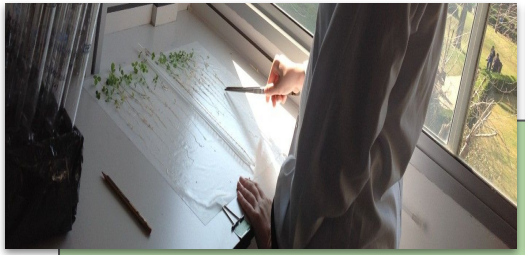
Maria del Saz Navarro

0000-0003-1551-5341
https://github.com/mariadsn
mariadsn89@gmail.com



Bióloga

- **Técnico de laboratorio Microbiología**
  U. Sevilla
- **Tecnico Bioinformática**
  Centro Andaluz de Biología del Desarrollo (CABD)
- **Pre-PhD Bioinformática (DATAi group)**
  U. Pablo de Olavide

Universidad Pablo de Olavide

Ctra. de Utrera, Sevilla

Intelligent Data Analysis (DATAi)

Members

Norberto Díaz Díaz
Carlos D. Barranco
Francisco A. Gómez Vela
Domingo S. Rodríguez Baena
Aurelio López Fernández
Juan José Díaz Montaña

Ph.D student

José A. Lagares
María del Saz Navarro
Francisco Gonzalez Cabanes

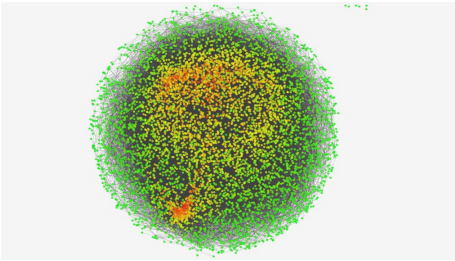
UNIVERSIDAD PABLO DE OLAVIDE SEVILLA

Universidad Internacional de Valencia

Máster Universitario en Big Data y Ciencia de Datos

7 Diciembre 2023
18:00h

*Técnicas de Procesamiento de datos de **Secuenciación de Alto Rendimiento***

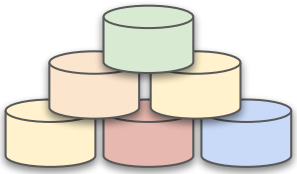


María del Saz
Pre-Ph.D

DATAI
Intelligent Data Analysis

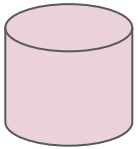
Minería de Datos

Procesamiento de Datos **BIOLÓGICOS**

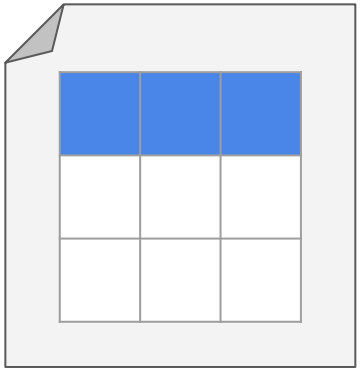


Recopilación

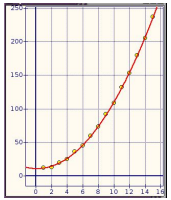
Datos



Selección, limpieza y transformación

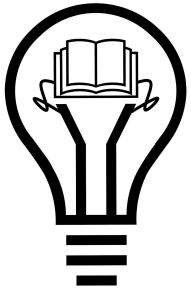


Data mining



Modelos

Interpretación y evaluación



Conocimiento

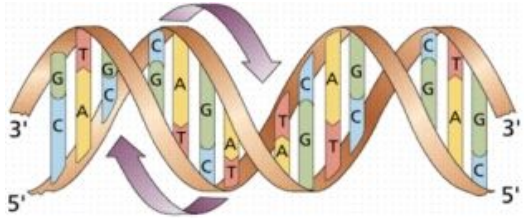
KDD es un proceso **sistemático** que busca:

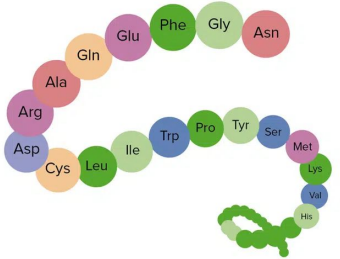
revelar **patrones**, **tendencias** y **relaciones significativas** en **grandes conjuntos de datos**

Fundamentos Básicos en Biología

1- Dogma Central de la Biología
2- Secuenciación de Segunda Generación

2- Conceptos / **Dogma central de la Biología Molecular**





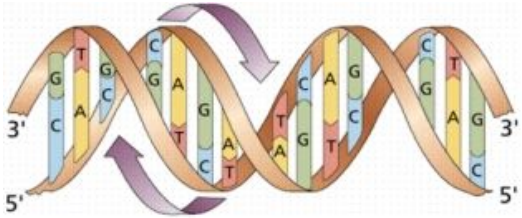
>Gen 1
TCATATTCGTTTTACGTTTGTCAAGCCTCAT
AGCCGCAGTTCGAACGTATACGCTCTCTGA
GTCAGACCTCAAATCGTAGCTCTACACAAT
TCTGTGAATTTTCCTTGTCGCGTGTGAAAC
ACTTCCAATAAAAACTCATATGGTGAGTACT
TTAAAAAAAAATCTAGTGAAATAATGCTGAA
AAGAAATTTGTGTGGGCAAAATTCAATGGG
CAAAAACGCGATGCGGCTTTTTCTCAAAAT
GGCGGCCGGCCTGCGTTTTTTCCTCAAAA
GTGATGACGTCATGCCTGTTTTTTTTTTTTG
TTCGCAATGAGGAATGGCTTTAAAAT…

