Taller RNA-seq

6 de Setiembre de 2022

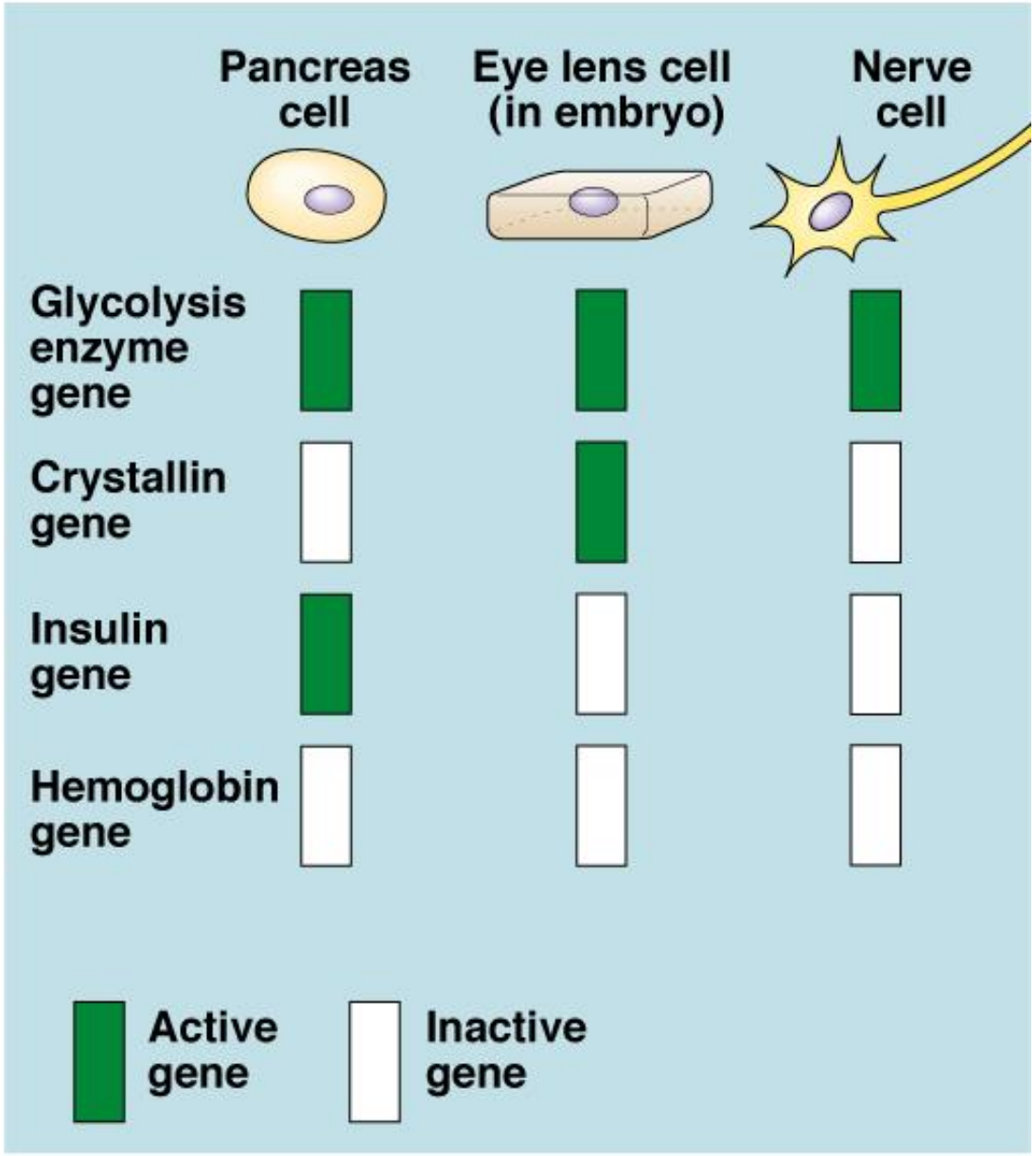
Joaquín Garat - joagarat@gmail.com
Hellen Schlueb - hellensch@fcien.edu.uy
Nicolas Papa Rodriguez - npapa@fcien.edu.uy
Maria Jose Arezo - maui@fcien.edu.uy
José Sotelo - sotelojos@gmail.com

Objetivo de este taller

Introducir la técnica de RNAseq y comenzar el análisis de datos generados a partir de peces anuales.

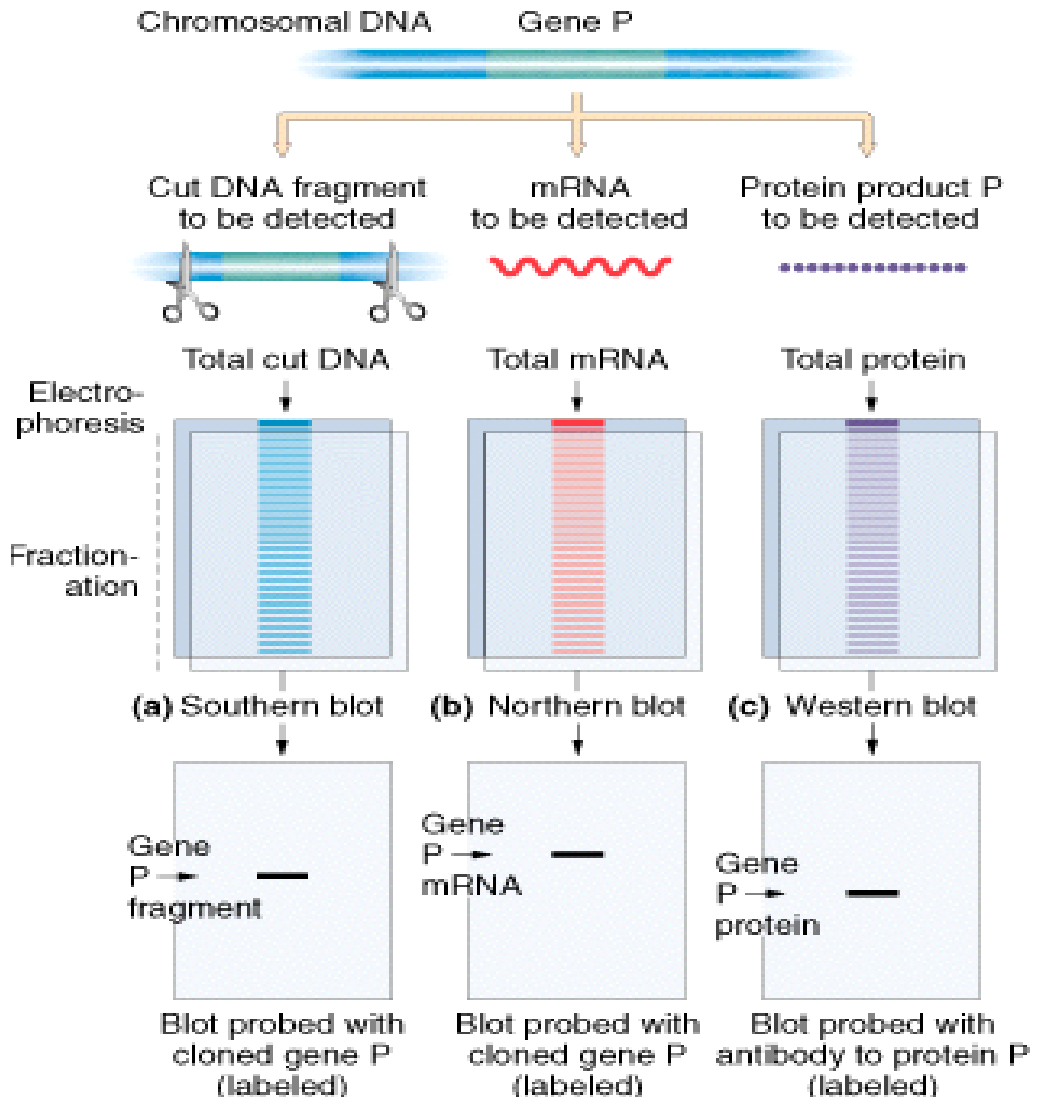
Expresión diferencial

- No todos los genes se expresan
  - existen distintos perfiles de ARNs y proteínas
- La expresión diferencial:
  - durante el desarrollo y diferenciación celular
  - en la homeostasis

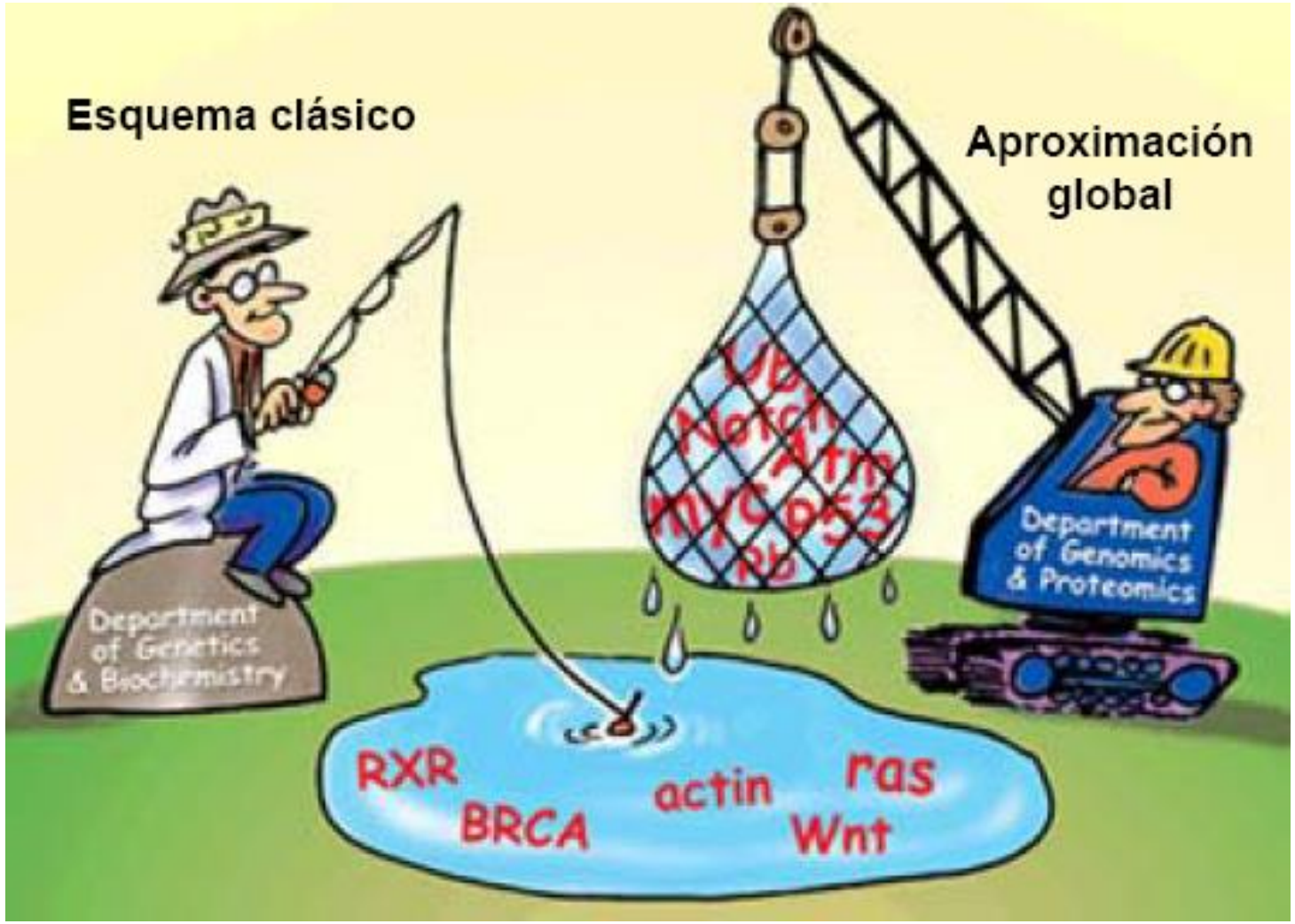


Estudios clásicos de expresión:

- Se focalizan en un número limitado de genes
  - *Northern blot*
  - RT PCR.

