Análisis de RNAseq con genóma de referencia

Elaborado para .

2023-08-10

Tabla de contenidos

Listado de Figuras

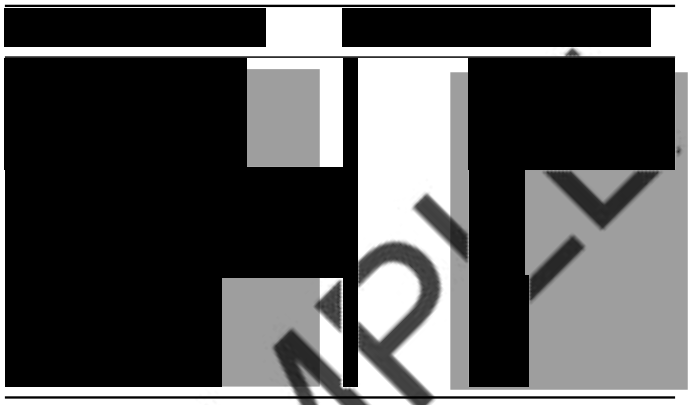
Listado de Tablas

1 Flujo de trabajo

1.1 Descripción del proyecto

Se recibieron lecturas de secuenciación illumina de las siguientes muestras:

Tabla 1.1: Descripción de las muestras

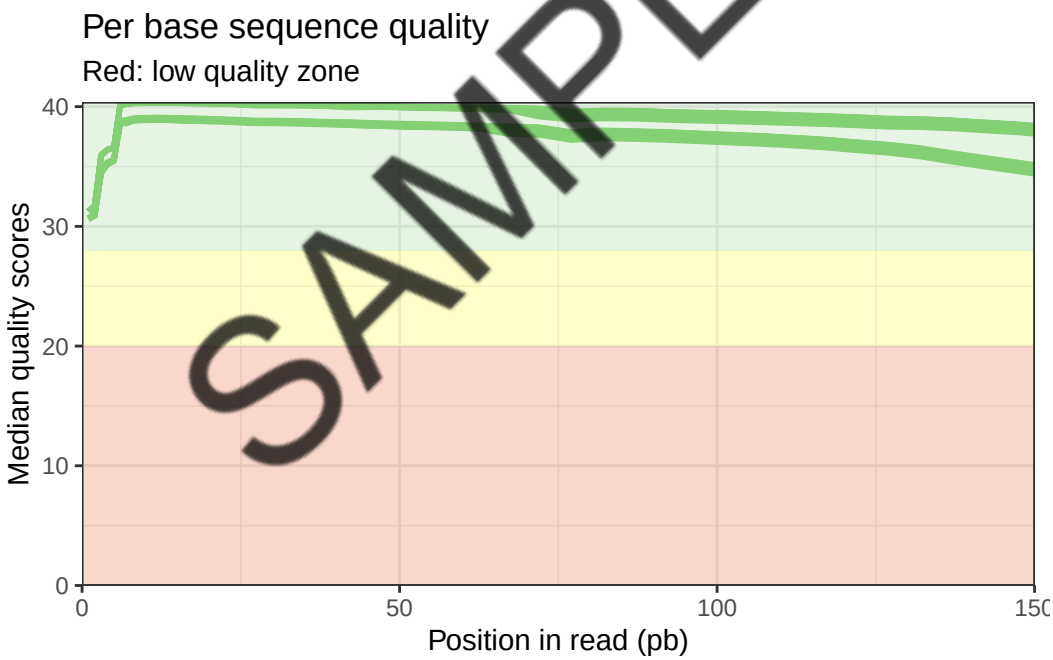


2 Control de calidad de lecturas crudas

Las secuencias con baja calidad, con errores técnicos o con presencia de adaptadores y/o contaminantes pueden afectar los análisis posteriores. La calidad de las lecturas se analizó utilizando el programa `FastQC v0.11.7` en con junto con `MultiQC v1.10.1`.

2.1 Histogramas de calidad de las secuencias

El eje **y** del gráfico muestra los valores de calidad de las secuencias. Mientras mas alto el valor, mas alta la probabilidad de que la lectura fue correcta. El color del fondo divide el gráfico en lecturas con calidad alta en verde, calidad aceptable en amarillo, y baja calidad en rojo. La calidad de las lecturas disminuirá en la mayoría de las plataformas conforme avanza la corrida, por lo que es común observar una disminución de la calidad hacia el final de la lectura.

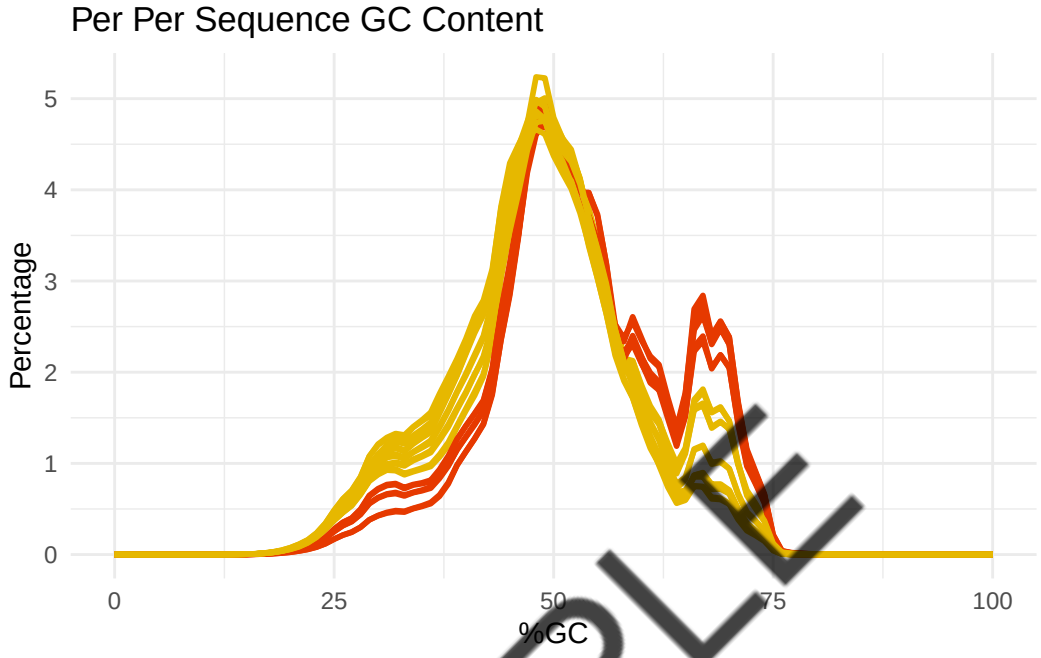


2.2 Porcentaje de GC por secuencia

El gráfico muestra el porcentaje de GC a lo largo de toda la lectura y lo contrasta con una distribución normal de GC modelada.

