

Análisis de expresión diferencial en *Prunus persica* utilizando DESeq2 y Expression Atlas.

Presentado por: Mariana Gutiérrez Tamayo

1. Descripción del estudio original.

Para el desarrollo de esta entrega, se seleccionó el artículo "*Transcriptional Responses in Root and Leaf of *Prunus persica* under Drought Stress Using RNA Sequencing*" propuesto por Ksouri et al. (2016). El estudio trata sobre la respuesta molecular y genética de *Prunus persica* ante el estrés por sequía y es un experimento de estrés abiótico controlado, utilizando un diseño de comparación de grupos. Dado que la sequía representa la principal amenaza ambiental para el sector agrícola global, superando a otros factores en la disminución del rendimiento de los cultivos, los autores se centraron en caracterizar la respuesta molecular de *P. persica* a 16 días de estrés hídrico utilizando la secuenciación de ARN (RNA-seq). Así entonces, su objetivo principal fue identificar y caracterizar los genes que se activan o desactivan de manera diferencial cuando la planta experimenta condiciones de escasez de agua, comparando las reacciones de las raíces y las hojas.

De acuerdo con (Arús et al., 2012) *Prunus persica* es el nombre científico que recibe el árbol frutal comúnmente conocido como duraznero o melocotonero. Genéticamente, se caracteriza por ser una especie diploide ($2n=2x=16$). Este árbol pertenece a la familia *Rosaceae* y es ampliamente cultivado gracias a su preciado fruto, el durazno o melocotón. Además, es el tercer árbol frutal con mayor tasa de cultivo a nivel mundial, por detrás de los manzanos y perales. Su relevancia radica en la alta concentración de compuestos bioactivos, como carotenoides y polifenoles que brindan un alto valor nutricional y prometedoras propiedades funcionales.

De esta manera la pregunta de investigación experimental tomada del artículo podría ser: ¿Cuáles son los genes y las vías metabólicas que se regulan diferencialmente (DEGs) en las raíces y las hojas de *Prunus persica* cuando se somete a un periodo de estrés hídrico controlado?

2. Diseño experimental.

Se trabajó con plantas de duraznero injertadas, analizando dos tejidos. El organismo de estudio fue *Prunus persica* L. Batsch, utilizando un injerto de la variedad *Catherina* sobre el portainjerto GF677, el cual fue seleccionado por su alta tolerancia a la sequía. El diseño experimental incluyó la comparación de dos grupos, el primero con plantas bien regadas (control) y el segundo, plantas con condiciones difíciles de acceso al agua (sequía) a las que se les administró el 80% de la evapotranspiración diaria, durante 16 días.

Se recolectaron muestras de dos tejidos cruciales para la respuesta a la sequía, la raíz (portainjerto GF677) y las hojas (variedad *Catherina*), con tres réplicas biológicas para cada tejido y condición, obteniendo así 12 librerías de ARN secuenciadas. La técnica principal utilizada fue la secuenciación de ARN de alto rendimiento (RNA-seq), lo que permitió cuantificar la expresión de genes en todo el genoma; mientras que los resultados se validaron mediante la reacción en cadena de la polimerasa con transcripción inversa cuantitativa (RT-qPCR). De esta manera, en la **Figura 1** se presenta el flujo de trabajo recopilado del artículo seleccionado, donde se ilustra el diseño experimental y las herramientas bioinformáticas utilizadas.