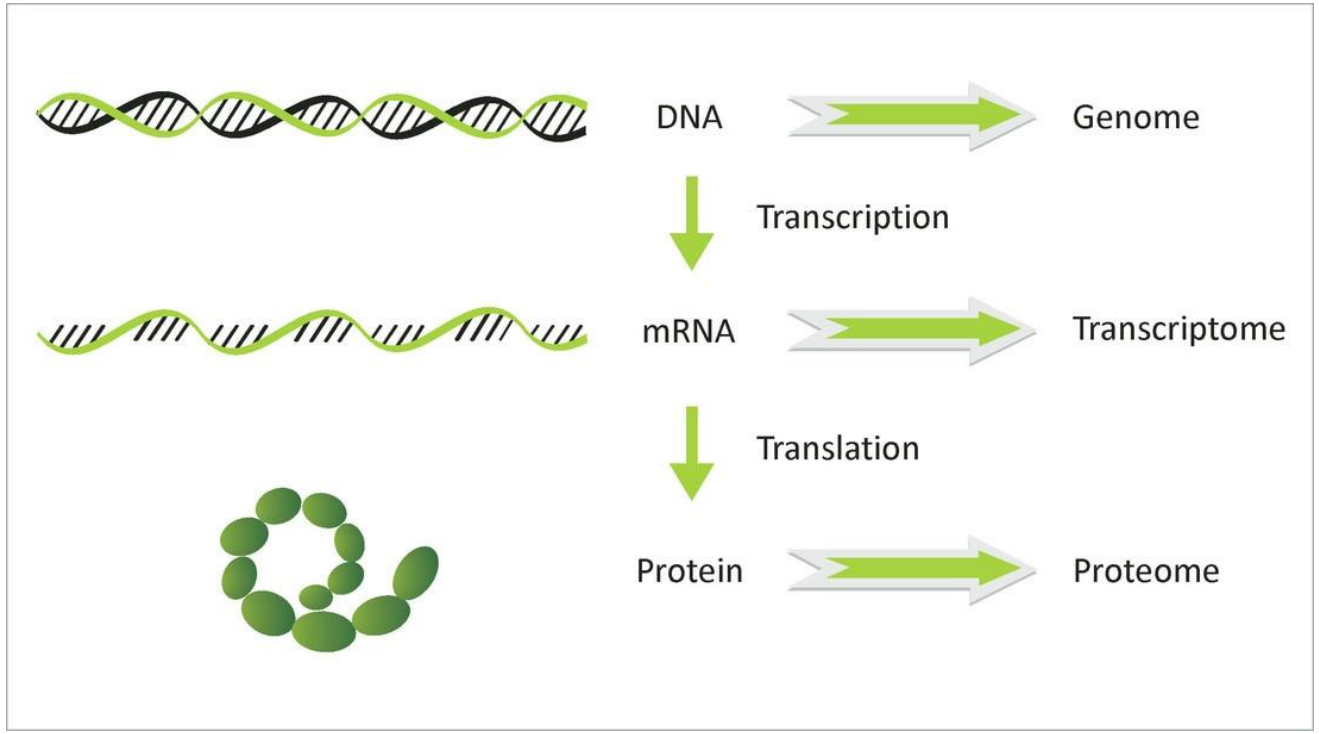


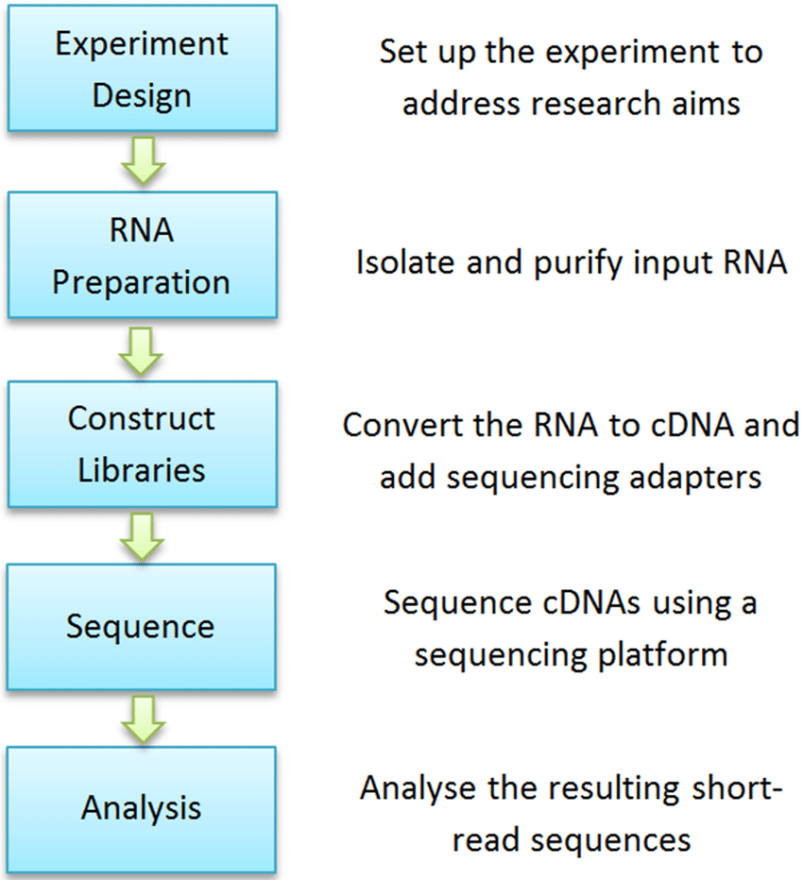
Ciencias “ómicas”



Qué es la transcriptómica?

- Estudio de la abundancia y secuencia de transcriptos presentes en una determinada muestra.

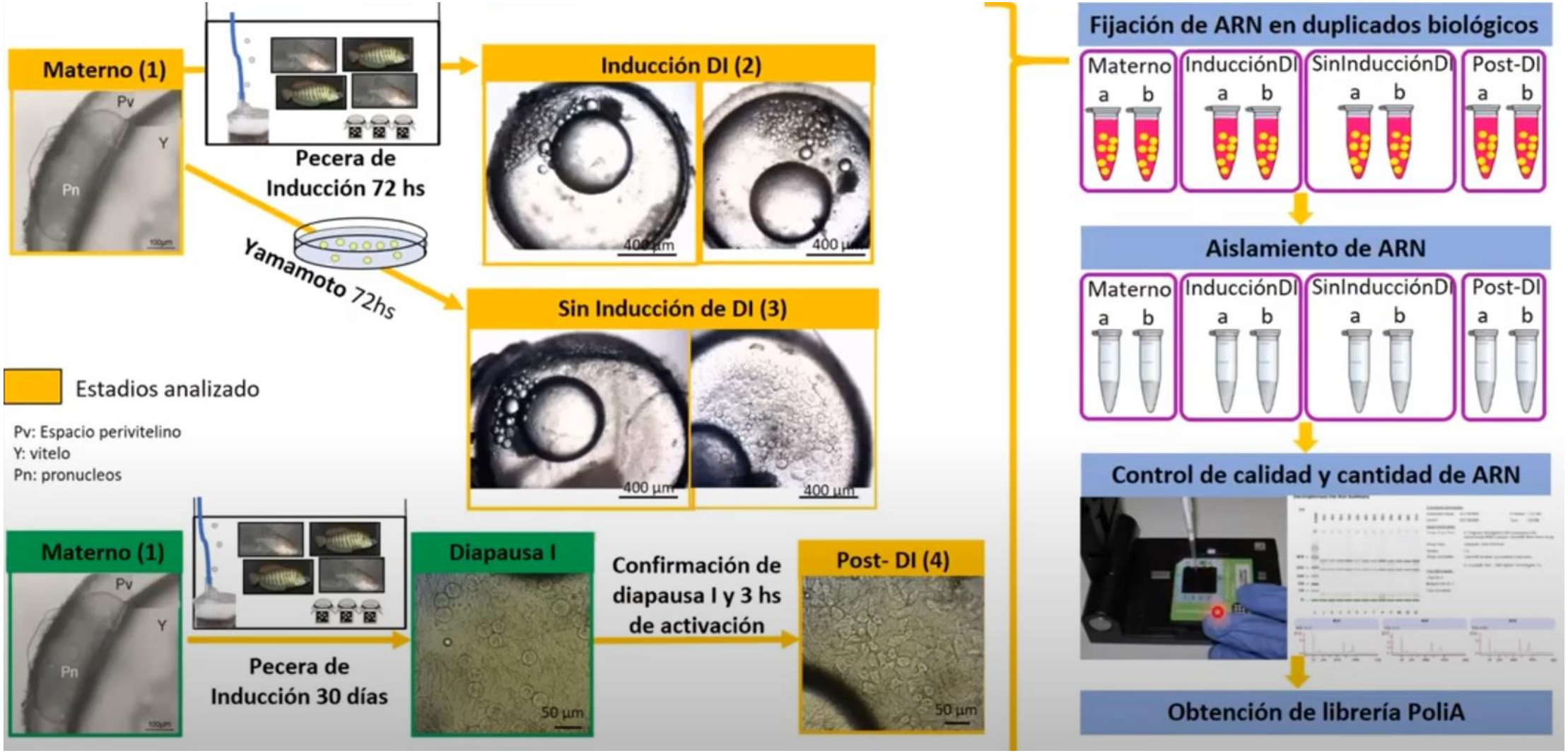


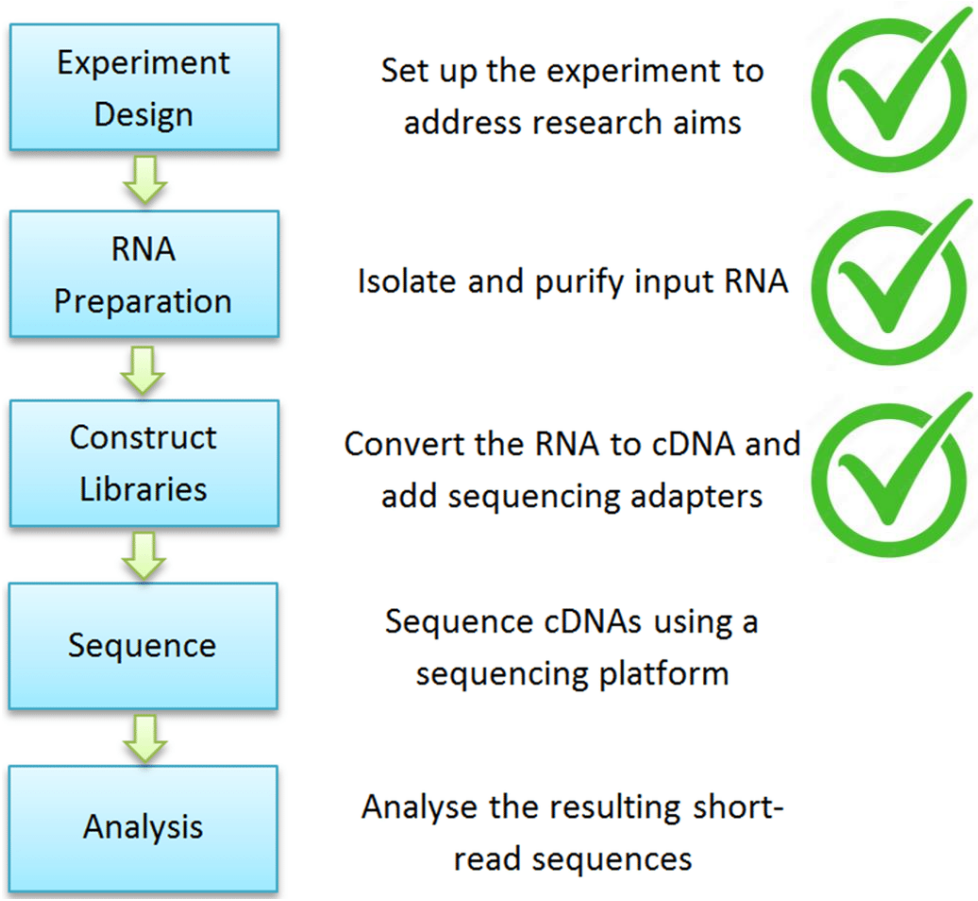


Diseño Experimental

- ¿Pregunta biológica?
- ¿Organismo con genoma secuenciado y anotado?
- ¿Estrategia de secuenciación? ¿Profundidad de secuenciación?
- Necesidad de réplicas
- ¿Recursos?