En una librería aleatoria normal, se esperaría tener una distribución normal del contenido de GC, donde el pico corresponde al contenido de GC del genoma subyacente. Dado que no se conoce el contenido de GC del

genoma, el contenido de GC modal se calcula a partir de los datos observados y se utiliza para construir una distribución de referencia. Una distribución de forma inusual podría indicar una biblioteca contaminada o algún otro tipo de subconjunto sesgado.



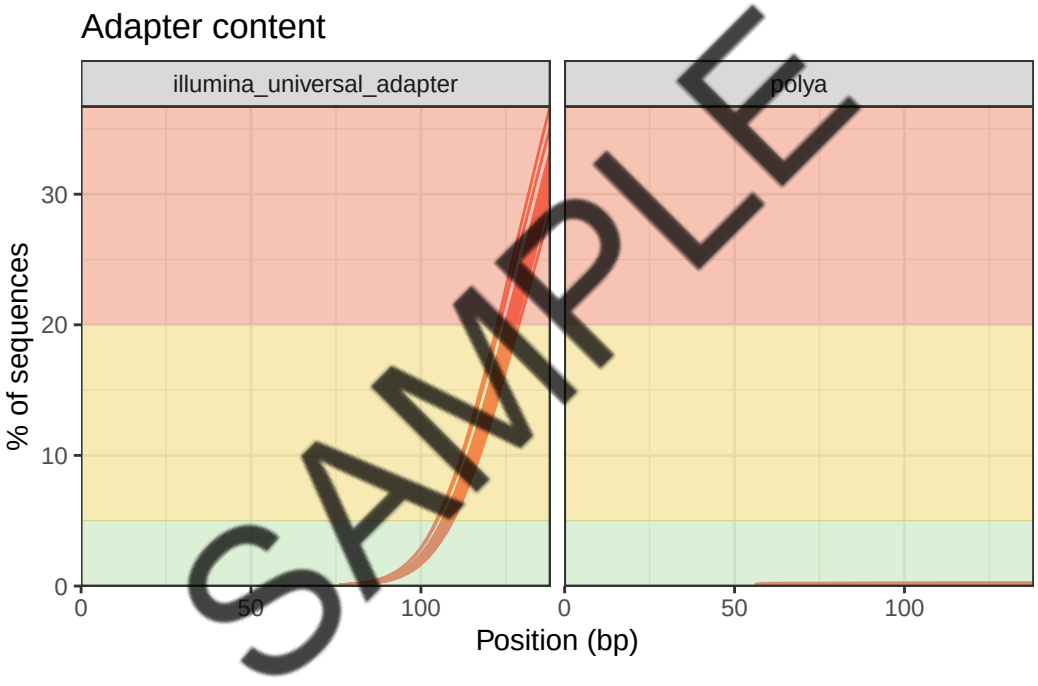
Advertencia

En estos datos se observan algunas librerías con un pico adicional alrededor del 70% de GC, lo que nos puede estar indicando contaminación de las muestras.

2.3 Presencia de adaptadores

Solamente las muestras con un porcentaje mayor al 0.1 % de contaminación con adaptadores se muestra.

Para ver el reporte detallado haz click aquí



3 Limpieza y filtrado de lecturas

Dado la presencia de adaptadores y la presencia de picos adicionales en el porcentaje de GC, las lecturas se procesaron utilizando el programa `Trimmomatic-0.36` (Bolger, Lohse, y Usadel 2014).

Trimmomatic

Trimmomatic es una herramienta que permite remover o cortar lecturas con baja calidad, así como eliminar adaptadores de las lecturas. Los parámetros utilizados fueron:

- ILLUMINACLIP: TruSeq3-PE-2.fa:2:30:10:2:True
- LEADING:5
- TRAILING:5
- SLIDINGWINDOW:4:15
- MINLEN:36

Contaminación ARNr

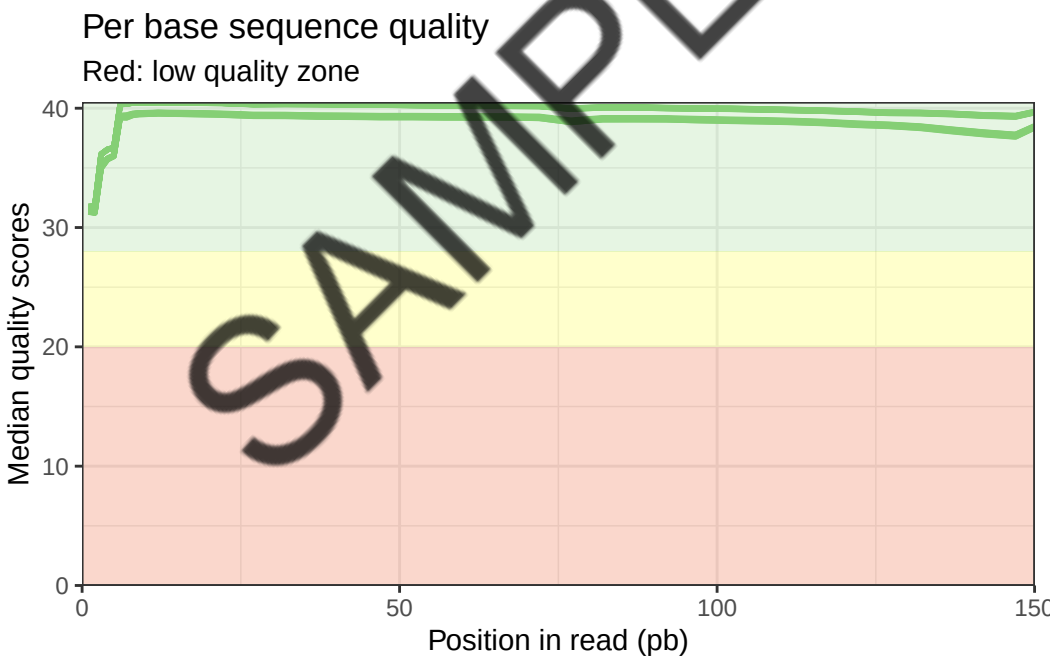
La distribución de picos en el %GC sugiere la contaminación con ARNr. Para tratar de minizar su efecto se realizó un filtrado con el programa SortMeRNA. Sin embargo, tras este proceso se perdió un imporntante porcentage de lecturas por lo que se decidió realizar todos los análisis subsiguientes con las lecturas completas.