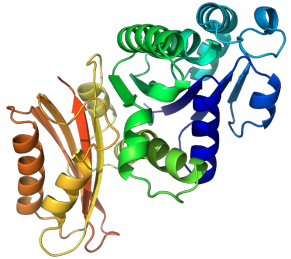
>ANRm 1
TCATATTCGTTTTACGTTTGTCAAGCCTCATAG
CCGGCAGTTCGAACGTATACGCTCTCTGAGTC
AGACCTCGAAATCGTAGCTCTACACAATTCTG
TGAATTTTCCTTGTCGCGTGTGAAACACTTCC
AATAAAAACTCAATATGCGTGAATGTATCTCTA
TCCATGTTGGTCAGGCTGGTGTCCAGATTGGA
AACGCCTGCTGGGAGCTCTACTGCTTGGAGC
ACGGCATCCAGCCCGATGGCCAGATGCCGTC
TGACAAGACCGTGGGCGGAGGTGATGACTCG
TTCAACACCTTCTTCAGCGAGACTGGAGCTGG
CAAGCAC…

>Proteína 1
MRECISIHVGQAGVQIGNACWELYCLEHGIQP
DGQMPSDKTVGGGDDSFNTFFSETGAGKHVP
RAVFVDLEPTVVDEVRTGTYRQLFHPEQLITGK
EDAANNYARGHYTIGKEIVDLVLDRIRKLADQC
TGLQGFLIFHSFGGGTGSGFTSLLMERLSVDY
GKKSKLEFAIYPAPQVSTAVVEPYNSILTTHTTL
EHSDCAFMVDNEAIYDICRRNLDIERPTYTNLN
RLIGQIVSSITASLRFDGALNVDLTEFQTNLVPY
PRIHFPLVTYAPVISAEKAYHEQLSVAEITNACF
EPANQMVKCDPRHGKYMACCMLYRGDVVPK
DVNAAIATIKTKRTIQFVDWCPTGFKVGINYQPP
TVVPGGDLAKVQRAVCM

DNA —Transcripción→ **RNA** —Traducción→ **Protein**

2- Conceptos / **Dogma central de la Biología Molecular**





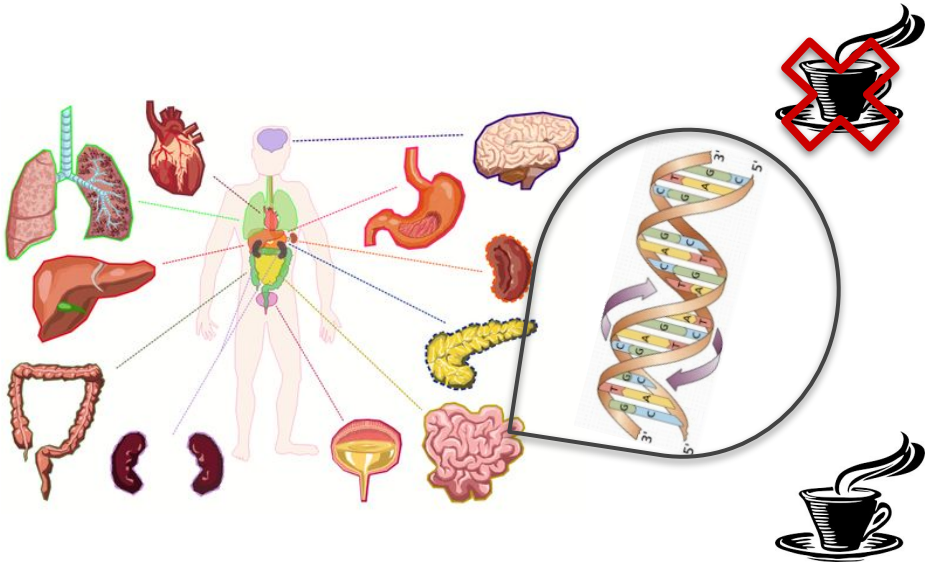
>Gen 1
TCATATTCGTTTTACGTTTGTCAAGCCTCAT
AGCCGCAGTTCGAACGTATACGCTCTCTGA
GTCAGACCTCAAATCGTAGCTCTACACAAT
TCTGTGAATTTTCCTTGTCGCGTGTGAAAC
ACTTCCAATAAAAACTCATATGGTGAGTACT
TTAAAAAAAAATCTAGTGAAATAATGCTGAA
AAGAAATTTGTGTGGGCAAAATTCAATGGG
CAAAAACGCGATGCGGCTTTTTCTCAAAAT
GGCGGCCGGCCTGCGTTTTTTCCTCAAAA
GTGATGACGTCATGCCTGTTTTTTTTTTTTG
TTCGCAATGAGGAATGGCTTTAAAAT…

