Pipeline

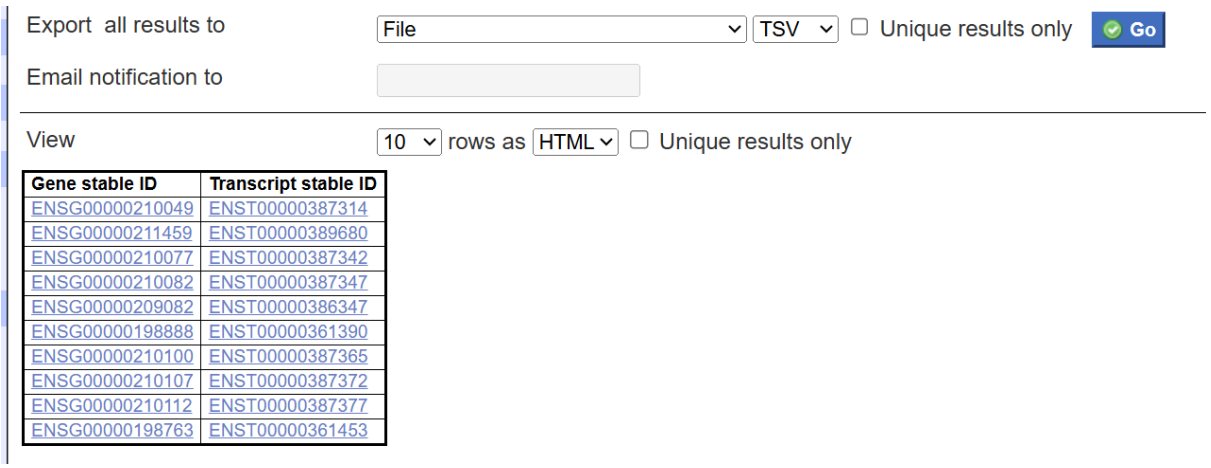
Objetivo: transformar datos crudos de secuenciación en información útil sobre qué genes y RNAs cambian entre sano y enfermo, y ver qué funciones biológicas podrían estar afectadas.

1) Primero vamos a descargar los archivos que se necesitan:

-Sano1_FastaQ-Paired
-Sano2_Fastaq-Paired
-Enfermo1_FastaQ-Paired
-Enfermo2_Fastaq-Paired
-go.obo. (http://geneontology.org/docs/download-ontology)
-goa_human_faf_gz (http://geneontology.org/docs/download-ontology)
-mart_export
-homo_sapiensGRCh38.cdna.all.fa (https://www.ensembl.org/Homo_sapiens/Info/Index)

Para descargar mart_export y prepararlo lo que hice fue

- ir a https://www.ensembl.org/biomart/martview/ef4797c3ff3bc128387eadfae4bc3868
- luego a biomart en el encabezado
- Selecciono ensembl genes 115
- Seleccione human genes (Grch38.p14)
- Ir a atribbutes
- Abro el apartado gene solo se deja el apartado gene stable idtranscrpit stable lo demas se desmarca
- Vamos a result y descargamos el archivo tsv.



Gene stable ID	Transcript stable ID
ENSG00000210049	ENST00000387314
ENSG00000211459	ENST00000389680
ENSG00000210077	ENST00000387342
ENSG00000210082	ENST00000387347
ENSG00000209082	ENST00000386347
ENSG00000198888	ENST00000361390
ENSG00000210100	ENST00000387365
ENSG00000210107	ENST00000387372
ENSG00000210112	ENST00000387377
ENSG00000198763	ENST00000361453

Una vez descargamos el archivo, en un excel cambiamos de lugar las columnas y eliminamos los encabezados si hace falta.