4 Control de calidad de las lecturas procesadas

Las secuencias con baja calidad, con errores técnicos o con presencia de adaptadores y/o contaminantes pueden afectar los análisis posteriores. La calidad de las lecturas se analizó utilizando el programa `FastQC v0.11.7` en con junto con `MultiQC v1.10.1`.

4.1 Histogramas de calidad de las secuencias

El eje **y** del gráfico muestra los valores de calidad de las secuencias. Mientras mas alto el valor, mas alta la probabilidad de que la lectura fue correcta. El color del fondo divide el gráfico en lecturas con calidad alta en verde, calidad aceptable en amarillo, y baja calidad en rojo. La calidad de las lecturas disminuirá en la mayoría de las plataformas conforme avanza la corrida, por lo que es común observar una disminución de la calidad hacia el final de la lectura.

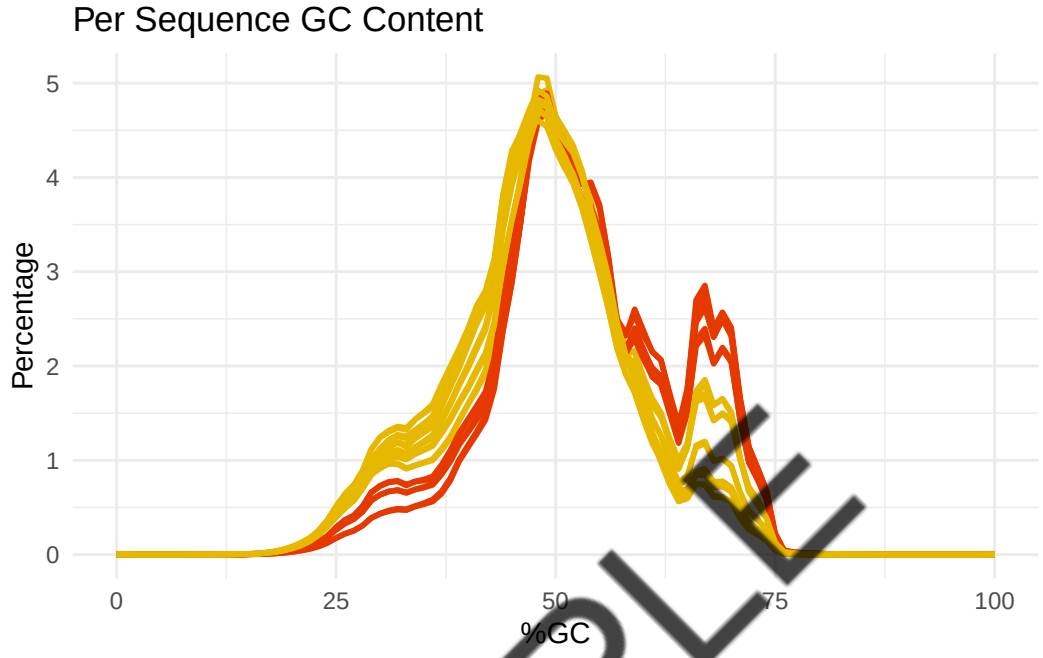


4.2 Porcentaje de GC por secuencia

El gráfico muestra el porcentaje de GC a lo largo de toda la lectura y lo contrasta con una distribución normal de GC modelada.

En una librería aleatoria normal, se esperaría tener una distribución normal del contenido de GC, donde el pico corresponde al contenido de GC del genoma subyacente. Dado que no se conoce el contenido de GC del

genoma, el contenido de GC modal se calcula a partir de los datos observados y se utiliza para construir una distribución de referencia. Una distribución de forma inusual podría indicar una biblioteca contaminada o algún otro tipo de subconjunto sesgado.



Advertencia

En estos datos se observan algunas librerías con un pico adicional alrededor del 70% de GC, lo que nos puede estar indicando contaminación de las muestras.