>ANRm 1
TCATATTCGTTTTACGTTTGTCAAGCCTCATAG
CCGGCAGTTCGAACGTATACGCTCTCTGAGTC
AGACCTCGAAATCGTAGCTCTACACAATTCTG
TGAATTTTCCTTGTCGCGTGTGAAACACTTCC
AATAAAAACTCAATATGCGTGAATGTATCTCTA
TCCATGTTGGTCAGGCTGGTGTCCAGATTGGA
AACGCCTGCTGGGAGCTCTACTGCTTGGAGC
ACGGCATCCAGCCCGATGGCCAGATGCCGTC
TGACAAGACCGTGGGCGGAGGTGATGACTCG
TTCAACACCTTCTTCAGCGAGACTGGAGCTGG
CAAGCAC…

>Proteína 1
MRECISIHVGQAGVQIGNACWELYCLEHGIQP
DGQMPSDKTVGGGDDSFNTFFSETGAGKHVP
RAVFVDLEPTVVDEVRTGTYRQLFHPEQLITGK
EDAANNYARGHYTIGKEIVDLVLDRIRKLADQC
TGLQGFLIFHSFGGGTGSGFTSLLMERLSVDY
GKKSKLEFAIYPAPQVSTAVVEPYNSILTTHTTL
EHSDCAFMVDNEAIYDICRRNLDIERPTYTNLN
RLIGQIVSSITASLRFDGALNVDLTEFQTNLVPY
PRIHFPLVTYAPVISAEKAYHEQLSVAEITNACF
EPANQMVKCDPRHGKYMACCMLYRGDVVPK
DVNAAIATIKTKRTIQFVDWCPTGFKVGINYQPP
TVVPGGDLAKVQRAVCM

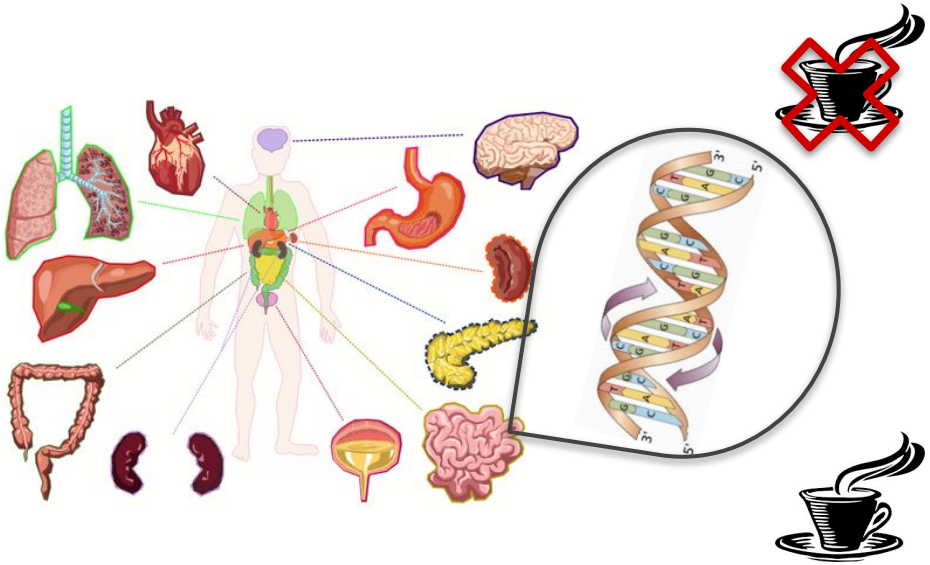
DNA —Transcripción→ **RNA** —Traducción→ **Protein**

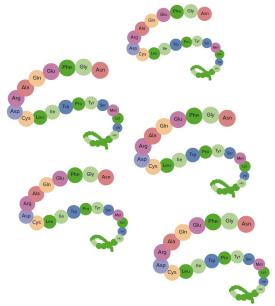
2- Conceptos / **Dogma central de la Biología Molecular**