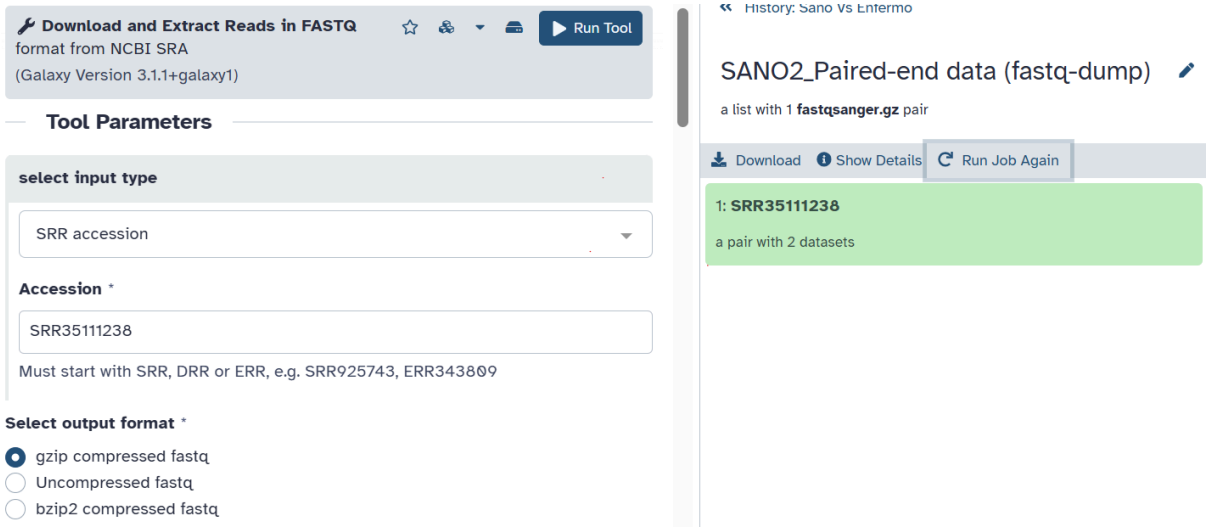
Para cargar los archivos en Galaxy

Vamos **Uploaded data**, cargamos los archivos manualmente descargados

- Entrar en https://usegalaxy.eu/
- Ir a Uploaded
- Elegir archivo local
- Se seleccionan todos los archivos mencionados anteriormente

.

De forma excepcional los FastaQ-Paired se pueden cargar con la herramienta **download and extract reads in FASTQ**



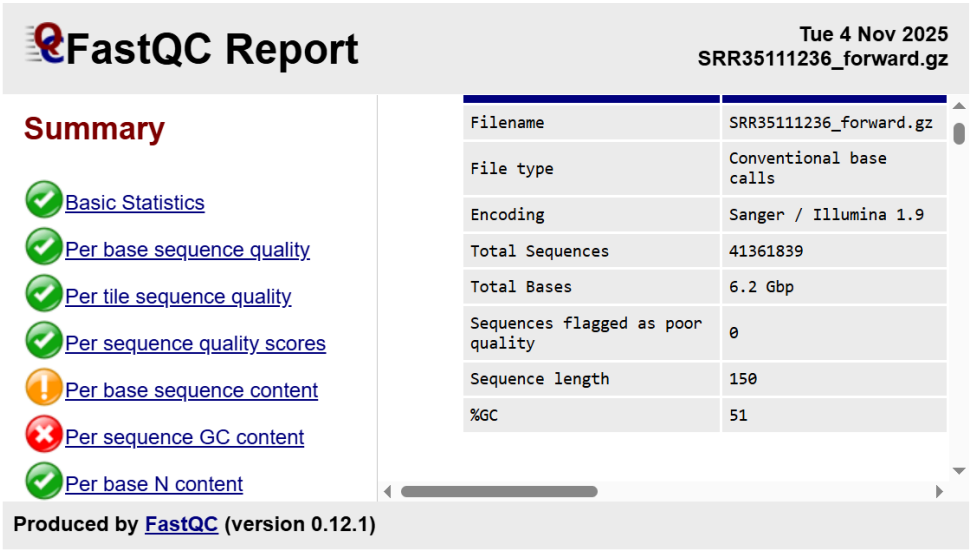
Los SRR para copiar y pegar son:

SANO_1: SRR35111236
SANO_ 2: SRR35111238
ENFERMO_1: SRR35111241
ENFERMO_ 2: SRR35111239

2) Evaluación de calidad de lecturas de mis RNA

Con la herramienta **FastaQC** es una herramienta de control de calidad para datos de secuenciación de alto rendimiento que proporciona un resumen rápido de la calidad de las secuencias brutas

como nuestros fasta son paired, tentremos que hacer 8 fastaqc, uno por cada reverse y forward

