



# Contagem, análise de expressão diferencial e ontologia gênica

Adriana Mércia Guaratini Ibelli

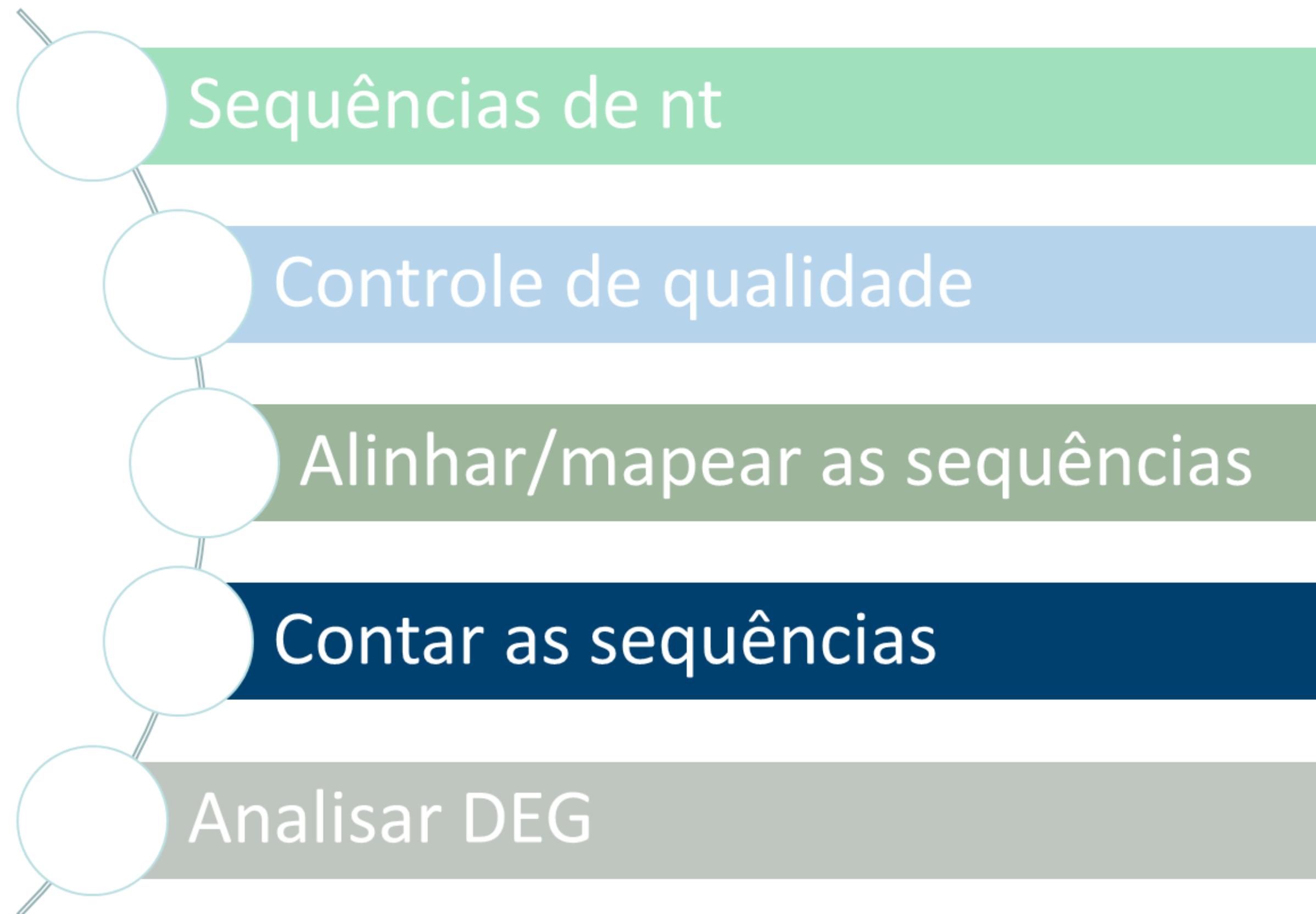
15 a 19 de maio de 2023



MINISTÉRIO DA  
AGRICULTURA E  
PECUÁRIA



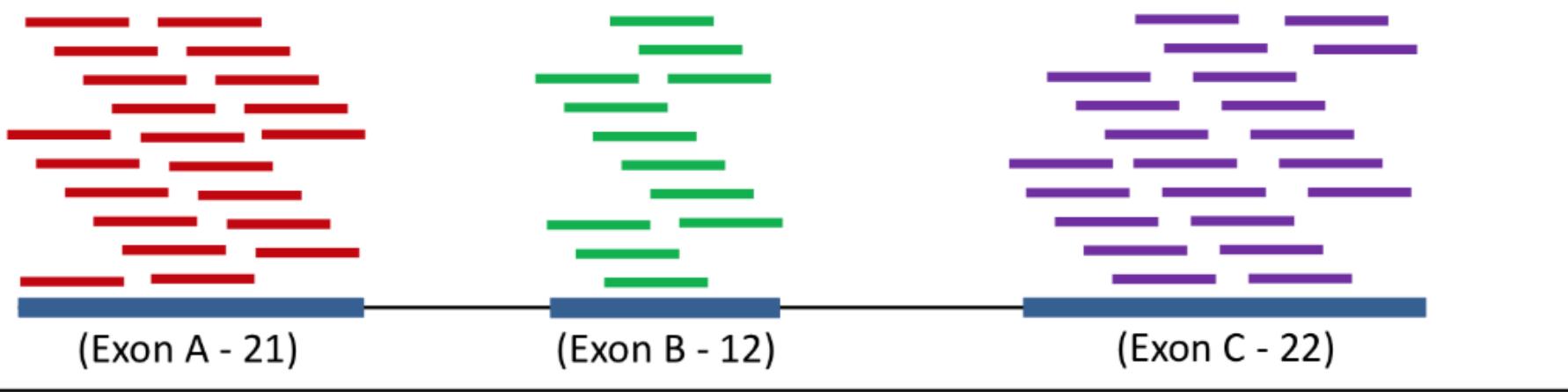
# Onde estamos ???



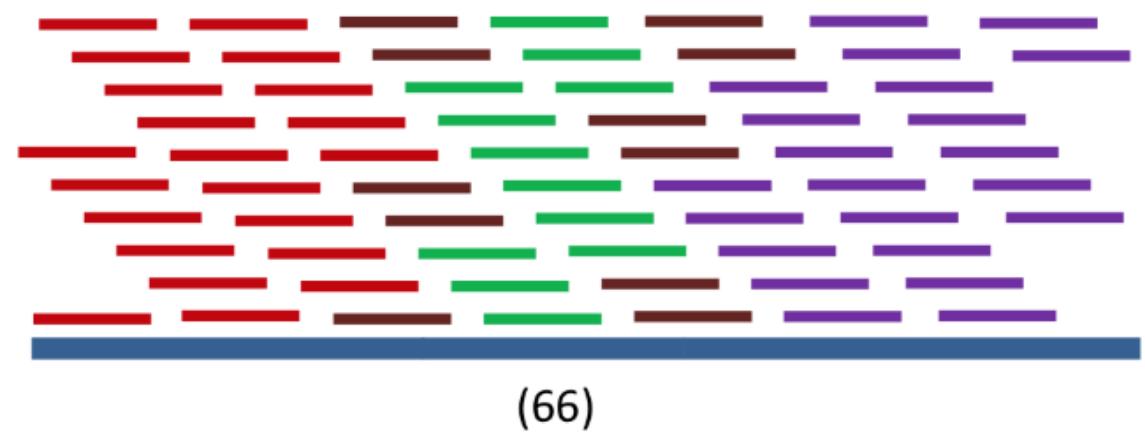
# Contagem das reads

Quanto > a contagem melhor !!

Alignment to Genome – one splice variant



Alignment to Transcriptome



- Genoma referência
- Quantidade de reads x tamanho do gene
- Qualidade do mapeamento

## Quantificação/contagem das reads por gene ou transcritos

**Importante saber como o programa trabalha com :**

- Overlapping das reads ( read completa vs parcial);
- Reads multi-mapeadas;
- Reads que fazem overlapping com muitos genes/features;
- Reads em ítrons;

**Ferramentas Populares de contagem:**

- Htseq-counts (Anders et al., 2014);
- FeatureCounts (Liao et al., 2014)

Genes

- RSEM (Li;Dewey, 2011)
- Cufflinks (Trapnell et al., 2012)

Transcritos

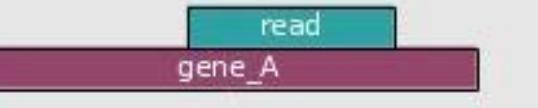
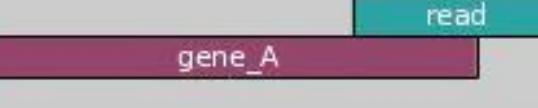
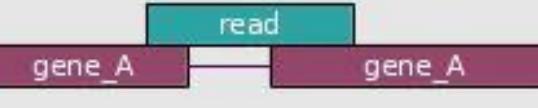
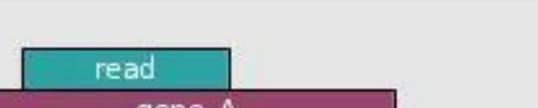
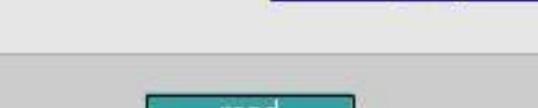
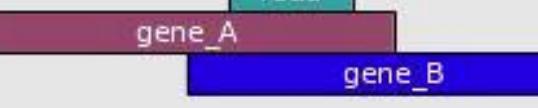


# htseq counts

STAR



htseq e RSEM para  
contagens

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

N_unmapped	85176	85176	85176
N_multimapping	64508	64508	64508
N_noFeature	263202	3389193	278428
N_ambiguous	193680	1704	50928
ENSSCG0000037372	30	3	27
ENSSCG0000027257	294	1	293
ENSSCG0000029697	207	2	205
ENSSCG0000027274	0	0	0
ENSSCG0000027726	59	0	59
ENSSCG0000033475	0	0	0
ENSSCG0000035944	2	2	0
ENSSCG000004010	42	0	42
ENSSCG000004009	0	0	0
ENSSCG0000030155	75	117	83
ENSSCG0000038931	39	11	153
ENSSCG000004008	4	0	4
ENSSCG000004012	3850	10	3840
ENSSCG0000035661	3	3	0
ENSSCG0000032192	0	0	0
ENSSCG000004013	56	0	56
ENSSCG0000037093	0	0	0
ENSSCG0000031703	0	0	0
ENSSCG000004017	0	0	0
ENSSCG000004016	0	0	0
ENSSCG000004018	242	1	241
ENSSCG0000038097	0	0	0
ENSSCG0000027404	9	0	9
ENSSCG0000022580	0	0	0
ENSSCG0000032403	6	0	6
ENSSCG0000037137	0	0	0
ENSSCG0000024562	2	0	2
ENSSCG0000034196	0	0	0
ENSSCG000004023	39	0	39
ENSSCG000004024	462	0	462
ENSSCG000004022	109	0	109
ENSSCG0000032916	90	0	90
ENSSCG0000040654	0	0	0
ENSSCG000004020	154	0	154
ENSSCG000004021	0	0	0
ENSSCG0000036759	0	0	0
ENSSCG0000034946	0	0	0
ENSSCG0000036609	0	0	0
ENSSCG0000036448	0	0	0
ENSSCG000004027	43	0	43
ENSSCG0000034772	0	0	0

HE27\_STAR\_ReadsPerGene.out.tab

## Tabela de contagens bruta



Mas a contagem é feita  
diretamente usando a quantidade  
bruta de reads ??



# Normalização de Contagens

**Normalização → Nos programas de expressão diferencial**

**Por que precisa ser feita uma normalização ???**

Nível de transcrição

Tamanho do gene

Profundidade do sequenciamento

Expressão de outros genes



# Metodologias de normalização

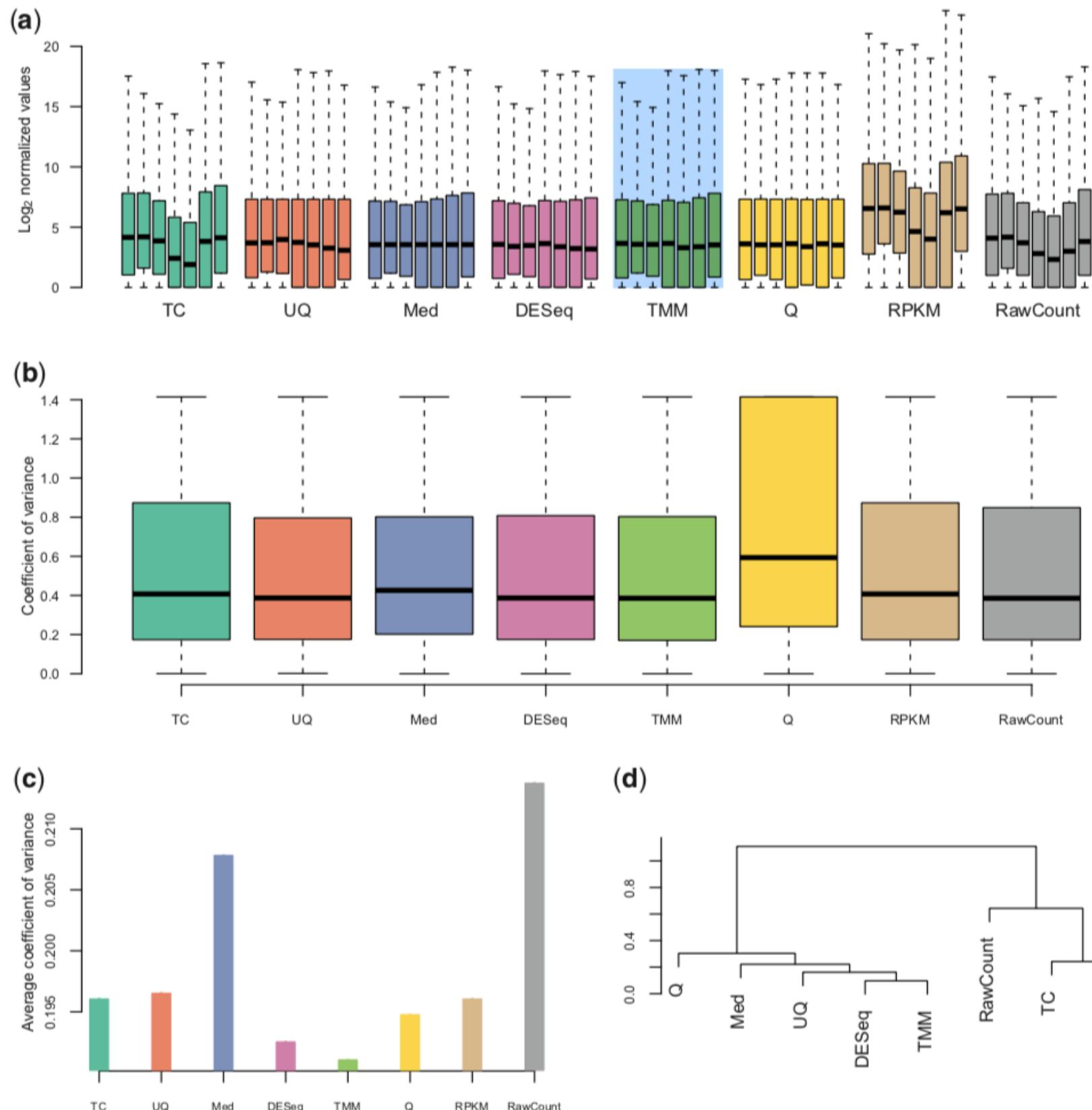
**Table 13:** Normalization methods for the comparison of gene read counts between different conditions. See, for example, Bullard et al. (2010) and Dillies et al. (2013) for comprehensive assessments of the individual methods.

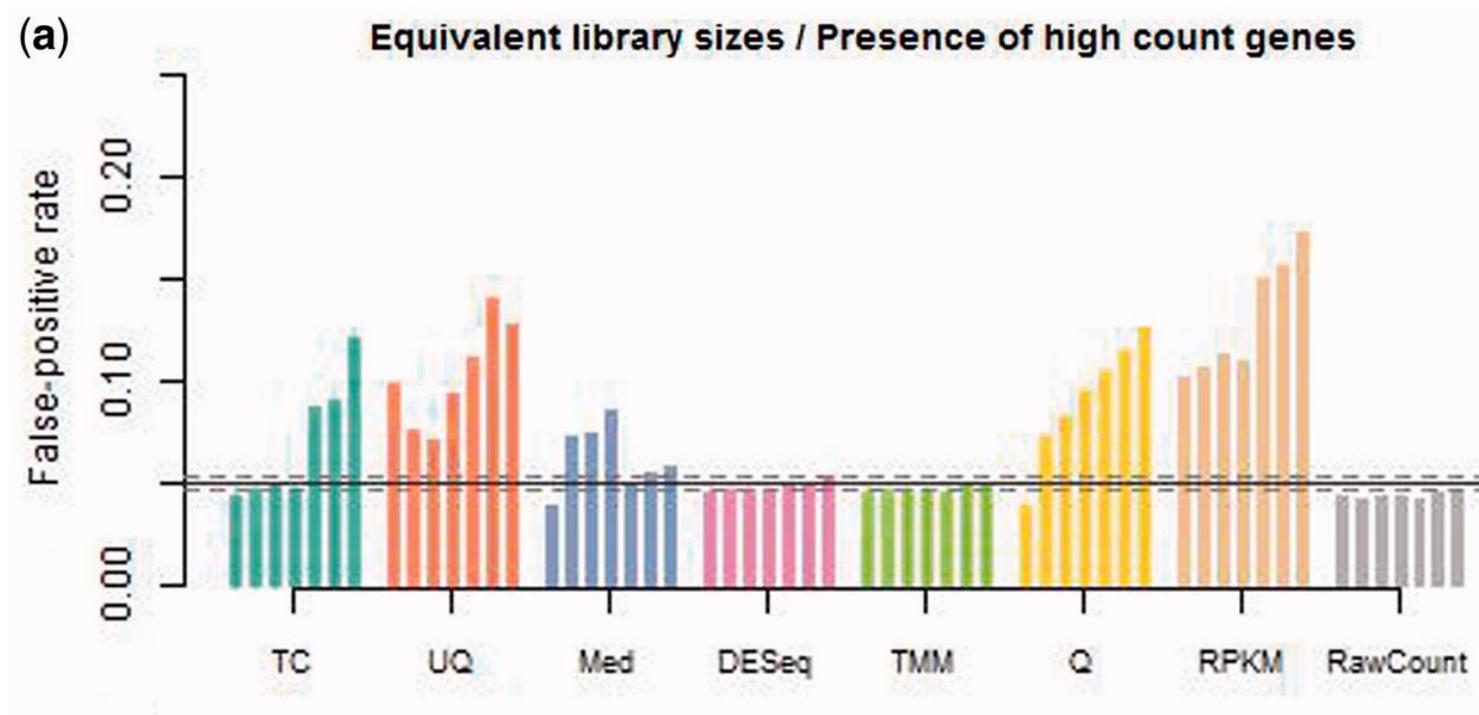
Name	Details	Comment
Total Count	All read counts are divided by the total number of reads (library size) and multiplied by the mean total count across all samples.	<ul style="list-style-type: none"> <li>biased by highly expressed genes</li> <li>cannot account for different RNA repertoire between samples</li> <li>poor detection sensitivity when benchmarked against qRT-PCR (Bullard et al. 2010)</li> </ul>
Counts Per Million	Each gene count is divided by the corresponding library size (in millions).	<ul style="list-style-type: none"> <li>see Total Count</li> </ul>
DESeq's size factor	<ol style="list-style-type: none"> <li>For each gene, the <b>geometric mean</b> of read counts across all samples is calculated.</li> <li>Every gene count is <b>divided by the geometric mean</b>.</li> <li>A sample's size factor is the <b>median of these ratios</b> (skipping the genes with a geometric mean of zero).</li> </ol>	<ul style="list-style-type: none"> <li>the size factor is applied to all read counts of a sample</li> <li>more robust than total count normalization</li> <li>implemented by the DESeq R library (<code>estimateSizeFactors()</code> function), also available in edgeR (<code>calcNormFactors()</code> function with option <code>method = "RLE"</code>)</li> <li>details in Anders and Huber (2010)</li> </ul>
Trimmed Mean of M-values (TMM)	<p>TMM is always calculated as the weighted mean of log ratios between two samples:</p> <ol style="list-style-type: none"> <li>Calculate gene-wise <math>\log_2</math> fold changes (= <b>M-values</b>):</li> <math display="block">M_g = \log_2\left(\frac{Y_{gk}}{N_k}\right)/\log_2\left(\frac{Y_{gk'}}{N_{k'}}\right)</math> <p>where <math>Y</math> is the observed number of reads per gene <math>g</math> in library <math>k</math> and <math>N</math> is the total number of reads.</p> <li><b>Trimming:</b> removal of upper and lower 30%.</li> <li><b>Precision weighing:</b> the inverse of the estimated variance is used to account for lower variance of genes with larger counts.</li> </ol>	<ul style="list-style-type: none"> <li>the size factor is applied to every sample's library size; normalized read counts are obtained by dividing raw read counts by the TMM-adjusted library sizes</li> <li>more robust than total count normalization</li> <li>implemented in edgeR via <code>calcNormFactors()</code> with the default <code>method = "TMM"</code></li> <li>details in Robinson and Oshlack (2010)</li> </ul>
Upper quartile	<ol style="list-style-type: none"> <li>Find the upper quartile value (top 75% read counts after removal of genes with 0 reads).</li> <li>Divide all read counts by this value.</li> </ol>	<ul style="list-style-type: none"> <li>similar to total count normalization, thus it also suffers from a great influence of highly-expressed DE genes</li> <li>can be calculated with edgeR's <code>calcNormFactors()</code> function (<code>method = "upperquartile"</code>)</li> </ul>