DNA —Transcripción→ **RNA** —Traducción→ **Protein**

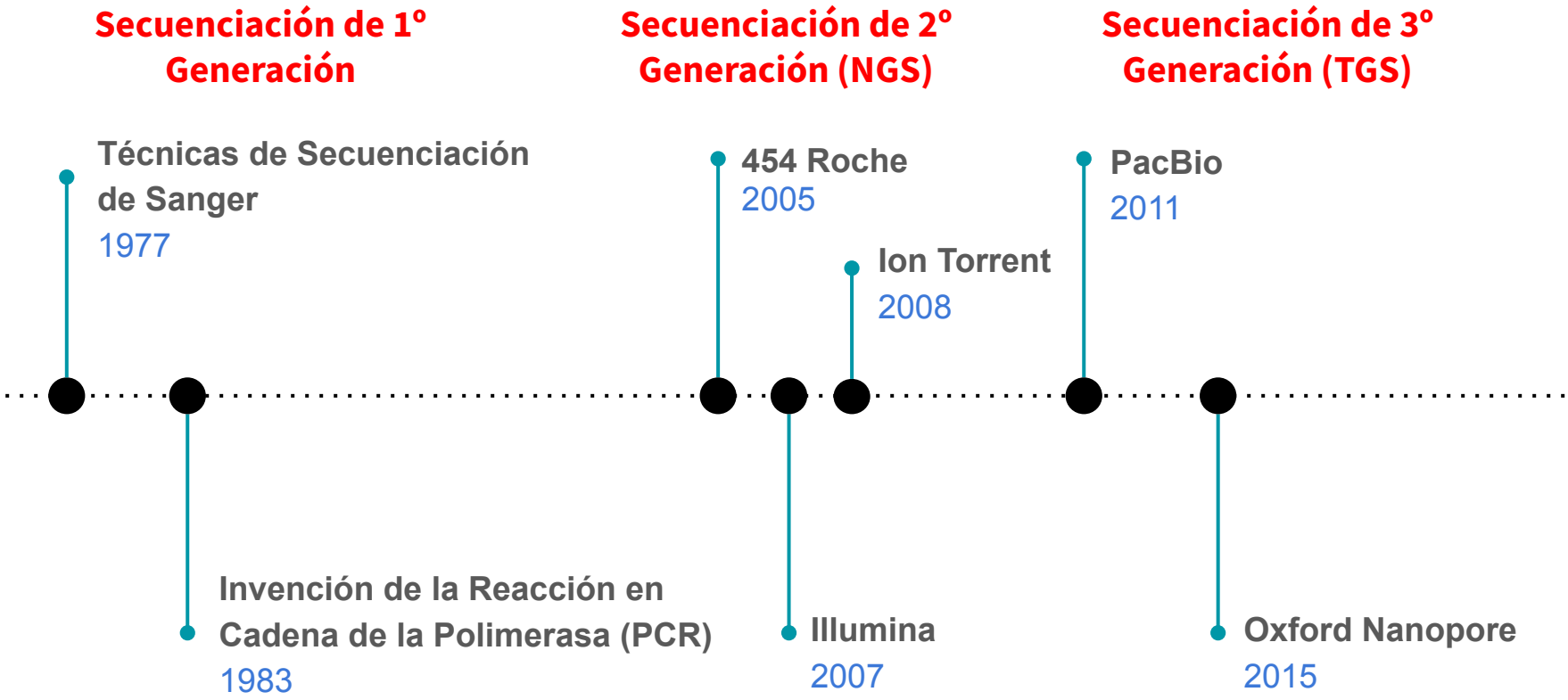
2- Conceptos / **Dogma central de la Biología Molecular**





DNA —Transcripción→ **RNA** —Traducción→ **Protein**

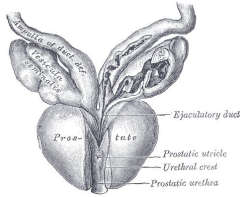
2- Conceptos / **Secuenciación de Segunda Generación**



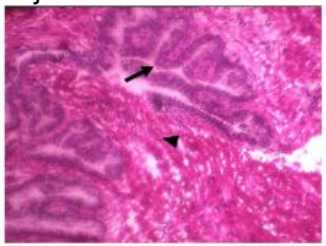
Investigación Biológica

1- Diseño Experimental
2- Fases del Estudio

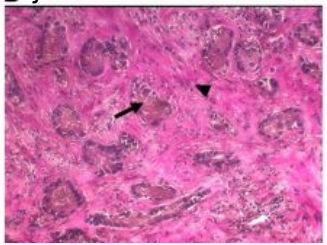
3- Investigación / **Diseño Experimental**

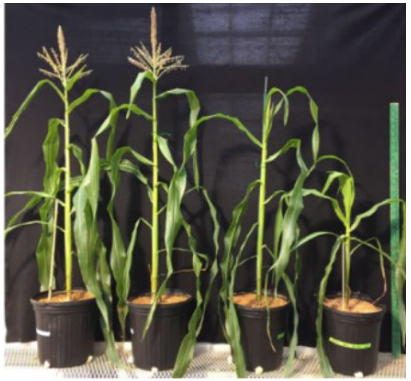


Tejido Prostático Normal

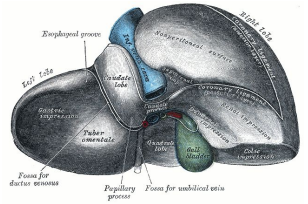


Tejido Prostático Tumoral

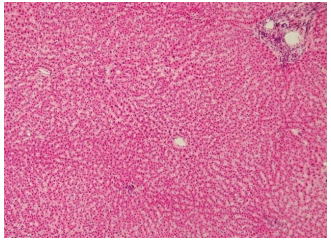




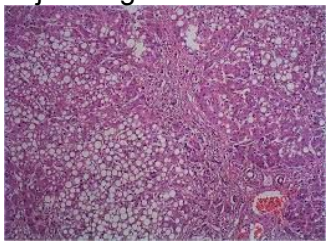
Ambiente Sequía



Tejido Hígado Normal



Tejido Hígado Cirrosis



3- Investigación / **Fases del Estudio**

Estudio transcriptómico basado en **RNA-Seq**

01 **Diseño Experimental**

02 Extracción del RNA y preparación

03 Secuenciación y obtención de lecturas

04 **Análisis matemático computacional de los datos**

05 Confirmación o validación biológica

Estudio transcriptómico basado en **RNA-Seq**

Diseño Experimental 01

02 Extracción del RNA y preparación

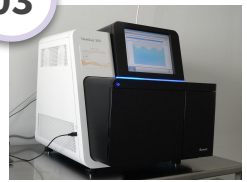
Secuenciación y obtención de lecturas 03