4.3 Presencia de adaptadores

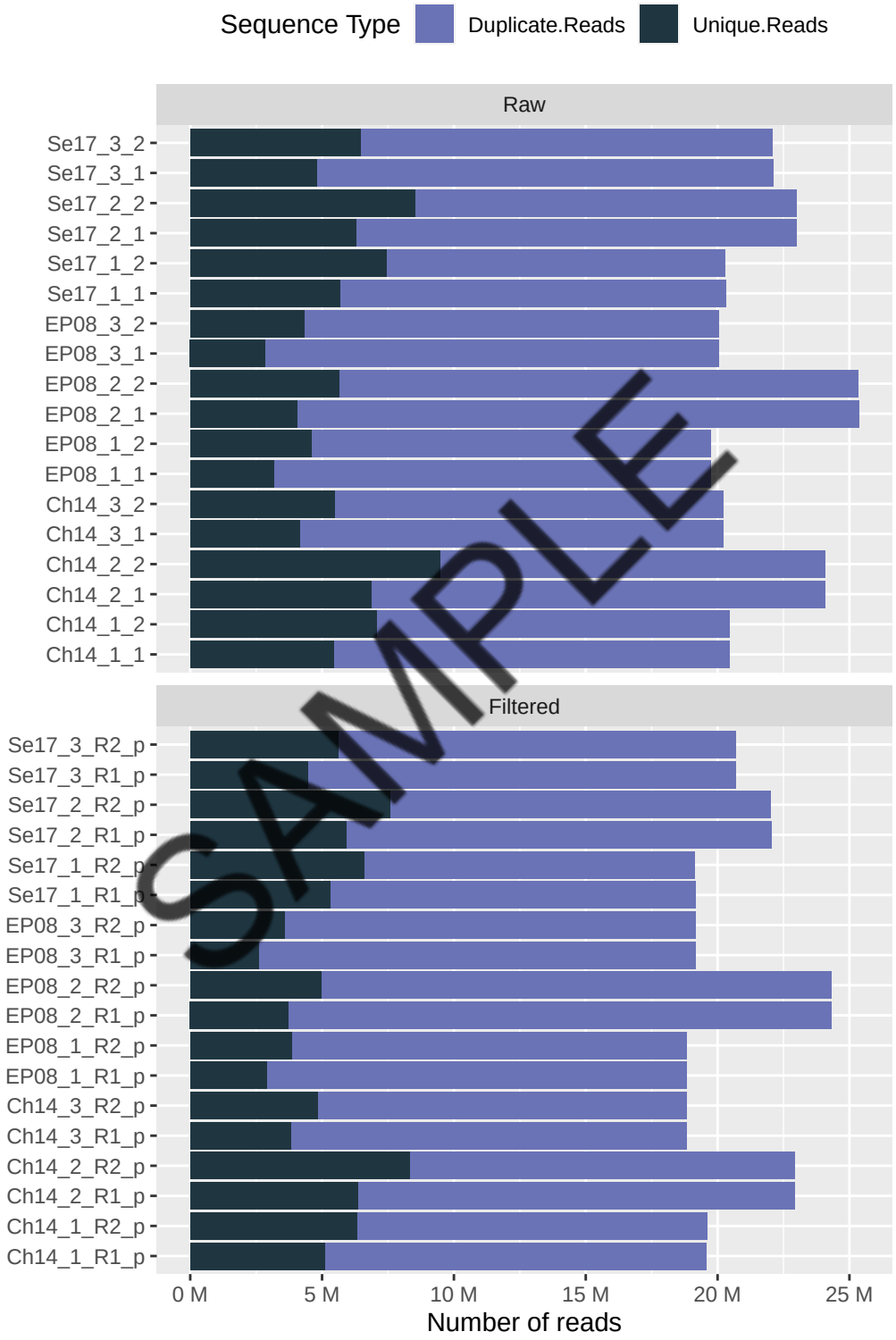
Solamente las muestras con un porcentaje mayor al 0.1 % de contaminación con adaptadores se muestra.

```
[1] "No adapters found"
```

No se encontraron adaptadores residuales

Para ver el reporte detallado haz click aquí

4.4 Comparativo de lecturas crudas y procesadas



5 Mapeo al genoma de referencia

Se utilizó el genoma de referencía GCA902806645v1 el cual tiene las siguientes caracteristicas:

Tabla 5.1: Estadisticos del genoma de referencia

	GenBank
Tamaño de genoma	647.9 Mb
Gaps entre scaffolds	0
Número de cromosomas	10
Número de scaffolds	236
Scaffold N50	58.5 Mb
Scaffold L50	5
Número de contigs	783
Contig N50	1.6 Mb
Contig L50	116
%GC	33
Nivel de ensamble	Cromosoma
Cobertura de genóma	70X

El mapeo de las lecturas al genoma de referencia fue realizado con el software Hisat2 (Kim et al. 2019) el cual esta diseñado específicamente para lecturas de RNA-seq y tiene un mejor desempeño que otros alineadores en cuanto a velocidad y tasa de alineamiento (p. ej. Musich, Cadle-Davidson, y Osier 2021).

5.1 Resultados de Hisat2

Para ver el reporte detallado haz click aquí

Tabla 5.2: Resultados de mapeo de lecturas al genoma de referencia

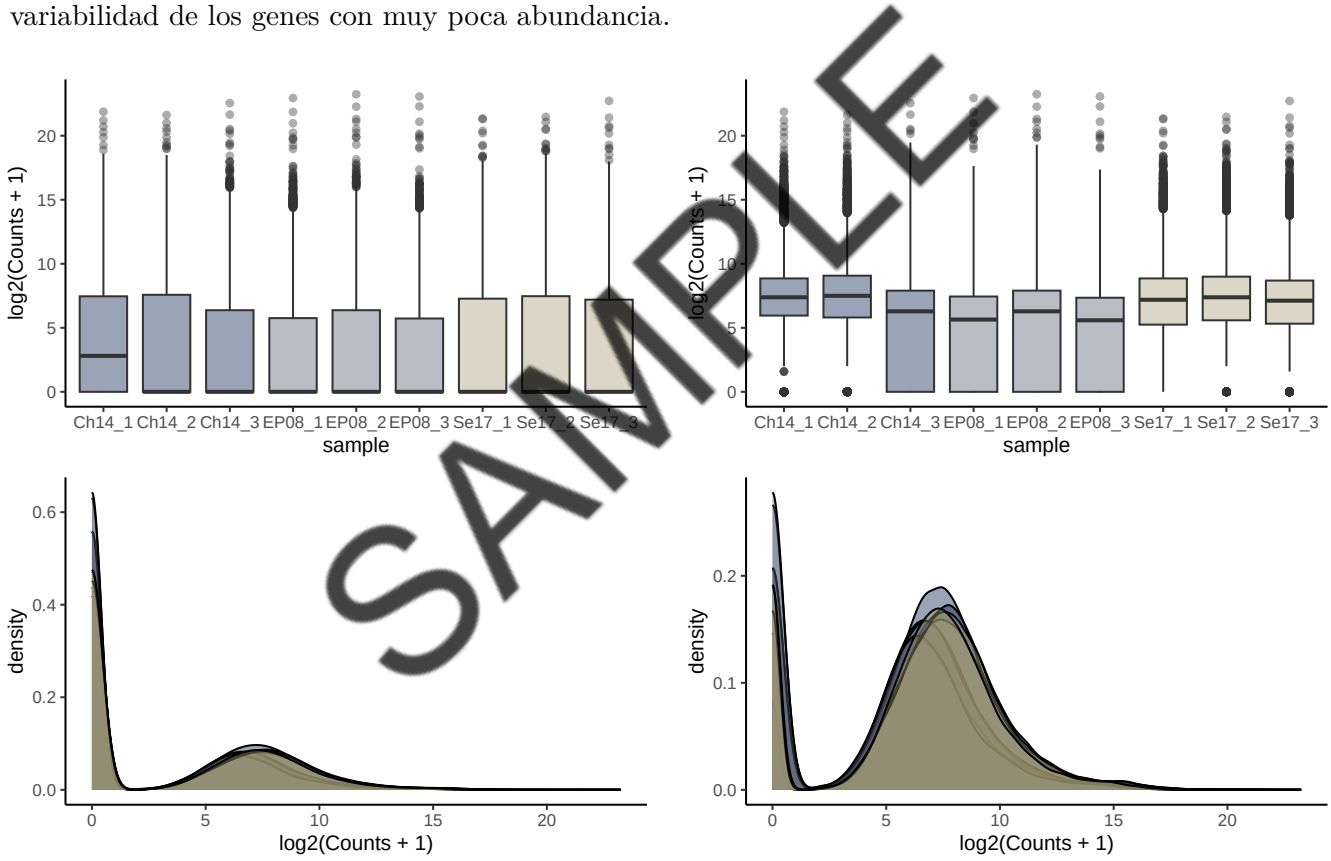
Sample	% Error Rate	M Reads Mapped	% Mapped	% Proper Pairs	% MQ0	M Total Seqs
Ch14_1	0.59	30.83	78.66	73.66	0.89	39.19
Ch14_2	0.67	34.89	76.06	70.71	0.96	45.87
Ch14_3	0.45	30.49	80.94	76.64	0.72	37.67
EP08_1	0.35	33.05	87.75	84.01	0.60	37.66
EP08_2	0.35	42.52	87.41	83.57	0.61	48.64
EP08_3	0.31	33.79	88.14	84.31	0.57	38.33
Se17_1	0.63	29.29	76.47	71.32	0.82	38.30
Se17_2	0.65	32.88	74.62	69.57	0.95	44.06
Se17_3	0.49	33.92	82.02	77.54	0.79	41.36