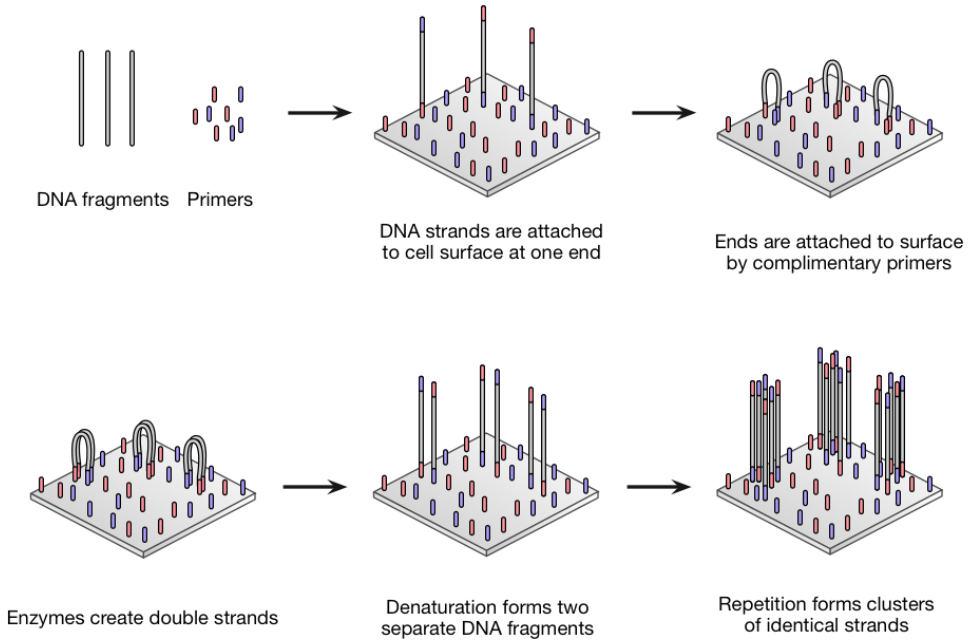
Expresión génica en la inducción de Diapausa en peces anuales

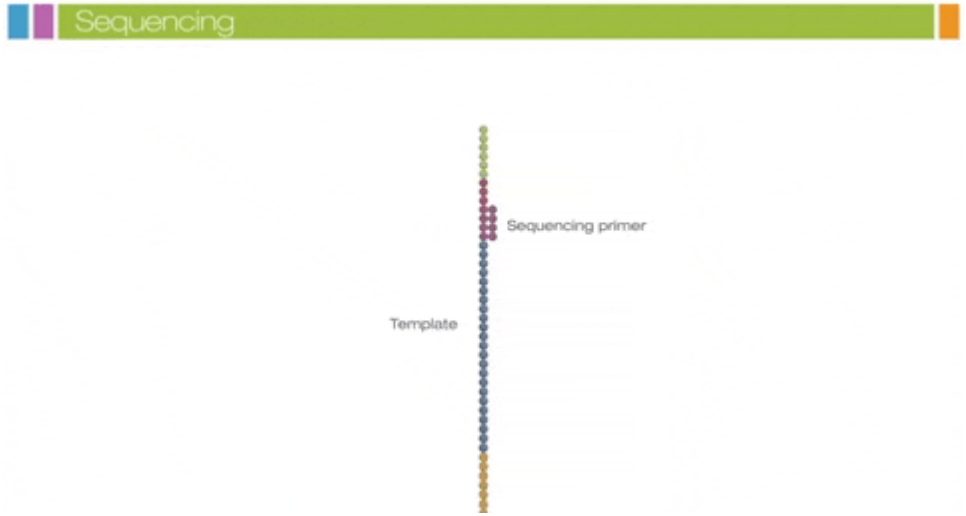


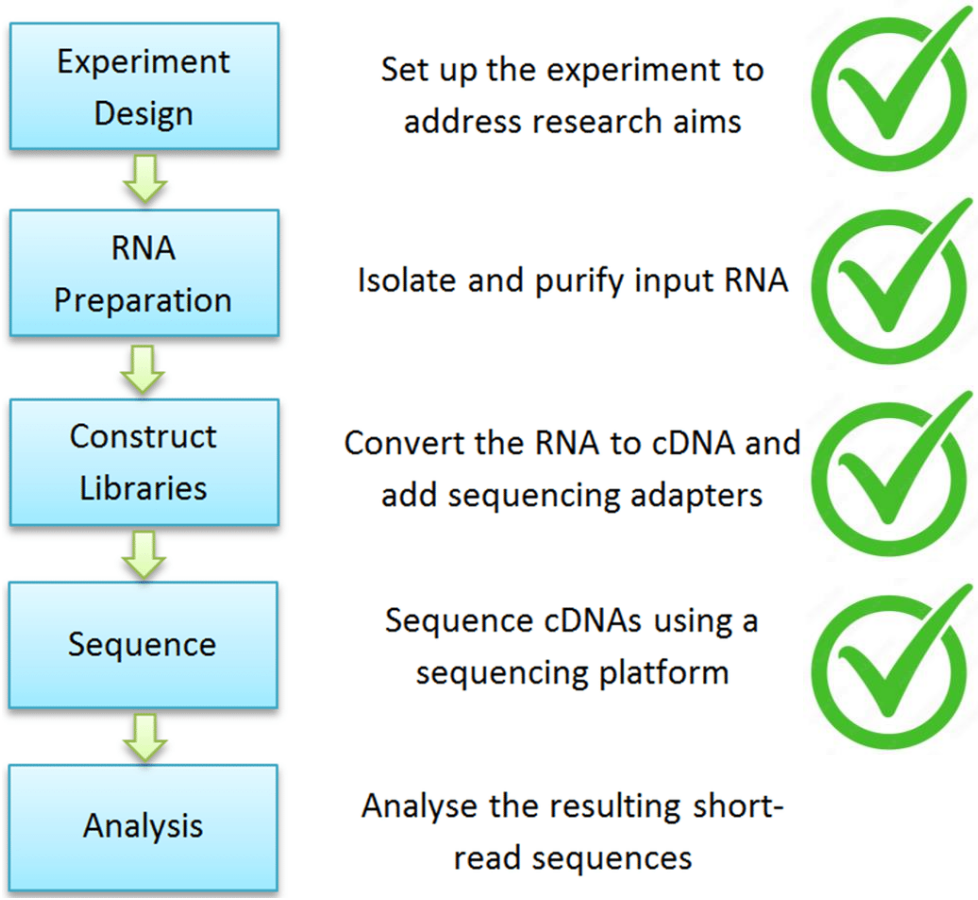

Experiment Design
Set up the experiment to address research aims
RNA Preparation
Isolate and purify input RNA
Construct Libraries
Convert the RNA to cDNA and add sequencing adapters
Sequence
Sequence cDNAs using a sequencing platform
Analysis
Analyse the resulting short-read sequences

Secuenciación Illumina

- i) Generación de clusters clonales
- ii) Escisión de hebra reverse, ciclos de secuenciación de hebra Forward con dNTPs marcados y bloqueados.
- iii) Escisión de hebra forward y secuenciación de hebra reverse