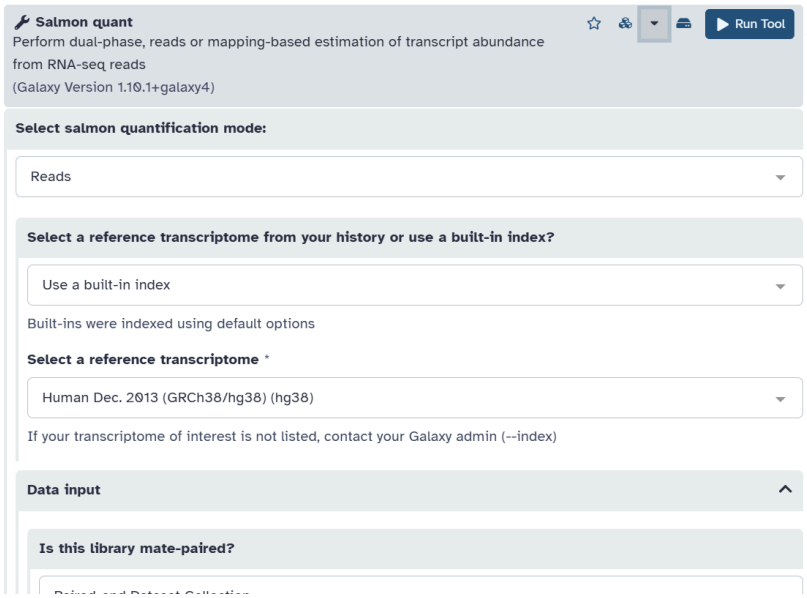


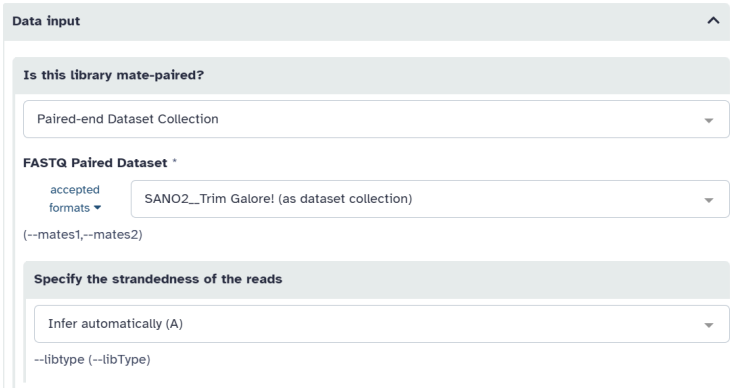
Nos dan en la mayoría los mismos errores de Per base sequence content y per sequence GC content y esto se arregla recortando adaptadores, emplearemos la herramienta **Trim Galore** corrigiendo los sesgos en composición de bases y la distribución anómala del contenido GC.

Ya tenemos los datos limpios

3) **Con la herramienta Salmón Quant** cuantificamos la expresión génica o transcripcional a partir de lecturas de RNA-Seq. Su función principal es estimar **cuántas lecturas** (reads) provienen de cada transcrito o gen, sin necesidad de hacer un alineamiento completo.

Entonces ponemos nuestros archivos de trim galore



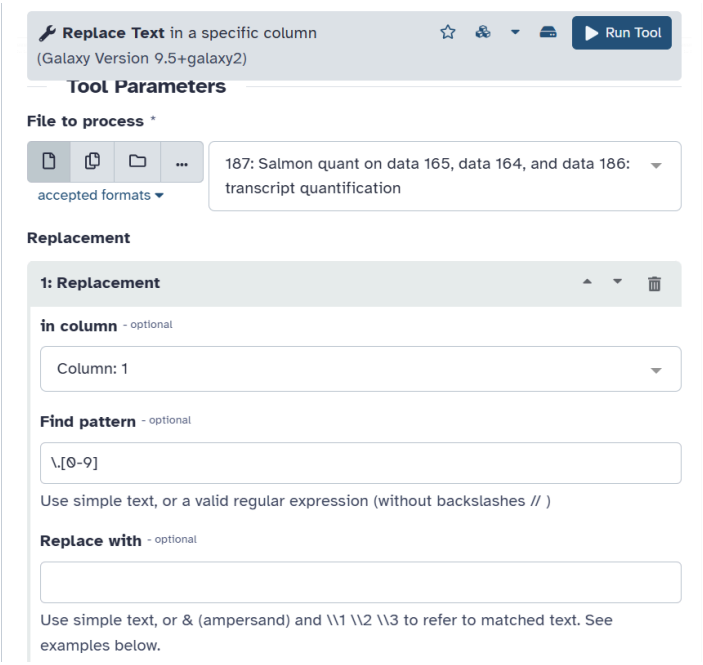


Nota: En lugar de human dec 2013 por predeterminado se usa el archivo homo_sapiens Grch, todas las demás opciones las dejamos en predeterminado

Se cambian los chr por enst **tiene sentido** si vas a cuantificar expresión de transcritos y hacer análisis funcional, porque ENST identifica **transcritos específicos** mientras que `chr` solo indica la posición genómica. Nos quedan asi nuestros Salmon