6 Estimación de la abundancia y Expresión diferencial

La cuantificación de la abundancia se realizó utilizando **Stringtie v2.2.1** siguiendo el protocolo descrito en (Pertea et al. 2016)

6.1 Pre-filtrado de los genes con poca abundancia

Se realizó un filtrado de genes con poca abundancia. Esto permite reducir el tamaño de los archivos y reduce la variabilidad de los genes con muy poca abundancia.



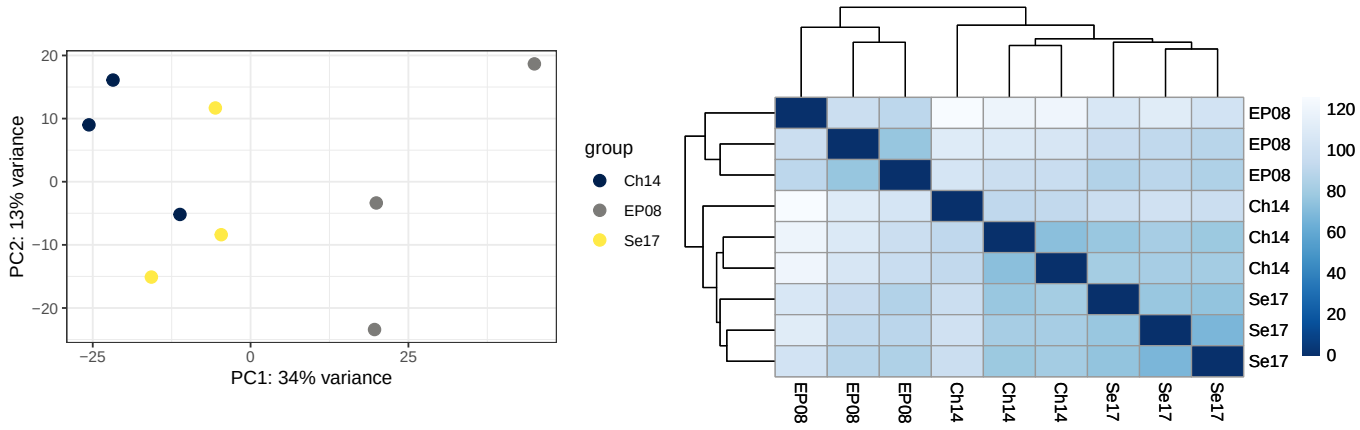
(a) Valores crudos

(b) Valores filtrados

Figura 6.1: Distribución de los valores de abundancia (log2) entre muestras

6.2 Control de calidad de las réplicas biológicas

Se realizó un PCA y una matriz de distancia para visualizar la distribución de las réplicas biológicas dentro de cada tratamiento.



(a) PCA con datos transformados (rlog)

(b) Análisis de distancia.

Figura 6.2: Agrupamiento de las replicas biológicas

6.3 Expresión diferencial con DESeq2

Se utilizó el programa **DESEq2 v.1.40.1** (Love, Huber, y Anders 2014) el cual se basa en una distribución binomial negativa (Gamma-Poisson) para analizar la expresión diferencial de genes, siguiendo el protocolo recomendado en la viñeta [^6.3.abundance_stringtie-1]: http://bioconductor.org/packages/devel/bioc/vignettes/DES expression-analysis

> **Nota**
>
> Para el análisis de expresión diferencial se evaluó la abundancia de cada tratamiento con respecto al promedio de los otros dos tratamientos. Es decir:
>
> - Se = Se - (EP + Ch)/2
> - Ep = EP - (Se + Ch)/2
> - Ch = Ch - (EP + Se)/2

```
[1] 1356
```

> **Importante**
>
> Para cada contraste, se consideró como diferencialmente expresados cuando un gen tuvo un cambio en la expresión (log2Fold change) > |0.5| y un valor P ajustado < 0.05)

6.4 Resultados del análisis de expresión diferencial

A continuación, e muestra un mapa de calor con todos los genes que se expresaron diferencialmente (P < 0.05).