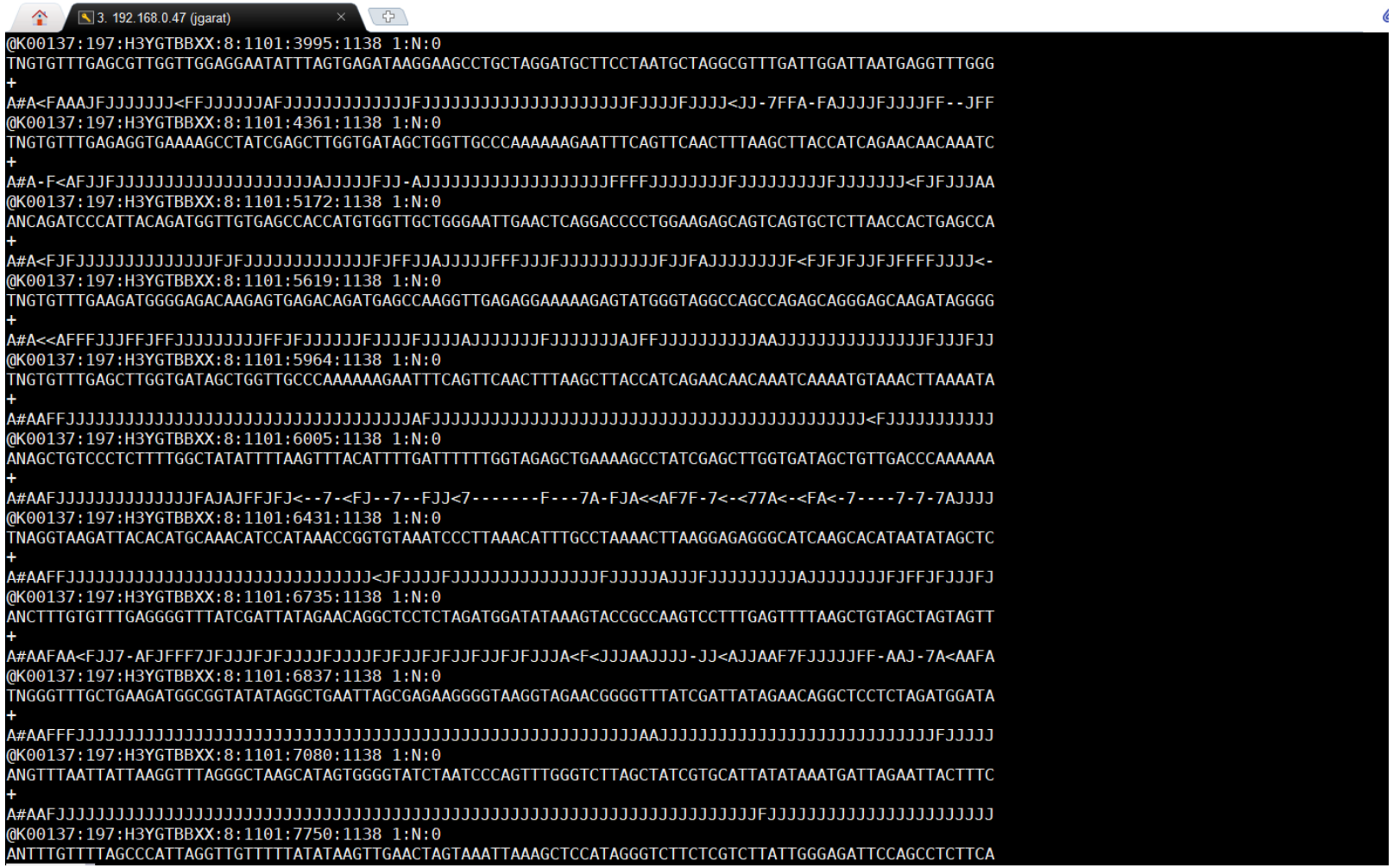




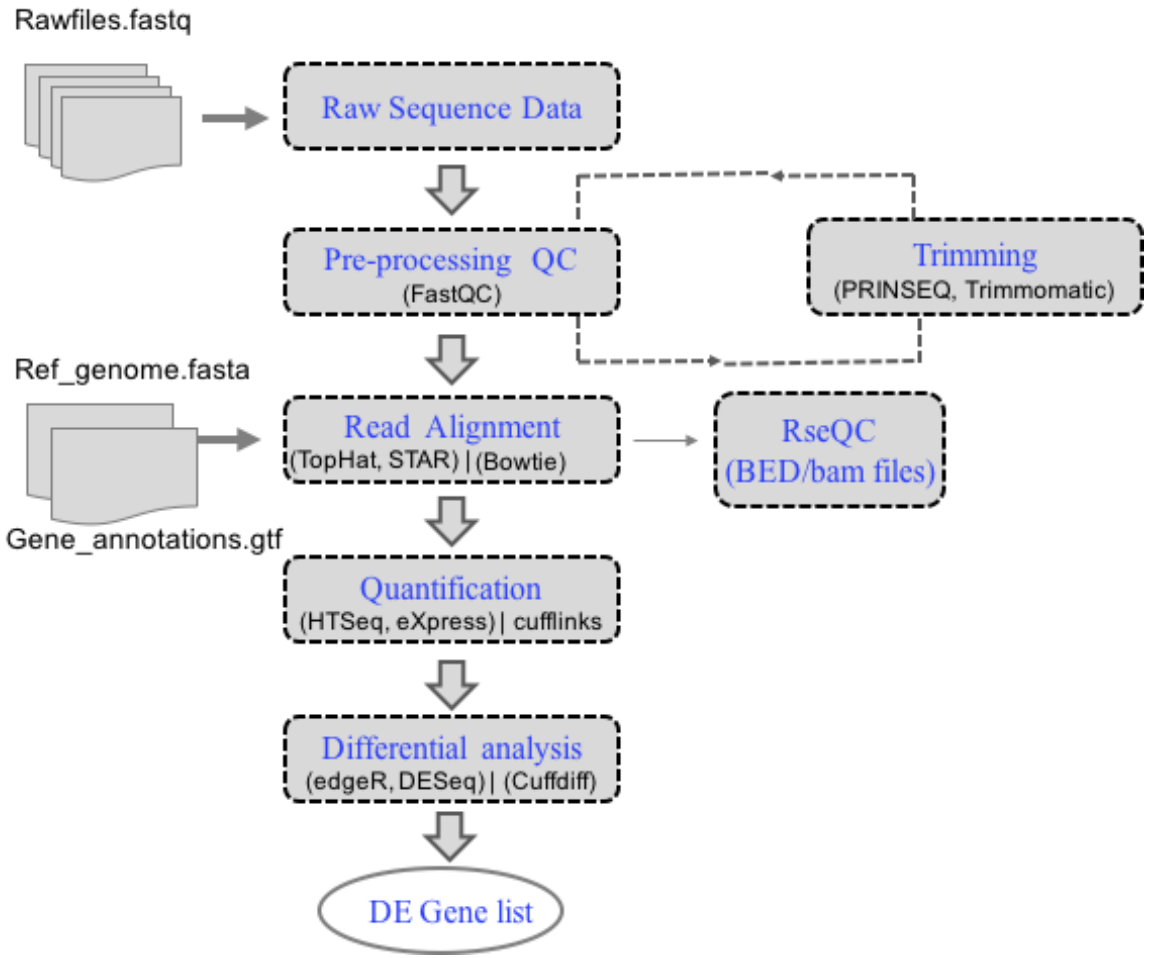

Experiment Design
Set up the experiment to address research aims
RNA Preparation
Isolate and purify input RNA
Construct Libraries
Convert the RNA to cDNA and add sequencing adapters
Sequence
Sequence cDNAs using a sequencing platform
Analysis
Analyse the resulting short-read sequences

Análisis de secuencias

- Usualmente se descargan las secuencias y se trabaja desde la línea de comando



Análisis de secuencias



¿ Ensamblado de Transcriptoma? (SOAPdenovo-trans, Oases, etc)

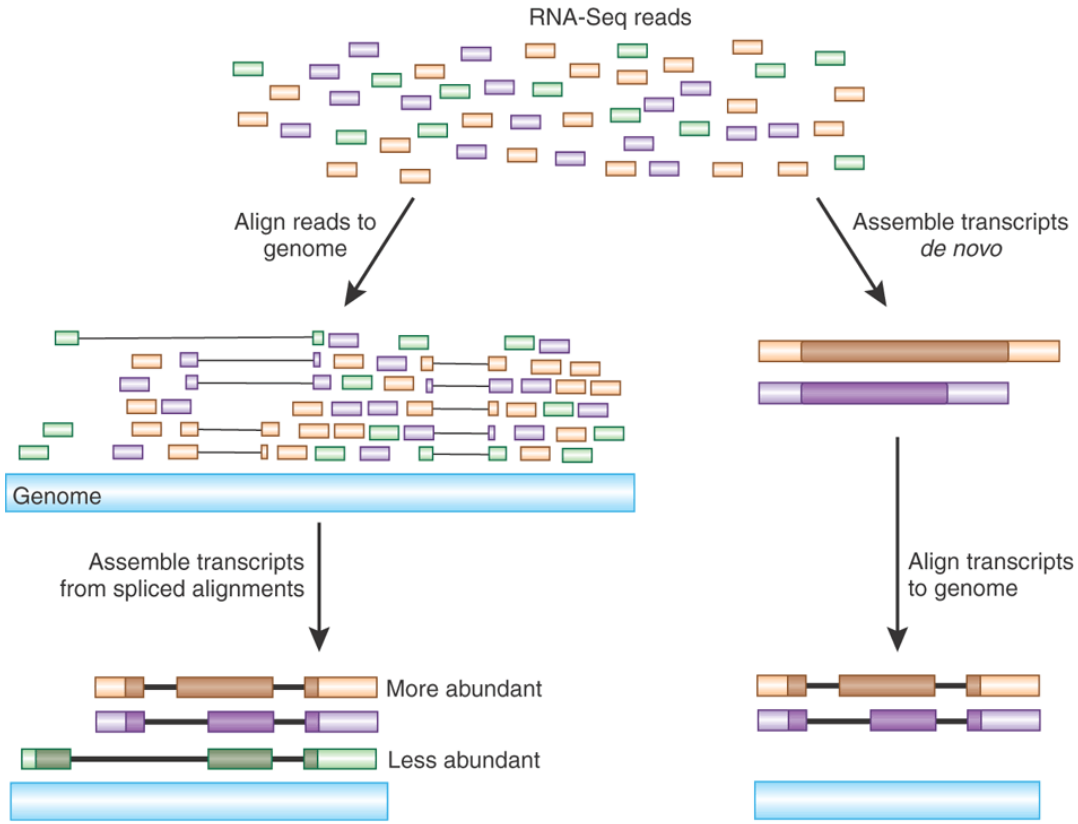
Pre-procesamiento de lecturas

- A efectos prácticos nos saltearemos algunos pasos (no sin antes discutirlos), y trabajaremos con una herramienta gráfica en la web.

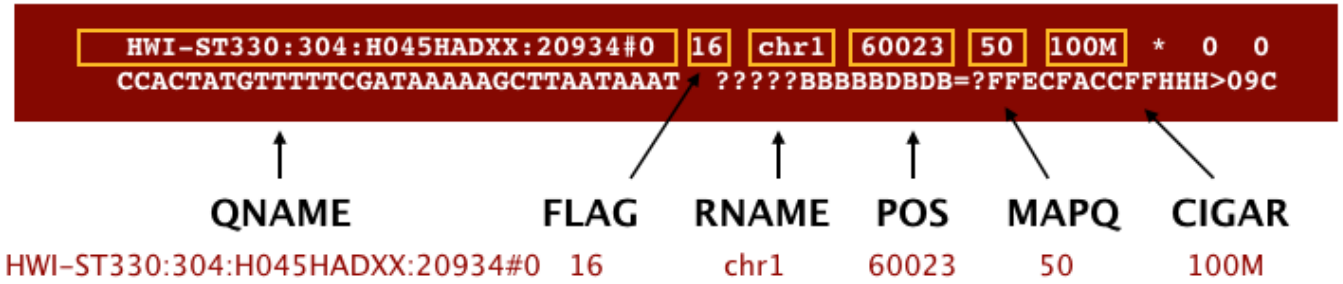
- Evaluación de calidad de las lecturas:
  - FASTQC

Descarguen el archivo FastQC del EVA y abranlo. Qué pueden ver?

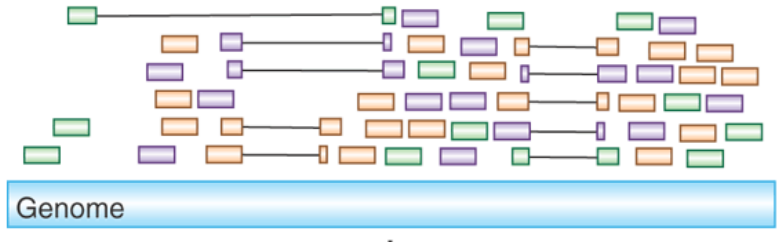
Alineamiento

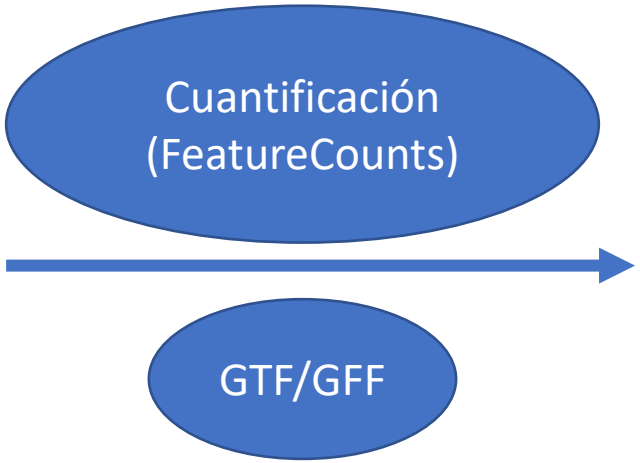


- Las lecturas se alinean contra una referencia con herramientas informáticas.



Cuantificación





Gen	Conteo M1_1	Conteo M1_2	…
Gen1	120	118	
Gen2	100	95	
…	…	…	
…	…	…	
…	…	…	

Tenemos una tabla de expresión, ¿y ahora?

- Descarguen la tabla de datos del EVA

Hay diversos programas que permiten realizar estos análisis, pero dado que no es un curso de genómica, seleccionamos uno que posee interfaz gráfica y es de fácil acceso, además de gratuito

Ge *et al. BMC Bioinformatics* (2018) 19:534
https://doi.org/10.1186/s12859-018-2486-6

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data

Steven Xijin Ge*, Eun Wo Son and Runan Yao

(Este artículo está disponible en la carpeta de práctico en el Drive del curso)

Hay diversos programas que permiten realizar estos análisis, pero dado que no es un curso de genómica, seleccionamos uno que posee interfaz gráfica y es de fácil acceso, además de gratuito

Ge *et al. BMC Bioinformatics* (2018) 19:534
https://doi.org/10.1186/s12859-018-2486-6

