1924991	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936860	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235778	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411355	Nrlp-GFP;Nrl-/-	GSM1924992	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936861	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235779	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411356	Nrlp-GFP;Nrl-/-	GSM1924993	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936862	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235780	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411357	Nrlp-GFP;Nrl-/-	GSM1924994	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936863	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235781	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411358	Nrlp-GFP;Nrl-/-	GSM1924995	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936864	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235782	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411359	Nrlp-GFP;Nrl-/-	GSM1924996	Illumina Genome An	SINGLE	cDNA	TRANS
SRR2936865	RNA-Seq	GCF_000001635.20	76	PRJNA301098	SAMN04235783	GFP positive retina c	GEO	public	sra	gs,ncbi,s3	gs.US,ncbi.public,s3	SRX1411360	Nrlp-GFP;Nrl-/-	GSM1924997	Illumina Genome An	SINGLE	cDNA	TRANS

Wild type
(16 samples: 36-51)

KO type
(14 samples: 52-65)

Installations

•**R (version 3.6.1)** : https://www.r-project.org/ For mac you can download binaries from : https://cran.r-project.org/bin/macosx/
It is possible to install R with homebrew: brew install r
For Windows you can follow instruction : https://cran.r-project.org/bin/windows/base/

•**R Studio:** https://rstudio.com/ . You can download installer for different OS from https://rstudio.com/products/rstudio/download/#download . With homebrew you can install with command : brew cask install rstudio

•After installing R, it is required to run this script to install **packages**:

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

Analyzing RNA-seq data with DESeq2

Michael I. Love, Simon Anders, and Wolfgang Huber

10/18/2019

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files. DESeq2 package version: 1.25.17

http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Thank You