Para ver el reporte detallado haz click aquí

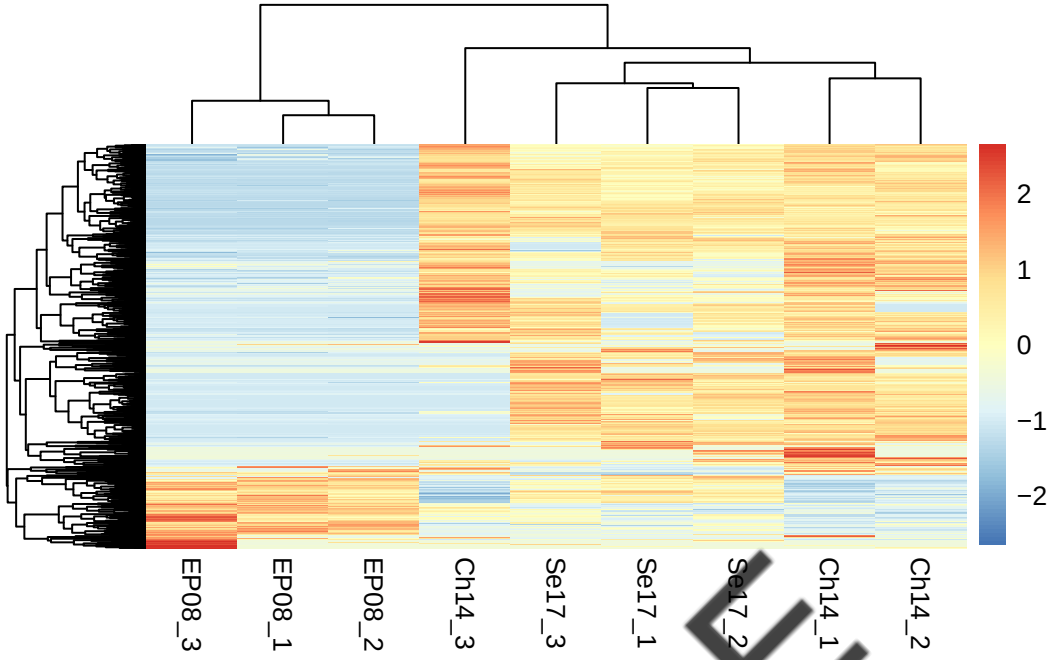
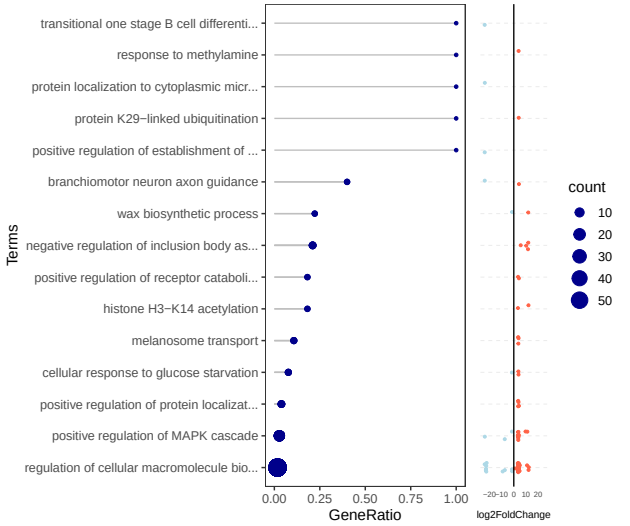


Figura 6.3: Heatmap con genes expresados diferencialmente

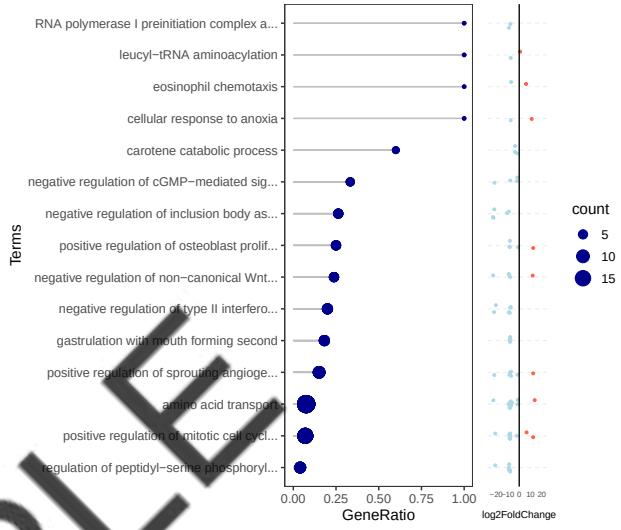
6.5 Enriquecimiento funcional

Se realizó un análisis de enriquecimiento funcional utilizando los términos de Gene Ontology (GO) con el programa TopGo (Alexa y Rahnenführer 2009).

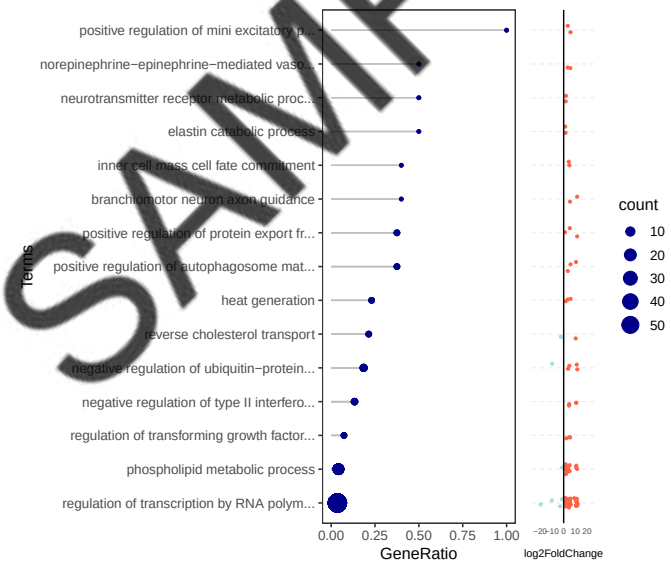
> **!** Importante
>
> Se utilizó el método de Sobre-representación (*Over-Representation Analysis*; ORA) el cual requiere un universo (*background*) y una lista de genes de interés. En este caso, el universo consiste en todos los genes anotados y la lista de genes corresponde a los genes expresados diferencialmente entre cada contraste



(a) Procesos Biológicos enriquecido[redacted]



(b) Procesos Biológicos enriquecido[redacted]



(c) Procesos Biológicos enriquecido[redacted]