BMC Bioinformatics

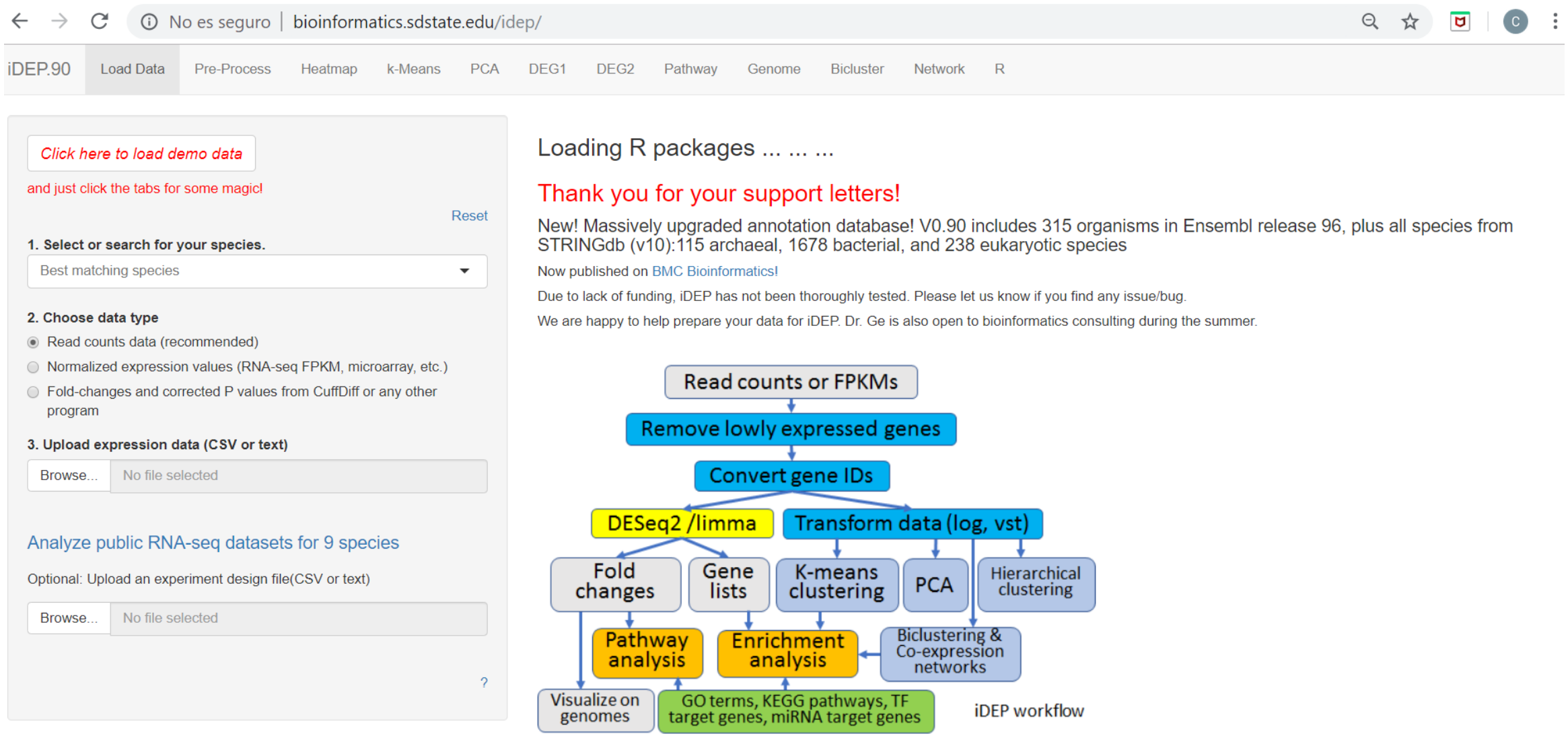
METHODOLOGY ARTICLE Open Access

iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data

Steven Xijin Ge*, Eun Wo Son and Runan Yao

(Este artículo está disponible en la carpeta de práctico en el Drive del curso)

http://bioinformatics.sdstate.edu/idep/



Formato de los datos

Sample names

Name columns carefully as iDEP parses column names to define sample groups. Replicates should be denoted by "_Rep1", "_Rep2", "_Rep3" at the end. Or it can be simply "_1", "_2", "_3". For example, *Control_1, Control_2, TreatmentA_1, TreatmentA_2, TreatmentB_1, TreatmentB_2*. The first part defines 3 sample groups that form the basis for differential expression analysis. All pair-wise comparisons are listed and analyzed. Also, avoid using a hyphen "-" or a dot "." in sample names. It affects the parsing of sample names. But underscore "_" is allowed.

More complex study designs can be represented by uploading a study design file. See below.

https://idepsite.wordpress.com/data-format/

No es seguro | bioinformatics.sdstate.edu/idep/

iDEP.90 | Load Data | Pre-Process | Heatmap | k-Means | PCA | DEG1 | DEG2 | Pathway | Genome | Bicluster | Network | R

Click here to load demo data

and just click the tabs for some magic!

Reset

1. Select or search for your species.

Mouse

2. Choose data type

- Read counts data (recommended)
- Normalized expression values (RNA-seq FPKM, microarray, etc.)
- Fold-changes and corrected P values from CuffDiff or any other program

3. Upload expression data (CSV or text)

Browse... tabla datos modulo 6.csv

Upload complete

Analyze public RNA-seq datasets for 9 species

Optional: Upload an experiment design file(CSV or text)

Browse... No file selected

Matched Species (genes)
Using selected species Mouse

?

tracking_id	Day0_1	Day0_2	Day0_3	Day2_1	Day2_2	Day2_3	Day3_1	Day3_2	Day3_3	Day4_1	Day4_2	Day4_3	Day4_4
NM_001011874	0.02	0.03	0.00	0.41	0.34	0.59	0.23	0.26	0.39	0.83	1.10	0.57	0.37
NM_001195662	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01
NM_011283	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
NM_001289467	0.00	0.06	0.09	0.05	0.00	0.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00
NM_001289464	0.11	0.00	0.00	0.00	0.07	0.00	0.25	0.32	1.34	0.70	1.05	0.63	0.72
NM_001289466	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00
NM_001289465	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NM_011441	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NR_033530	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NM_025300	11.60	11.49	9.35	12.52	13.69	11.86	4.74	5.15	5.85	6.61	7.24	7.01	6.76
NM_001177658	20.72	19.21	20.17	21.77	24.06	18.58	12.33	12.03	12.12	9.50	9.26	9.28	9.45
NM_008866	78.75	80.90	80.18	60.71	66.85	57.23	46.37	48.84	48.48	38.51	39.20	42.50	42.01
NM_001290372	0.13	0.03	0.16	0.00	2.49	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.59
NM_021374	0.00	0.00	0.00	0.89	1.08	1.28	0.00	0.48	0.00	0.62	0.26	0.72	0.00
NM_001177795	0.00	0.08	0.00	1.92	0.00	0.00	0.41	0.00	0.56	0.00	0.45	0.00	0.00
NM_001204371	0.00	0.01	0.01	0.41	0.02	0.02	0.00	0.01	0.00	0.01	0.00	0.01	0.05
NM_011011	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NM_133826	35.06	0.01	0.02	0.00	0.19	0.18	0.06	0.10	0.12	0.05	0.04	0.03	0.04
NM_010342	0.01	0.00	0.00	0.18	0.00	40.59	39.40	41.64	0.15	50.43	0.15	44.23	0.16
NM_001195732	0.00	41.64	40.78	40.42	43.12	0.00	0.05	0.00	38.69	0.05	49.11	0.20	43.68

Loading R packages ... ...

Done. Ready to load data files.

Normalización

Keep genes with minimal counts per million (CPM) in at least n libraries:

Min. CPM: 0.5

n libraries: 1

Transform counts data for clustering & PCA.

- ○ VST: variance stabilizing transform
- ○ rlog: regularized log (slow)
- ◉ EdgeR: log2(CPM+c)

Pseudo count c: 4

Missing values imputation:

Gene median

Plot one or more genes

Search processed data

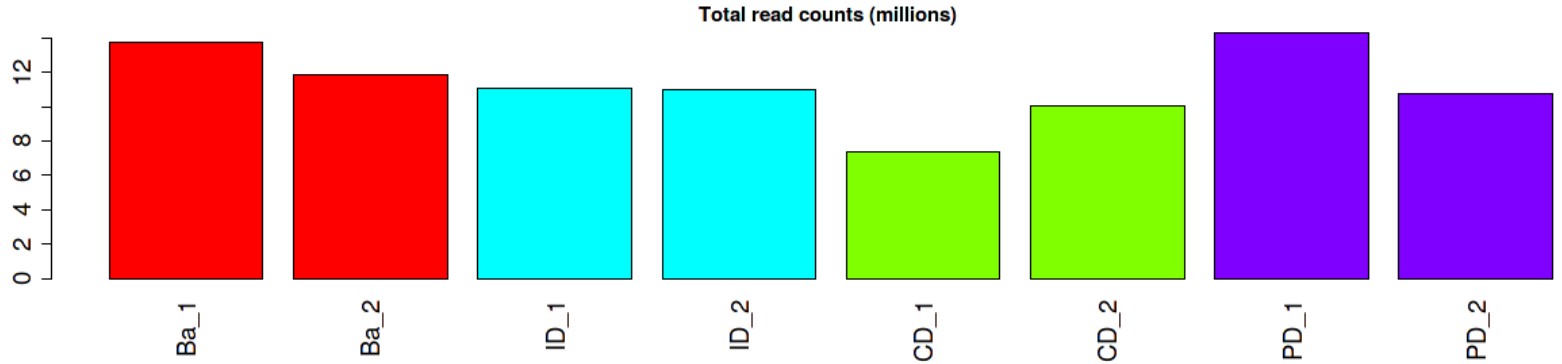
☐ Do not convert gene IDs to Ensembl.

Processed data

Converted counts data

High-resolution figure

Aspect ratios of figures can be adjusted by changing the width of browser window.

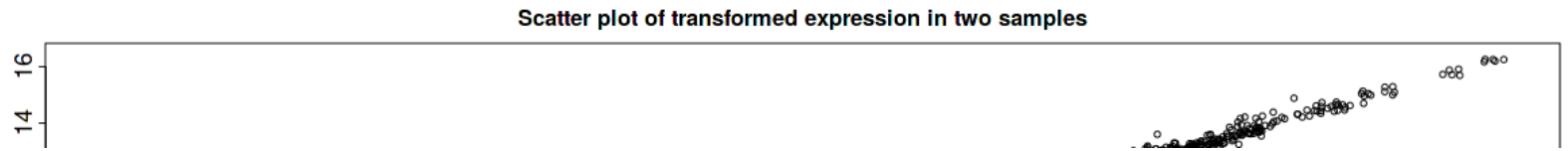


Select a sample for x-axis

Ba_1

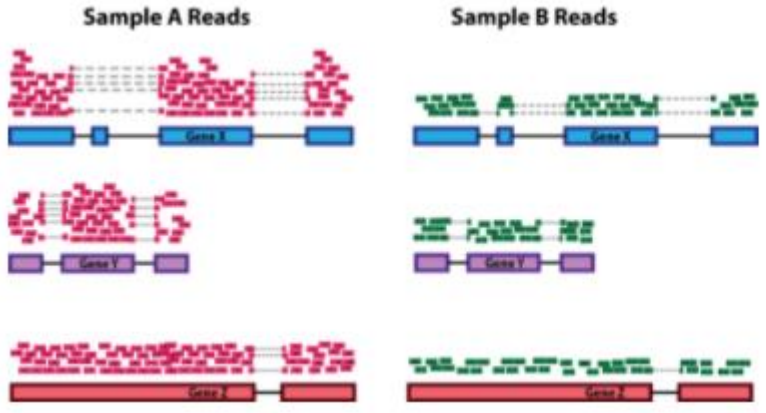
Select a sample for y-axis

Ba_2



Normalización

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**

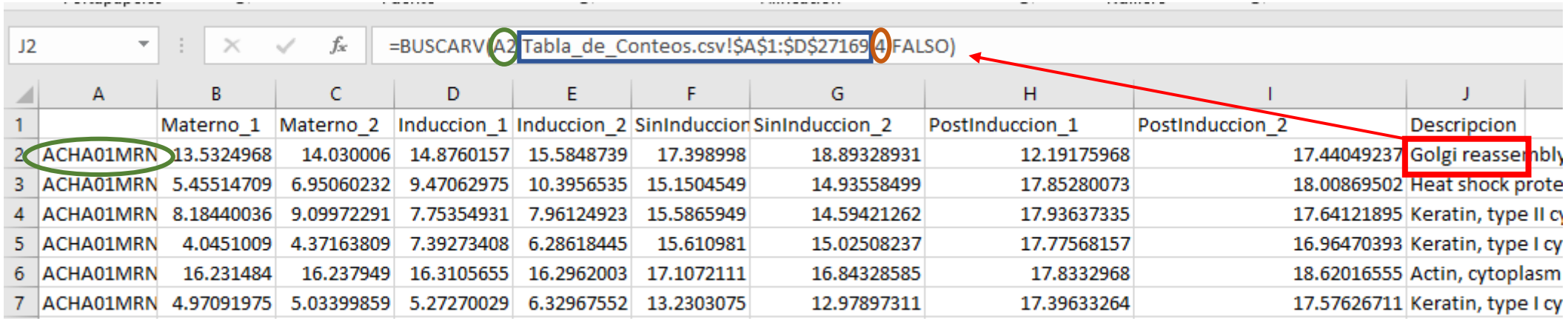


Cómo podríamos saber cuál es el gen más expresado en una determinada muestra?

$$CPM = \frac{counts}{TotalSampleCounts/1E6}$$

$$RPKM = \frac{CPM}{GeneLength(kb)}$$

Cómo cruzamos tablas en Excel para recuperar la descripción de la Proteína?