Column 1	Column 2	Column 3	Column 4	Column 5
Name	Length	EffectiveLength	TPM	NumReads
ENST00000622028.1	353	150.915	0.242076	1.000
ENST00000632585.1	408	198.472	11.407375	62.000
ENST00000632205.1	532	314.485	1049.555058	9038.814
ENST00000632891.1	380	173.807	229.310360	1091.431
ENST00000633250.1	445	232.019	16.555641	105.190
ENST00000631614.2	398	189.585	96.269341	499.802
ENST00000633273.1	409	199.366	325.719562	1778.288
ENST00000632088.2	392	184.308	1772.653497	8946.948
ENST00000632502.1	398	189.585	446.877156	2320.054

En los enst los .(numero) hay que eliminarlos por que sino hay herramientas de galaxy que se confunden para eso usamos o excel o la herramienta **Replace text**

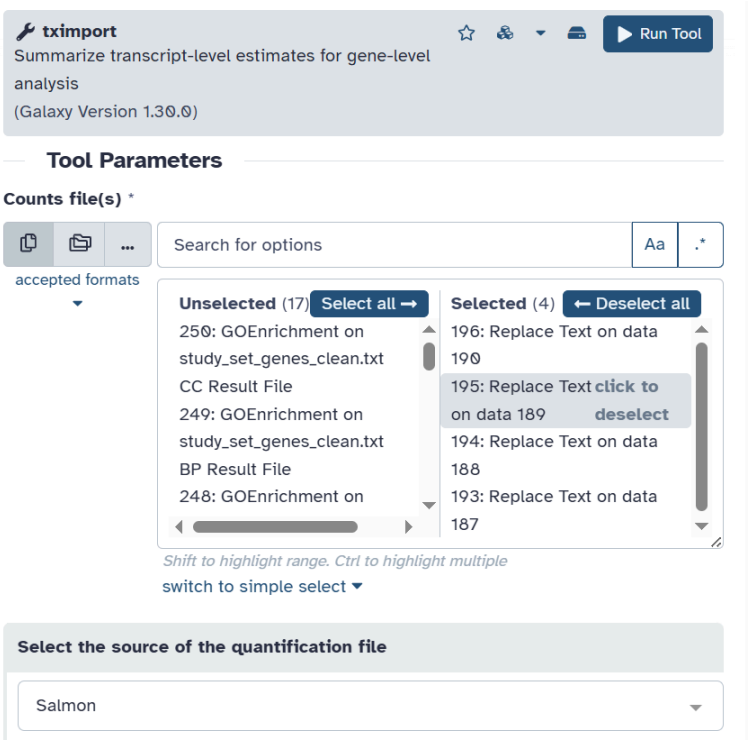


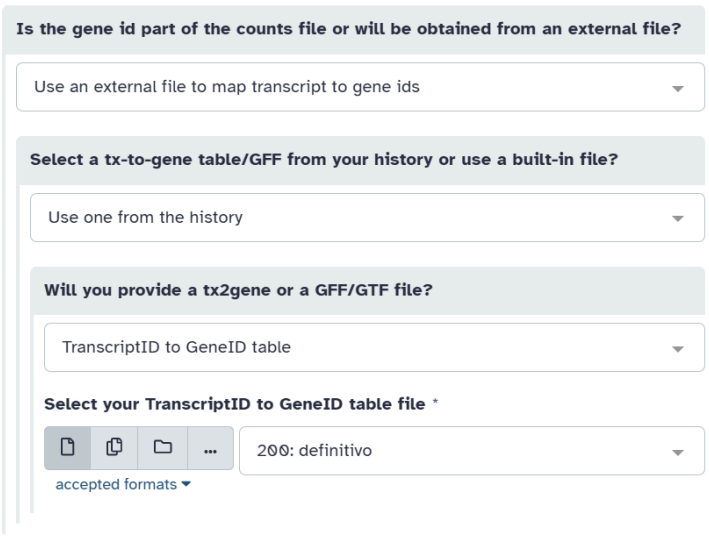
Column 1	Column 2	Column 3	Column 4	Column 5
Name	Length	EffectiveLength	TPM	NumReads
ENST00000622028	353	150.915	0.242076	1.000
ENST00000632585	408	198.472	11.407375	62.000
ENST00000632205	532	314.485	1049.555058	9038.814
ENST00000632891	380	173.807	229.310360	1091.431
ENST00000633250	445	232.019	16.555641	105.190
ENST00000631614	398	189.585	96.269341	499.802
ENST00000633273	409	199.366	325.719562	1778.288
ENST00000632088	392	184.308	1772.653497	8946.948
ENST00000632502	398	189.585	446.877156	2320.054
ENST00000632887	442	229.260	802.443523	5037.898
ENST00000632011	378	172.077	231.131919	1089.153

Resultado final

4) Ponemos todos los salmon quant en tx impot

Tengo un archivo de Salmon con abundancias a nivel de transcrito (ENST). Quiero convertir estas abundancias a nivel de gen (ENSG) utilizando una tabla de transcript-to-gene, y obtener la expresión total por gen usando los valores TPM con la herramienta **`tximport`**