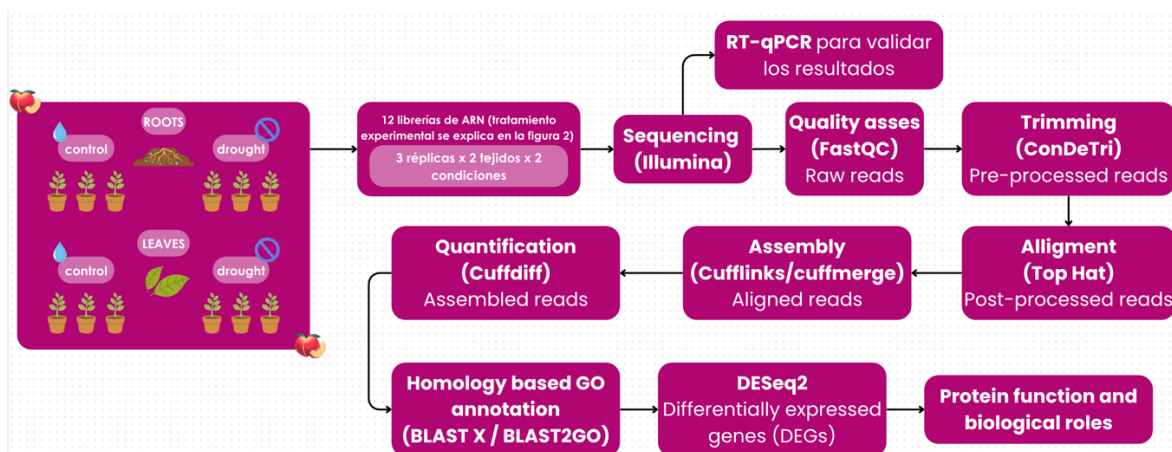


Figura 1. Flujo de trabajo seguido por el artículo para el tratamiento de las 12 librerías de ARN. Construido por Mariana Gutiérrez Tamayo a partir de la información encontrada en el artículo seleccionado.

3. Experimento RNA-seq e implementación de DESeq2.

El experimento de RNA-seq fue la técnica utilizada para generar los datos genéticos del estudio. Tal como se mencionó antes, se extrajo el ARN total de las muestras de raíz y hojas, y luego el ARN se purificó para obtener solo el ARN mensajero (ARNm), que fue fragmentado para continuar con la síntesis del ADN complementario (ADNc) que fue preparado para la secuenciación. Las librerías de ADNc se secuenciaron utilizando la tecnología *Illumina HiSeq 2000*. Una vez obtenidas las secuencias de reads, estas se alinearon con el genoma de referencia del duraznero *Prunus persica*. Después del mapeo, se realizó el conteo de lecturas por gen, es decir, cuántas veces se leyó cada gen en cada una de las 12 muestras y, finalmente dicho conteo fue la entrada de datos crudos para el análisis de expresión diferencial. Lo anterior, coincide con el flujo de trabajo presentado en la **Figura 1** y se complementa con el flujo del experimento RNA-seq presentado en la **Figura 2**.

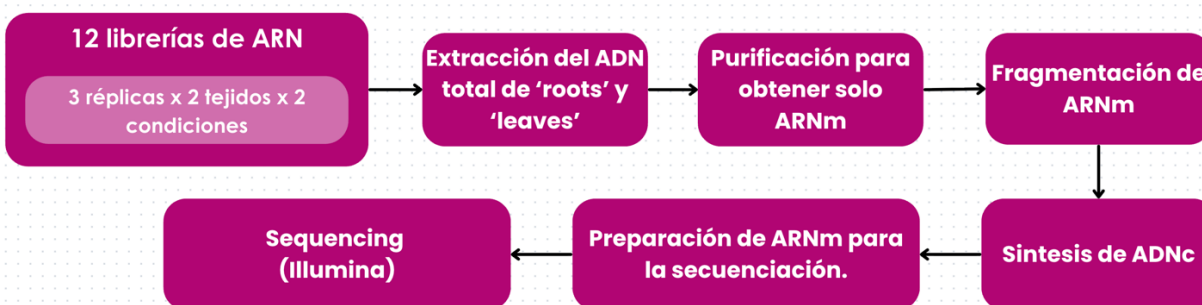


Figura 2. Flujo de trabajo seguido por el artículo para el experimento de trabajo de la técnica RNA-seq. Construido por Mariana Gutiérrez Tamayo a partir de la información encontrada en el artículo seleccionado.

El análisis de expresión diferencial se llevó a cabo utilizando el paquete de R, *DESeq2*, el cual es una herramienta bioestadística que modela la variación en los datos de conteo de las lecturas genéticas, utilizando una distribución binomial negativa, que permite la correcta estimación de la dispersión de los datos. *DESeq2* primero realiza una normalización de los datos de conteo, siendo esto un paso esencial porque la cantidad total de lecturas varía entre las diferentes librerías secuenciadas, llevando a conclusiones erróneas. En este sentido, normalizar ajusta los datos para que las comparaciones de expresión génica sean válidas. El propósito de usar *DESeq2* es identificar genes cuya expresión ha

cambiado significativamente entre las condiciones experimentales (Control vs. Sequía). Aunque en el artículo se trabaja con ambos tejidos, para este caso se eligieron solamente las raíces, es decir, *Raíces bajo Sequía vs. Control*.