**14:** Normalization methods for the comparison of gene read counts within the same sample.

ae	Details	Comment
:M (reads silobase of s per on mapped s)	<ol style="list-style-type: none"> <li>For each gene, count the number of reads mapping to it (<math>X_i</math>).</li> <li>Divide that count by: the length of the gene, <math>l_i</math>, in base pairs divided by 1,000 multiplied by the total number of mapped reads, <math>N</math>, divided by <math>10^6</math>.</li> </ol> $RPKM_i = \frac{X_i}{\left(\frac{l_i}{10^3}\right)\left(\frac{N}{10^6}\right)}$	<ul style="list-style-type: none"> <li>introduces a bias in the per-gene variances, in particular for low expressed genes (Oshlack and Wakefield 2009)</li> </ul>
:M gments per base...)	<ol style="list-style-type: none"> <li>Same as RPKM, but for paired-end reads:</li> <li>The number of fragments (defined by two reads each) is used.</li> </ol>	<ul style="list-style-type: none"> <li>implemented in DESeq2's <code>fPKM()</code> function</li> </ul>
I	<p>Instead of normalizing to the total library size, TPM represents the abundance of an individual gene <math>i</math> in relation to the abundances of the other transcripts (e.g., <math>j</math>) in the sample.</p> <ol style="list-style-type: none"> <li>For each gene, count the number of reads mapping to it and divide by its length in base pairs (= counts per base).</li> <li>Multiply that value by 1 divided by the sum of all counts per base of every gene.</li> <li>Multiply that number by <math>10^6</math>.</li> </ol> $TPM_i = \frac{X_i}{l_i} * \frac{1}{\sum_j \frac{X_j}{l_k}} * 10^6$	<ul style="list-style-type: none"> <li>details in Wagne et al. (2012)</li> </ul>

# Impacto da metologia de contagem





Contagem total e FPKM → Não devem ser usadas para DE

Summary of comparison results for the seven normalization methods under consideration

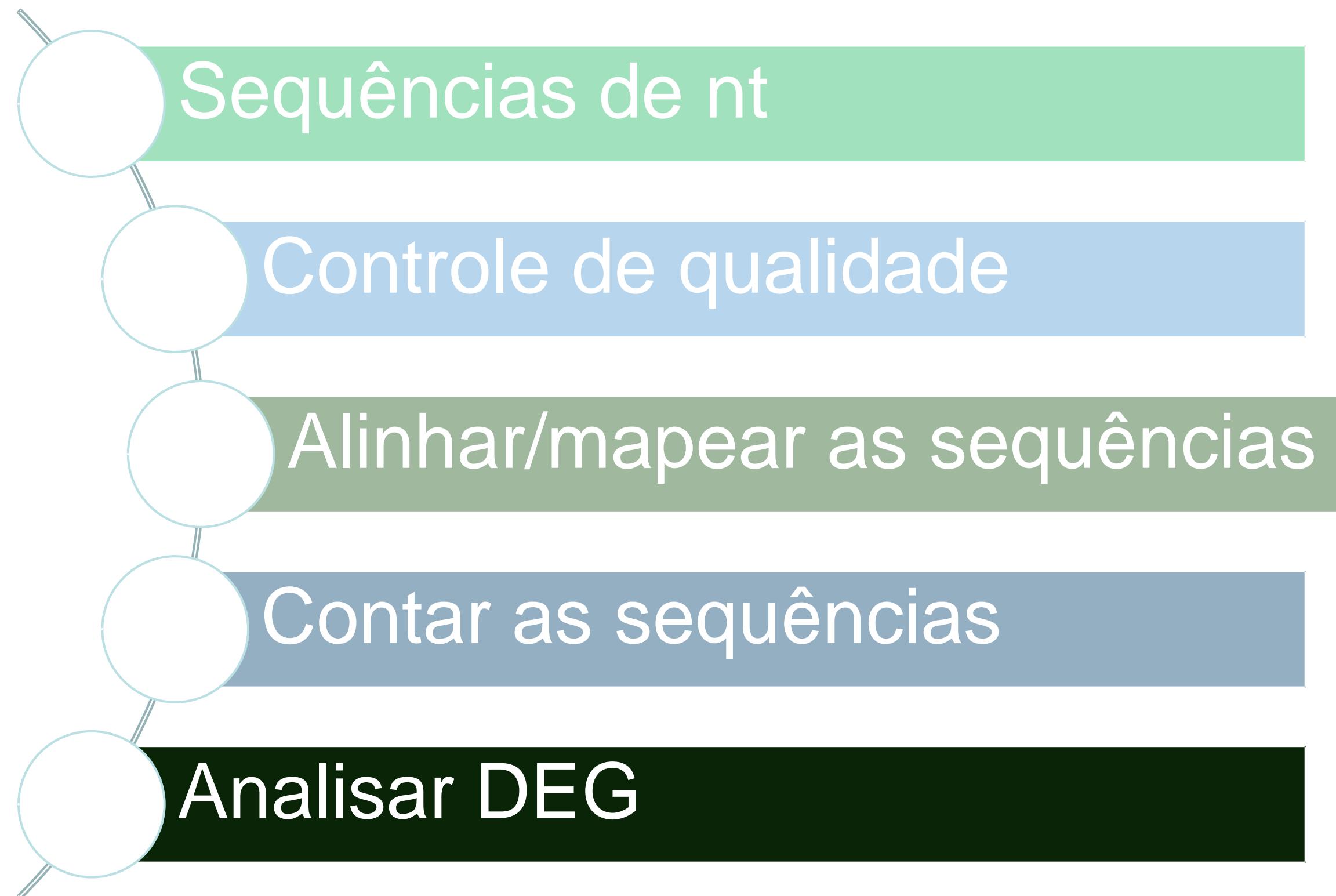
Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	-	+	+	-	-
UQ	++	++	+	++	-
Med	++	++	-	++	-
<b>DESeq</b>	++	++	++	++	++
<b>TMM</b>	++	++	++	++	++
Q	++	-	+	++	-
RPKM	-	+	+	-	-

A ‘-’ indicates that the method provided unsatisfactory results for the given criterion, while a ‘+’ and ‘++’ indicate satisfactory and very satisfactory results for the given criterion.

DeSeq e TMM (EdgeR) → mais robustas

# DeSeq e TMM (EdgeR) → recomendados para DEG

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; <b>NOT for within sample comparisons or DE analysis</b>
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; <b>NOT for DE analysis</b>
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; <b>NOT for between sample comparisons or DE analysis</b>
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 <b>DE analysis</b> ; <b>NOT for within sample comparisons</b>
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 <b>DE analysis</b>



Sequências de nt

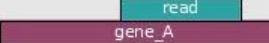
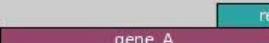
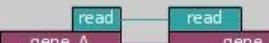
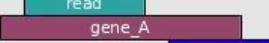
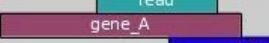
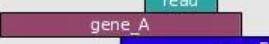
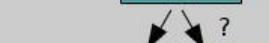
Controle de qualidade

Alinhar/mapear as sequências

Contar as sequências

Analisar DEG

# Análise de expressão gênica diferencial (DEG)

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
 	gene_A	no_feature	gene_A
 	gene_A	no_feature	gene_A
 	gene_A	gene_A	gene_A
 	gene_A	gene_A	gene_A
 	ambiguous (both genes with --nonunique all)	gene_A	gene_A
 	ambiguous (both genes with --nonunique all)		
 	alignment_not_unique (both genes with --nonunique all)		

# DEG

## Objetivos:

- 1- Estimar a magnitude da expressão entre 2 ou mais condições;
- 2- Estimar a significância das diferenças;

DEG → distribuição Binomial Negativa, Poisson, etc.

Semelhante a análises de microarranjos

# DEG

Efeitos específicos da amostras são normalizados:

Específico da amostra:  
Profundidade do sequenciamento  
Composição do RNA

Conteúdo de GC  
Tamanho do gene

# Ferramentas de análise



**Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

**ShrinkSeq**

**NoiSeq**

**baySeq**

**Vsf**

**Voom**

**SAMseq**

**TSPM**

**DESeq**

**EBSeq**

**NBPSeq**

**edgeR**

**limma**

**+ other (not-R)  
including CuffDiff**

# DEG

Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
<b>Seq. depth normalization</b>	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
<b>Dispersion estimate</b>	Cox-Reid approximate conditional inference with focus on maximum <i>individual</i> dispersion estimate	Cox-Reid approximate conditional inference moderated towards the <i>mean</i>	squeezes gene-wise residual variances towards the global variance	
<b>Assumed distribution</b>	Neg. binomial	Neg. binomial	<i>log</i> -normal	Neg. binomial
<b>Test for DE</b>	Wald test (2 factors); LRT for multiple factors	exact test for 2 factors; LRT for multiple factors	<i>t</i> -test	<i>t</i> -test
<b>False positives</b>	Low	Low	Low	High
<b>Detection of differential isoforms</b>	No	No	No	Yes
<b>Support for multi-factored experiments</b>	Yes	Yes	Yes	No
<b>Runtime (3-5 replicates)</b>	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

**Table 2 Summary of the main observations**

		TSPM	
DESeq	<ul style="list-style-type: none"> <li>- Conservative with default settings. Becomes more conservative when outliers are introduced.</li> <li>- Generally low TPR.</li> <li>- Poor FDR control with 2 samples/condition, good FDR control for larger sample sizes, also with outliers.</li> <li>- Medium computational time requirement, increases slightly with sample size.</li> </ul>		<ul style="list-style-type: none"> <li>- Overall highly sample-size dependent performance.</li> <li>- Liberal for small sample sizes, largely unaffected by outliers.</li> <li>- Very poor FDR control for small sample sizes, improves rapidly with increasing sample size. Largely unaffected by outliers.</li> <li>- When all genes are overdispersed, many truly non-DE genes are among the ones with smallest p-values. Remedied when the counts for some genes are Poisson distributed.</li> <li>- Medium computational time requirement, largely independent of sample size.</li> </ul>
edgeR	<ul style="list-style-type: none"> <li>- Slightly liberal for small sample sizes with default settings. Becomes more liberal when outliers are introduced.</li> <li>- Generally high TPR.</li> <li>- Poor FDR control in many cases, worse with outliers.</li> <li>- Medium computational time requirement, largely independent of sample size.</li> </ul>	voom / vst	<ul style="list-style-type: none"> <li>- Good type I error control, becomes more conservative when outliers are introduced.</li> <li>- Low power for small sample sizes. Medium TPR for larger sample sizes.</li> <li>- Good FDR control except for simulation study <math>B_0^{4000}</math>. Largely unaffected by introduction of outliers.</li> <li>- Computationally fast.</li> </ul>
NBPSeq	<ul style="list-style-type: none"> <li>- Liberal for all sample sizes. Becomes more liberal when outliers are introduced.</li> <li>- Medium TPR.</li> <li>- Poor FDR control, worse with outliers. Often truly non-DE genes are among those with smallest p-values.</li> <li>- Medium computational time requirement, increases slightly with sample size.</li> </ul>	baySeq	<ul style="list-style-type: none"> <li>- Highly variable results when all DE genes are regulated in the same direction. Less variability when the DE genes are regulated in different directions.</li> <li>- Low TPR. Largely unaffected by outliers.</li> <li>- Poor FDR control with 2 samples/condition, good for larger sample sizes in the absence of outliers. Poor FDR control in the presence of outliers.</li> <li>- Computationally slow, but allows parallelization.</li> </ul>
		EBSeq	<ul style="list-style-type: none"> <li>- TPR relatively independent of sample size and presence of outliers.</li> <li>- Poor FDR control in most situations, relatively unaffected by outliers.</li> <li>- Medium computational time requirement, increases slightly with sample size.</li> </ul>

## edgeR (empirical analysis of DGE in R)

Considera distribuição binomial negativo

## limma (Linear Models for Microarray Data)

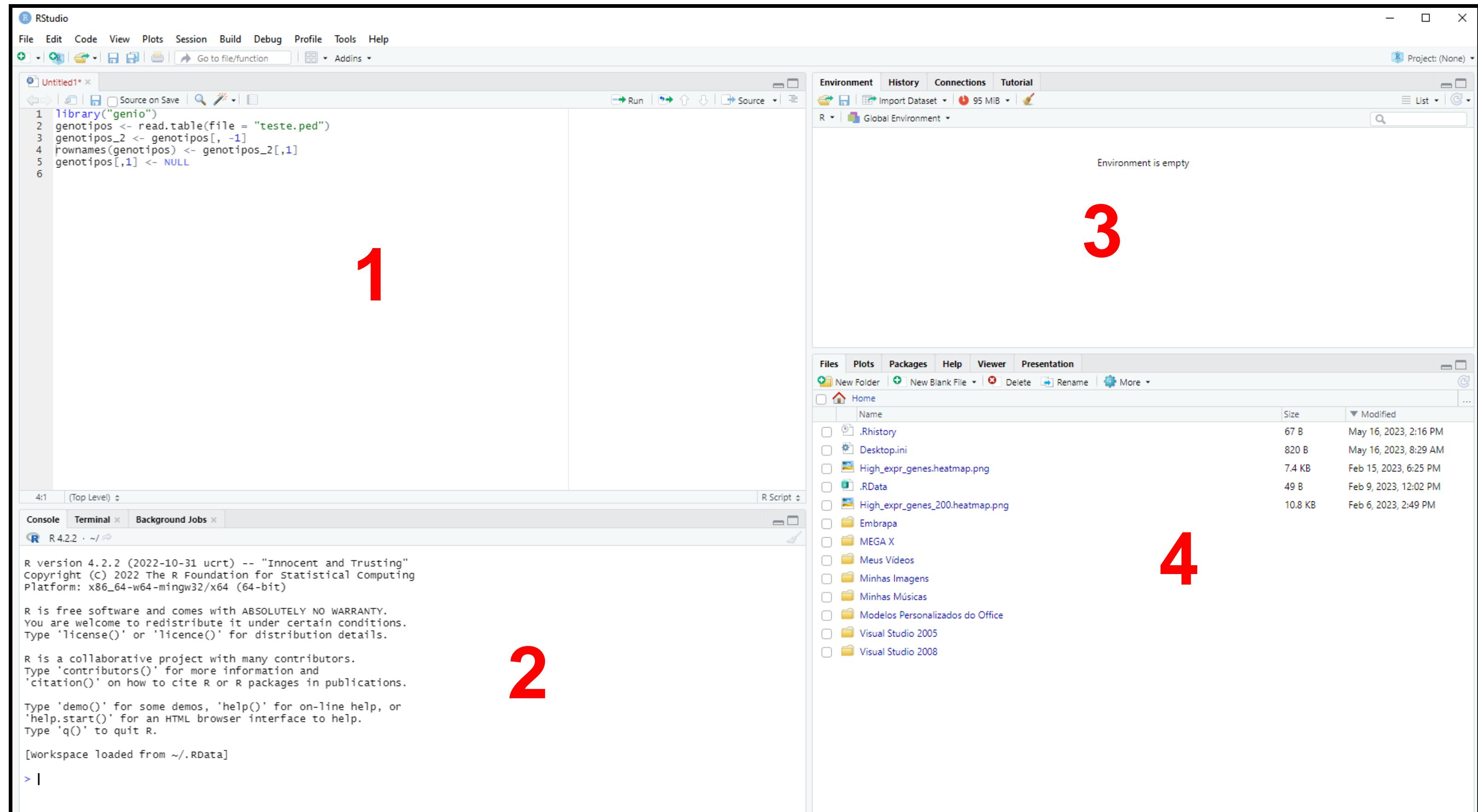
Log-normal

### Pressupostos:

- Cada amostra é sequenciada e as reads são mapeadas apropriadamente no genoma referência.
- O tamanho esperado da contagem: tamanho da biblioteca e a abundância relativa do gene



# No R e Rstudio



# No R e Rstudio

The screenshot shows the Bioconductor website homepage. At the top, there's a dark blue header bar with the Bioconductor logo and navigation links for Home, Install, Help, Developers, and About. A search bar is also present. The main content area has several sections: 'About Bioconductor' (with a mission statement and Docker note), 'Bioc2023 Conference' (mentioning a hybrid conference from August 2-4, 2023), 'Important Notice!' (about branch renaming), 'Install' (with links to software packages and R installation), 'Learn' (with links to courses, training, and support), 'Use' (with links to annotation, experiment packages, and books), and 'Develop' (with links to developer resources and package guidelines). A 'News' sidebar on the left lists recent updates.