J2 | =BUSCARV(A2;Tabla_de_Conteos.csv!$A$1:$D$27169;4;FALSO)

	A	B	C	D	E	F	G	H	I	J
1		Materno_1	Materno_2	Induccion_1	Induccion_2	SinInduccion	SinInduccion_2	PostInduccion_1	PostInduccion_2	Descripcion
2	ACHA01MRN	13.5324968	14.030006	14.8760157	15.5848739	17.398998	18.89328931	12.19175968	17.44049237	Golgi reassembly
3	ACHA01MRN	5.45514709	6.95060232	9.47062975	10.3956535	15.1504549	14.93558499	17.85280073	18.00869502	Heat shock prote
4	ACHA01MRN	8.18440036	9.09972291	7.75354931	7.96124923	15.5865949	14.59421262	17.93637335	17.64121895	Keratin, type II cy
5	ACHA01MRN	4.0451009	4.37163809	7.39273408	6.28618445	15.610981	15.02508237	17.77568157	16.96470393	Keratin, type I cy
6	ACHA01MRN	16.231484	16.237949	16.3105655	16.2962003	17.1072111	16.84328585	17.8332968	18.62016555	Actin, cytoplasm
7	ACHA01MRN	4.97091975	5.03399859	5.27270029	6.32967552	13.2303075	12.97897311	17.39633264	17.57626711	Keratin, type I cy

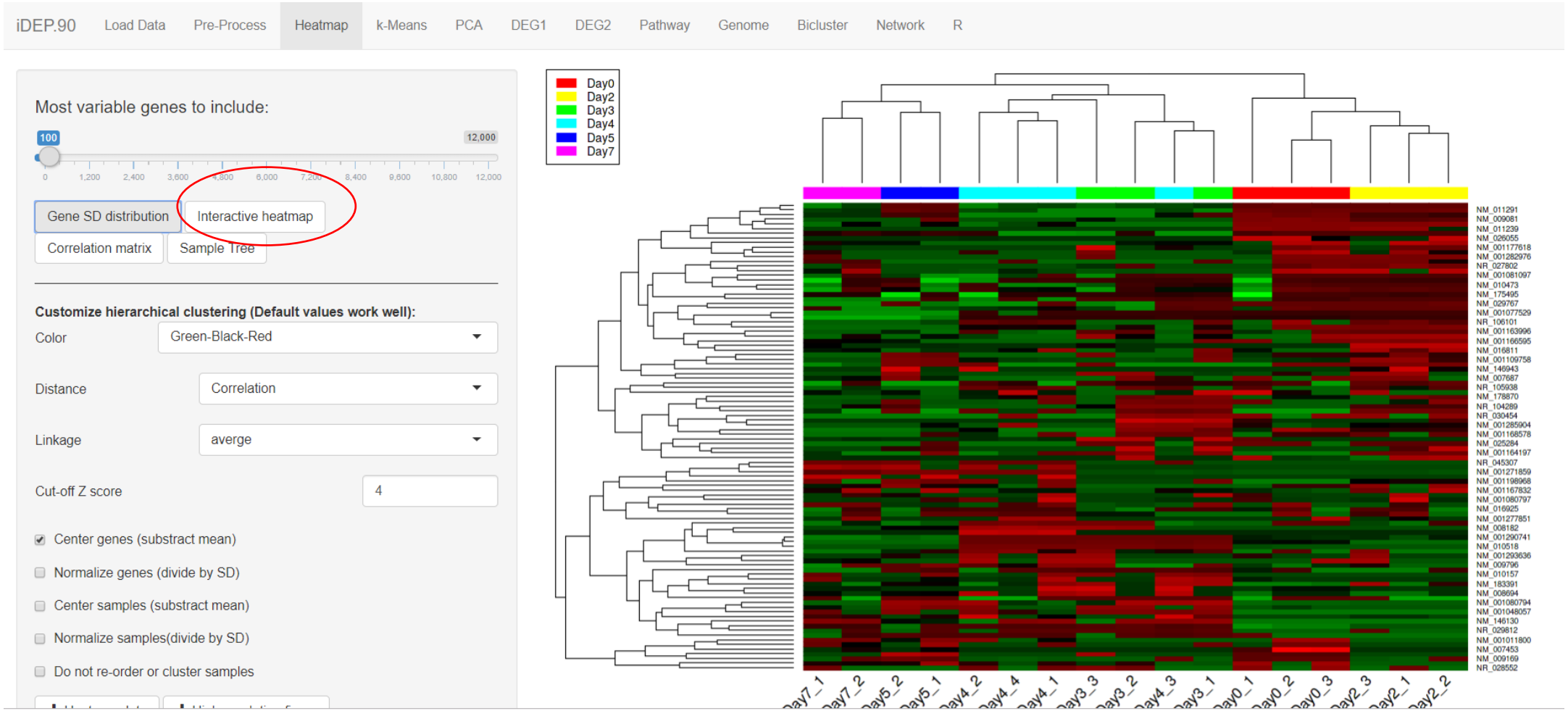
A.charrua Id	Largo de Ger	UniProtId	Descripción	Materno_1	Materno_2	Induccion_1	Induccion_2	SinInduccion	SinInduccion	PostInduccic	PostInducci
ACHA01MRN	738	Q5XFQ6	Transcriptio	12	6	11	4	7	8	9	
ACHA01MRN	1093	Q90WY4	Alpha-2A ad	0	0	0	0	1	0	0	
ACHA01MRN	1344	Q6NY64	Serine/thre	5520	4380	2294	1488	1021	1443	965	53
ACHA01MRN	714	Q7ZV80	Survival of r	1567	1247	353	245	1159	1559	723	16
ACHA01MRN	1093	Q07802	Transcriptio	39	12	56	43	20	36	32	1
ACHA01MRN	1947	G3V909	Cyclic AMP-	1101	846	962	820	542	808	671	28
ACHA01MRN	2502	Q8BM13	Noelin-2	2	2	10	19	14	10	24	
ACHA01MRN	865	Q96P48	Arf-GAP wit	253	148	81	110	50	108	26	2
ACHA01MRN	660	Q8WZ64	Arf-GAP wit	202	171	38	49	33	57	20	2
ACHA01MRN	4844	A0JN92	Up-regulato	3208	3269	416	237	848	1337	521	2

Celda compartida entre dos tablas que usamos como base para obtener datos de otra tabla

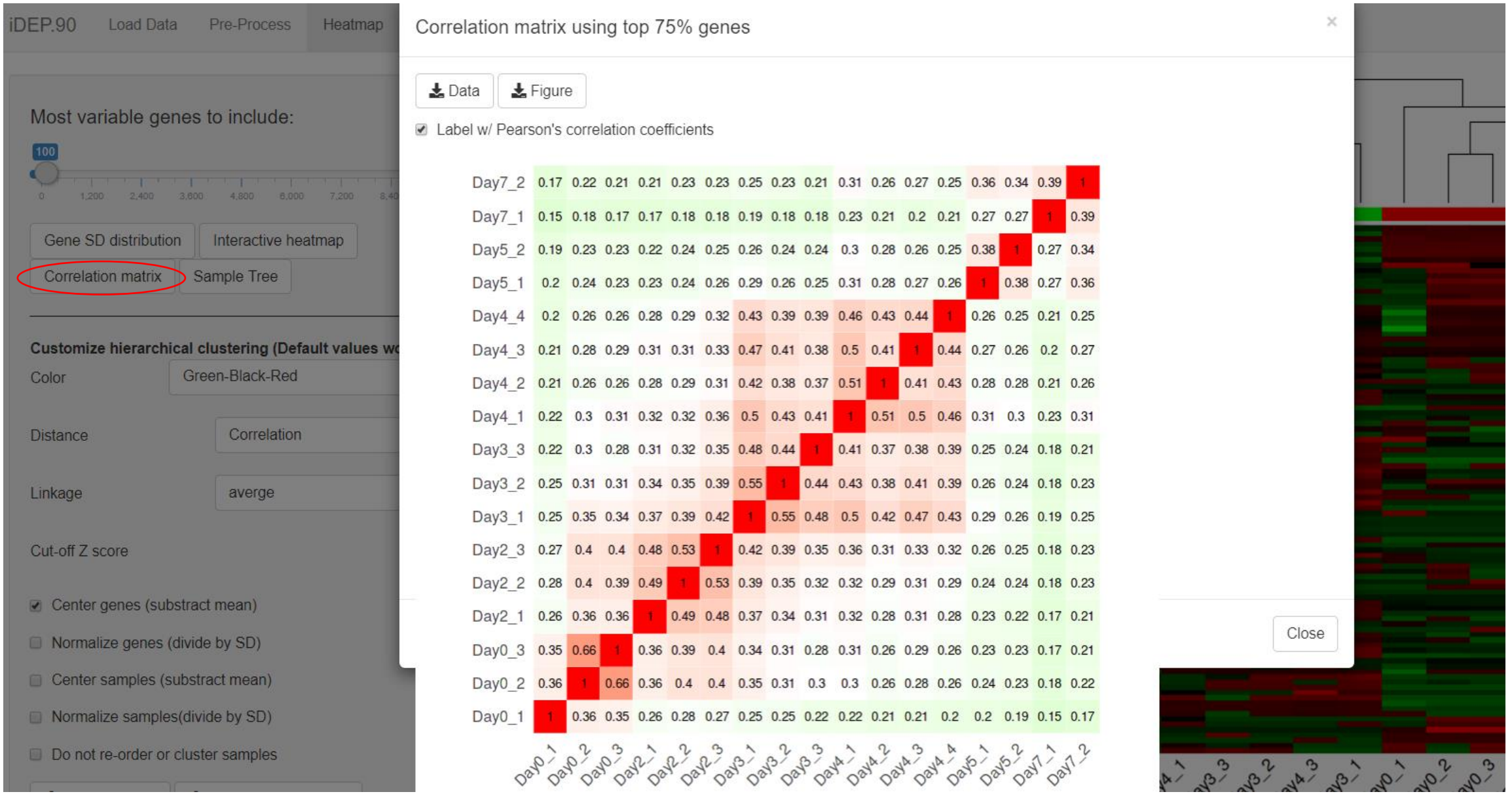
Nombre de la tabla de la cual se quiere obtener información, y rango de celdas en los que buscar (incluir celdas compartidas y celdas deseadas)

Número de Columna del rango seleccionado anteriormente en el cual se encuentra la celda que deseo