Tras el modelado, la herramienta calcula dos métricas clave para la expresión diferencial; el `Log2 Fold Change (Log2FC)` que mide la magnitud del cambio en la expresión de un gen entre los dos grupos y el valor *p* ajustado (`False Discovery Rate - FDR`) que estima la significancia estadística del cambio. Para el caso del estudio de referencia, se consideró un gen como diferencialmente expresado (DEG) si tenía un $\text{Log2FC} \geq |1|$ y un $\text{FDR} \leq 0.05$, indicando probabilidad de que el cambio sea aleatorio es muy baja. Por otra parte, el contraste biológico implementado en el análisis de expresión diferencial del artículo define el `Log2FC` como el logaritmo de la razón entre la expresión de la condición de estrés (sequía), utilizada como numerador y grupo de interés, y la condición control (riego normal), utilizada como denominador y grupo de referencia. La expresión mencionada se muestra en la ecuación (1).

$$\text{Log2FC} = \text{Log}_2 \left(\frac{\text{Expresión}_{\text{sequía}}}{\text{Expresión}_{\text{control}}} \right) \quad (1)$$

Entonces, un valor de $\text{Log2FC} \geq |1|$ indica que la expresión del gen ha sido inducida (sobreexpresión) por el estrés hídrico, sugiriendo un papel en la respuesta de defensa o tolerancia a la sequía. Por el contrario, un valor fuera del rango estipulado representa represión (subexpresión) del gen, lo que puede reflejar una estrategia de ahorro de energía o un daño inducido por la sequía. La elección de este contraste es fundamental, ya que facilita la identificación directa de los genes candidatos a la tolerancia al medir su cambio con respecto al estado de normalidad. Entonces, con la finalidad de garantizar las mismas condiciones utilizadas en el artículo, para la réplica del *DESeq2* a partir de los datos crudos, se utilizó el mismo contraste, un *p*-valor ajustado (FDR) de ≤ 0.05 , y un valor de corte para el `Log2 Fold Change` $\geq |1|$.

4. Resultados principales.

Al ejecutar la librería *DESeq2* en R, el *output* principal fueron los datos de expresión génica diferencial (DEGs), arrojados en formato *.csv* que proporcionaron una fila por cada gen analizado, con métricas como el `LFC` que cuantifica la magnitud y dirección del cambio (*Up* o *Down*) y el *p*-valor ajustado utilizado para determinar la significancia estadística. Una vez filtrados los genes significativos mediante las condiciones de $\text{LFC} \geq |1|$ y $\text{padj} \leq 0.05$, se generó una tabla comparativa que cruzó dichos DEGs con el archivo de resultados descargado de *Expression Atlas*, buscando emparejar el ID de los genes. Finalmente, se utilizaron comandos en la terminal para cuantificar los genes obtenidos (totales, Up-regulated y Down-regulated) tanto en la corrida local como en los resultados de *Atlas*, para así validar la replicación del análisis.

Nota: los archivos de salida luego de correr el script de R se pueden encontrar en el GitHub correspondiente a esta entrega, mientras que el header de la tabla que compara ambos resultados se encuentran en la **Tabla 1**. Por su parte, en la **Tabla 2**, se muestran los resultados obtenidos con los comandos utilizados para analizar los datos obtenidos; dichos comandos, también se encuentran en disponibles en el GitHub en el archivo README. Además, en la **Gráfica 1** se muestra el volcano plot obtenido con la corrida local.