## Software de bioinformática e dados biológicos



<https://bioconductor.org/>

# No R e Rstudio

## Instalar pacotes

**Arquivo: instalar\_pacotes.R**

### Instalar pacote individual do Bioconductor:

```
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")

BiocManager::install("Glimma")
```

### Instalar pacote individual do CRAN:

```
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")

BiocManager::install("Glimma")
```

# Iniciando as análises

#Fazer um arquivo.txt com o design do informação das amostras:

Sample_ID	Animal_ID	Tissue	Treatment
T17N	T17N	osso	controle
T42N	T42N	osso	controle
T44N	T44N	osso	controle
T16A	T16A	osso	afetado
T47A	T47A	osso	afetado
T49A	T49A	osso	afetado

# Ler o arquivo dentro do Rstudio e incluir informação dos arquivos de contagem

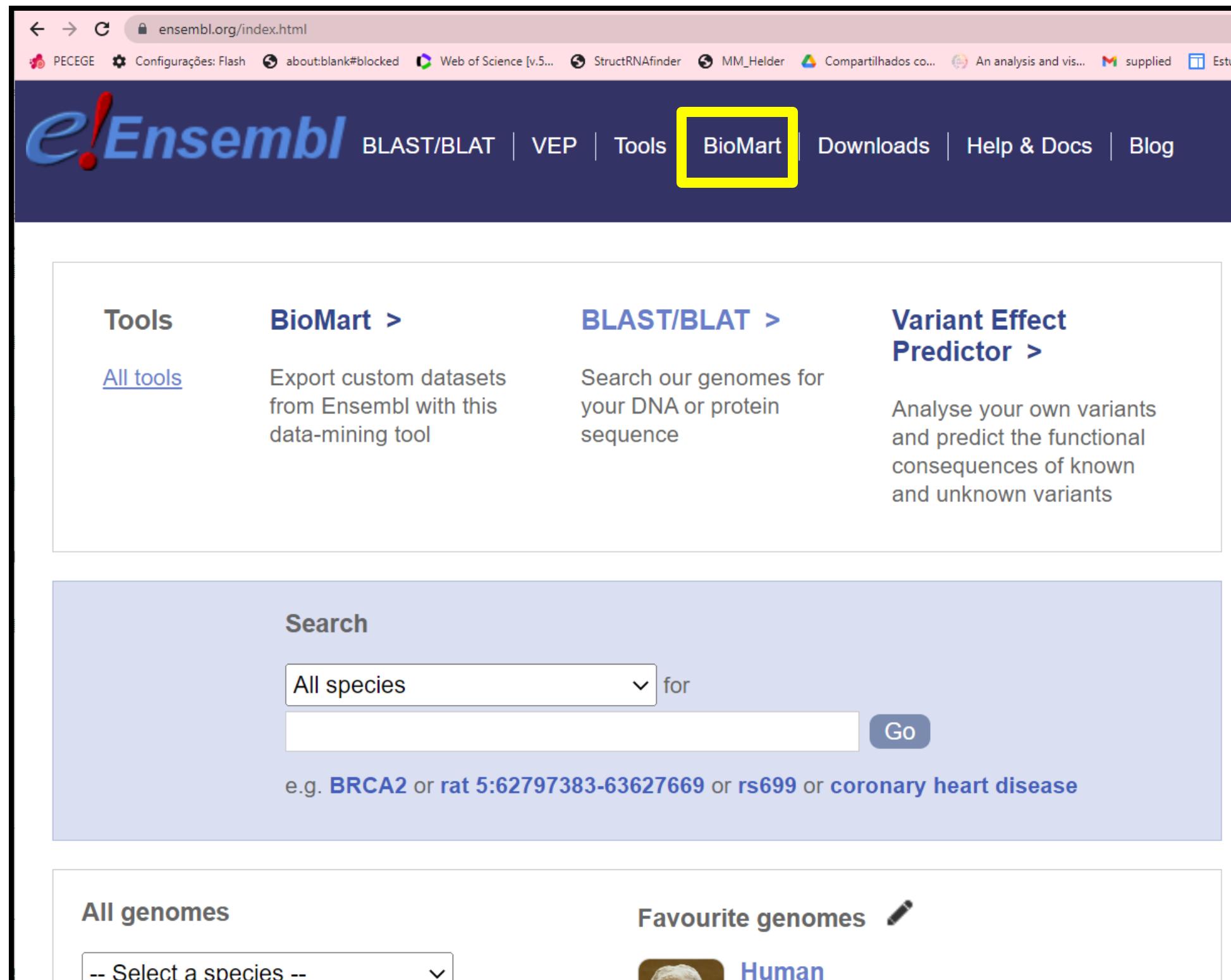
#pode ser feito diretamente no R, mas sempre fazemos um arquivo .txt

```
$ samples <- read.table("amostras.txt", header=T, as.is=T)
$ files <- c("04-GeneCountsSTAR/T16A_ReadsPerGene.counts", "04-GeneCountsSTAR/T17N_ReadsPerGene.counts", "04-GeneCountsSTAR/T42N_ReadsPerGene.counts",
  "04-GeneCountsSTAR/T44N_ReadsPerGene.counts", "04-GeneCountsSTAR/47A_ReadsPerGene.counts", "04-GeneCountsSTAR/T49A_ReadsPerGene.counts")
```

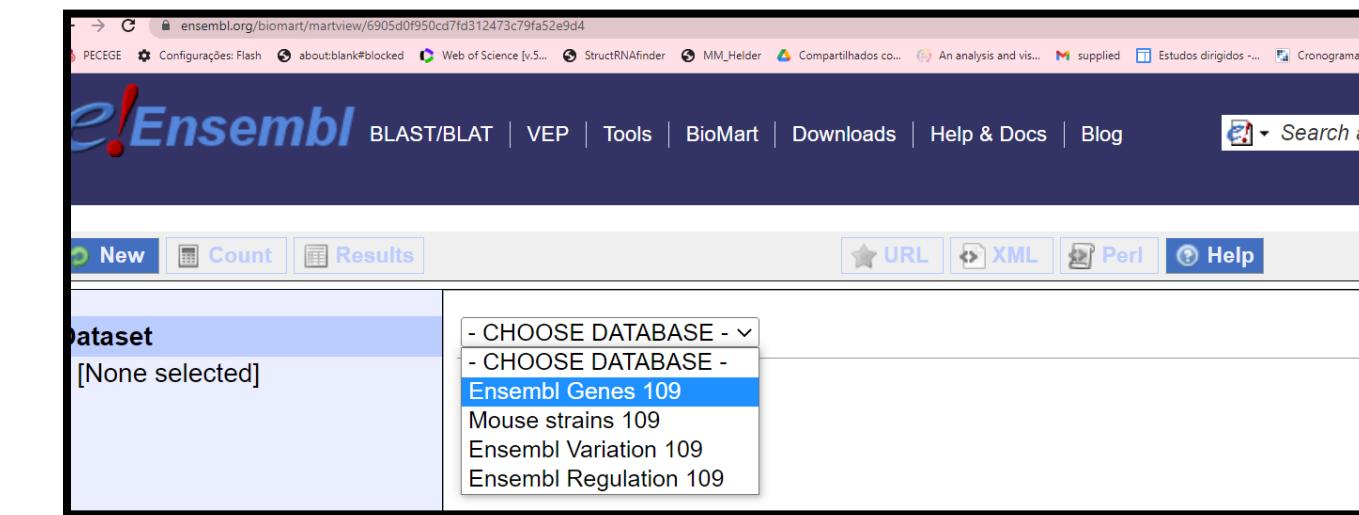
```
$ amostras <- mutate(amostras, files)
```

	Sample_ID	Animal_ID	Tissue	Treatment	files
1	T16A	T16A	osso	afetado	04-GeneCountsSTAR/T16A_ReadsPerGene.counts
2	T17N	T17N	osso	controle	04-GeneCountsSTAR/T17N_ReadsPerGene.counts
3	T42N	T42N	osso	controle	04-GeneCountsSTAR/T42N_ReadsPerGene.counts
4	T44N	T44N	osso	controle	04-GeneCountsSTAR/T44N_ReadsPerGene.counts
5	T47A	T47A	osso	afetado	04-GeneCountsSTAR/47A_ReadsPerGene.counts
6	T49A	T49A	osso	afetado	04-GeneCountsSTAR/T49A_ReadsPerGene.counts

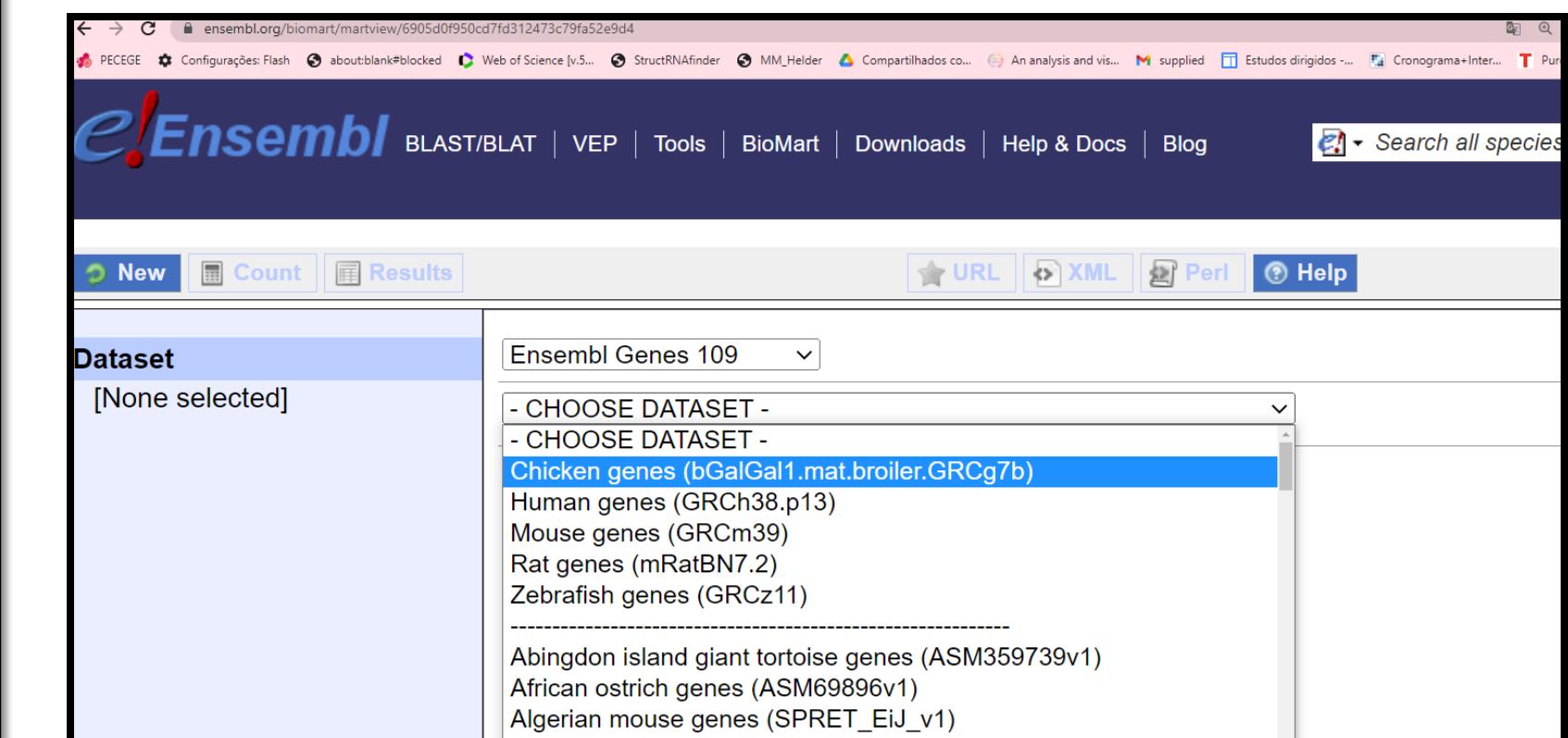
# #Preparar os arquivos de anotação do biomart no site do Ensembl e no R



The screenshot shows the Ensembl homepage with a yellow box highlighting the "BioMart" link in the top navigation bar. Below the navigation, there are four main tool sections: Tools, BioMart >, BLAST/BLAT >, and Variant Effect Predictor >. The BioMart section includes a "All tools" link and a search interface for "All species" and "for". A note below provides examples like BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease. At the bottom, there are links for "All genomes" and "Favourite genomes".



The screenshot shows the Ensembl BioMart interface. The "Dataset" dropdown menu is open, showing options: - CHOOSE DATABASE - (selected), Ensembl Genes 109, Mouse strains 109, Ensembl Variation 109, and Ensembl Regulation 109.



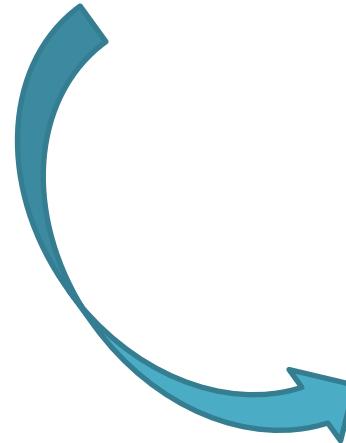
The screenshot shows the Ensembl BioMart interface. The "Dataset" dropdown menu is open, showing options: - CHOOSE DATASET - (selected), Chicken genes (bGalGal1.mat.broiler.GRCg7b) (selected), Human genes (GRCh38.p13), Mouse genes (GRCm39), Rat genes (mRatBN7.2), Zebrafish genes (GRCz11). Below the dropdown, other datasets are listed: Abingdon island giant tortoise genes (ASM359739v1), African ostrich genes (ASM69896v1), Algerian mouse genes (SPRET\_EiJ\_v1), and All genomes (GRCh38).

## #Chamar as contagens e incluir informações na DGE-list

```
```{r, results='hide', warning=FALSE}
#incluindo nome dos files no dataframe:
d <- readDGE(files, header = F)
d$counts <- d$counts[order(rownames(d$counts)),]
class(d)
dim(d$counts)
d.full <- d # objeto backup, caso queria voltar ao original

#Para ver se da match os dois dataframes na mesma ordem, pois counts e anotacao tem que estar na mesma
#ordem
match(genes$ensembl_gene_id, row.names(d$counts))
summary(match(genes$ensembl_gene_id, row.names(d$counts)))

```
```



edgeR e limma

```
```{r}
colnames(d) <- sampleNames
head(d)

d$samples$group <- group

d$genes <- genes
d$genes
```

# Avaliando as amostras

## Contagem por milhão

Baseada no tamanho da biblioteca sequenciada

```
> y$samples
```

	group	lib.size	norm.factors
Sample1	1	10880519	1
Sample2	1	9314747	1
Sample3	1	11959792	1
Sample4	2	7460595	1
Sample5	2	6714958	1

CPM de 1 corresponde a 6-7 na menor amostra sequenciada

# RPKM/FPKM e TPM

## RPKM (Reads Per Kilobase Million)

$$RPKM \text{ of a gene} = \frac{\text{Number of reads mapped to a gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads from given library} \times \text{gene length in bp}}$$

Single-end

## FPKM (Fragments Per Kilobase Million)

Paired-end

## CPM (Copies Per Million)

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

Não leva em consideração  
tamanho do gene

## TPM (Transcripts Per Million)

$$TPM = 10^6 * \frac{\text{reads mapped to transcript / transcript length}}{\text{Sum (reads mapped to transcript / transcript length)}}$$

Melhor para verificar a proporção  
das reads em cada gene

## TPM

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5kb	Gene A	4.71	3.99	5.57
2kb	Gene B	5.29	6	4.426

9.99                    9.99                    9.99

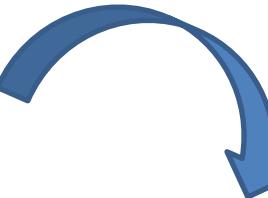
## FPKM

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5kb	Gene A	6.66	5.55	14.16
2kb	Gene B	7.5	8.33	11.25

14.16                    13.88                    25.41

# Genes pouco expressos

Filtrar os genes com poucas reads



Mas.. O que considerar pouco expresso ?

```
$ keep = rowSums(cpm > 1) >= 5 & !is.na(keep)
$ counts = counts[keep,]
$ colnames(counts) = samples$SAMPLE_ID
$ dim(counts)
```

Não há um consenso

Depende de cada experimento

```
> keep <- rowSums(cpm(y) > 0.5) >= 2
> table(keep)
```

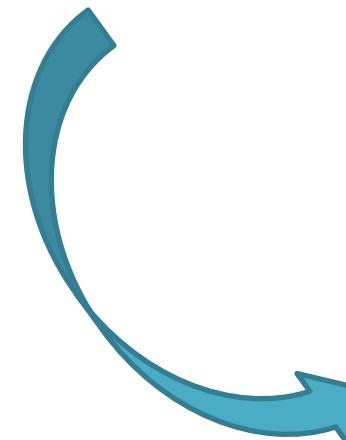
Sempre fazer em cpm

# Genes pouco expressos

```
keep.exprs <- filterByExpr(x, group=group)
x <- x[keep.exprs,, keep.lib.sizes=FALSE]
dim(x)
```



**edgeR**



<b>y</b>	matrix of counts, or a <b>DGEList</b> object, or a <b>SummarizedExperiment</b> object.
<b>design</b>	design matrix. Ignored if <b>group</b> is not <b>NULL</b> .
<b>group</b>	vector or factor giving group membership for a oneway layout, if appropriate.
<b>lib.size</b>	library size, defaults to <b>colSums(y)</b> .
<b>min.count</b>	numeric. Minimum count required for at least some samples.
<b>min.total.count</b>	numeric. Minimum total count required.
<b>large.n</b>	integer. Number of samples per group that is considered to be “large”.
<b>min.prop</b>	numeric. Minimum proportion of samples in the smallest group that express the gene.
<b>...</b>	any other arguments. For the <b>DGEList</b> and <b>SummarizedExperiment</b> methods, other arguments will be passed to the default method. For the <b>default</b> method, other arguments are not currently used.

# Normalizando para o tamanho das bibliotecas

```
x <- calcNormFactors(x, method = "TMM")
x$samples$norm.factors
## [1] 0.894 1.025 1.046 1.046 1.016 0.922 0.996 1.086 0.984
```

Scaling factor



Não altera o tamanho das bibliotecas



edgeR e limma → TMM