04 **Análisis matemático computacional de los datos**

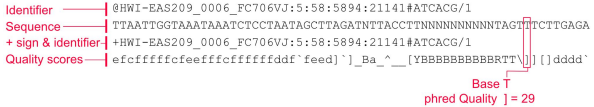
Confirmación o validación biológica 05

02

03



.fastq



Estudio transcriptómico basado en **RNA-Seq**

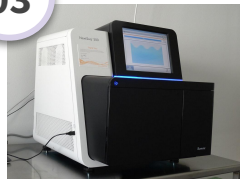
Diseño Experimental 01

02 Extracción del RNA y preparación

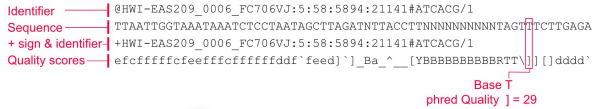
Secuenciación y obtención de lecturas 03

04 **Análisis matemático computacional de los datos**

Confirmación o validación biológica 05



.fastq



04

4.1 Quality control — 4.2 Data preprocessing — 4.3 Inferential statistics — 4.4 Exploratory analysis

05

Estudio transcriptómico

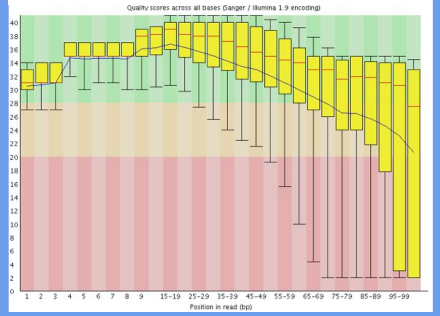
1- .fastq
2- Objetivo

4- Workshop / **Objetivo**

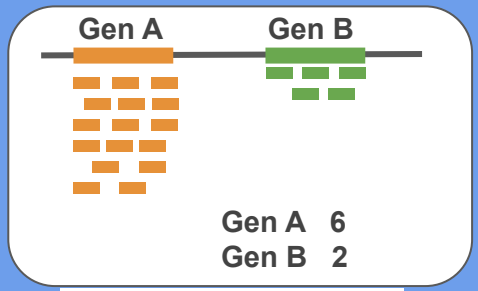
https://github.com/mariadsn/GeneExpressionAnalysis_Seminar

4.1 Quality control

FastQC, FastQScreen, FASTX



4.2 Data preprocessing

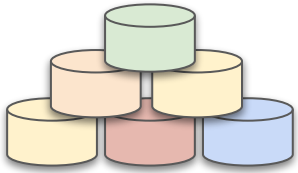


STAR, Bowtie2, HISAT2

4.3 Inferential statistics

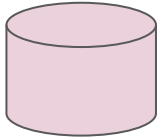
4.4 Exploratory analysis

4- Workshop / **Objetivo**



Recopilación

Datos



TCGA-BRCA

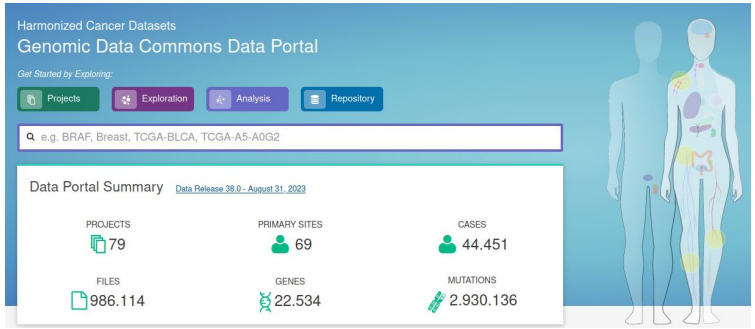
NCBI

Gene Expression Omnibus (GEO)

Noviembre 2023

Organism	Platforms	Samples
Homo sapiens	6,257	3,585,720
Mus musculus	2,746	2,035,017
Drosophila melanogaster	402	115,391
Arabidopsis thaliana	416	94,502
Saccharomyces cerevisiae	639	90,187