NOTA- EL denifitivo, es el mart_text pero le cambie el nombre luego de intercambiar las columnas

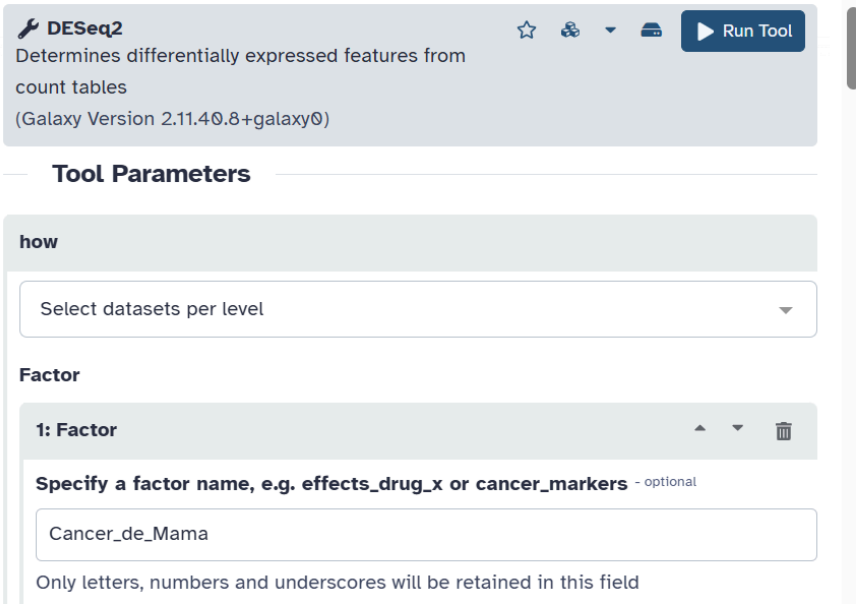
Column 1	Column 2	Column 3	Column 4	Column 5
Replace Text on data 190	Replace Text on data 189	Replace Text on data 188	Replace Text on data 187	
ENSG00000000003	1202.354	1202.354	1301.071	1301.068
ENSG00000000005	2	2	79	79
ENSG00000000419	386.033	385.854	146.915	146.917
ENSG00000000457	581.751	581.746	338.808	338.91
ENSG00000000460	1260.847	1260.846	164.999	165
ENSG00000000938	243.758	243.759	100.993	100.994
ENSG00000000971	2173.629	2173.636	5756.096	5756.16
ENSG00000001036	1794.7	1794.713	1441.336	1441.194
ENSG00000001084	949.048	949.049	701	700.998

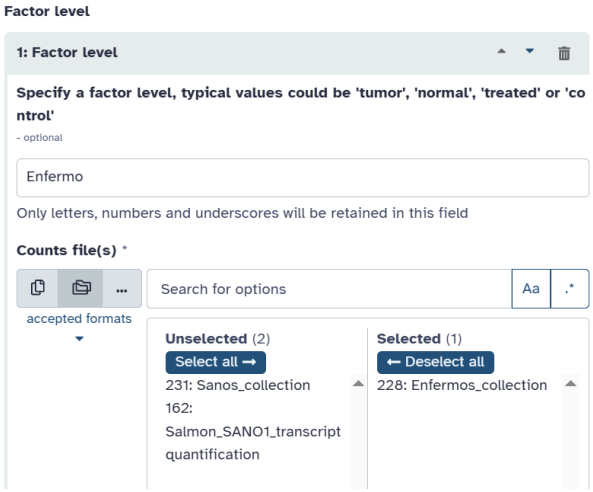
nos queda algo asi.Con excel hice que se viera asi, cambie los encabezados elimine los decimales para que deseq pueda leerlos bien