```
calcNormFactors(object,
                 method = c("TMM", "TMMwsp", "RLE", "upperquartile", "none"),
                 refColumn = NULL, logratioTrim = .3, sumTrim = 0.05, doWeighting = TRUE,
                 Acutoff = -1e10, p = 0.75, ...)
```

# Como o edgeR faz a normalização das amostras ??

Genes	Amostras	
	T16A	T17N
ENSGALG00010000002	0	0
ENSGALG00010000003	2832	3255
ENSGALG00010000004	0	0
ENSGALG00010000005	2084	2129
ENSGALG00010000006	0	0
ENSGALG00010000007	10000	17142
ENSGALG00010000008	0	0
ENSGALG00010000009	0	0
ENSGALG00010000010	0	0
ENSGALG00010000012	0	0
ENSGALG00010000013	0	0
ENSGALG00010000014	0	0
ENSGALG00010000015	0	0
ENSGALG00010000016	0	0
ENSGALG00010000017	106566	135908
ENSGALG00010000018	0	0
Total de reads	121482	158434



Genes	Amostras	
	T16A	T17N
ENSGALG00010000003	2832	3255
ENSGALG00010000005	2084	2129
ENSGALG00010000007	10000	17142
ENSGALG00010000017	106566	135908
Total de reads	121482	158434

1. Remove os genes pouco expressos

2. Escolhe uma amostra referência

Scaling fator → amostra mais mediana

Amostras		Amostras	
T16A	T17N	T16A	T17N
2832/121482	3255/158434	0,023	0,021
2084/121482	2129/158434	0,017	0,013
10000/121482	17142/158434	0,082	0,108
10324/121482	12844/158434	0,877	0,858

- 2.1 Calcula o quartil de 75%  
2.2. Faz a média dos quartis de 75%  
2.3. Escolhe a amostra que tem o valor mais próximo da média dos quartis 75%

Média Q75% = 0,095

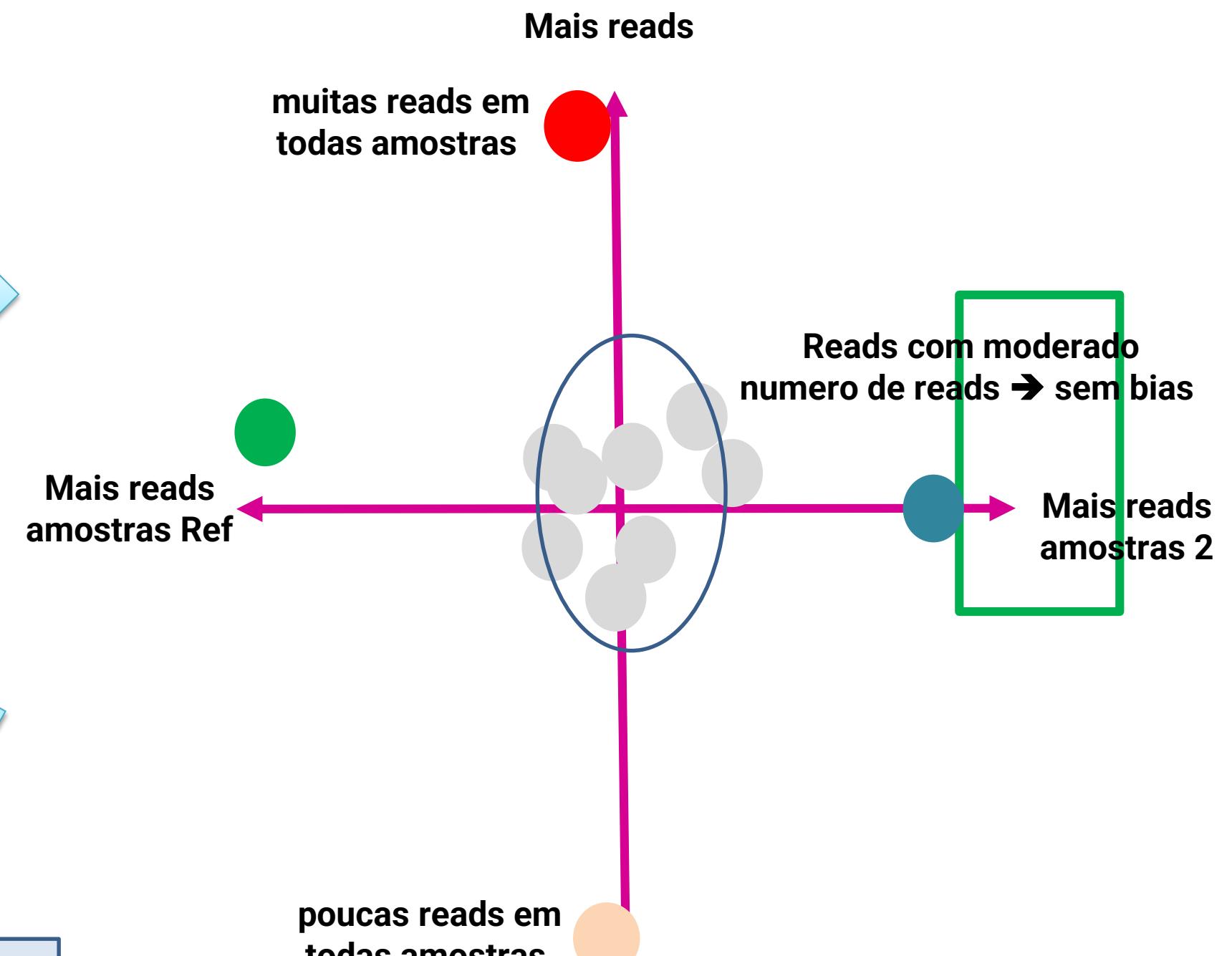
AMOSTRA  
REFERÊNCIA

# Como o edgeR faz a normalização das amostras ??

## 3. Escolhe genes para fazer o scaling factor

Amostras	
T16A	T17N
2832/121482	3255/158434
2084/121482	2129/158434
10000/121482	17142/158434
10324/121482	12844/158434

Amostras	
T16A	T17N
0,023	0,021
0,017	0,013
0,082	0,108
0,877	0,858



$$\log_2\left(\frac{\text{ref, gene 1}}{\text{amostra 2, gene 1}}\right)$$

# Como o edgeR faz a normalização das amostras ??

## 3. Escolhe genes para fazer o scaling factor

$$\log_2\left(\frac{\text{ref, gene 1}}{\text{amostra 2, gene 1}}\right)$$



Remove genes com log2 Inf



**Tabela 1 → genes tendenciosos**

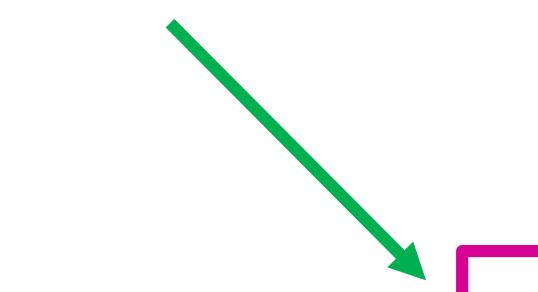
$$\left(\frac{\log_2(\text{ref, gene 1}) + \log_2(\#2, \text{gene 1})}{2}\right)$$



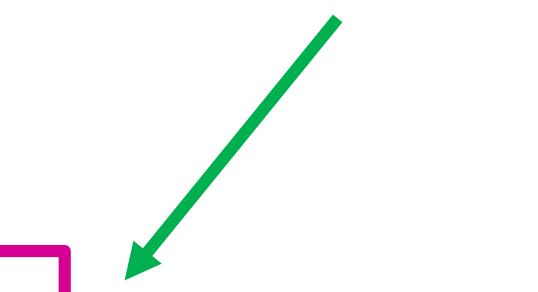
Calcula média geométrica dos logs para todos os genes e amostras



**Tabela 2 → Verifica genes altamente expressos e pouco expressos**



Ordena em ordem crescente



Filtre 30% dos genes de cada extremo

**Genes comuns → scaling factors**

Filtre 5% dos genes de cada extremo

# Como o edgeR faz a normalização das amostras ??

Genes comuns → scaling factors



4. Calcula a média ponderada das razões de log2 remanescentes



“WEIGHTED TRIMMED MEAN OF THE LOG2 RATIOS”



“Trima” ou filtra os genes extremos



DIMINUI O EFEITO DE OUTLIERS

# Como o edgeR faz a normalização das amostras ??

Genes que passaram no filtro



4. Calcula a média ponderada das razões de log2 remanescentes



Genes com mais reads → mais peso



Genes pouco expressos → > variância

	Sample #1	Sample #2	$\log_2(\text{ratio})$
Gene #1	202	101	1
Gene #2	204	101	1.01
Gene #3	206	101	1.02
Gene #4	2	1	1
Gene #5	4	1	2
Gene #6	6	1	2.6

5. Calcula a média ponderada das razões de log2

"normal numbers"

$2^{\text{média ponderada das } \log_2 \#2} = \text{scaling factor}$



faz para todas amostras



Valores são centrados em torno de 1



Centered scaling factor

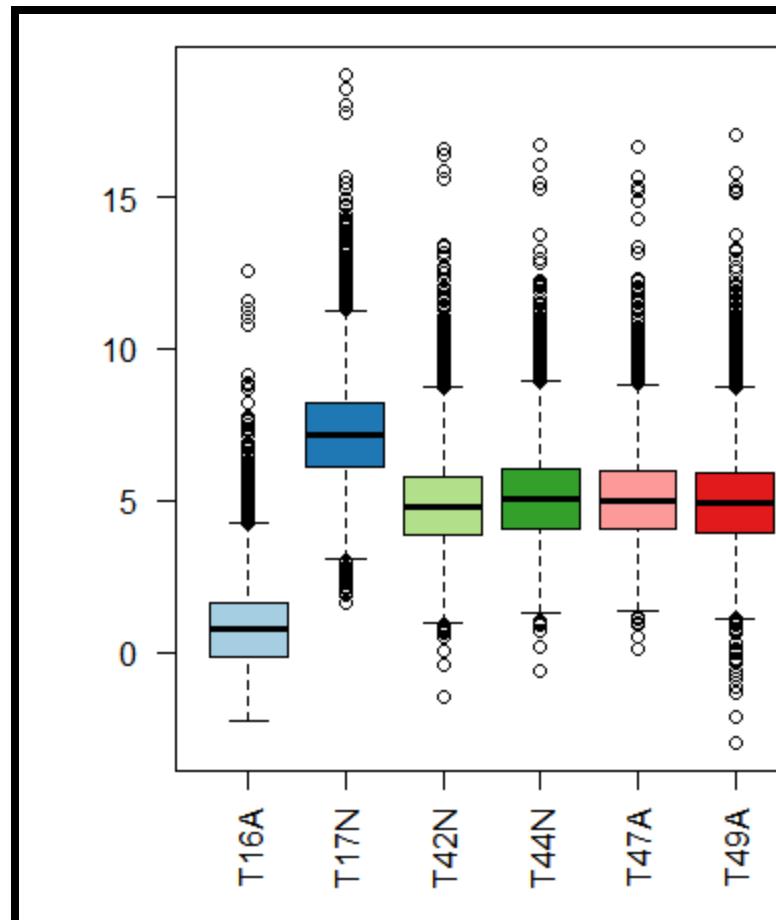


$$\frac{\text{raw scaling factor}}{\text{média geométrica da amostras}}$$

## ANTES

```
> d$samples
```

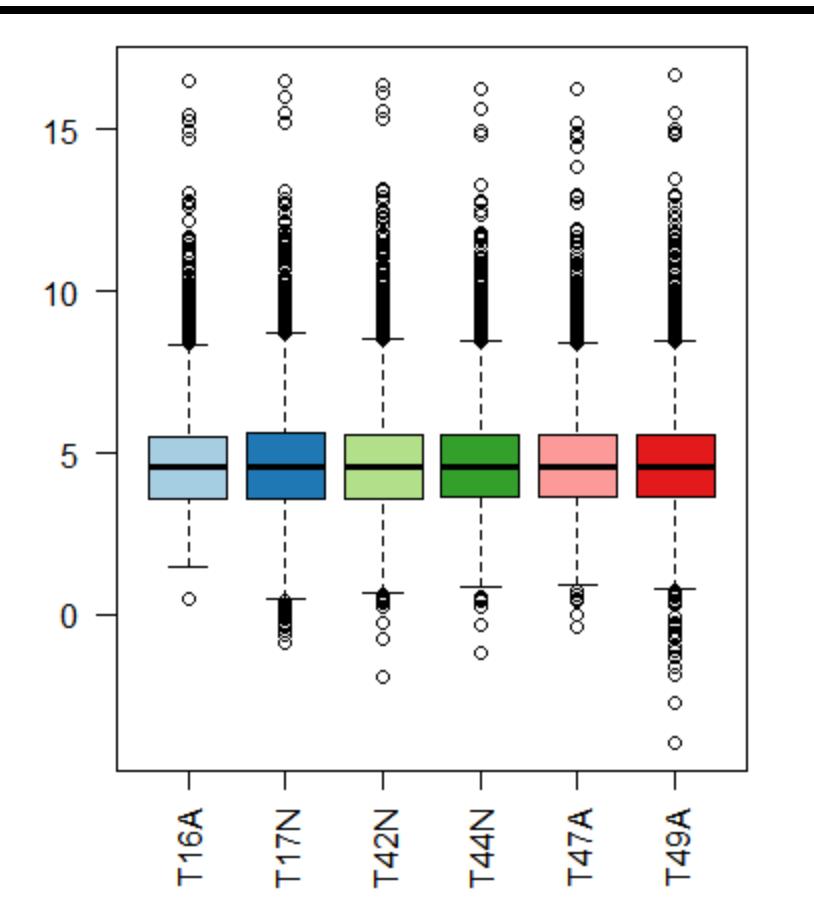
files	group	lib.size	norm.factors
T16A 04-GeneCountsSTAR/T16A_ReadsPerGene.counts	afetado	11447263	1
T17N 04-GeneCountsSTAR/T17N_ReadsPerGene.counts	controle	17097830	1
T42N 04-GeneCountsSTAR/T42N_ReadsPerGene.counts	controle	13150041	1
T44N 04-GeneCountsSTAR/T44N_ReadsPerGene.counts	controle	21096609	1
T47A 04-GeneCountsSTAR/T47A_ReadsPerGene.counts	afetado	15119253	1
T49A 04-GeneCountsSTAR/T49A_ReadsPerGene.counts	afetado	18723518	1



## DEPOIS

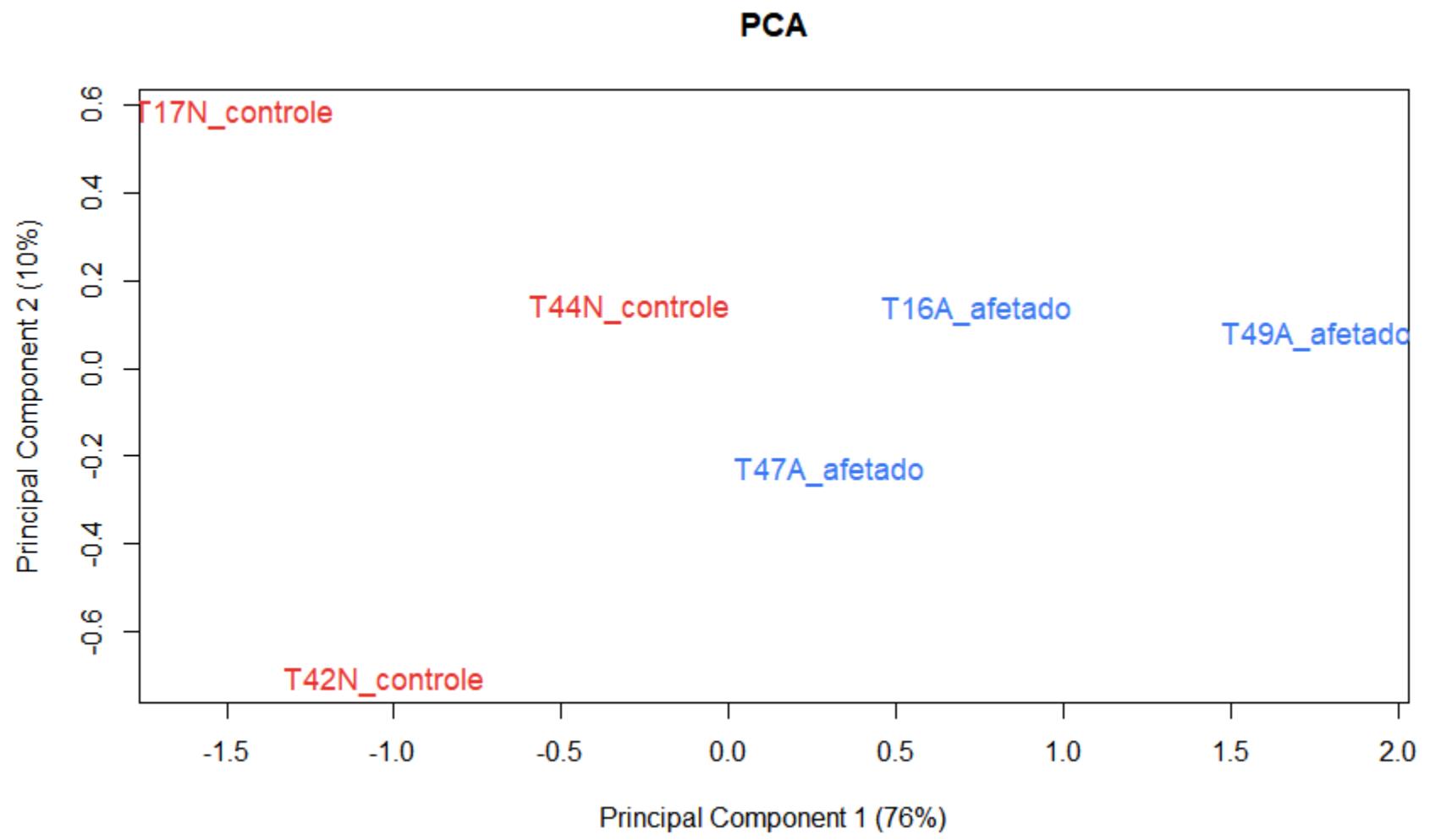
```
> d$samples
```

files	group	lib.size	norm.factors
T16A 04-GeneCountsSTAR/T16A_ReadsPerGene.counts	afetado	11447263	1.0067830
T17N 04-GeneCountsSTAR/T17N_ReadsPerGene.counts	controle	17097830	0.9470692
T42N 04-GeneCountsSTAR/T42N_ReadsPerGene.counts	controle	13150041	0.9366670
T44N 04-GeneCountsSTAR/T44N_ReadsPerGene.counts	controle	21096609	1.0818763
T47A 04-GeneCountsSTAR/T47A_ReadsPerGene.counts	afetado	15119253	1.0604755
T49A 04-GeneCountsSTAR/T49A_ReadsPerGene.counts	afetado	18723518	0.9759306

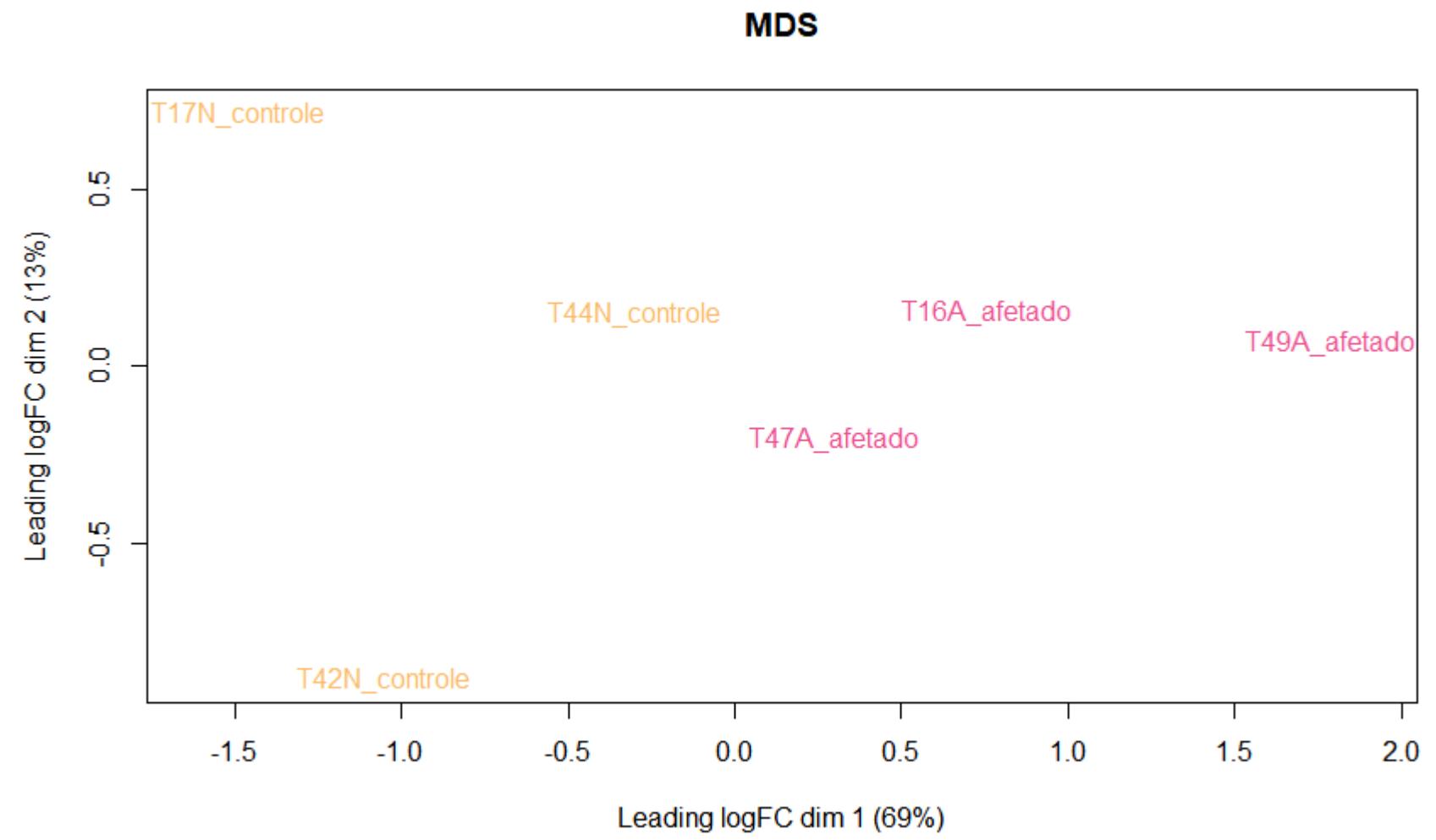


# Verificando a distribuição das amostras

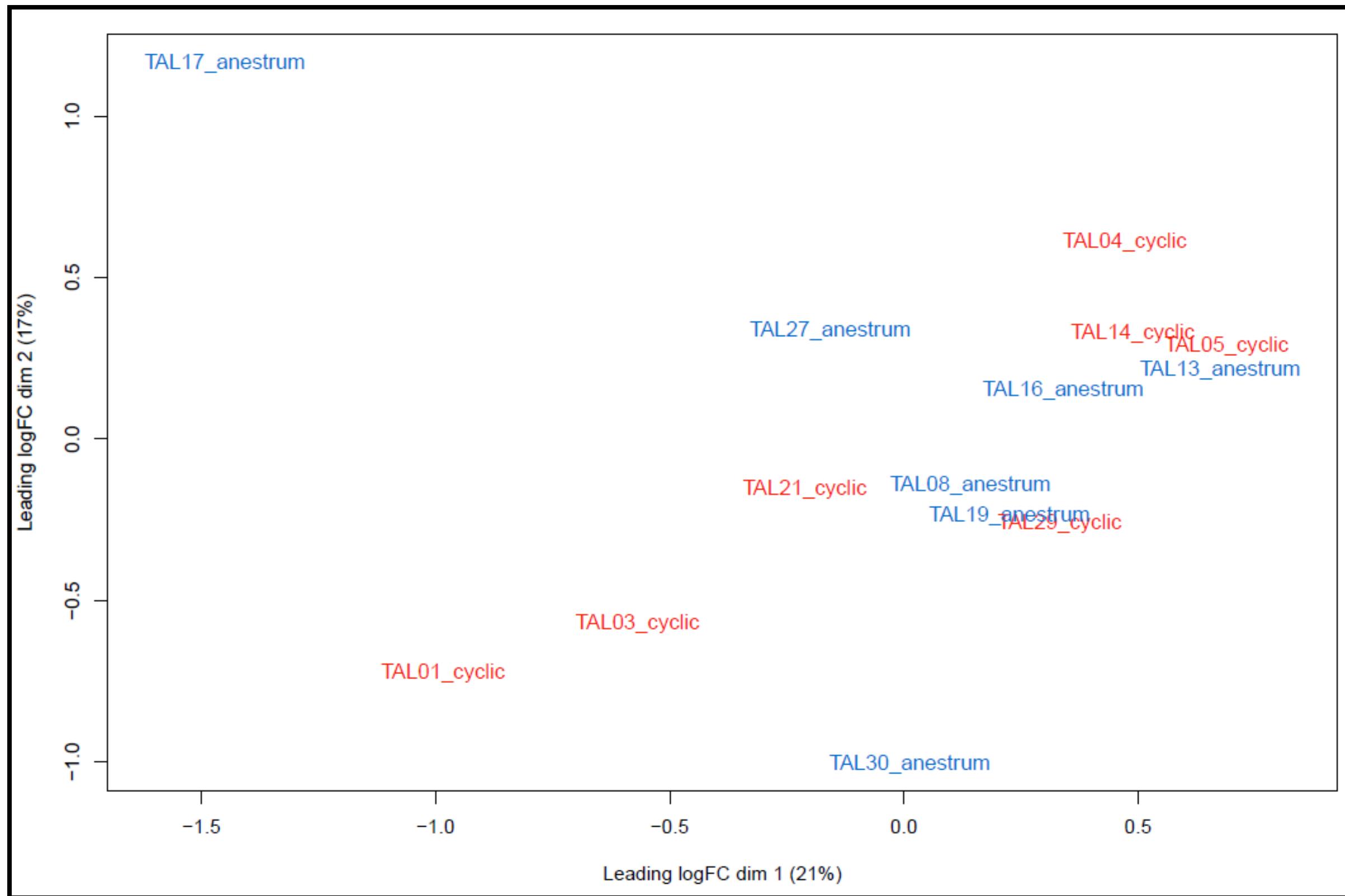
**PCA**



**MDS**



## E neste caso ????



# Criando a matriz de design e contrastes

Várias formas de fazer → modelo de análise