Heatmap


iDEP.90
Load Data
Pre-Process
Heatmap
k-Means
PCA
DEG1
DEG2
Pathway
Genome
Bicluster
Network
R
Most variable genes to include:
100
12,000
0
1,200
2,400
3,600
4,800
6,000
7,200
8,400
9,600
10,800
12,000
Gene SD distribution
Interactive heatmap
Correlation matrix
Sample Tree
Customize hierarchical clustering (Default values work well):
Color
Green-Black-Red
Distance
Correlation
Linkage
averge
Cut-off Z score
4
Center genes (substract mean)
Normalize genes (divide by SD)
Center samples (substract mean)
Normalize samples(divide by SD)
Do not re-order or cluster samples
Day0
Day2
Day3
Day4
Day5
Day7


iDEP.90
Load Data
Pre-Process
Heatmap
k-Means
PCA
DEG1
DEG2
Pathway
Genome
Bicluster
Network
R
Most variable genes to include:
100
0 1,200 2,400 3,600 4,800 6,000 7,200 8,400
Gene SD distribution
Interactive heatmap
Correlation matrix
Sample Tree
Customize hierarchical clustering (Default values wor
Color
Green-Black-Red
Distance
Correlation
Linkage
averge
Cut-off Z score
Center genes (substract mean)
Normalize genes (divide by SD)
Center samples (substract mean)
Normalize samples(divide by SD)
Do not re-order or cluster samples
Heatmap with hierarchical clustering tree
Most variable genes to include:
100
12,000
0 1,200 2,400 3,600 4,800 6,000 7,200 8,400 9,600 12,000
Mouse over to see gene names. To zoom, click and drag up or downward and release.
value
10
5
0
-5
-10


iDEP.90
Load Data
Pre-Process
Heatmap
Most variable genes to include:
100
Gene SD distribution
Interactive heatmap
Correlation matrix
Sample Tree
Customize hierarchical clustering (Default values wo
Color
Green-Black-Red
Distance
Correlation
Linkage
averge
Cut-off Z score
Center genes (substract mean)
Normalize genes (divide by SD)
Center samples (substract mean)
Normalize samples(divide by SD)
Do not re-order or cluster samples
Correlation matrix using top 75% genes
Data
Figure
Label w/ Pearson's correlation coefficients
Day7_2 0.17 0.22 0.21 0.21 0.23 0.23 0.25 0.23 0.21 0.31 0.26 0.27 0.25 0.36 0.34 0.39 1
Day7_1 0.15 0.18 0.17 0.17 0.18 0.18 0.19 0.18 0.18 0.23 0.21 0.2 0.21 0.27 0.27 1 0.39
Day5_2 0.19 0.23 0.23 0.22 0.24 0.25 0.26 0.24 0.24 0.3 0.28 0.26 0.25 0.38 1 0.27 0.34
Day5_1 0.2 0.24 0.23 0.23 0.24 0.26 0.29 0.26 0.25 0.31 0.28 0.27 0.26 1 0.38 0.27 0.36
Day4_4 0.2 0.26 0.26 0.28 0.29 0.32 0.43 0.39 0.39 0.46 0.43 0.44 1 0.26 0.25 0.21 0.25
Day4_3 0.21 0.28 0.29 0.31 0.31 0.33 0.47 0.41 0.38 0.5 0.41 1 0.44 0.27 0.26 0.2 0.27
Day4_2 0.21 0.26 0.26 0.28 0.29 0.31 0.42 0.38 0.37 0.51 1 0.41 0.43 0.28 0.28 0.21 0.26
Day4_1 0.22 0.3 0.31 0.32 0.32 0.36 0.5 0.43 0.41 1 0.51 0.5 0.46 0.31 0.3 0.23 0.31
Day3_3 0.22 0.3 0.28 0.31 0.32 0.35 0.48 0.44 1 0.41 0.37 0.38 0.39 0.25 0.24 0.18 0.21
Day3_2 0.25 0.31 0.31 0.34 0.35 0.39 0.55 1 0.44 0.43 0.38 0.41 0.39 0.26 0.24 0.18 0.23
Day3_1 0.25 0.35 0.34 0.37 0.39 0.42 1 0.55 0.48 0.5 0.42 0.47 0.43 0.29 0.26 0.19 0.25
Day2_3 0.27 0.4 0.4 0.48 0.53 1 0.42 0.39 0.35 0.36 0.31 0.33 0.32 0.26 0.25 0.18 0.23
Day2_2 0.28 0.4 0.39 0.49 1 0.53 0.39 0.35 0.32 0.32 0.29 0.31 0.29 0.24 0.24 0.18 0.23
Day2_1 0.26 0.36 0.36 1 0.49 0.48 0.37 0.34 0.31 0.32 0.28 0.31 0.28 0.23 0.22 0.17 0.21
Day0_3 0.35 0.66 1 0.36 0.39 0.4 0.34 0.31 0.28 0.31 0.26 0.29 0.26 0.23 0.23 0.17 0.21
Day0_2 0.36 1 0.66 0.36 0.4 0.4 0.35 0.31 0.3 0.3 0.26 0.28 0.26 0.24 0.23 0.18 0.22
Day0_1 1 0.36 0.35 0.26 0.28 0.27 0.25 0.25 0.22 0.22 0.21 0.21 0.2 0.2 0.19 0.15 0.17
Day0_1 Day0_2 Day0_3 Day2_1 Day2_2 Day2_3 Day3_1 Day3_2 Day3_3 Day4_1 Day4_2 Day4_3 Day4_4 Day5_1 Day5_2 Day7_1 Day7_2
Close

K-means

- K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems.
- It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand. The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

K-means

iDEP.90 | Load Data | Pre-Process | Heatmap | k-Means | PCA | DEG1 | DEG2 | Pathway | Genome | Bicluster | Network | R

Most variable genes to include

0 2,000 12,000

0 1,200 2,400 3,600 4,800 6,000 7,200 8,400 9,600 10,800 12,000

Number of Clusters

2 4 20

2 4 6 8 10 12 14 16 18 20

Re-Run | How many clusters? | Gene SD distribution | t-SNE map

Normalize by gene:

Mean center

Enriched TF binding motifs

K-means data | High-resolution figure

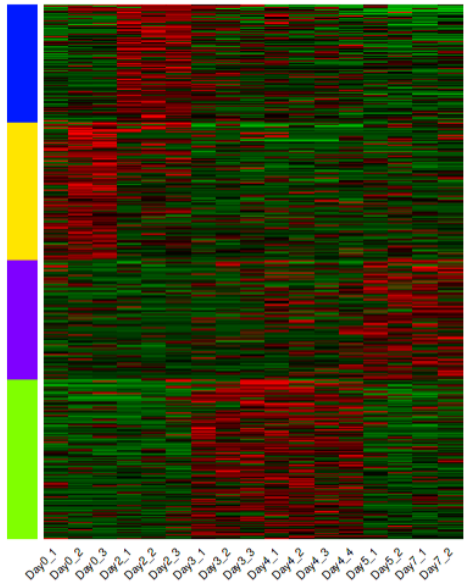
Pathway database

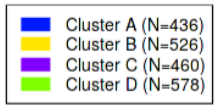
GO Biological Process

Remove redudant genesets

Visualize enrichment | Enrichment details

?



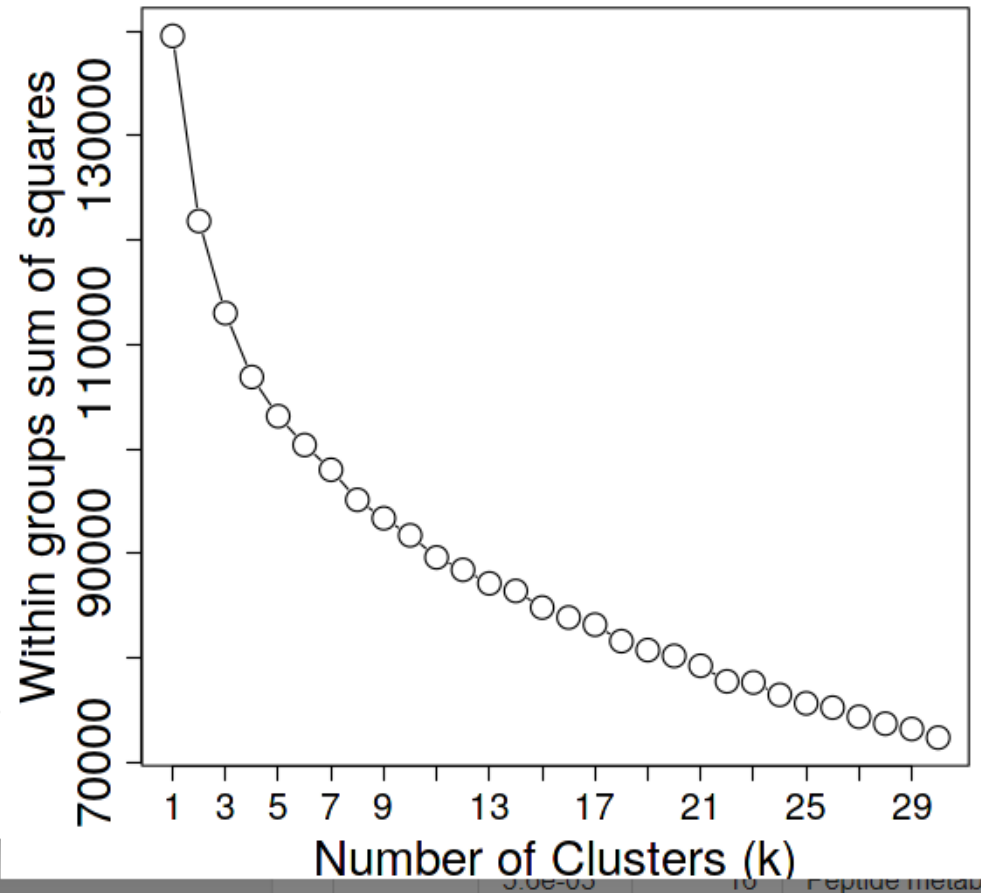


Enriched pathways for each cluster

Cluster	adj.Pval	nGenes	Pathways
B	1.7e-09	38	Translation
	2.4e-09	41	Amide biosynthetic process
	4.1e-09	48	Cellular protein-containing complex assembly
	4.6e-09	41	Peptide metabolic process

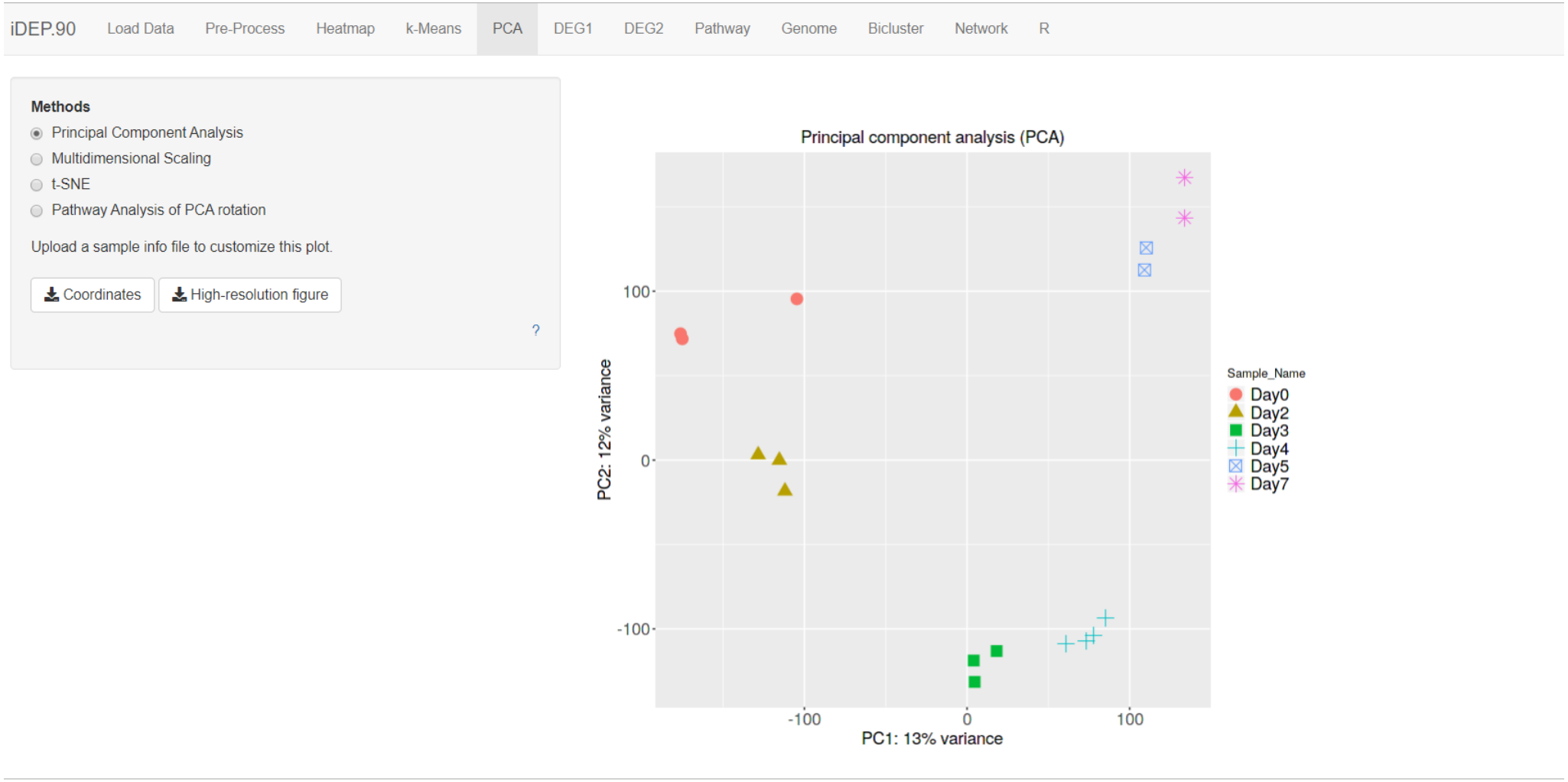
Determining the number of clusters (k)