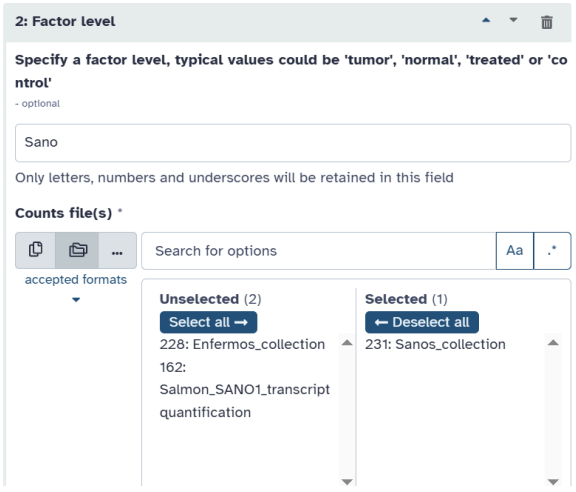
Column 1	Column 2	Column 3	Column 4	Column 5
Gene_ID	enfermo2	enfermo1	sanos2	sanos1
ENSG00000000003	1202354	1202354	1301071	1301068
ENSG00000000005	2	2	79	79
ENSG00000000419	386033	385854	146915	146917
ENSG00000000457	581751	581746	338808	33891
ENSG00000000460	1260847	1260846	164999	165
ENSG00000000938	243758	243759	100993	100994
ENSG00000000971	2173629	2173636	5756096	575616
ENSG00000001036	17947	1794713	1441336	1441194
ENSG00000001084	949048	949049	701	700998
ENSG00000001167	1664356	1664277	621506	621502
ENSG00000001460	76862	76848	15447	154379

5) Deseq2
Ahora para el deseq2 necesito hacer 2 datasets uno sano y el otro enfermo. Con sus archivos consecuentes. Entonces de mi tximport lo modifico en excel creando 4 archivos por separado de solo 2 columnas Gene_ID y los counts

txt1(Gene_Id SANO 1) y (Gene_Id SANO 2) van dentro del dataset_collection_SANO
txt2(Gene_Id ENFERMO1) y (Gene_Id ENFERMO2) van dentro del_dataset_collection ENFERMO


DESeq2
Determines differentially expressed features from count tables
(Galaxy Version 2.11.40.8+galaxy0)
Run Tool
Tool Parameters
how
Select datasets per level
Factor
1: Factor
Specify a factor name, e.g. effects_drug_x or cancer_markers - optional
Cancer_de_Mama
Only letters, numbers and underscores will be retained in this field


Factor level
1: Factor level
Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'
- optional
Enfermo
Only letters, numbers and underscores will be retained in this field
Counts file(s) *
accepted formats
Search for options
Aa
.*
Unselected (2)
Select all
231: Sanos_collection
162: Salmon_SANO1_transcript quantification
Selected (1)
Deselect all
228: Enfermos_collection


2: Factor level
Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'
- optional
Sano
Only letters, numbers and underscores will be retained in this field
Counts file(s) *
accepted formats
Search for options
Aa
.*
Unselected (2)
Select all
228: Enfermos_collection
162: Salmon_SANO1_transcript quantification
Selected (1)
Deselect all
231: Sanos_collection

las demas opciones predeterminado

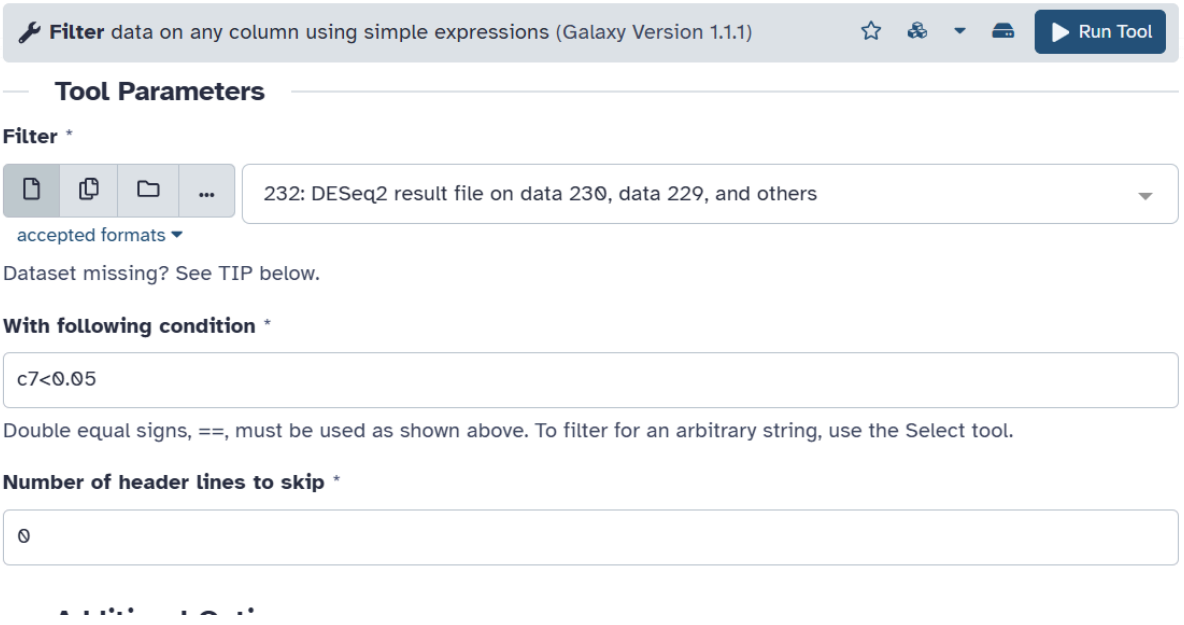
nos debe quedar una tabla de estas columnas