```
{r}
design <- model.matrix(~0+group)
colnames(design) <- gsub("group", "", colnames(design))
rownames(design) <- sampleNames

contr.matrix <- makeContrasts("Afetado-Normal"=afetado-normal, levels=design)
```

F1000Research  
F1000Research 2020, 9:1444 Last updated: 18 JUL 2022  
Check for updates

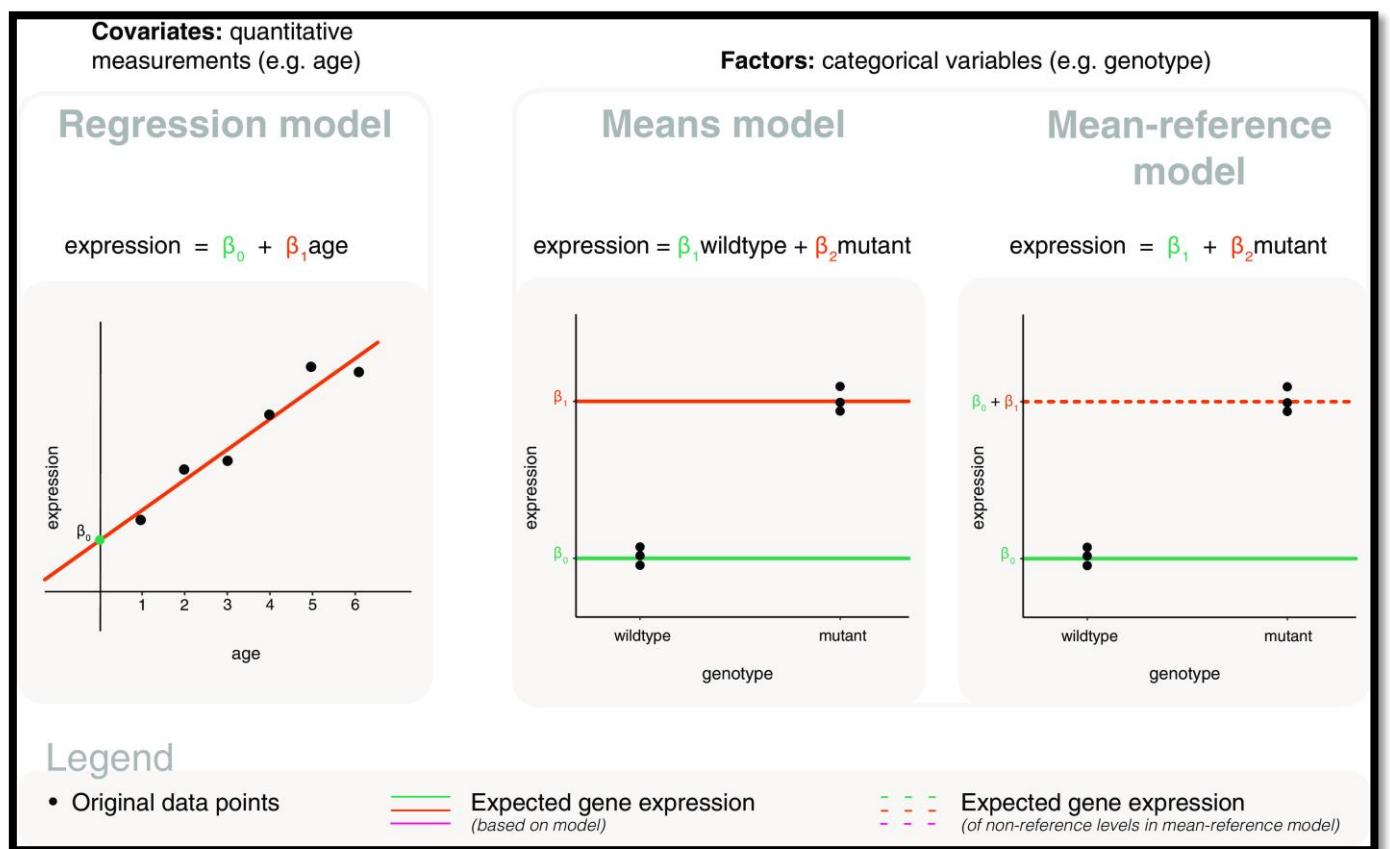
METHOD ARTICLE  
A guide to creating design matrices for gene expression experiments [version 1; peer review: 2 approved]

Charity W. Law<sup>1,2</sup>, Kathleen Zeglinski<sup>1,3</sup>, Xueyi Dong<sup>1,2</sup>, Monther Alhamdoosh<sup>1,3</sup>, Gordon K. Smyth<sup>1,4</sup>, Matthew E. Ritchie<sup>1,2,4</sup>

<sup>1</sup>The Walter and Eliza Hall Institute of Medical Research, Parkville, 3052, Australia  
<sup>2</sup>Department of Medical Biology, The University of Melbourne, Parkville, 3010, Australia  
<sup>3</sup>Research and Development, CSL Limited, Bio21 Institute, Parkville, 3010, Australia  
<sup>4</sup>School of Mathematics and Statistics, The University of Melbourne, Parkville, 3010, Australia

v1 First published: 10 Dec 2020, 9:1444  
<https://doi.org/10.12688/f1000research.27893.1>  
Latest published: 10 Dec 2020, 9:1444  
<https://doi.org/10.12688/f1000research.27893.1>

Abstract  
Differential expression analysis of genomic data types, such as RNA-seq experiments, use linear models to determine the size and direction of the changes in gene expression. For RNA-seq, there are several established software packages for this purpose accompanied with analysis pipelines that are well described. However, there are two crucial steps in the analysis process that can be a stumbling block for many – the set up of an appropriate model via design matrices and the set up of comparisons of interest via contrast matrices. These steps are particularly troublesome because an extensive catalogue for design and contrast matrices does not currently exist. One would usually search for example case studies across different platforms and mix and match the advice from those sources to suit the dataset they have at hand. This article guides the reader through the basics of how to set up design and contrast matrices. We take a practical approach by providing code and graphical representation of each case study, starting with simpler examples (e.g. models with a single explanatory variable) and moving onto more complex ones (e.g. interaction models, mixed effects models, higher order time series and cyclical models). Although our work has been written specifically with a limma-style pipeline in mind, most of it is also applicable to other software packages for differential



Explanatory variables	Design matrix	Section
age	model.matrix(~age) model.matrix(~0+age)	Covariates: With intercept Covariates: Without intercept
group <i>HEALTHY, SICK</i>	model.matrix(~group) model.matrix(~0+group)	Factors: With intercept Factors: Without intercept

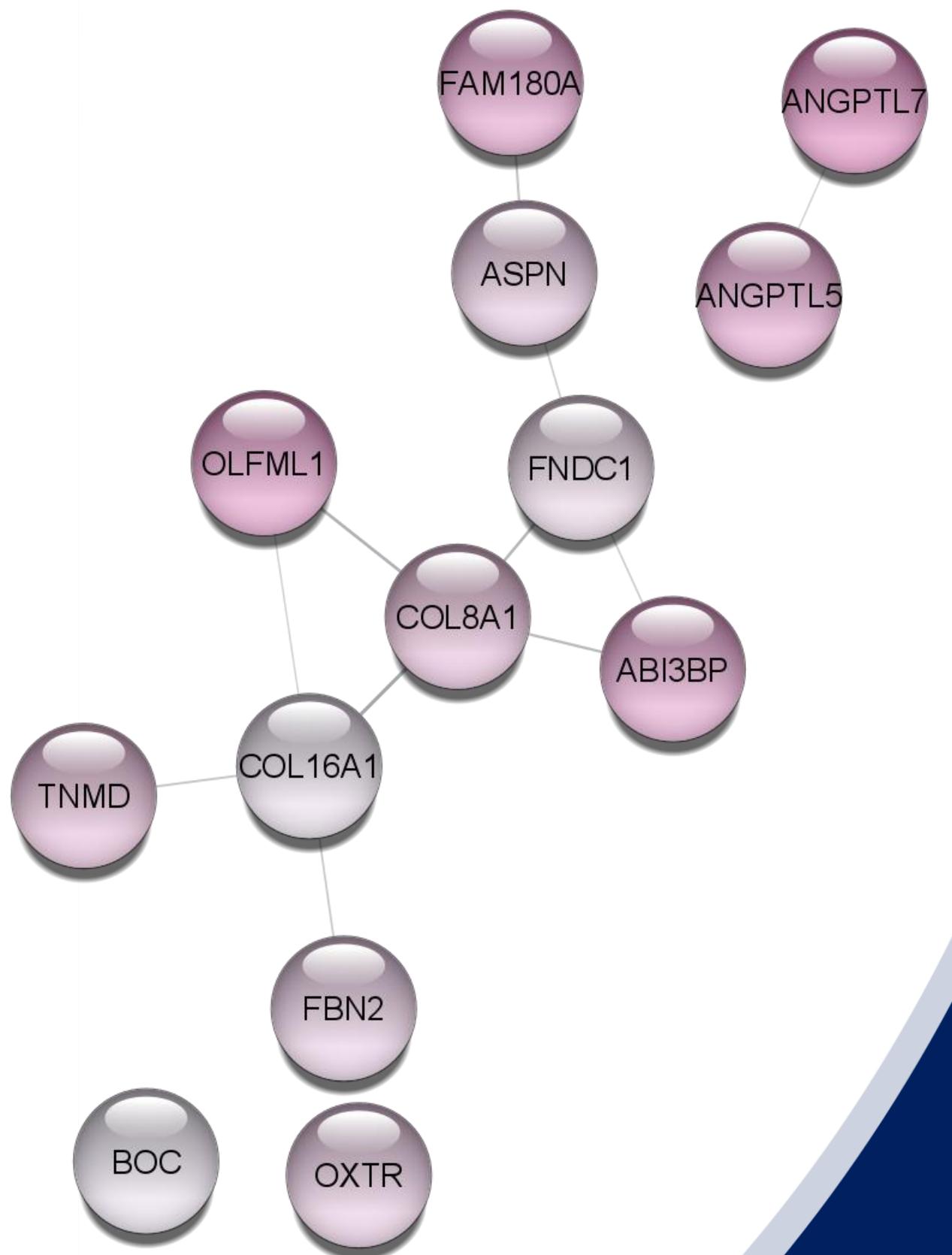
# Correção de múltiplos testes

Testes dos genes não é independente → tem que considerar na análise

## Benjamini and Hochberg False Discovery Rate

Common Methods Ranked by Stringency	Mais falso negativos
Bonferroni	
Bonferroni Step-Down	
Westfall and Young Permutation	
Benjamini and Hochberg False Discovery Rate	Mais falsos positivos
None	

# Análises Funcionais



# Achei meus genes DE, e agora???

UM MUNDO DE OPÇÕES SE ABRE!!!!!!

Quais genes eu vou estudar ?  
Quantos genes tenho que avaliar?  
O que esses genes fazem?  
De onde vieram ????  
E onde vivem??



# Análises funcionais

## MÉTODOS DE ENRIQUECIMENTOS MAIS COMUNS:

Análise de enriquecimento fucional

Análise de vias metabólicas

Análise de Redes Gênicas

Integração com outros dados, .....

Over-representation analysis (ORA)

- ↳ Genes diferencialmente expressos
- ↳ Genes expressos (*background*)

Functional class scoring (FCS)

- ↳ Genes diferencialmente expressos + níveis de expressão (logFC)
- ↳ Genes expressos (*background*)
- ↳ Gene Set Enrichment Analysis (GSEA)

# Análises funcionais

## Ontologias gênicas

 GENEONTOLOGY  
Unifying Biology

About    Ontology    Annotations    Downloads    Help

## THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

 InterPro

Classification of protein families

Home    ▶ Search    ▶ Browse    ▶ Results    Release notes    Download    ▶ Help    ▶ About

 InterPro 94.0  
9 May 2023

### Classification of protein families

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

▼ Citing InterPro  
If you find InterPro useful, please cite the reference that describes this work:

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileshi ML, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A. *InterPro in 2022*. *Nucleic Acids Research*, Nov 2022, (doi: 10.1093/nar/gkac993)

## Vias metabólicas

KEGG    Databases    Tools    Auto annotation    Kanehisa Lab

 KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

KEGG2    PATHWAY    BRITE    MODULE    KO    GENES    COMPOUND    NETWORK    DISEASE    DRUG

Select prefix:   Enter keywords:  Go    Help

[ New pathway maps ]

 reactome

About    Content    Docs    Tools    Community    Download

Find Reactions, Proteins and Pathways

e.g. O95631, NTN1, signaling by EGFR, glucose    Go!

 Pathway Browser

Visualize and interact with Reactome biological pathways

 Analysis Tools

Merges pathway identifier mapping, over-representation, and expression analysis

 ReactomeFIViz

Designed to find pathways and network patterns related to cancer and other types of diseases

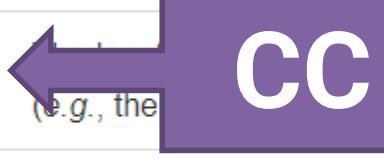
 Documentation

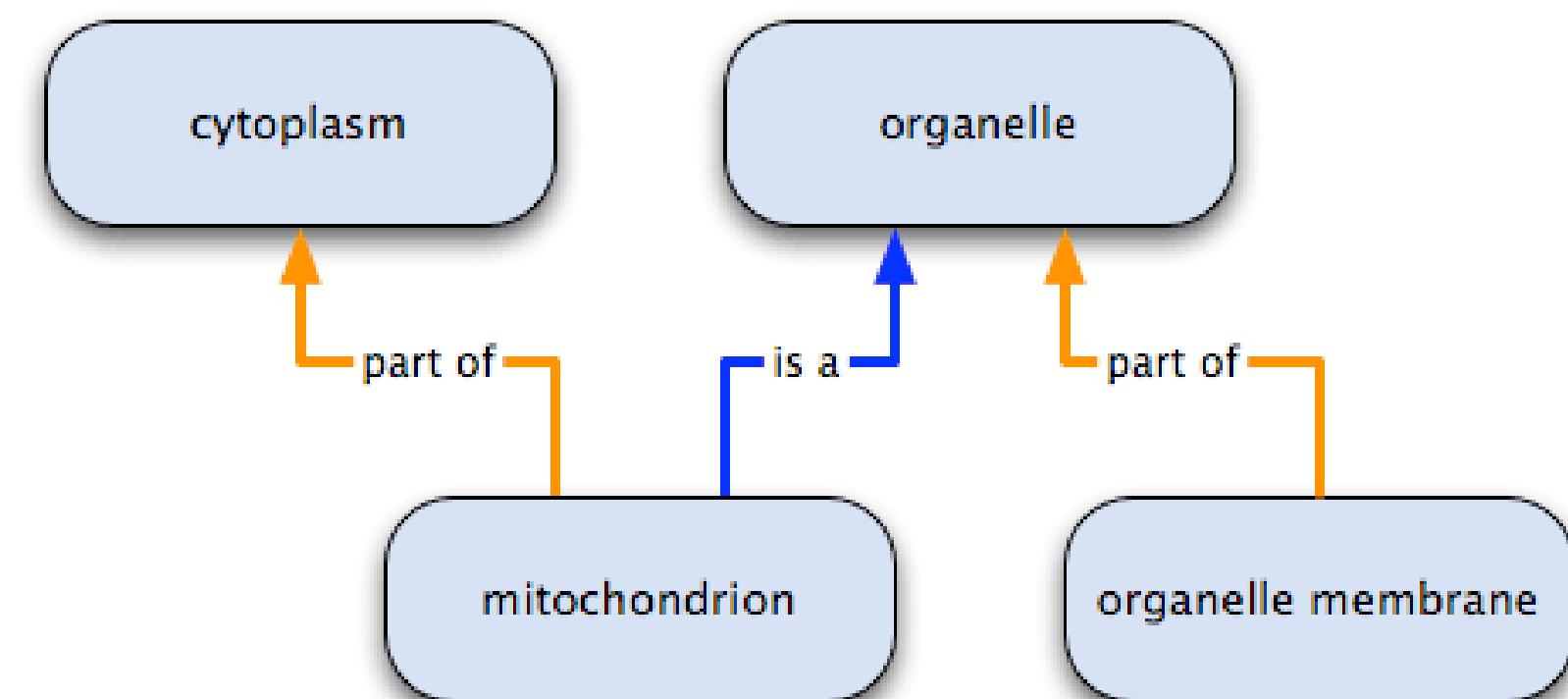
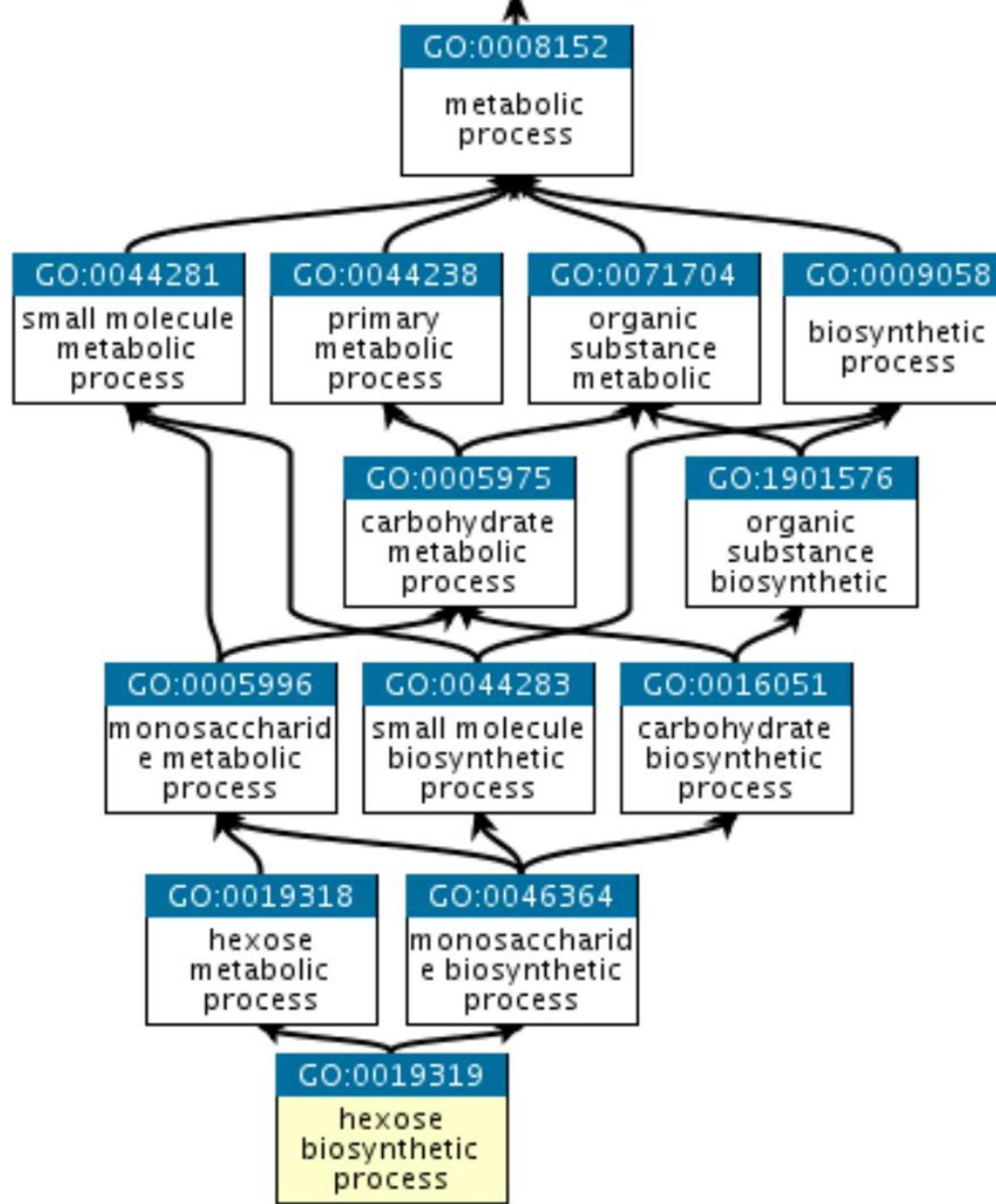
Information to browse the database and use its principal tools for data analysis

# Ontologias gênicas

## Gene Ontology overview

An ontology is a formal representation of a body of knowledge within a given domain. Ontologies usually consist of a set of classes (or terms or concepts) with [relations](#) that operate between them. The Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects:

<b>Molecular Function</b>	 <b>MF</b> Example: <i>represor</i>	Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as "catalysis" or "transport". GO molecular function terms represent the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally can be performed by individual gene products (i.e. a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products. Broad terms are <a href="#">catalytic activity</a> and <a href="#">transporter activity</a> ; examples of narrower functional terms are <a href="#">adenylate cyclase activity</a> or <a href="#">Toll-like receptor binding</a> . To avoid confusion between gene product names and their molecular functions, GO molecular functions are often appended with the word "activity" (a <i>protein kinase</i> would have the GO molecular function <i>protein kinase activity</i> ).
<b>Cellular Component</b>	 <b>CC</b> Example: <i>(e.g., the nucleus)</i>	Cellular structures in which a gene product performs a function, either cellular compartments (e.g., <a href="#">mitochondrion</a> ), or stable macromolecular complexes of which they are parts (e.g., the ribosome). In the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy.