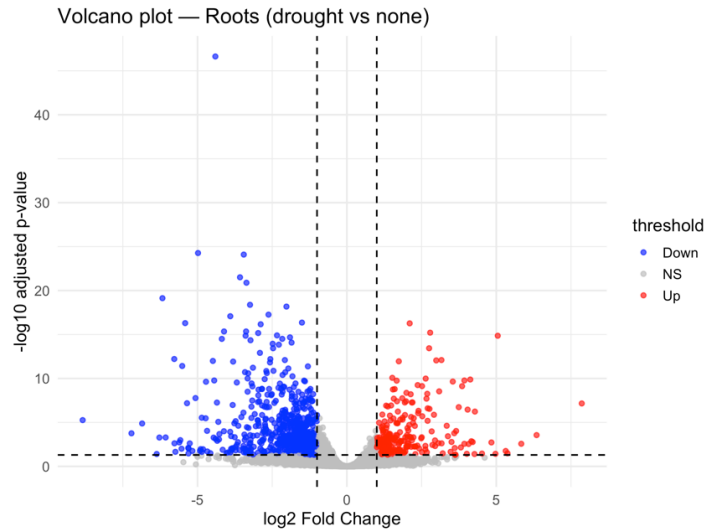
Expresión diferencial de genes - Transcriptómica

Tabla 1. Resultados comparativos entre los archivos de corrida local y los relacionados con el artículo descargados desde *Expression Atlas*. Generado a partir de un script y adaptado por Mariana Gutiérrez Tamayo.

Column1	lfc R	lfc_atla	lfc_differenc	padj R	padj_atla	Direction	Direction_atla	Direction_matc
PRUPE_4G035600	-8,844684916	-2	6,844684916	5,4511E-06	0,001108904	Down	Down	TRUE
PRUPE_1G036600	-6,366576677	-0,8	5,566576677	0,039133804	0,330887333	Down	Down	TRUE
PRUPE_1G341900	-7,211705821	-1,8	5,411705821	0,000173051	0,003737163	Down	Down	TRUE
PRUPE_8G110900	-6,854044953	-1,7	5,154044953	1,31972E-05	0,007860717	Down	Down	TRUE
PRUPE_7G152100	5,375327945	1,2	4,175327945	0,032598346	0,093296962	Up	Up	TRUE
PRUPE_8G074700	-5,303819553	-1,2	4,103819553	0,033958985	0,093366983	Down	Down	TRUE
PRUPE_3G277500	7,861394137	3,8	4,061394137	6,94167E-08	1,36141E-13	Up	Up	TRUE
PRUPE_1G088500	6,348987076	2,3	4,048987076	0,000276499	0,00010773	Up	Up	TRUE
PRUPE_4G066200	5,836510964	1,8	4,036510964	0,002691708	0,003755563	Up	Up	TRUE
PRUPE_6G175500	-5,535444914	-1,5	4,035444914	0,009851454	0,024159632	Down	Down	TRUE

Tabla 2. Resumen comparativo entre los datos de corrida local y los relacionados con el artículo descargados desde *Expression Atlas*. **Construido por:** Mariana Gutiérrez Tamayo.

Característica	R – DESeq2	Expresión Atlas
Genes totales	1298	
Direction_Match = TRUE	972	
Direction_Match = FALSE	326	
Conteo Up	195	392
Conteo Down	678	906
Conteo NS	326	0



Gráfica 1. Volcano plot con el \log_2 foldChange vs el adjusted p-value. Generado por el script de R.

5. Interpretación de diferencias.

En este caso en la **Gráfica 1** se muestran los DEGs para el contraste *Raíces bajo Sequía vs. Control*. Los puntos rojos tienen $LFC > +1$ y $p_{adj} < 0.05$ y corresponden a los genes sobre expresados, mientras que los puntos azules tienen $LFC < -1$ y $p_{adj} < 0.05$ y son los sub expresados. Los puntos más altos en el gráfico representan los genes con mayor certeza estadística, es decir, con los p_{adj} más bajos y se encuentran mayormente de color rojo (sobre expresados). Esta distribución asimétrica evidencia que, si bien la sequía induce la represión de algunos procesos metabólicos para el ahorro de energía, la reacción dominante de la raíz es la inducción de genes asociados a la tolerancia y a la defensa. El volcano plot entonces muestra que la raíz puede ser el principal órgano de respuesta al estrés hídrico.