NIH NATIONAL CANCER INSTITUTE



Expression Atlas

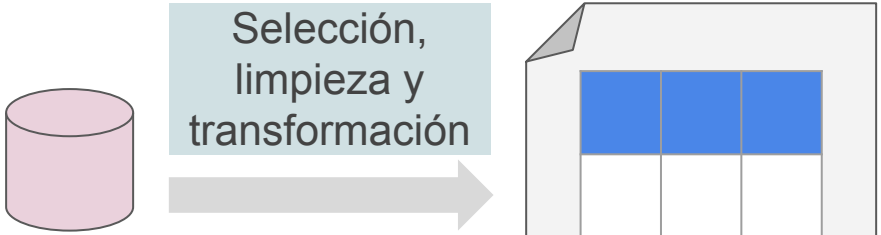
ENA European Nucleotide Archive

4- Workshop / **Objetivo**

Objetivos del Workshop

- **Anotación funcional** de genes
- Obtención de **biomarcadores**

4- Workshop / **Objetivo**



- select counts
- remove low counts
- log2
- nombre informativo de columnas

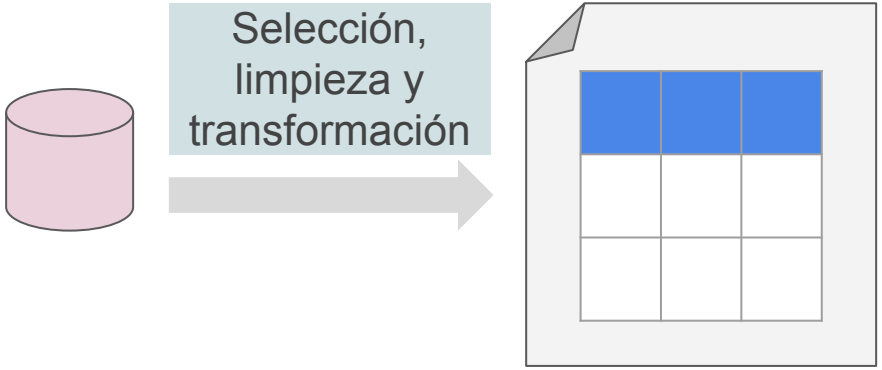
attr_name	Normal_breast_1	Normal_breast_2	Normal_breast_3	Normal_brea
ENSG00000000005.6	9.082149041353870	9.184875342908280	8.897845456005510	12.0161118
ENSG00000002726.21	4.247927513443590	4.700439718141090	3.584962500721160	3.584962500
ENSG00000003096.14	11.17617314910750	10.82257083095020	12.25296103639370	10.6329951
ENSG00000003137.8	9.417852514885590	11.27204652438990	7.787902559391430	10.7919768
ENSG00000003249.15	8.618385502258610	8.954196310386870	9.192292814470770	9.084808380
ENSG00000004776.13	10.01680828768660	10.62935662007960	9.296916206879290	11.1953722
ENSG00000004838.14	7.409390936137700	6.321928094887360	5.285402218862250	7.607330310
ENSG00000004846.16	10.19229281447080	8.257387842692650	8.960001932068080	7.813781190
ENSG00000004848.8	2.807354922057600	2.807354922057600	2.807354922057600	
ENSG00000004939.15	4	2	2.321928094887360	6.475733430
ENSG00000006118.15	9.878050912728540	9.997179480937620	9.259743263690780	9.209453360
ENSG00000006210.7	13.81878187227540	14.26590800923110	15.46103173308550	13.680799000
ENSG00000007174.18	7.149747111950468	7.800898999203000	7	7.118941070
ENSG00000007314.12	9.569855606833095	10.57931593758000	8.511752653767380	10.625708800
ENSG00000007372.24	2	1	2	2.807354922
ENSG00000007968.7	6.392317422778760	5.882643049361840	6.700439718141090	5.832890014
ENSG00000008735.14	8.487840033823050	8.344295907915820	7.686500527183220	7.189824550
ENSG00000010282.15	5.672425349715000	3.807354922057600	4.584962500721160	2.807354922
ENSG00000010319.7	13.17617314910750	13.16490692667570	11.12734954105910	13.810872800
ENSG00000010438.17	2.807354922057600	4.906890595608520	4.954196310386870	3.459431618
ENSG00000011426.11	8.769184342776930	7.098033082960530	8.931521846075700	6.727920454

Normal breast

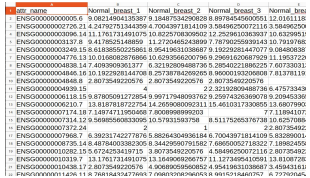
attr_name	Tumor_breast_1	Tumor_breast_2	Tumor_breast_3	Tumor_brea
ENSG00000000005.6	1	6.507794640198700	1	5.285402218
ENSG00000002726.21	12.14847658217830	6.149747119504680	7.044394119358450	5.614709844
ENSG00000003096.14	6.643856189774720	8.738092259620490	9.002815015607050	9.038918980
ENSG00000003137.8	8.044394119358450	7.918863237274590	7.228818690495880	7.700439718
ENSG00000003249.15	11.10656294044490	10.84549005094440	10.60917873814200	10.81297940
ENSG00000004776.13	4.459431618637300	6.965784284660290	6.727920454563200	7.930733330
ENSG00000004838.14	6.584962500721160	8.703903573444660	7.924812503605780	9.216745850
ENSG00000004846.16	1	4.700439718141090	4.459431618637300	6.977279923
ENSG00000004848.8	3	3.700439718141090	1.584962500721160	4.857980995
ENSG00000004939.15	1	1	1	
ENSG00000006118.15	14.05570609716130	10.14847658217830	12.27815903143720	13.020806300
ENSG00000006210.7	11.08547245918130	12.54303182022550	11.16679075071820	11.213711750
ENSG00000007174.18	1.584962500721160	5.930737337562890	4.247927513443590	5.781359713
ENSG00000007314.12	5.392317422778760	7.339850002884620	5.672425349197150	8.016802830
ENSG00000007372.24	7.451211111183230	2.584962500721160	4.754887502163470	2.321928094
ENSG00000007968.7	12.67970000576920	9.566054038171090	9.894817763307940	7.294620748
ENSG00000008735.14	10.37937836707130	10.67771946416410	11.77396336843360	10.375039430
ENSG00000010282.15	2	1	0	
ENSG00000010319.7	9.381542951184580	11.39874369193820	10.49785183695110	11.095397020
ENSG00000010438.17	1.584962500721160	0	2.807354922057600	
ENSG00000011426.11	11.60316787853440	10.91886232727460	10.33663793900260	9.197216690