<b>Biological Process</b>	 <b>BP</b> Example: <i>pyrimidine nucleotide metabolism</i>	The larger processes, or 'biological processes' are composed of one or more molecular activities. Examples of broad biological process terms are <a href="#">DNA repair</a> or <a href="#">signal transduction</a> . Examples of more specific terms are <a href="#">pyrimidine nucleotide metabolism</a> or <a href="#">glucose transmembrane transport</a> . Note that a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would describe a pathway.



# BASES DE DADOS MAIS COMUNS

## Ontologias gênicas

## Vias metabólicas

KEGG      Databases      Tools      Auto annotation      Kanehisa Lab



**KEGG PATHWAY Database**  
Wiring diagrams of molecular interactions, reactions and relations

KEGG2    PATHWAY    BRITE    MODULE    KO    GENES    COMPOUND    NETWORK    DISEASE    DRUG

Select prefix:      Enter keywords:

[ New pathway maps ]



About    Content    Docs    Tools    Community    Download

Find Reactions, Proteins and Pathways  
e.g. O95631, NTN1, signaling by EGFR, glucose

 Pathway Browser  
Visualize and interact with Reactome biological pathways

 Analysis Tools  
Merges pathway identifier mapping, over-representation, and expression analysis

 ReactomeFIViz  
Designed to find pathways and network patterns related to cancer and other types of diseases

 Documentation  
Information to browse the database and use its principal tools for data analysis

# BASES DE DADOS MAIS COMUNS

DAVID

david.ncifcrf.gov

PECEGE Configurações: Flash about:blank#blocked Web of Science [v.5... StructRNAtinder MM\_Helder Compartilhados

## DAVID Bioinformatics Resources

Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service About DAVID A

### Overview

The Database for Annotation, Visualization and Integrated Discovery (DAVID) provides a comprehensive set of functional annotation tools for investigators to understand the biological meaning behind large lists of genes. These tools are powered by the comprehensive DAVID Knowledgebase built upon the DAVID Gene concept which pulls together multiple sources of functional annotations. For any given gene list, DAVID tools are able to:

## Annotation Summary Results

Help and Tool Manual

Current Gene List: demolist1

145 DAVID IDs

Current Background: Homo sapiens

Check Defaults

[Clear All](#)

### Disease (2 selected)

<input type="checkbox"/> DISGENET	64.8%	94	<a href="#">Chart</a>	
<input type="checkbox"/> GAD_DISEASE	80.7%	117	<a href="#">Chart</a>	
<input type="checkbox"/> GAD_DISEASE_CLASS	80.7%	117	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> OMIM_DISEASE	34.5%	50	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> UP_KW_DISEASE	30.3%	44	<a href="#">Chart</a>	

### Functional\_Annotations (5 selected)

<input checked="" type="checkbox"/> UP_KW_BIOLOGICAL_PROCESS	62.8%	91	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> UP_KW_CELLULAR_COMPONENT	84.8%	123	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> UP_KW_MOLECULAR_FUNCTION	62.1%	90	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> UP_KW_PTM	80.7%	117	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> UP_SEQ_FEATURE	93.8%	136	<a href="#">Chart</a>	

### Gene\_Ontology (3 selected)

<input type="checkbox"/> GOTERM_BP_1	92.4%	134	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_BP_2	92.4%	134	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_BP_3	92.4%	134	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_BP_4	91.0%	132	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_BP_5	89.0%	129	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_BP_ALL	92.4%	134	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> GOTERM_BP_DIRECT	92.4%	134	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_BP_FAT	91.7%	133	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_CC_1	95.2%	138	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_CC_2	94.5%	137	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_CC_3	93.8%	136	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_CC_4	91.0%	132	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_CC_5	87.6%	127	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_CC_ALL	95.2%	138	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> GOTERM_CC_DIRECT	95.2%	138	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_CC_FAT	91.7%	133	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_MF_1	91.7%	133	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_MF_2	91.0%	132	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_MF_3	86.9%	126	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_MF_4	82.1%	119	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_MF_5	67.6%	98	<a href="#">Chart</a>	
<input type="checkbox"/> GOTERM_MF_ALL	91.7%	133	<a href="#">Chart</a>	
<input checked="" type="checkbox"/> GOTERM_MF_DIRECT	91.7%	133	<a href="#">Chart</a>	

# BASES DE DADOS MAIS COMUNS

## PANTHER

The mission of the PANTHER knowledgebase is to support biomedical and other research by providing **comprehensive information about the evolution of protein-coding gene families**, particularly protein phylogeny, function and genetic variation impacting that function. [Learn more](#)

PANTHER selected as a [Global Core Biodata Resource](#). [Click](#) for more details.

search keyword

Home About Data Version Tools API/Services Publications Workspace Downloads FAQ/Help/Tutorial Login Register Contact us

Current Release: PANTHER 17.0 | 15,619 family phylogenetic trees | 143 species | News Whole genome function views

Gene List Analysis Browse Sequence Search Genetic Variant Impact Keyword Search

Please refer to our article in [Nature Protocols](#) for detailed instructions on how to use this page.

**Help Tips**

Steps:

- 1. Select list and list type to analyze
- 2. Select Organism
- 3. Select operation

[Using enhancer data](#)

1. Enter ids and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.

Enter IDs:   
Supported IDs  
Upload IDs:  Nenhum arquivo escolhido  
[File format](#)

Please [login](#) to be able to select lists from your workspace.

Select List Type:  ID List  
 Previously exported text search results  
 Workspace list  
 PANTHER Generic Mapping  
 ID's from Reference Proteome Genome  
Organism for id list   
 VCF File Flanking region 20 Kb

2. Select organism.

Homo sapiens  
Mus musculus  
Rattus norvegicus  
Gallus gallus  
Danio rerio

3. Select Analysis.

Functional classification viewed in gene list  
 Functional classification viewed in graphic charts  
 Bar chart  
 Pie chart  
 Statistical overrepresentation test  
 Statistical enrichment test

# BASES DE DADOS MAIS COMUNS

## MSigdb

The screenshot shows the homepage of the Molecular Signatures Database (MSigDB). At the top, there's a navigation bar with links for GSEA Home, Downloads, Molecular Signatures Database (which is the active tab), Documentation, Contact, and Team. Below the navigation is a logo for "MSigDB Molecular Signatures Database" featuring a blue icon of three overlapping circles. The main content area has a pink header box containing text about funding support. To the left is a sidebar with links for MSigDB Home, Human Collections, Mouse Collections, and Help. The main content area includes sections for Overview, Human Collections, and a grid of gene set types labeled H through C8.