El análisis comparativo demuestra una buena consistencia biológica entre los resultados de la corrida de R y el archivo de *Expression Atlas*, donde la **Tabla 2** proporciona evidencia de que la replicación del análisis fue exitosa. Respecto a la dirección de la regulación, los genes clasificados como *Up-regulated* en el resultado del *Atlas* mantienen un LFC con el mismo signo en el análisis local, al igual que los genes *Down-regulated*. Lo anterior confirma que para ambos casos se utilizó el mismo contraste, sin embargo, los conteos no son iguales porque algunos de los genes de la replicación fueron clasificados como no significativos (zona gris del volcano plot y Conteo NS en la **Tabla 2**). En cuanto a la magnitud del cambio, se observa una diferencia de aproximadamente 6 unidades en los valores absolutos del LFC para el primer gen de la tabla y la diferencia de magnitud tiende a encontrarse en ese rango. Finalmente, la significancia estadística, muestra que los p-valores de los genes más regulados son bajos en ambas columnas, lo que garantiza que la capacidad para distinguir el cambio biológico real del ruido estadístico es similar entre la réplica local y el resultado del Atlas. Los cambios encontrados pueden deberse a que la estimación del LFC fue modelada de forma diferente al estudio, pero el hecho de que la dirección y la significancia de los genes clave sean reproducibles valida la conclusión biológica del estudio original.

6. Interpretación biológica del contraste.

El contraste biológico implementado en el análisis, representado por la comparación entre las plantas con estrés por sequía y las plantas de control, sirve como referencia para medir el impacto de la falta de agua a nivel molecular. En el estudio se aísla el efecto del tratamiento, permitiendo identificar aquellas rutas moleculares y genes que son específicamente activados o desactivados por la falta de agua. El artículo de (Ksouri et al., 2016) plantea que la raíz (portainjerto GF677) exhibe una respuesta mucho más robusta e intensa, con 500 genes diferencialmente expresados, lo que es más del doble de los 236 DEGs encontrados en la hoja (*Catherina*). En ambos tejidos, fue más fuerte la sobreexpresión de genes, indicando una estrategia de adaptación y defensa de las plantas. Específicamente, se identificó la regulación clave de genes involucrados en rutas hormonales, como la citoquinina y el ácido abscísico (ABA), además de la activación de factores de transcripción de las familias NAC y AP2/ERF, los cuales son interruptores esenciales en la señalización del estrés. La raíz, además, mostró reprogramación del metabolismo para el ajuste energético, es decir, que durante la sequía la planta tiende a apagar algunos genes, pero enciende otros intentando reducir el gasto energético y “cerrar” procesos fisiológicos que consuman agua. Por ejemplo, de acuerdo con el artículo seleccionado, la activación de la biosíntesis de fenilpropanoides es un mecanismo clave de defensa y aclimatación, ya que es una vía precursora de compuestos estructurales como la lignina, que fortalece la pared celular, y de metabolitos secundarios protectores como los flavonoides.

7. Conclusiones.

Las raíces del portainjerto GF677 reaccionan al estrés hídrico activando diversas rutas de tolerancia, lo que confirma su rol clave como amortiguador del estrés por sequía. La respuesta obtenida es consistente con el papel de la raíz como el primer tejido vegetal en percibir el estrés hídrico, mientras que la hoja, al ser el tejido para el control de la pérdida de agua, muestra una más relacionada con la conservación de recursos y la defensa estructural. El estudio también mostró que *Prunus persica* tiene una compleja red regulatoria involucrada en la respuesta a la sequía, incluyendo proteínas asociadas con la transducción de señales, regulación de la transcripción, regulación hormonal y homeostasis redox.

REFERENCIAS

Arús, P., Verde, I., Sosinski, B., Zhebentyayeva, T., and Abbott, A. G. (2012). The peach genome. *Tree Genet. Genomes* 8, 531–547. doi: 10.1007/s11295-012-0493-8

Ksouri, N., Jiménez, S., Wells, C. E., Contreras-Moreira, B., & Gogorcena, Y. (2016). Transcriptional responses in root and leaf of *prunus persica* under drought stress using RNA sequencing. *Frontiers in Plant Science*, 7. <https://doi.org/10.3389/fpls.2016.01715>