Tumor breast

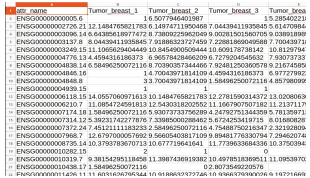
4- Workshop / **Objetivo**



Normal breast



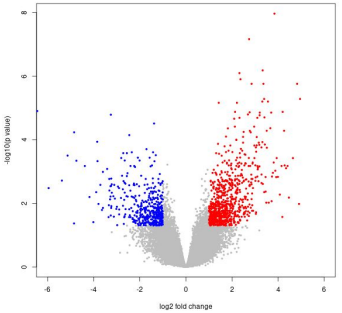
Tumor breast



Tools

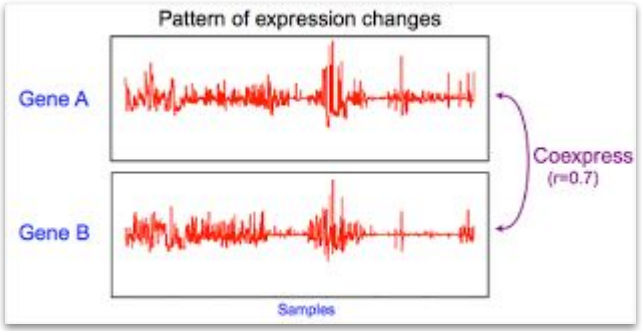
- **DESeq**
- edgeR
- limma-voom
- NOISeq
- eXpress

Expresión diferencial de genes

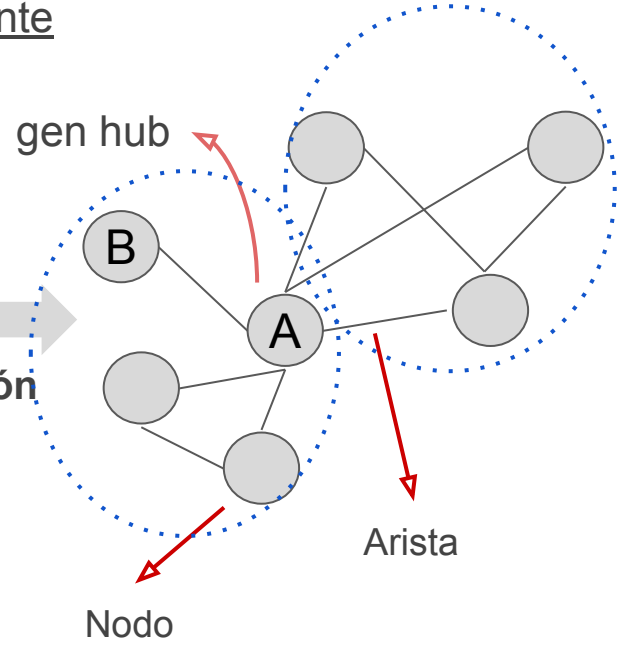


4- Workshop / **Objetivo**

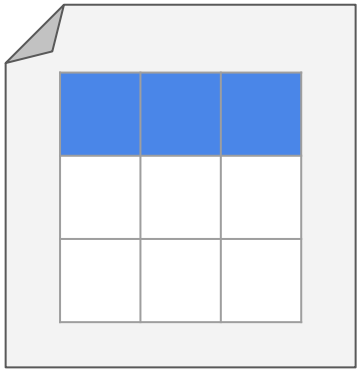
Genes con **patrones de expresión similares** (co-expresados)
están <u>asociados funcionalmente</u>