**We need your help: Update on GSEA/MSigDB funding support**

Last November we submitted a proposal to NCI's Information Technology for Cancer Research (ITCR) program for the continued funding of GSEA and MSigDB. Unfortunately, our proposal was not funded in this round, but we were encouraged to resubmit for the next one. This funding is critical for our continuing support and enhancement of the GSEA-MSigDB resource.

For our original submission many of you sent us emails of support, an important requirement for these grants. We now ask for your help again. We would greatly appreciate a short email message from you describing how the resource has been of value to your work and any concerns you may have about its continued availability.

Please send us your message of support to [gsea-los@broadinstitute.org](mailto:gsea-los@broadinstitute.org) on or before **Monday June 5, 2023**.

Thanks in advance for your help and support.  
The GSEA/MSigDB Team.

**UC San Diego**

**BROAD INSTITUTE**

**Overview**

The Molecular Signatures Database (MSigDB) is a resource of tens of thousands of annotated gene sets for use with GSEA software, divided into Human and Mouse collections. From this web site, you can

- ▶ **Examine** a gene set and its annotations. See, for example, the [HALLMARK\\_APOPTOSIS](#) human gene set page.
- ▶ **Browse** gene sets by name or collection.
- ▶ **Search** for gene sets by keyword.
- ▶ **Investigate** gene sets:
  - ▶ **Compute overlaps** between your gene set and gene sets in MSigDB.
  - ▶ **Categorize** members of a gene set by gene families.
  - ▶ **View the expression profile** of a gene set in a provided public expression compendia.
  - ▶ Investigate the gene set in the online **biological network repository** [NDEX](#)
- ▶ **Download** gene sets.

**Human Collections**

<b>H</b> <b>hallmark gene sets</b> are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.	<b>C5</b> <b>ontology gene sets</b> consist of genes annotated by the same ontology term.
<b>C1</b> <b>positional gene sets</b> corresponding to human chromosome cytogenetic bands.	<b>C6</b> <b>oncogenic signature gene sets</b> defined directly from microarray gene expression data from cancer gene perturbations.
<b>C2</b> <b>curated gene sets</b> from online pathway databases, publications in PubMed, and knowledge of domain experts.	<b>C7</b> <b>immunologic signature gene sets</b> represent cell states and perturbations within the immune system.
<b>C3</b> <b>regulatory target gene sets</b> based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.	<b>C8</b> <b>cell type signature gene sets</b> curated from cluster markers identified in single-cell sequencing studies of human tissue.
<b>C4</b> <b>computational gene sets</b> defined by mining large collections of cancer-oriented microarray data.	

**License Terms**

# Ferramentas/Softwares

Não seguro | bioinformatics.sdstate.edu/go/

PECEGE Configurações: Flash about:blank#blocked Web of Science [v.5... StructRNAfinder MM\_Helper Compartilhados co... An analysis and vis... M supplied Estudos dirigidos ... Cronograma+Inter... T P

## ShinyGO 0.77

Select or search your species:

- Best matching species
- Demo genes

Reset

5/1/2023: ShinyGO 0.80 release in testing mode. Thanks to Jenny's hardwork, we update to Ensembl release 107 which includes protists and 1 bacteria. We also included 14,094 species from STRING-DB 11.5.

We urgently need your emails of support

We are working on a grant proposal (due May 31st) to redevelop, improve, and maintain ShinyGO. If you briefly state your general needs including major findings, we can use it as a support letter. Also include any features requests such as multiple gene sets. Without your help, we will not be able to continue this project.

Jan. 19, 2023: Thanks to a user's feedback, we found a serious bug in ShinyGO 0.76. As some genes are represented by multiple IDs in calculating enrichment. We believe this is fixed. If you pasted Ensembl gene IDs to ShinyGO 0.76 between April 4, 2022 and Jan. 19, 2023, please double check your results with other tools such as G:profiler, Enrichr, STRING-db, and DataMiner.

If this server is busy, please use a mirror sever <http://ge-lab.org/go/> hosted by NSF-funded JetStream2.

Email Jenny for questions, suggestions or data contributions. Follow Dr Ge on Twitter for updates.

Feb. 11, 2022: Like ShinyGO but your genome is not covered? Customized ShinyGO is now available. Its database includes species not covered by ShinyGO. To add a new species/genome, fill in this Form.

### A graphical tool for gene enrichment analysis

Just paste your gene list to get enriched GO terms and other pathways for over 420 plant and animal species, based on annotations. An additional 5000 genomes (including bacteria and fungi) are annotated based on STRING-db (v.11). In addition, it also produces hierarchical clustering trees and networks summarizing overlapping terms/pathways, protein-protein interaction networks, gene expression networks, and many other example outputs below:

digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/

Configurações: Flash about:blank#blocked Web of Science [v.5... StructRNAfinder MM\_Helper Compartilhados co... An analysis and vis... M supplied Estudos dirigidos ... Cronograma+Inter... T P

BIOINFORMATICS POWERED BY INGENUITY Alcibia BIOBASE OmicSoft

English 日本語

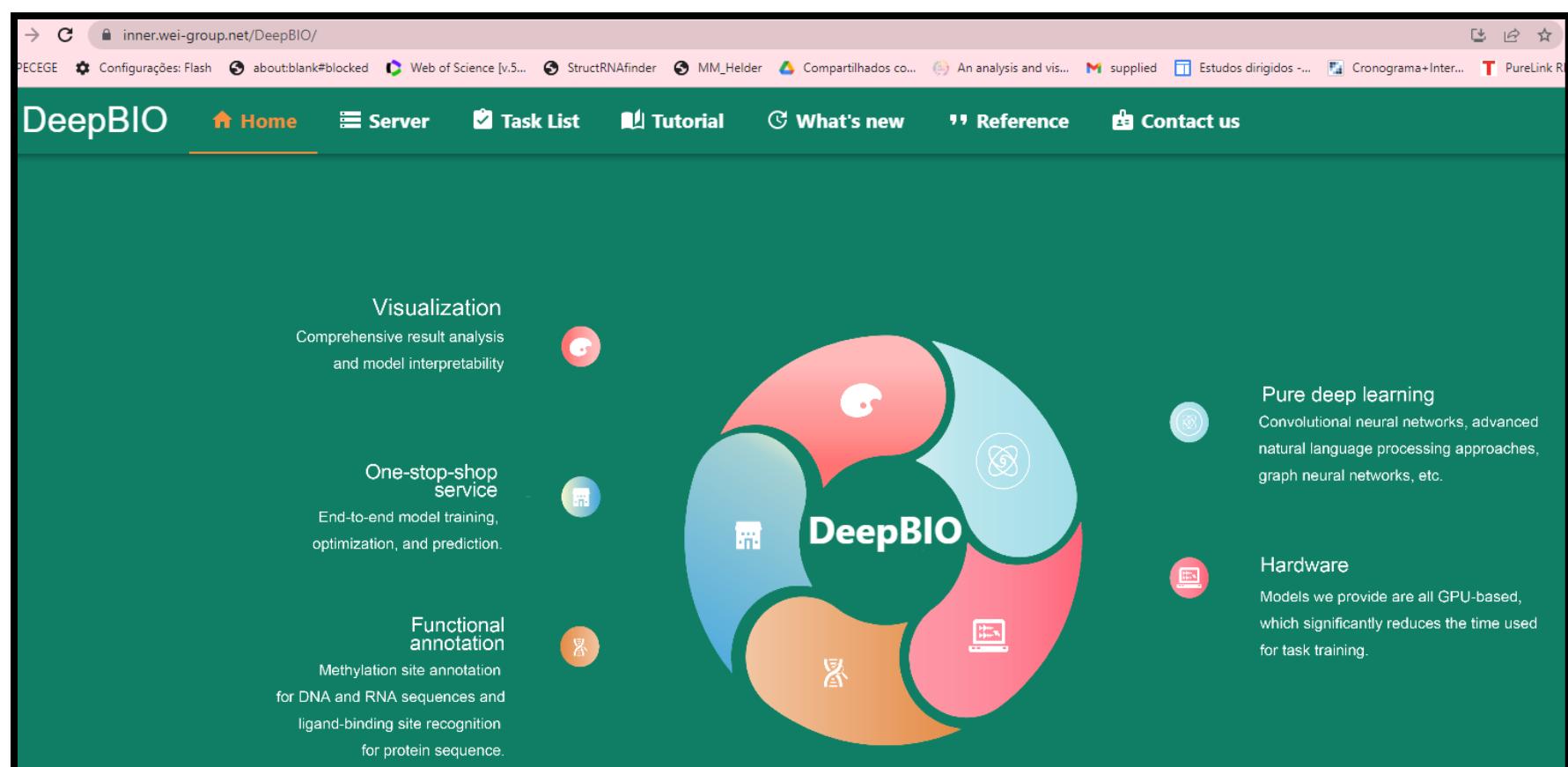
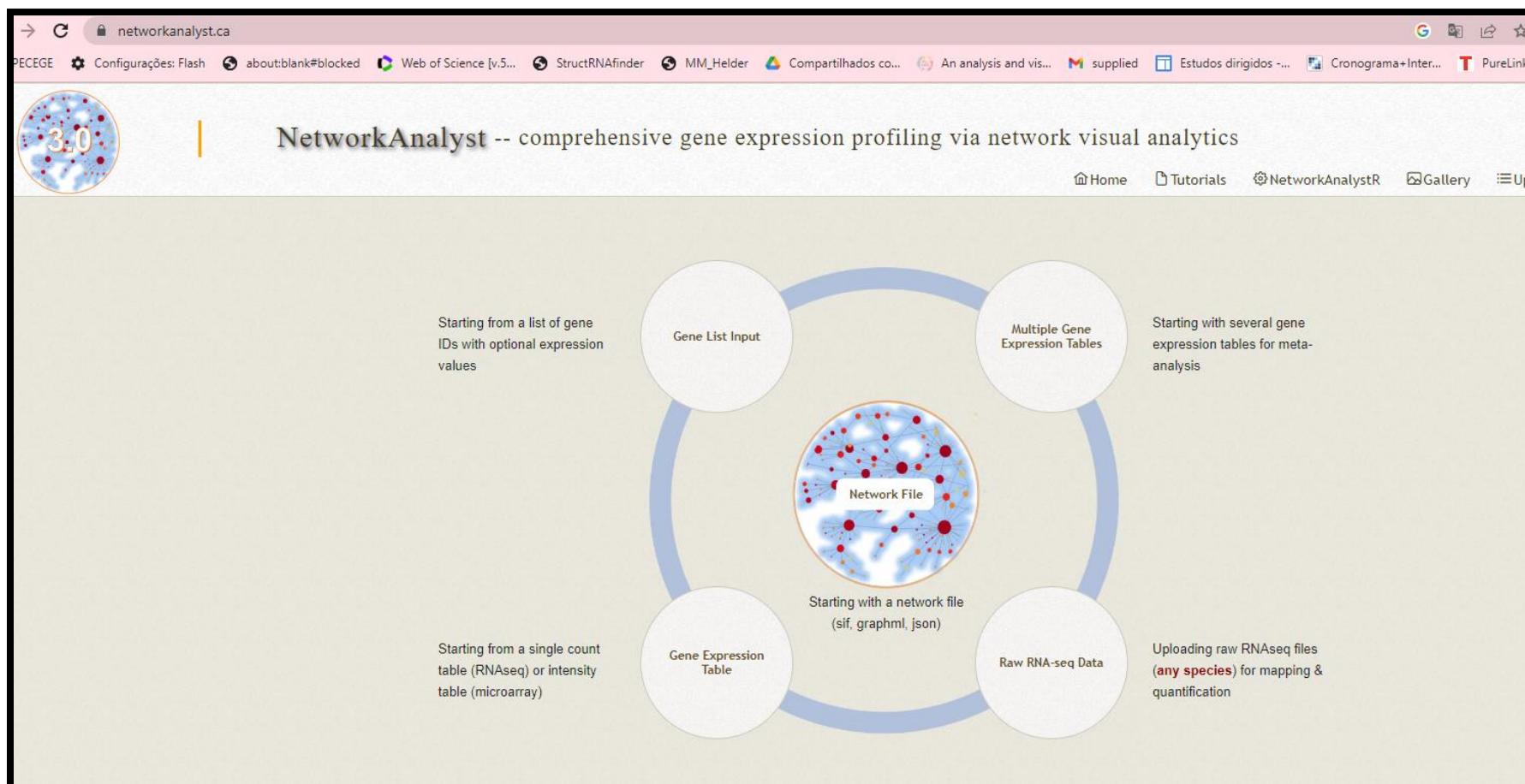
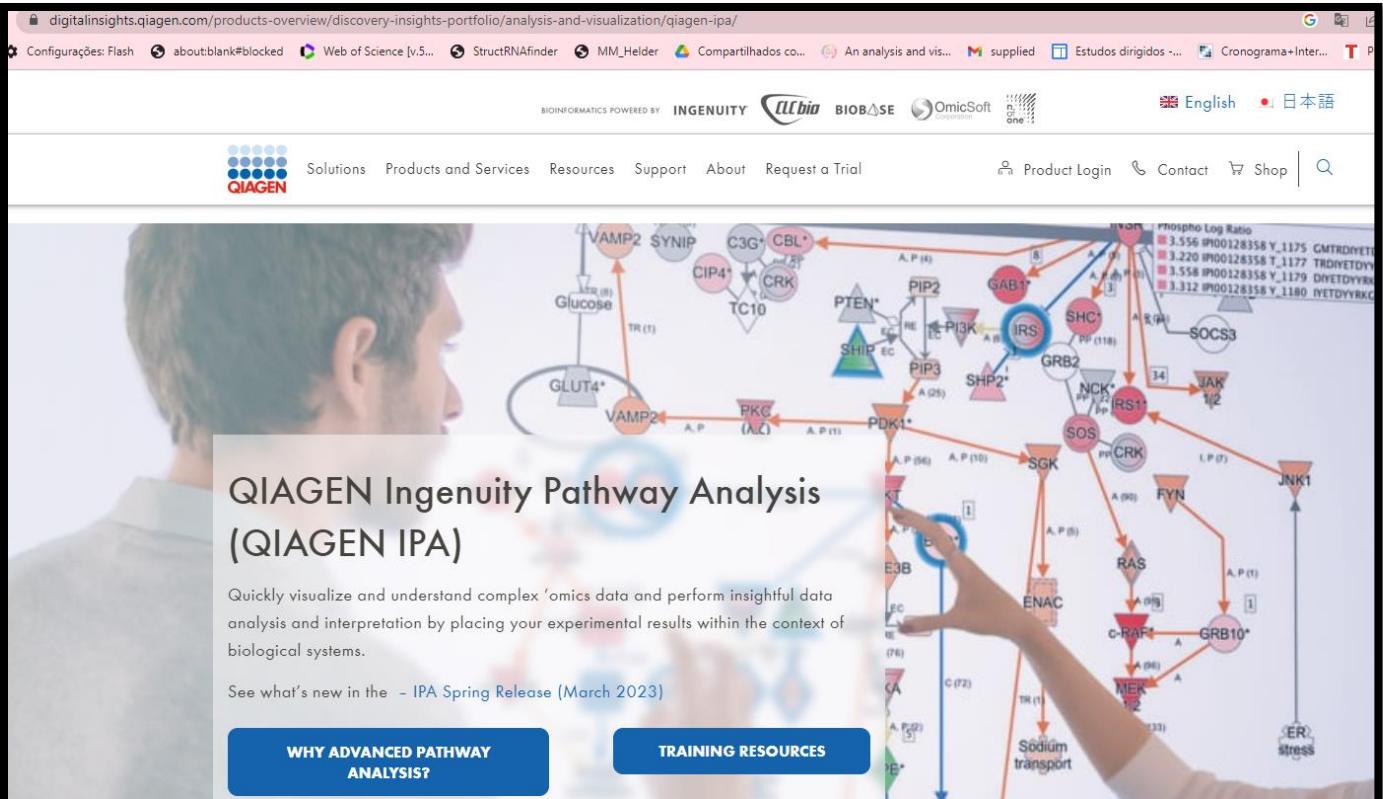
Solutions Products and Services Resources Support About Request a Trial Product Login Contact Shop

## QIAGEN Ingenuity Pathway Analysis (QIAGEN IPA)

Quickly visualize and understand complex 'omics data and perform insightful data analysis and interpretation by placing your experimental results within the context of biological systems.

See what's new in the - IPA Spring Release (March 2023)

WHY ADVANCED PATHWAY ANALYSIS? TRAINING RESOURCES



 **Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home    Install    Help

Home » Bioconductor 3.17 » Software Packages » clusterProfiler

## clusterProfiler

platforms all rank 41 / 2229 support 12 / 21 in Bioc 12 years  
build ok updated < 1 month dependencies 128

DOI: [10.18129/B9.bioc.clusterProfiler](https://doi.org/10.18129/B9.bioc.clusterProfiler)

A universal enrichment tool for interpreting omics data

Bioconductor version: Release (3.17)

This package supports functional characteristics of both coding and non-coding genomics data for thousands of species with up-to-date gene annotation. It provides a universal interface for gene functional annotation from a variety of sources and thus can be applied in diverse scenarios. It provides a tidy interface to access, manipulate, and visualize enrichment results to help users achieve efficient data interpretation. Datasets obtained from multiple treatments and time points can be analyzed and compared in a single run, easily revealing functional consensus and differences among distinct conditions.