Following the elbow method, one should choose k so that adding another cluster does not substantially reduce the within groups sum of squares. Wikipedia



Close

PCA (Principal components Analysis)
Análisis de componentes principales



DEG (Differentially Expressed Genes)

iDEP.90 | Load Data | Pre-Process | Heatmap | k-Means | PCA | DEG1 | DEG2 | Pathway | Genome | Bicluster | Network | R

Identifying Differential Expressed Genes (DEGs). See next tab for details.
Using the limma package

FDR cutoff: 0,1

Min fold change: 2

Select factors & comparisons

Venn Diagram

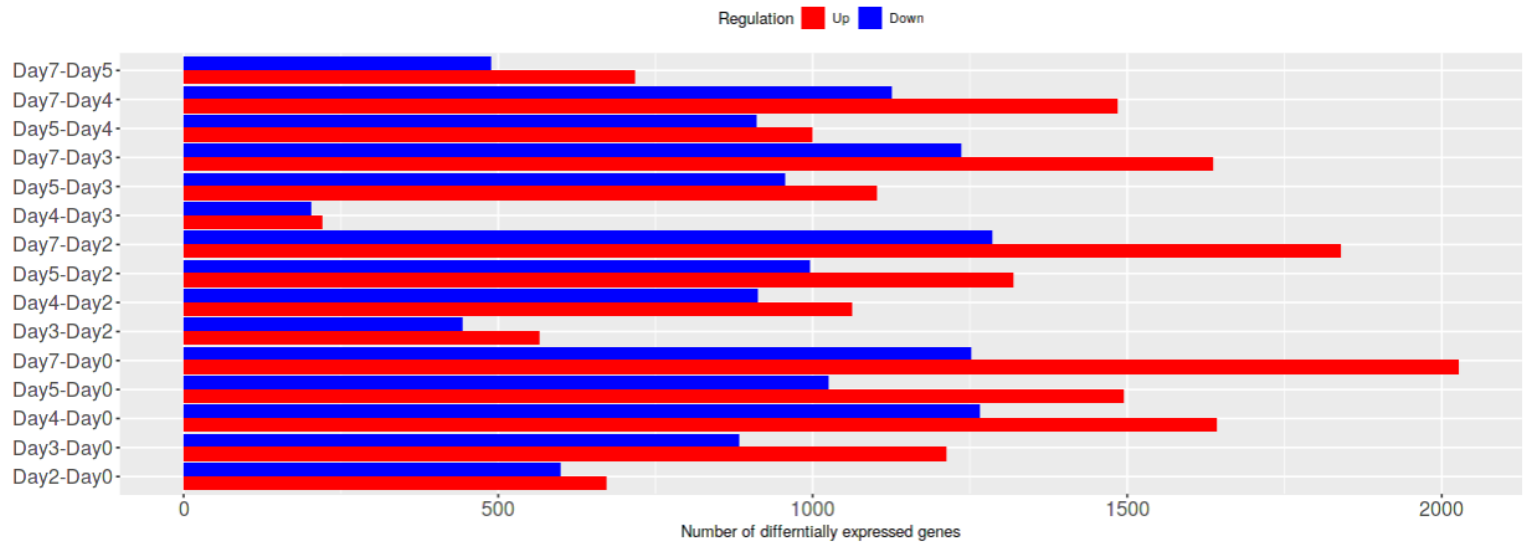
Gene lists

FDR & fold-changes for all genes

Figure

6 sample groups detected.

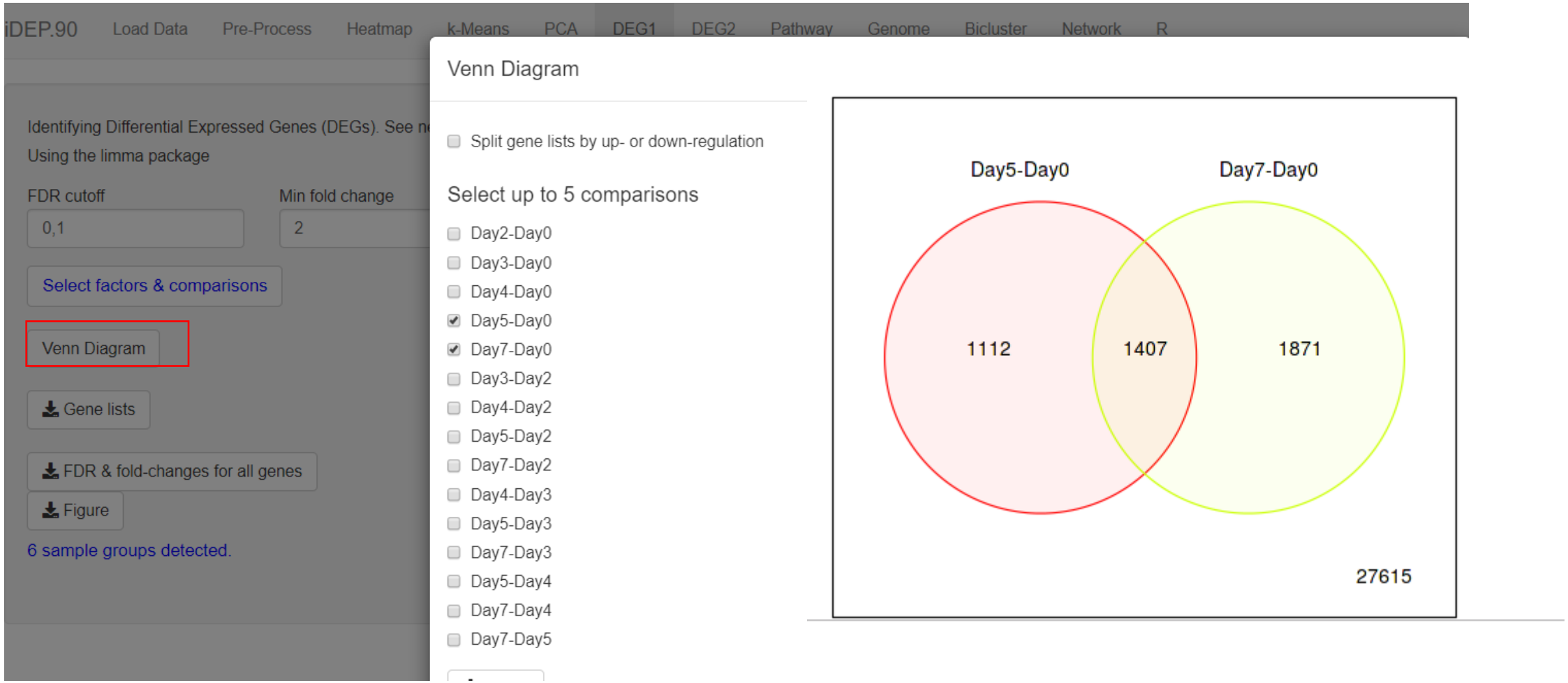
?



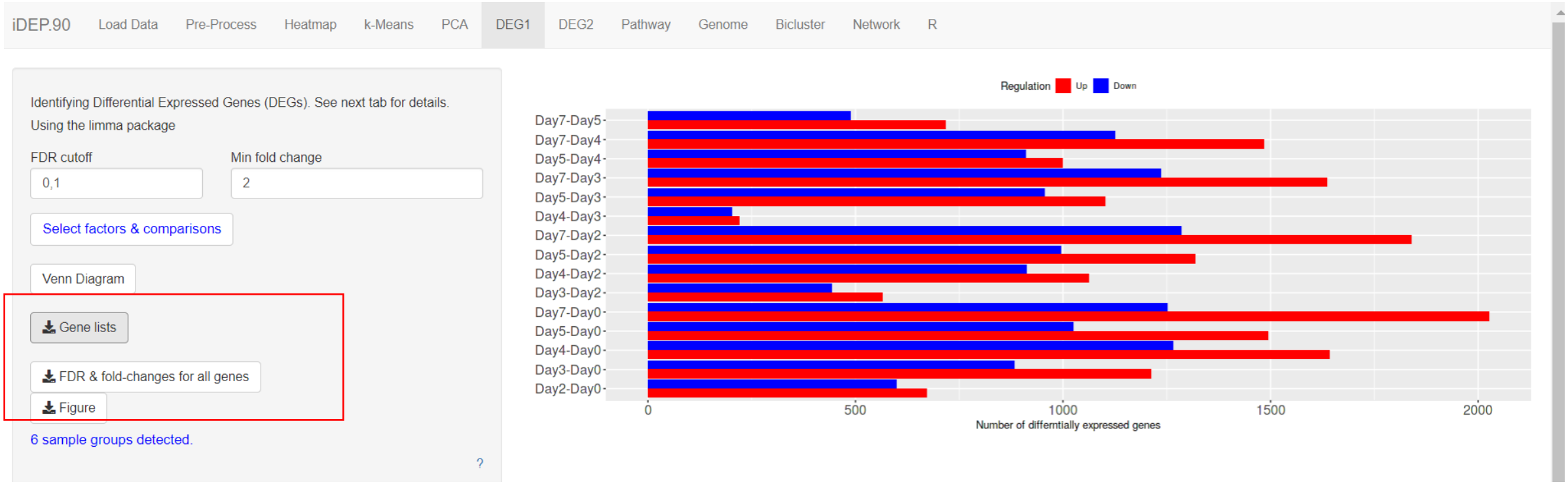
Numbers of differentially expressed genes for all comparisons. "B-A" means B vs. A. Interaction terms start with "I:"

Comparisons	Up	Down
Day7-Day5	716	488
Day7-Day4	1484	1124
Day5-Day4	998	909
Day7-Day3	1635	1235
Day5-Day3	1101	956
Day4-Day3	220	201

Podemos representar la información con un diagrama de Venn



Podemos descargar la lista de genes



Examine the results of DEGs for each comparison

Select a comparison to examine. "A-B" means A vs. B (See heatmap). Interaction terms start with "I:"

Day2-Day0

Volcano Plot | MA Plot | Scatter Plot

TF binding motifs in promoters | Gene list & data

High-resolution figure

Enrichment analysis for DEGs:

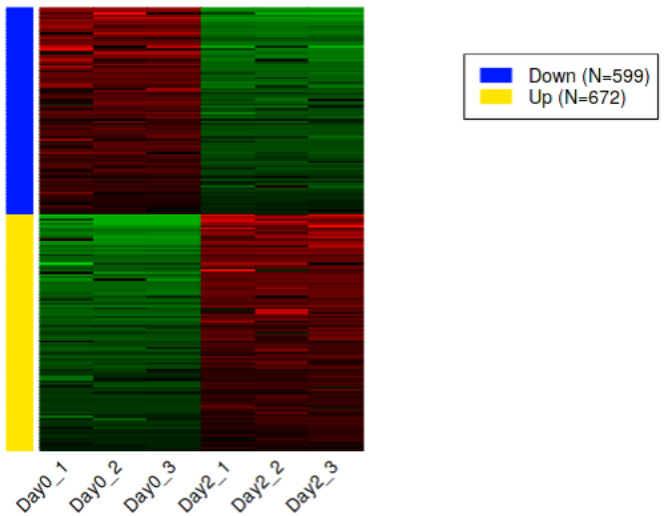
GO Biological Process

Enrichment tree | Enrichment network | Enrichment details

Enrichment using STRING API

Also try ShinyGO

?



Enriched pathways in DEGs for the selected comparison:

Direction	adj.Pval	nGenes	Pathways
Down regulated	5.3e-07	154	Nucleic acid metabolic process
	4.9e-06	106	Cellular component biogenesis
	2.6e-05	117	Organelle organization
	3.0e-05	133	RNA metabolic process
	8.2e-05	41	RNA processing
	1.6e-04	138	Macromolecule biosynthetic process
	5.7e-04	25	Ribonucleoprotein complex biogenesis