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj

Nos importa el p-value para seleccionar **solo los geness que realmente cambian**. Esa lista “limpia” es la que vamos a usar para GO enrichment, porque así detectamos **procesos biológicos realmente sobre-representados** y no ruido estadístico

6) Filtro

A continuación, extraemos aquellos genes cuya expresión se expresan de manera estadísticamente significativa (DE) seleccionando aquellos cuyo valor de p ajustado sea menor o igual a 0,05.



Cabe señalar que de ese filtro pasamos de tener casi 38633 a 4890 genes

Con excel o la herramienta **cut o coloums** , eliminamos las columnas de tal modo que solo nos quede con Gene_Id y P-value es importante que no tenga encabezados

7) Mapeo de Gene_ID

Antes de realizar el análisis de GO Enrichment, fue necesario asegurarse de que los identificadores de los genes coincidieran con los usados en el archivo GAF de referencia. Idealmente, se hubiera utilizado un archivo GAF que ya contuviera los mismos IDs de Ensembl que los genes del estudio, para evitar pasos de mapeo adicionales.

En este caso, el archivo de “gene production” que se obtuvo estaba en UniProt, mientras que los IDs de los genes diferenciales eran de Ensembl. Por ello, se exportó el archivo desde Galaxy en formato TXT y se procesó en Colab utilizando herramientas informáticas para convertir los IDs de Ensembl a UniProt. Durante este procedimiento, se perdieron 239 filas; sin embargo, esto **no afectó significativamente** el análisis final, ya que la mayoría de los genes diferenciales se conservaron.

Bibliotecas

```
import pandas as pd
import time
!pip install mygene
import mygene
```

Funciones

```
df = pd.read_csv("/content/Galaxy239-[Cut on data 238].tabular",
sep="\t", header=None)

mg = mygene.MyGeneInfo()
mapped_ids = []

block_size = 500
for i in range(0, len(df), block_size):
    block = df[0].iloc[i:i+block_size].tolist()  # solo la columna
GeneID
    res = mg.querymany(block, scopes='ensembl.gene', fields='uniprot,
symbol', species='human')

    # Tomar UniProt si existe, sino HGNC
    for x in res:
        up_id = None
        if 'uniprot' in x and isinstance(x['uniprot'], dict):
            for k in ['Swiss-Prot', 'TrEMBL']:
                if k in x['uniprot']:
                    v = x['uniprot'][k]
                    if isinstance(v, list):
```

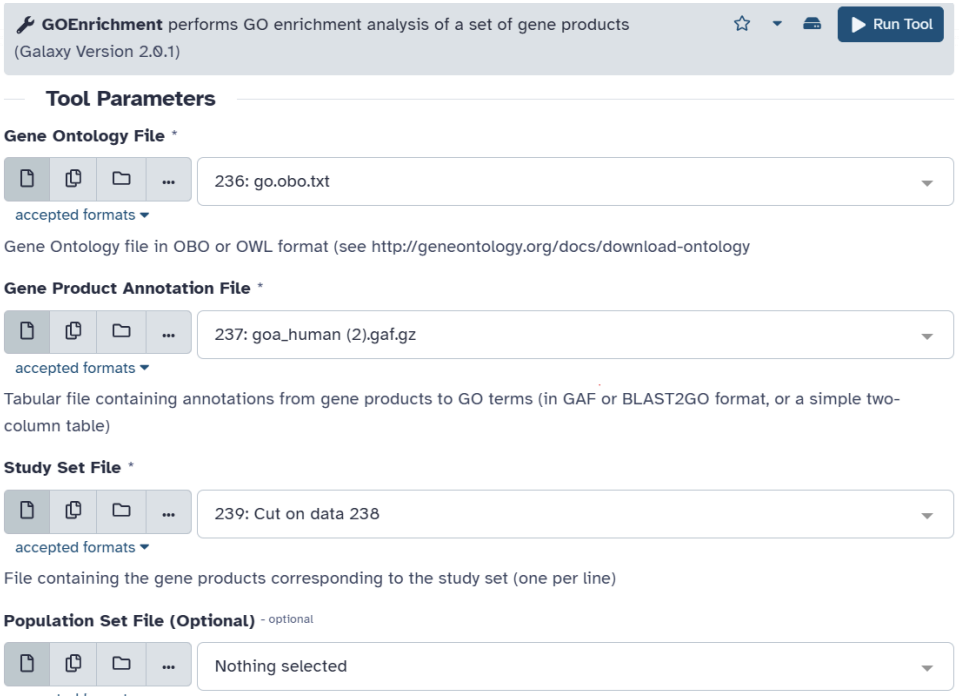
```
                        up_id = v[0]
                    else:
                        up_id = v
                    break
        if not up_id:
            up_id = x.get('symbol', None)
        mapped_ids.append(up_id)

    print(f"Bloque {i}-{i+len(block)} mapeado")
    time.sleep(0.2)  # pequeño descanso para no saturar la API

# Reemplazar primera columna con UniProt/HGNC
df[0] = mapped_ids

# Guardar archivo listo para GO enrichment
df.to_csv("/content/study_set_genes.txt", sep="\t", index=False,
header=False)
print("Archivo listo para GO enrichment guardado como
study_set_genes.txt")
```