Author: Guangchuang Yu [aut, cre, cph] , Li-Gen Wang [ctb], Erqiang Hu [ctb], Xiao Luo [ctb], Meijun Chen [ctb], Giovanni Dall'Olio [ctb], Wanqian Wei [ctb], Chun-Hui Gao [ctb] 

Maintainer: Guangchuang Yu <[guangchuangyu@gmail.com](mailto:guangchuangyu@gmail.com)>

Citation (from within R, enter `citation("clusterProfiler")`):

blast2go.com

Home    Blast2GO    OmicsBox    Support

 biobam  
BIOINFORMATICS SOLUTIONS

## Blast2GO

Functional Genomics Made Easy

[Request Trial](#)

 Request a Free Trial  
Experience a full-featured

 Subscribe Now!  
Subscribe to OmicsBox for all

 Download  
Blast2GO/OmicsBox

 Tutorials, Videos and News  
Browse the BioBam Blog for all



# Ferramentas/Softwares para redes gênicas

Version: 11.5

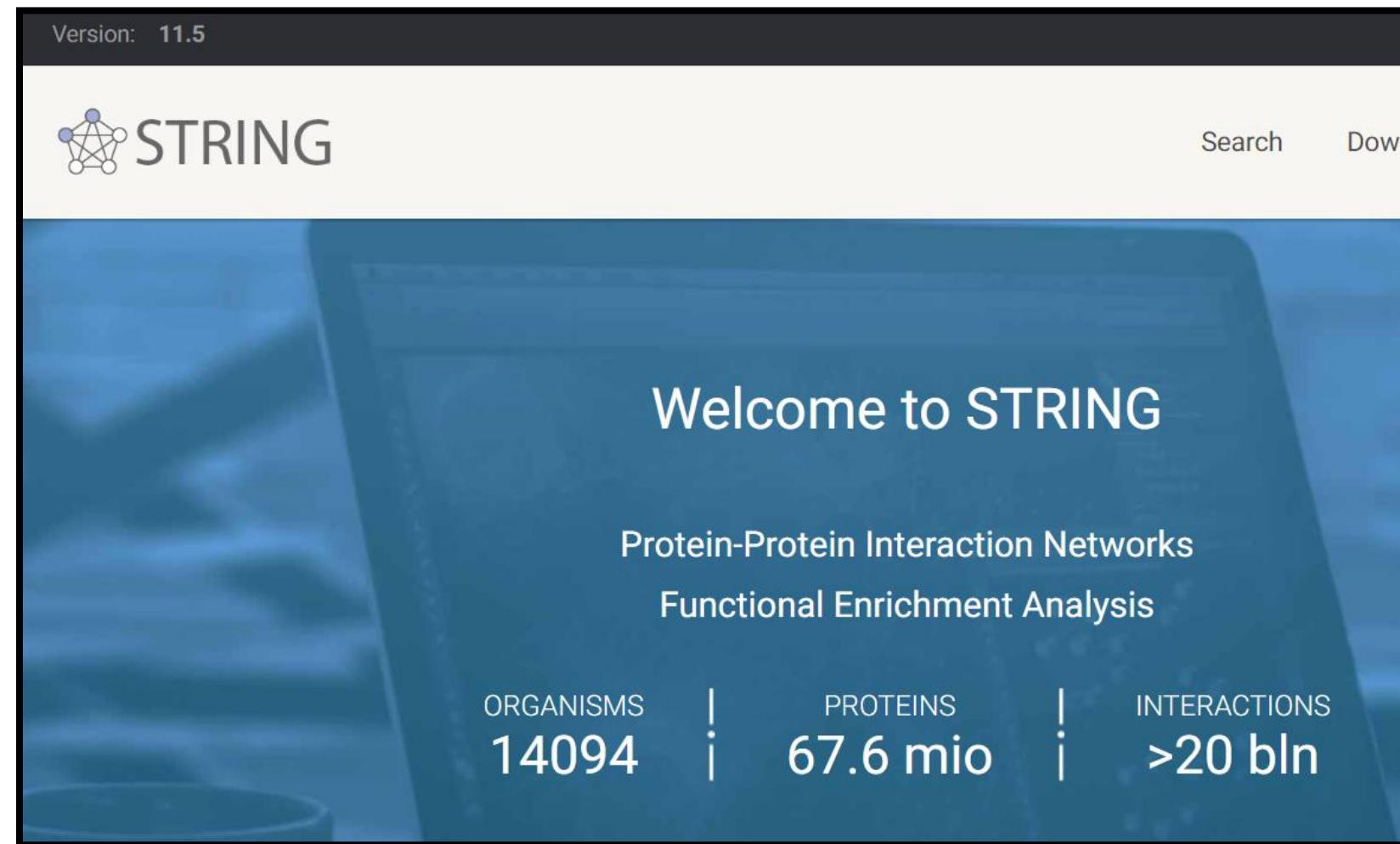
## STRING

Welcome to STRING

Protein-Protein Interaction Networks  
Functional Enrichment Analysis

ORGANISMS | PROTEINS | INTERACTIONS

14094 | 67.6 mio | >20 bln



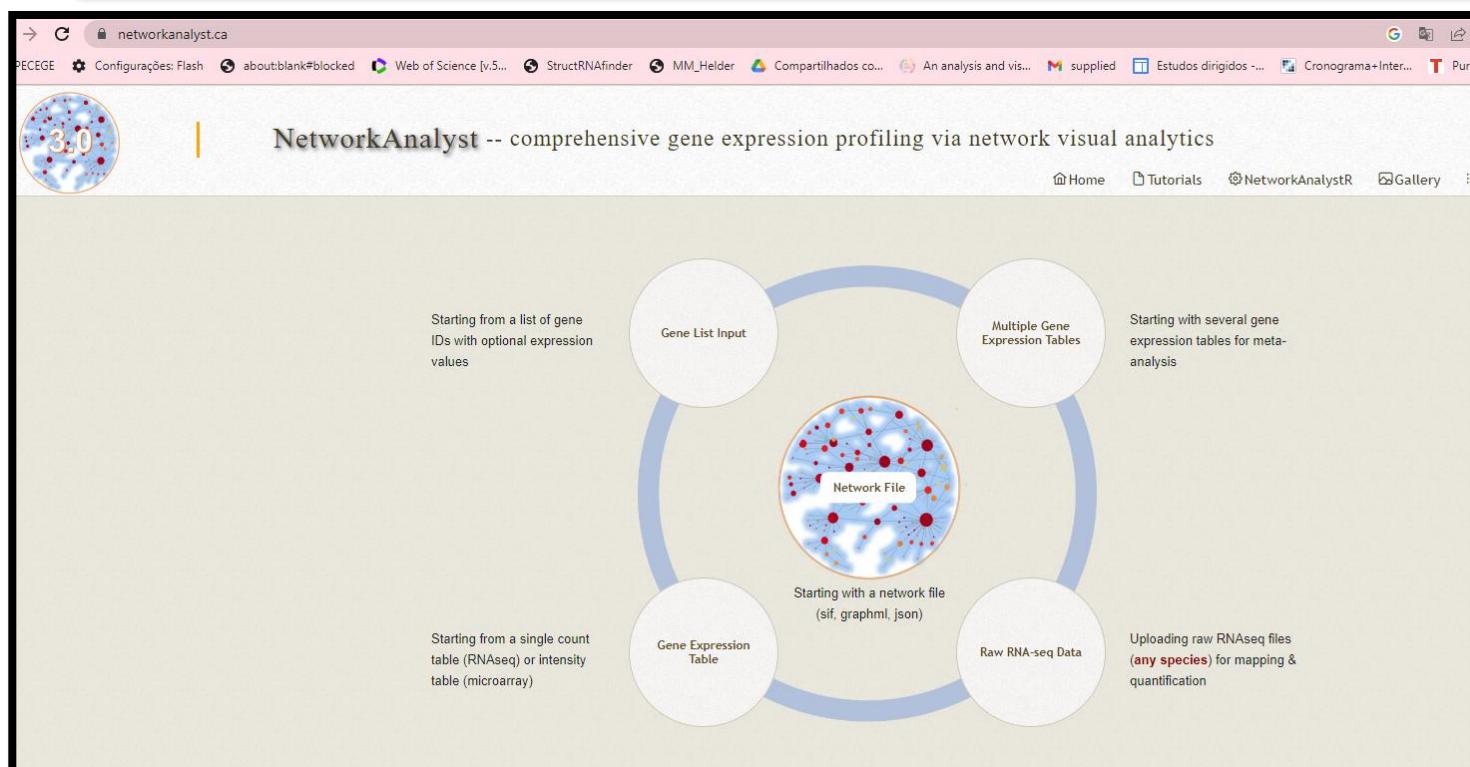
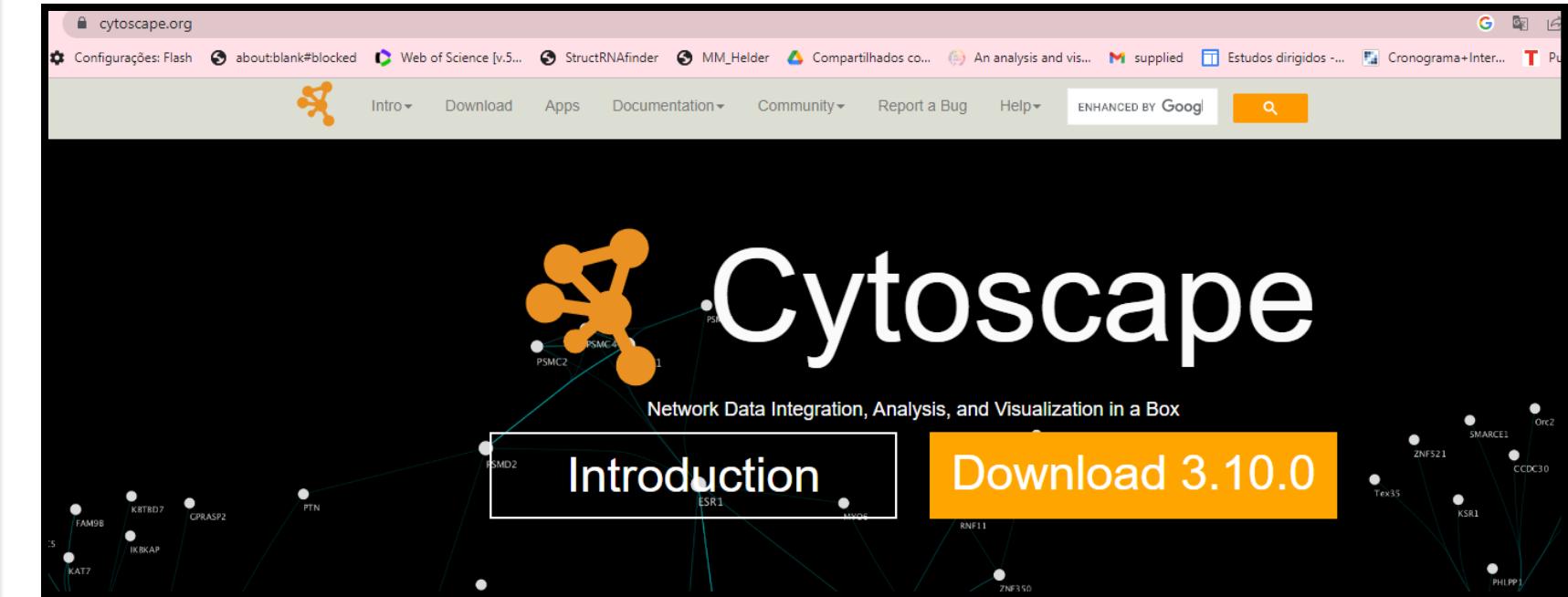
cytoscape.org

# Cytoscape

Network Data Integration, Analysis, and Visualization in a Box

Introduction

Download 3.10.0

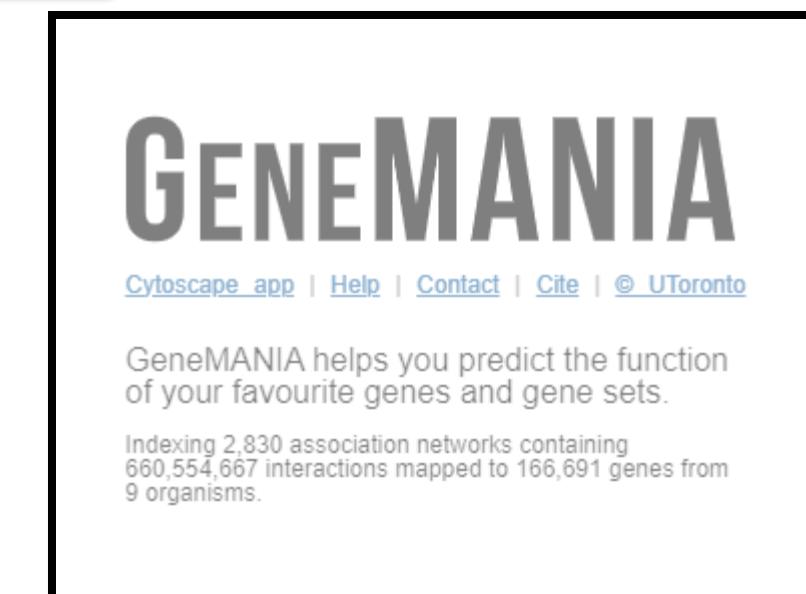


# GENEMANIA

Cytoscape app | Help | Contact | Cite | @ UToronto

GeneMANIA helps you predict the function of your favourite genes and gene sets.

Indexing 2,830 association networks containing 660,554,667 interactions mapped to 166,691 genes from 9 organisms.



genenetwork.org

## Select and Search

Species: Mouse (Mus musculus, mm10)

Group: BXD Family

Type: Hippocampus mRNA

Dataset: Hippocampus Consortium M430v2 (Jun06) PDNN

Get Any:

Combined:

Tutorials

Webinars & Courses

Tutorials: Training materials in HTML, PDF and video formats

Documentation

In-person courses, live webinars and webinar recordings

Online manuals, handbooks, fact sheets and FAQs

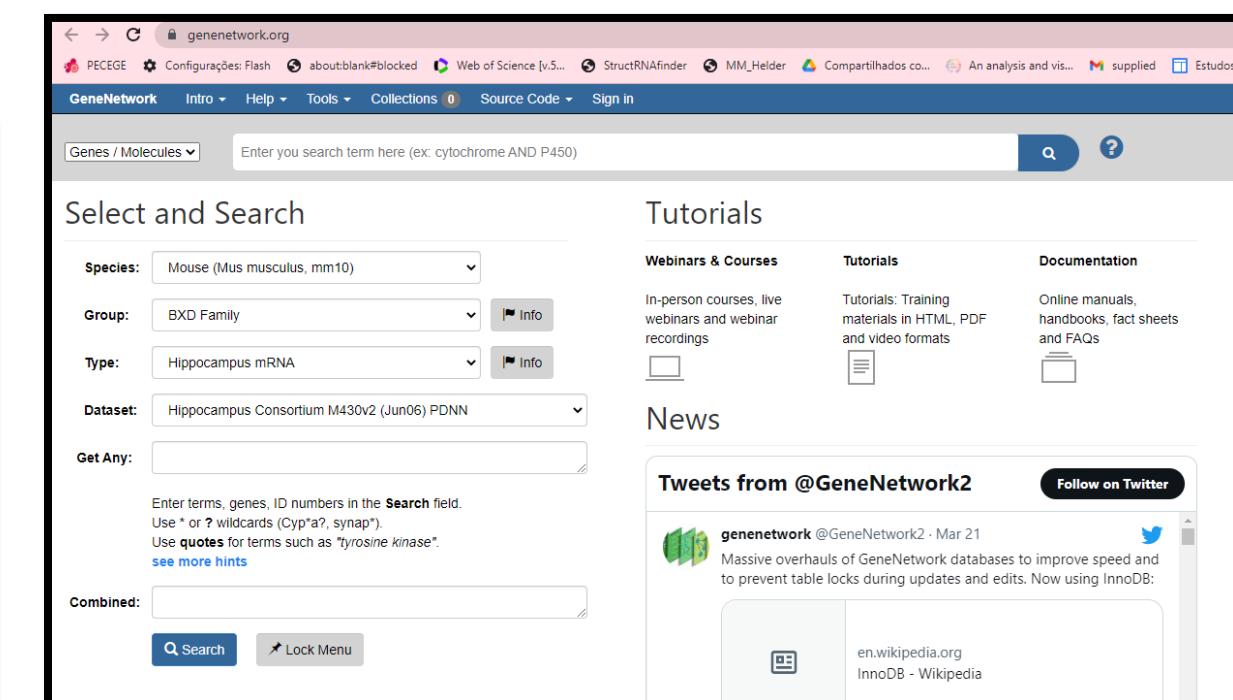
News

Tweets from @GeneNetwork2

genenetwork @GeneNetwork2 · Mar 21

Massive overhauls of GeneNetwork databases to improve speed and to prevent table locks during updates and edits. Now using InnoDB:

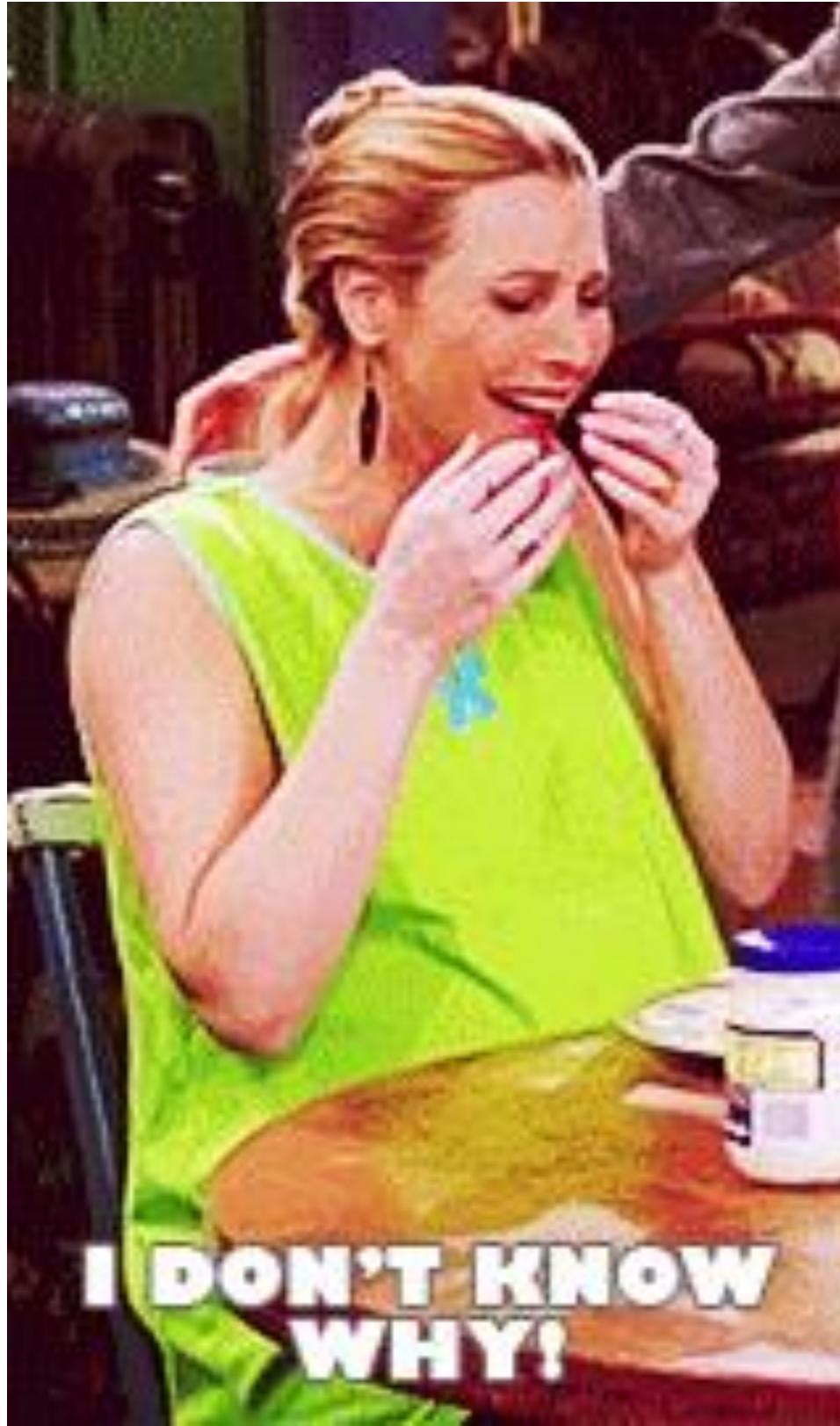
en.wikipedia.org  
InnoDB - Wikipedia



# E quais ferramentas usar ???



# E quais ferramentas usar ???



**OXFORD**

*Briefings in Bioinformatics*, 2023, 24(1), 1–17  
<https://doi.org/10.1093/bib/bbac529>  
Review

## The hitchhikers' guide to RNA sequencing and functional analysis

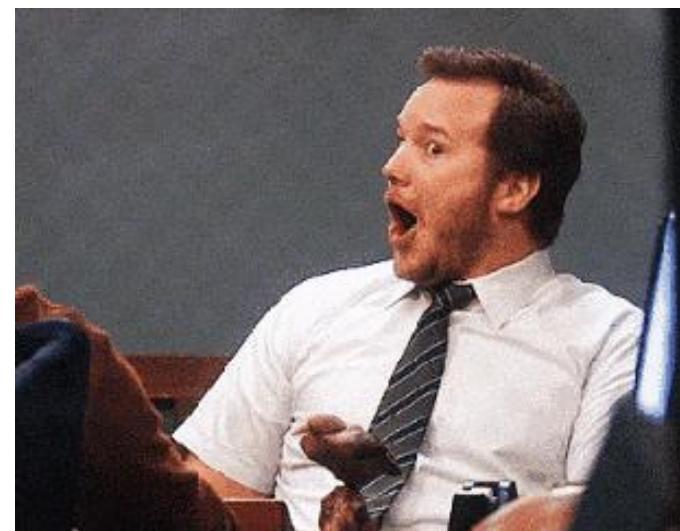
Jiung-Wen Chen, Lisa Shrestha, George Green, André Leier and Tatiana T. Marquez-Lago

Corresponding author: Tatiana T. Marquez-Lago, Department of Genetics, University of Alabama at Birmingham, School of Medicine, Birmingham, AL, USA.  
Tel.: +1 (205) 9343194. E-mail: tmarquez@uab.edu

### Abstract

DNA and RNA sequencing technologies have revolutionized biology and biomedical sciences, sequencing full genomes at very high speeds and reasonably low costs. RNA sequencing (RNA-Seq) enables transcript identification and quantification once sequencing has concluded. Researchers can be easily overwhelmed with questions such as how to go from raw data to expression (DE), pathway analysis and interpretation. Several pipelines and procedures have been developed to this end, but there is no unique way to perform RNA-Seq analysis; it usually follows these steps: 1) raw reads quality check, 2) alignment to a reference genome, 3) aligned reads' summarization according to an annotation file, 4) DE analysis and 5) gene set enrichment analysis. Each step requires researchers to make decisions, and the wide variety of options and volumes of data often lead to interpretation challenges. There also seems to be insufficient guidance on how best to use this information and derive actionable knowledge from transcription experiments. In this paper, we explain RNA-Seq steps, outline differences and similarities of different popular options, as well as advantages and disadvantages. We also discuss RNA analysis, multi-omics, meta-transcriptomics and the use of artificial intelligence methods complementing the tools available to researchers. Lastly, we perform a complete analysis from raw reads to DE and functional enrichment analysis, illustrating how results are not absolute truths and how algorithmic decisions can greatly impact results and interpretation.

**Keywords:** RNA sequencing, differential expression, functional analysis, machine learning, multi-omics



## PLOS COMPUTATIONAL BIOLOGY

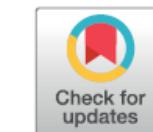
### RESEARCH ARTICLE

## Urgent need for consistent standards in functional enrichment analysis

Kaumadi Wijesooriya <sup>1</sup>, Sameer A. Jadaan <sup>2</sup>, Kaushalya L. Perera <sup>1</sup>, Tanuveer Kaur <sup>1</sup>, Mark Ziemann <sup>1</sup>\*

<sup>1</sup> Deakin University, School of Life and Environmental Sciences, Geelong, Australia, <sup>2</sup> College of Health and Medical Technology, Middle Technical University, Baghdad, Iraq

\* [m.ziemann@deakin.edu.au](mailto:m.ziemann@deakin.edu.au)



### OPEN ACCESS

**Citation:** Wijesooriya K, Jadaan SA, Perera KL, Kaur T, Ziemann M (2022) Urgent need for consistent standards in functional enrichment analysis. PLoS Comput Biol 18(3): e1009935. <https://doi.org/10.1371/journal.pcbi.1009935>

**Editor:** Melissa L. Kemp, Georgia Institute of Technology and Emory University, UNITED STATES

**Received:** December 7, 2021

**Accepted:** February 18, 2022

**Published:** March 9, 2022

### Abstract

Gene set enrichment tests (a.k.a. functional enrichment analysis) are among the most frequently used methods in computational biology. Despite this popularity, there are concerns that these methods are being applied incorrectly and the results of some peer-reviewed publications are unreliable. These problems include the use of inappropriate background gene lists, lack of false discovery rate correction and lack of methodological detail. To ascertain the frequency of these issues in the literature, we performed a screen of 186 open-access research articles describing functional enrichment results. We find that 95% of analyses using over-representation tests did not implement an appropriate background gene list or did not describe this in the methods. Failure to perform p-value correction for multiple tests was identified in 43% of analyses. Many studies lacked detail in the methods section about the tools and gene sets used. An extension of this survey showed that these problems are not associated with journal or article level bibliometrics. Using seven independent RNA-seq datasets, we show misuse of enrichment tools alters results substantially. In conclusion, most published functional enrichment studies suffered from one or more major flaws, highlighting the need for stronger standards for enrichment analysis.

# Obrigada!

adriana.lbelli@embrapa.br

 @dricaibelli



MINISTÉRIO DA  
AGRICULTURA E  
PECUÁRIA

GOVERNO FEDERAL  
**BRASIL**  
UNIÃO E RECONSTRUÇÃO