Red de Co-expresión génica

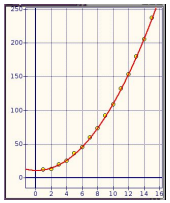


4- Workshop / **Objetivo**

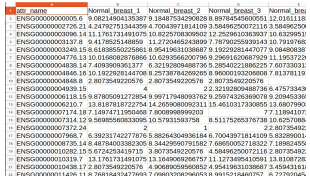


Data mining

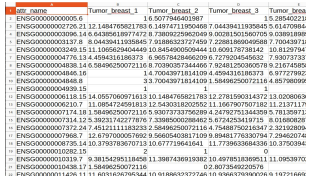
Modelos



Normal breast



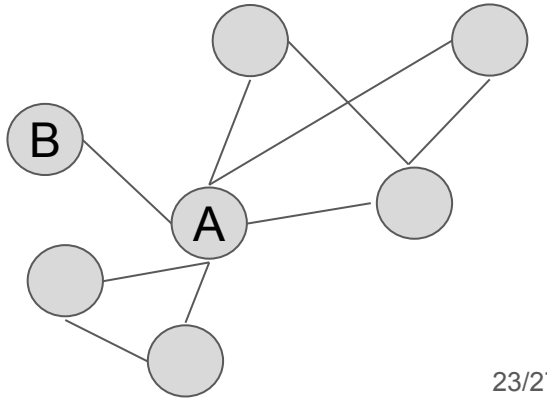
Tumor breast



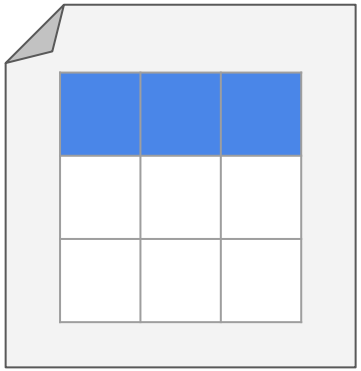
Tools

- **EnGNEt**
- WGCNA
- NetMiner
- DiffCoEx
- RNASeqNet

Redes de coexpresión

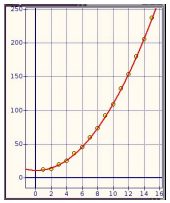


4- Workshop / **Objetivo**

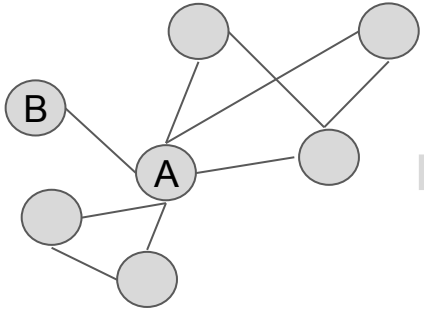


Data mining

Modelos



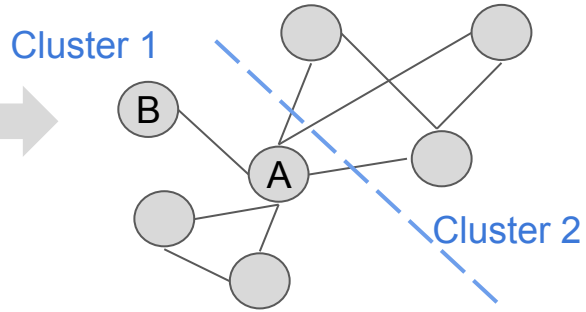
Redes de coexpresión



Tools

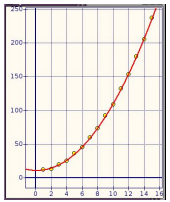
- **Community cluster (GLay)**
- K-means cluster
- MCODE cluster
- Cluster Fizzifer

Clusterización de la red

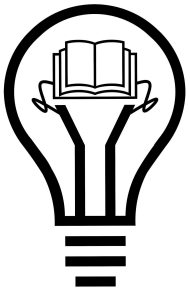


4- Workshop / **Objetivo**

Modelos

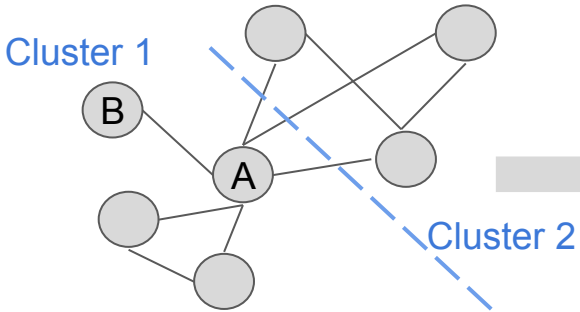


Interpretación y evaluación



Conocimiento

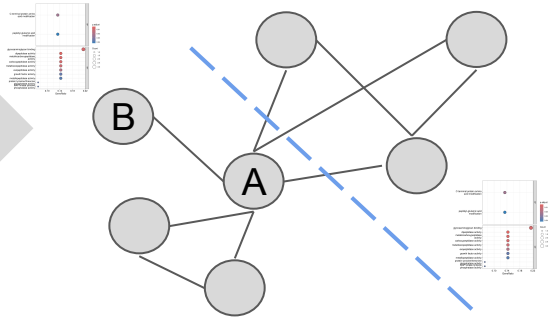
Redes de coexpresión



Tools

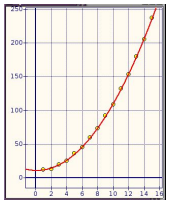
- **clusterProfiler** (GO y KEGG)
- DAVID
- GOstats
- GSEA

Enriquecimiento funcional

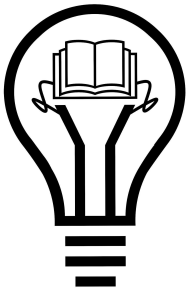


4- Workshop / **Objetivo**

Modelos

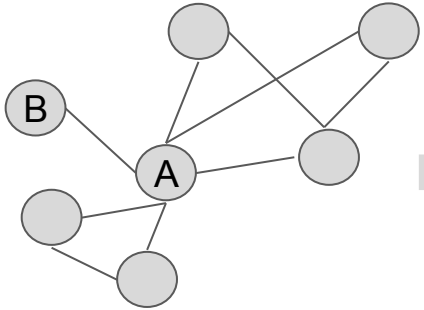


Interpretación y evaluación



Conocimiento

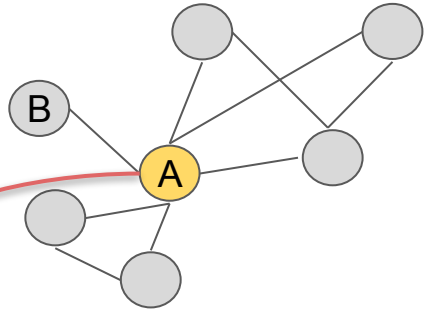
Redes de coexpresión



Tools

- **Degree centrality**
- Betweenness centrality

Enriquecimiento funcional



gen hub

4- Workshop / **.fastq**

https://github.com/mariadsn/GeneExpressionAnalysis_Seminar

Quality control 4.1

4.2 **Data preprocessing**

Inferential statistics 4.3

4.4 **Exploratory analysis**