Eso nos devolverá el mismo archivo pero cambiando el geneid, al uniprot que requiere el go enrichment. Que va al study set file del GO Enrichment

Column 1	Column 2
P19835	8.08837059334456e-27
P12273	3.41891301131638e-26
IGHG2	5.990370611271561e-25
Q9UKZ9	1.413181029658l3e-21
P31350	1.7887311397883002e-21
Q13296	2.43404479281576e-21
Q96AQ7	2.43404479281576e-21
Q07325	2.91264520711611e-21
O14625	3.93251295267576e-21
P02778	3.949620319727501e-21
P58417	9.58956436874908e-21
O43692	1.04095458576991e-20

8) Go enrichment

El **GO enrichment** (o enriquecimiento de Gene Ontology) es un análisis bioinformático que sirve para identificar qué funciones biológicas, procesos celulares o componentes celulares están **sobrerrepresentados** en un conjunto de genes de interés, en comparación con lo que se esperaría por azar en todo el genoma o transcriptoma.



y finalmente los resultados

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8
GO Term	Study #	Study Freq.	Pop. Freq.	p-value	q-value	name	gene products
GO:0032502	1132	32%	17%	7.05E-124	7.16E-120	developmental process	P30405,Q9Y265,Q9N
GO:0051179	925	26%	14%	1.92E-85	4.88E-82	localization	Q9UMX9,Q9Y265,Q9N
GO:0050896	1123	31%	19%	9.11E-84	1.54E-80	response to stimulus	P30405,Q9Y265,Q9H
GO:0048856	743	21%	11%	8.48E-82	1.23E-78	anatomical structure development	Q92618,P30405,Q9NY
GO:0032501	894	22%	12%	4.75E-71	4.34E-	multicellular	Q9UMX9,Q9H6Q3,Q9

La tabla es extensa, aproximadamente 4800 filas pero pone los más importantes al principio

9) Resultados

Que podemos observar…

Lo importante para entender los resultados es que la columna siete te muestra el valor de referencia estadístico y la columna 3 lo que realmenste saco

Cell population proliferation

Este término indica que los genes asociados a la proliferación celular están sobrerrepresentados en el dataset de cáncer de mama, lo cual es consistente con la biología tumoral, donde las células cancerosas proliferan más rápidamente que las células normales. Aumenta de 1.6% a 4.1%

Cell cycle process

La proporción de genes relacionados con el ciclo celular aumenta de 2.7% a 5.4%. En el contexto del cáncer, estos genes suelen estar sobrerrepresentados, lo que refleja un crecimiento tumoral activo y desregulación de los mecanismos que controlan la división celular.

Apoptosis

Uno de los rasgos clásicos de las células cancerosas es justamente evadir la apoptosis. En tumores de mama, muchos genes que bloquean la muerte celular se expresan más, permitiendo que las células tumorales sigan proliferando a pesar de daños o señales de control. Estos genes están sobrerrepresentados (5.3% observados vs 2.9% esperados), indica que en tu muestra hay una mayor activación de vías que inhiben la apoptosis.

Stress cell

Las células tumorales están sometidas a estrés constante, por ejemplo: Hipoxia (falta de oxígeno) Estrés metabólico por rápido crecimiento. Daño en ADN por mutaciones o radiación interna. Para sobrevivir, el tumor activa vías de respuesta al estrés que permiten que las células continúen proliferando y evadiendo la apoptosis.
Por eso tiene sentido que está sobrerrepresentado: 18% observado vs 10% esperado.

Conclusión del análisis de GO Enrichment

En conjunto, estos hallazgos resaltan patrones biológicos esperables en cáncer de mama, confirmando que el análisis identifica procesos relevantes y consistentes con la literatura científica.