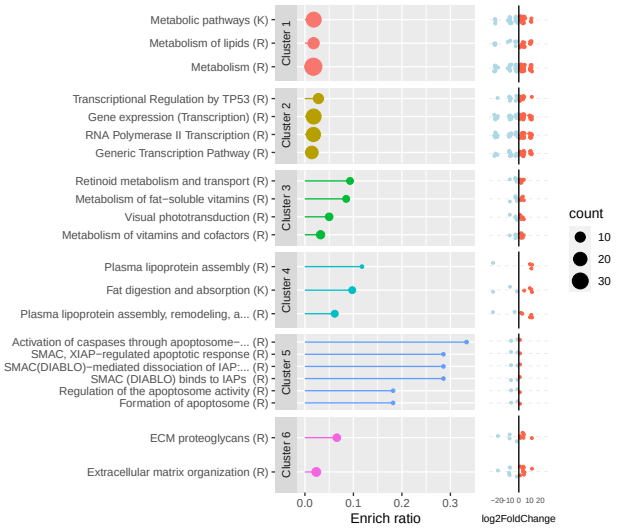
Figura 6.4: Enriquecimiento de Procesos Biológicos de Gene Ontology (GO)

7 Rutas metabolicas enriquecidas

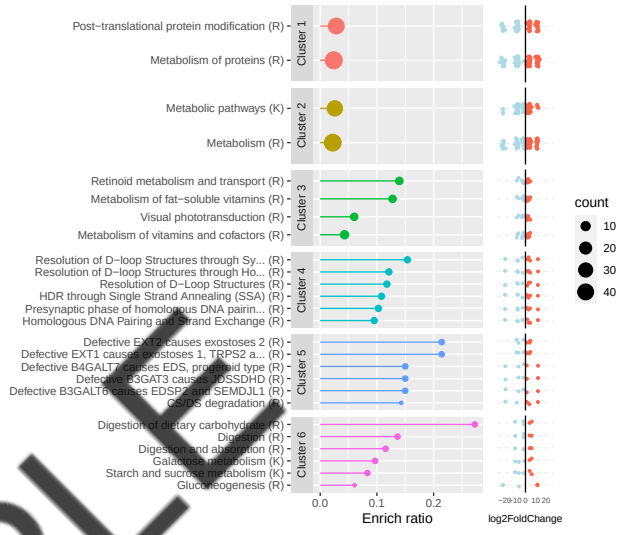
A partir de la lista de genes expresados diferencialmente para cada localidad, se realizó el enriquecimiento de rutas metabolicas utilizando el programa en linea Kobas V3.0 (Bu et al. 2021).

Para el análisis se realizaron los siguientes pasos:

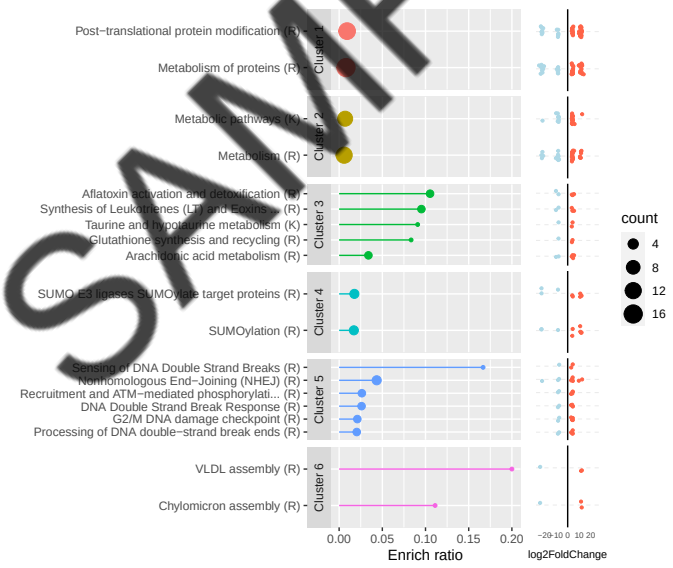
- Se utilizó el modulo de enriquecimiento y se utilizarón los nombres de los genes en formato `Gene symbol`.
- Se utilizaron las bases de datos `KEGG pathway` y `Reactome` de *Homo sapiens* ya que la base de *Crassostrea gigas* arroja muy pocas rutas metabólicas y carece de resultados de Reactome
- Finalmente ese utilizó el modulo de agrupamiento (*Clustering*) de Kobas para agrupar aquellas rutas metabolicas similares entre ellas



(a) Rutas metabólicas enriquecidas



(b) Rutas metabólicas enriquecidas en



(c) Rutas metabólicas enriquecidas en

Figura 7.1: Enriquecimiento de rutas metabolicas

8 Patrones de expresión de genes de interés

A partir de los procesos biológicos y las rutas metabólicas enriquecidas, se seleccionaron ciertos procesos que potencialmente tengan relación con las diferencias encontradas con los niveles de contaminantes encontrados en cada sitio.

Se seleccionaron los siguientes términos:

- *Phospholipid metabolic process (GO:0006644)*: The chemical reactions and pathways involving phospholipids, any lipid containing phosphoric acid as a mono- or diester.
- *Positive regulation of MAPK cascade (GO:0043410)*: Any process that activates or increases the frequency, rate or extent of signal transduction mediated by the MAPK cascade
- *CS/DS degradation (R-HSA-2024101)*: degradation of chondroitin sulfate and dermatan sulfate. Complete degradation of glycoproteins is required to avoid build up of glycosaminoglycan fragments which can cause lysosomal storage diseases
- *ECM-receptor interaction (hsa04512)*: The extracellular matrix (ECM) consists of a complex mixture of structural and functional macromolecules and serves an important role in tissue and organ morphogenesis and in the maintenance of cell and tissue structure and function. Specific interactions between cells and the ECM are mediated by transmembrane molecules, mainly integrins and perhaps also proteoglycans, CD36, or other cell-surface-associated components. These interactions lead to a direct or indirect control of cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis. In addition, integrins function as mechanoreceptors and provide a force-transmitting physical link between the ECM and the cytoskeleton. Integrins are a family of glycosylated, heterodimeric transmembrane adhesion receptors that consist of noncovalently bound alpha- and beta-subunits.
- *Activation of caspases through apoptosome-mediated cleavage (R-HSA-111459)*: Procaspase-3 and 7 are cleaved by the apoptosome
- *Digestion of dietary carbohydrate (R-HSA-189085)*: Carbohydrate is a major component of the human diet, and includes starch (amylose and amylopectin) and disaccharides such as sucrose, lactose, maltose and, in small amounts, trehalose. The digestion of starch begins with the action of amylase enzymes secreted in the saliva and small intestine, which convert it to maltotriose, maltose, limit dextrins, and some glucose. Digestion of the limit dextrins and disaccharides, both dietary and starch-derived, to monosaccharides - glucose, galactose, and fructose - is accomplished by enzymes located on the luminal surfaces of enterocytes lining the microvilli of the small intestine (Van Beers et al. 1995).
- *Cytokine Signaling in Immune system (R-HSA-1280215)*: Cytokines are small proteins that regulate and mediate immunity, inflammation, and hematopoiesis. They are secreted in response to immune stimuli, and usually act briefly, locally, at very low concentrations. Cytokines bind to specific membrane receptors, which then signal the cell via second messengers, to regulate cellular activity.

! Importante

Dado el análisis de expresión diferencial realizado, un gen se consideró como sobre-expresado en cada localidad cuando su valor de expresión se incrementa con respecto al promedio de las otras dos localidades. De igual manera, un gen se considera como sub-expresado cuando su valor en cada localidad disminuye con respecto al promedio de las otras dos localidades.

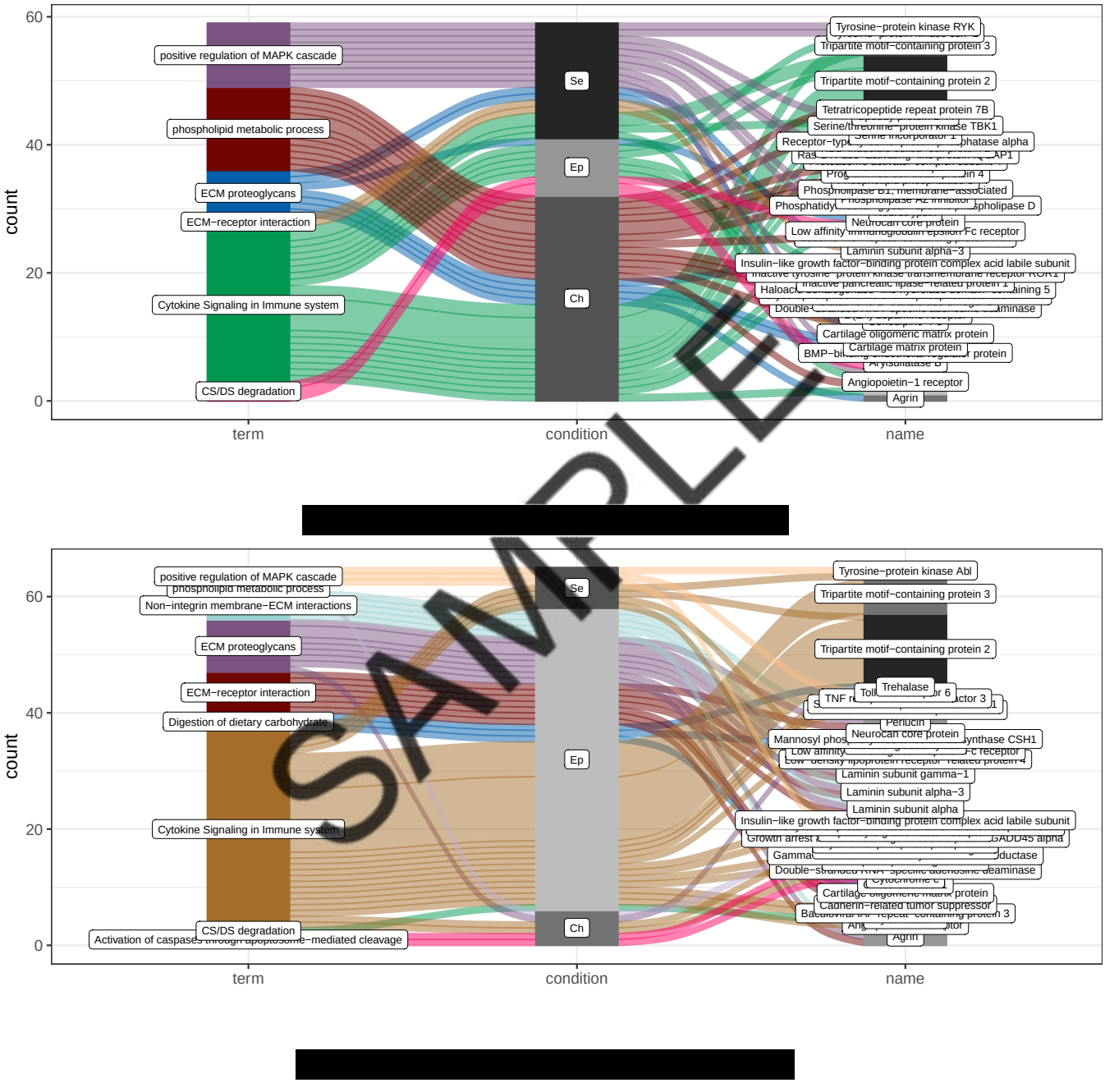


Figura 8.1: Relación de genes expresados diferencialmente en procesos y rutas de interés

9 Referencias

Alexa, Adrian, y Jörg Rahnenführer. 2009. «Gene set enrichment analysis with topGO». *Bioconductor Improv* 27: 1-26.

Bolger, Anthony M, Marc Lohse, y Bjoern Usadel. 2014. «Trimmomatic: a flexible trimmer for Illumina sequence data». *Bioinformatics* 30 (15): 2114-20.

Bu, Dechao, Haitao Luo, Peipei Huo, Zhihao Wang, Shan Zhang, Zihao He, Yang Wu, et al. 2021. «KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis». *Nucleic Acids Research* 49 (W1): W317-25. https://doi.org/10.1093/nar/gkab447.

Kim, Daehwan, Joseph M Paggi, Chanhee Park, Christopher Bennett, y Steven L Salzberg. 2019. «Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype». *Nature biotechnology* 37 (8): 907-15.

Love, Michael I, Wolfgang Huber, y Simon Anders. 2014. «Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2». *Genome biology* 15 (12): 1-21.

Musich, Ryan, Lance Cadle-Davidson, y Michael V Osier. 2021. «Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider». *Frontiers in Plant Science*, 692.

Pertea, Mihaela, Daehwan Kim, Geo M Pertea, Jeffrey T Leek, y Steven L Salzberg. 2016. «Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown». *Nature protocols* 11 (9): 1650-67.